Missing Data Methods for Observational Health Dataset

Gang Cheng

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:
Yen-Chi Chen, Chair
Kwun Chan
Thomas Richardson

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

Missing Data Methods for Observational Health Dataset

Gang Cheng

Chair of the Supervisory Committee:
Yen-Chi Chen
Department of Statistics

This dissertation is motivated by missing data problems arising from two observational health datasets. The first dataset is created by the SWOG study that linked medicare claims to a prostate cancer prevention trial dataset. The second dataset is a diabetes EHR dataset that contains longitudinal measurements of diabetes patients for 11 years.

For the first dataset, we are interested in estimating the long-term effect of a treatment. In a time-to-event setting, medicare claims are linked to clinical trial data to extend the follow-up period for trial participants. This allows the estimation of the long-term effect that cannot be estimated by clinical trial data alone. However, such data linkages are often incomplete for various reasons. We formulate incomplete linkages as a missing data problem with careful considerations of the relationship between the linkage status and the missing data mechanism. We propose a conditional linking at random (CLAR) assumption and an inverse probability of linkage weighting (IPLW) partial likelihood estimator. We show that our IPLW partial likelihood estimator is consistent and asymptotically normal.

For the second dataset, the longitudinal measurements for diabetes patients are subject to nonmonotone missingness. The conventional ignorability and missing-at-random (MAR) conditions are unlikely to hold for nonmonotone missing data and data analysis can be very challenging with few complete data. We introduce the available complete-case missing value (ACCMV) assumption for handling nonmonotone and missing-not-at-random (MNAR)

problem. Our ACCMV assumption is applicable to dataset with a small set of complete observations and we show that the ACCMV assumption leads to nonparametric identification of the distribution for the variables of interest. We further propose an inverse probability weighting estimator, a regression adjustment estimator and a multiply-robust estimator for estimating a parameter of interest. Asymptotic and efficiency theories of the proposed estimators are studied. We further illustrate the applicability of our method by applying it to the diabetes EHR dataset.

Finally, we consider the problem of trajectory recovery. Repeated measurements collected from individuals naturally form a long trajectory and the length of the trajectory creates additional difficulty for modeling and computation. We introduce a block-Markov type assumption to handle such missing data problems. We prove that our assumption leads to nonparametric identification of the joint distribution of the trajectory. Based on this assumption, we are able to decompose trajectories into multiple missing blocks and thus greatly reduce both the computation and modeling complexity. For modeling purposes, we further propose a model-based assumption, which allows us to use both linear models and flexible machine learning models to impute missing values. We further illustrate the applicability of our method by applying it to the diabetes EHR dataset.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to thank my PhD advisor, Yen-Chi Chen. This dissertation would not be possible without his mentorship and guidance. I learned a lot from Yen-Chi during my Ph.D study and he is always patient, supportive and insightful. Getting a Ph.D is pretty tough and Yen-Chi's support and encouragement has made this a lot easier.

I would also like to thank Yingqi Zhao for introducing me to the SWOG study and the diabetes dataset, of which this dissertation is built on. I would like to thank Gary Chan, Thomas Richardson and Jing Tao for serving on my supervisory committee and offering many helpful advice. I thank Alex Luedtke for being my academic advisor. I am grateful to Scott Emerson and Noah Simon for encouraging me to pursue a Ph.D when I was a master student. I am also grateful to the faculty at the statistics and biostatistics department for many wonderful courses.

I would like to thank Tracy Pham, Ellen Reynolds and Kristine Chan for the many help during my Ph.D study. I would also thank my friends at UW, who have made my life at UW a lot more fun.

Finally, I am grateful to my parents, for always understanding and supporting me. I would also like to thank Xinyuan for the love and support during this tough journey.

DEDICATION

to my family

Chapter 1

INTRODUCTION

The increasing availability of electronic health data has created both opportunities and challenges. The SWOG study recently linked medicare claims to the prostate cancer prevention trial (PCPT) dataset to extend the follow-up period for trial participants. PCPT showed that seven years of finasteride[1] reduced PC risk by 25% (Thompson et al., 2003). However, it is unclear if seven years' of trial follow-up suffices to determine the maximum benefit of the treatment. Further, the reduced risk of prostate cancer for subjects receiving finasteride might not be maintained after finasteride discontinuation (Unger et al., 2018). The linkage of medicare claims extends the follow-up periods up to a maximum of 20 years compared to 7 years by the PCPT and allows the estimation of long-term effect of finasteride. However, data linkages are incomplete. About 75% of the participants were linked to the medicare claims and thus a quarter of the participants missed their information after the end of the clinical trial.

In chapter 2, we propose a novel conditional linking at random (CLAR) assumption for the linkage mechanism to deal with the incomplete linkage problem. This allows us to develop an inverse probability weighting (IPW) type estimator to estimate the long-term effect with Cox model. We further provide the asymptotic theories for our estimator, allowing for the inclusion of time-dependent covariates, which appears to be new in the literature. We further consider a couple of alternative approaches and carefully compare it to the IPW type estimator we propose. This work is presented in Cheng et al. (2022).

Chapters 3 and 4 are motivated by a diabetes electronic health records (EHR) dataset.

---

[1] a treatment that inhibits the development of potent androgen that fuels the malignancy of prostate cancer.

This dataset contains quarterly longitudinal measurements of diabetes patients for 11 years, from 2003 to 2013. However, EHR data poses significant challenges. For example, a patient's information is recorded only if and when they visit a clinic. This naturally leads to non-monotone missing data when a patient reappeared after one or more missed visits. Further, the missing patterns are often associated with the underlying missing measurements. For example, sicker patients are likely to visit the clinic often and have fewer missing values, while healthier patients are likely to miss visits and thus have more missing values. This suggests that the missing mechanism is also missing not at random (MNAR). Thus, it is common to have nonmonotone and MNAR data for the longitudinal measurments contained in the EHR dataset.

To handle the missing data under MNAR, there are two common frameworks: the selection model and the pattern mixture model. Let $X \in \mathbb{R}^d$ be the study variables and $R \in \{0,1\}^d$ be a binary vector indicating the response pattern such that $R_i = 1$ if variable X_i is observed. The selection model framework attemps to model the selection probability $P(R = r|X)$ such that this selection probability is identifiable from the data. A common approach to identify the selection probability is via assuming a parametric model and some additional structures so that we can estimate the selection probability from the data. The pattern mixture model framework instead focuses on modeling the extrapolation density $P(X_{\bar{r}}|X_r, R = r)$, where $X_{\bar{r}} = (X_i : r_i = 0)$ are the unobserved variables and $X_r = (X_i : r_i = 1)$ are the observed variables.

With nonmonotone missing data, modeling the selection probability $P(R = r|x)$ is a very challenging task even under missing at random (MAR) assumption (Robins and Gill, 1997; Sun and Tchetgen Tchetgen, 2018). Thus, we focus on pattern mixture framework that identifies the extrapolation densities. Under the pattern mixture framework, previous work have considered the complete-case missing value (CCMV) restriction, which assumes that $P(X_{\bar{r}}|X_r, R = r) = P(X_{\bar{r}}|X_r, R = 1_d)$. Thus, CCMV identifies the extrapolation density with the set of complete observations $R = 1_d$. This may be suitable when the size of the complete data is not too small. However, we only have a very small set of complete

observations in the diabetes EHR dataset. For example, for the first year's data, complete cases only account for 5% of the observations in our dataset.

In Chapter 3, we focus on the first year's data and propose the available complete-case missing value (ACCMV) assumption for handling nonmonotone missing data that is MNAR. Our assumption is suitable for analyzing datasets with few complete cases. We propose several estimators and study the efficiency theory and asymptotic thoeries.

In Chapter 4, we consider the problem of trajectory recovery. Longitudinal measurements naturally form long trajectories and the length of the trajectory creates additional difficulties for modeling and computation. We then propose a block-Markov type assumption that decomposes the trajectory into multiple missing blocks. We prove that our assumption leads to nonparametric identification of the joint distribution of the trajectory. For modeling purposes, we modify our assumption to a model-based version that assumes a multivariate normal distribution for each block. We propose to esitmate the multivariate normal distribution with linear models or other flexible nonparametric/machine learning models. Further, we use multiple imputation to obtain multiple completed trajectories and estimate the uncertainty with bootstrap.

Chapter 2

LONG-TERM EFFECT ESTIMATION WHEN COMBINING CLINICAL TRIAL AND OBSERVATIONAL FOLLOW-UP DATASETS

2.1 Introduction

With the increasing availability of electronic health data, combining experimental and observational datasets has been widely applied in public health research (Warren et al., 2002; Gilbert et al., 2018). In a time-to-event setting, we consider the setup when data from a clinical trial is combined with an observational follow-up dataset, such as electronic health records or administrative claims. Clinical trials often study the effect of a particular treatment for a fixed period of time and it might not be long enough to determine the maximum benefit of the treatment. In contrast, an observational dataset such as medicare claims naturally extends the follow-up period for clinical trial participants at minimal cost. This enables the estimation of the long-term effect for the treatment after the clinical trial. To combine the observational follow-up dataset with the clinical trial data, records belonging to the same individual can be linked with unique identifiers from both datasets. We use Cox model (Cox, 1972) to define the long-term effect as the parameter for treatment when participants from the clinical trial are linked to an observational follow-up dataset.

For a real data example, the Prostate Cancer Prevention Trial (PCPT) was previously launched to examine whether finasteride[1] could prevent the development of prostate cancer (PC). PCPT showed that seven years of finasteride reduced PC risk by 25% (Thompson et al., 2003). However, it was unclear if seven years' of trial follow-up sufficed to determine the

[1]a treatment that inhibits the development of potent androgen that fuels the malignancy of prostate cancer

maximum benefit of the treatment. Further, the reduced risk of prostate cancer for subjects receiving finasteride might not be maintained after finasteride discontinuation (Unger et al., 2018). A later study linked medicare claims to the clinical records for participants in PCPT with their social security numbers (SSN) (Unger et al., 2018) to estimate the long-term effect of finasteride on prostate cancer (PC) development. In this example, PCPT is the clinical trial and medicare claim is the observational follow-up dataset. Medicare claims extend the follow-up periods up to a maximum of 20 years compared to 7 years by the PCPT. Thus, we can observe more diagnosis times of PC within the medicare claims dataset.

However, not every participant in the clinical trial can be linked to the observational dataset. For the PCPT-medicare example, some participants might not be willing to share their SSNs or they may be enrolled in health maintenance organization (HMO) and medicare claims are not applicable to HMO individuals (Unger et al., 2018). With incomplete linkages, survival outcomes in the observational dataset might be missing for some participants. For a participant censored in the PCPT, meaning he was not diagnosed with PC within the clinical trial, his survival outcome in the observational dataset would be missing if he is unlinked. On the other hand, if a participant was diagnosed with PC within the clinical trial, his survival outcome has been already observed within the clinical trial and the linkage to the observational follow-up dataset is in fact not necessary. This suggests that the missingness of survival outcome depends both on the linkage status and whether a participant was censored in the clinical trial or not.

To deal with the missing survival outcomes, a complete-case analysis that only includes linked participants will ignore all the unlinked participants with observed survival outcomes within the clinical trial. However, simply adding those unlinked participants to the complete-case analysis will also cause biased estimate. Essentially this would lead to the missingness of the survival outcomes to depend on itself and the missingness is then missing not at random (MNAR). To properly incorporate those unlinked participants with observed survival outcomes, we choose to model the linkage probability directly. As we discussed above, participants who miss the survival outcomes in the observational dataset are those who are

censored in the clinical trial and unlinked. Hence we take a missing data perspective and propose a novel conditional linking at random (CLAR) assumption for the linkage mechanism. More specifically, we assume that for participants who are censored in the clinical trial, linkages are independent of the survival outcomes after conditioning on their covariates vectors, such as social economic status or other clinical factors. No linkage assumptions are made for those participants uncensored in the clinical trial. Under the CLAR assumption, we can then weight each participant appropriately and obtain unbiased estimates for the long-term effect.

As we use Cox model to define the long-term effect, we develop an inverse probability of linkage weighting (IPLW) partial likelihood estimator. We prove the asymptotic normality and consistency of our IPLW partial likelihood estimator. Our approach allows inclusion of time-dependent covariates for more flexibility. While there has been plenty work (Binder, 1992; Robins, 1993; Lin, 2000; Qi et al., 2005) on proving the asymptotic convergence for an inverse probability weighting (IPW) type partial likelihood estimator when there are only time-independent covariates, their proof cannot be easily generalized to the case when there are time-dependent covariates (Breslow and Wellner, 2007). To this end, we establish an IPLW empirical process weak convergence result that builds on the work in Saegusa and Wellner (2013) and borrow the techniques from Lin and Wei (1989) to extend the theoretical results to include time-dependent covariates.

Related work. There has been an increasing amount of work on combining different datasets and studying the treatment effect on long-term outcomes in causal inference (Rosenman et al., 2018, 2020; Kallus and Mao, 2020; Athey et al., 2020). All these works focus on using experimental and observational datasets that contain different set of individuals, which is different from our setup. IPW has also been widely applied in the survival analysis setting (Binder, 1992; Robins et al., 1994; Robins and Finkelstein, 2000; Hernán et al., 2000; Lin, 2000; Qi et al., 2005; Tsiatis, 2007; Breslow and Wellner, 2007; Saegusa and Wellner, 2013). Robins and Finkelstein (2000) applied inverse probability of censoring weights to estimate Cox model that adjusts for dependent censoring by utilizing data collected on time-

dependent prognostic factors. IPW has also been applied for Cox models with two-phase stratified sampling under right censoring (Binder, 1992; Lin, 2000; Breslow and Wellner, 2007), while Saegusa and Wellner (2013) further studied the problem of two-phase sampling for the Cox model under interval censoring with IPW. Our approach is different from all previous works as we also allow for time-dependent covariates.

Outline. In Section 2.2, we provide background and notations required for our methodological developments. We also introduce several alternative approaches. We introduce our main IPLW estimator in Section 2.3 and provide theoretical justifications. We conduct simulation studies in Section 2.4 to illustrate the validity of the proposed method. We apply our approach to the SWOG prevention trial in Section 2.5. We further compared our IPLW estimator to an alternative approach in Section 2.6. In Section 2.7, we conclude this chapter and point out some possible future directions.

2.2 Background and Notations

We first consider the oracle setting that all participants from the clinical trial are linked. We make a "no gap" assumption such that there is no gap between a participant's last recorded date within the clinical trial and the start date of the observational follow-up dataset. This "no gap" assumption eliminates the possibility of interval censoring in which a participant is diagnosed with the event of interest while not under observation. For simplicity, we make this "no gap" assumption to focus on the right censoring problem and we discuss how to relax this no gap assumption in Appendix A.6.

Time is measured since enrollment in the clinical trial and we align each patient's enrollment time to the same origin. We define T as the failure time, C_1 as the censoring time within the clinical trial and $Q = I(T \leqslant C_1)$ as the censoring indicator for the clinical trial. We use a constant τ_1 to denote the end time of clinical trial. Possible reasons for censoring in the clinical trial include loss to follow-up and administrative censoring. In contrast, the length of observational follow-up dataset is often determined by the data availability and also vary from person to person. We set a constant τ_2 with $\tau_2 > \tau_1$ as the common end

time for observational follow-up dataset and assume that there are a significant proportion of participants at risk after τ_2. Thus, we are interested in estimating the long-term effect on survival up to time τ_2 using data from clinical trial records and observational follow-up. Similarly, we define C_2 as the censoring time in the observational follow-up dataset. Possible reasons for censoring in the observational follow-up include short coverages such that a participant is not covered long enough by the observational dataset, administrative censoring where a participant is event-free and covered by observational follow-up until τ_2.

We use $C = \max(C_1, C_2)$ to denote the actual censoring time and let $\widetilde{T} = \min\{T, C\}$ denote the actual observed time and $\Delta = I(T \leqslant C)$ be the censoring indicator throughout the entire follow-up period. Let $\mathbf{X} \in \mathbb{R}^p$ denote baseline characteristics, clinical factors and treatment assignment. We also use A to denote the treatment assignment when necessary. We make the independent censoring assumption:

$$T \perp\!\!\!\perp (C_1, C_2)|\mathbf{X}$$

Thus, T is conditionally independent of $\max(C_1, C_2)$ given $\mathbf{X}$.

2.2.1 Cox model and Long-term effect

We use the Cox model to define the long-term effect and we allow for the possibility of model-misspecification. We now discuss the parameter for the long-term effect. Assume that there are n participants in the clinical trial and they are all linked to the observational follow-up dataset, so $(\mathbf{X}_i, Q_i, \widetilde{T}_i, \Delta_i)$ for $i = 1, \ldots, n$ are all observed. The Cox model assumes that the hazard function has the following form:

$$\lambda(t|\mathbf{X}, \boldsymbol{\beta}_0) = \lambda_0(t) \exp(\beta_1 A + \boldsymbol{\beta}_0'^T \mathbf{X}_{-A}) \tag{2.1}$$

with $\boldsymbol{\beta}_0' \in \mathbb{R}^{p-1}$ and $\mathbf{X}_{-A}$ is the covariate vector excluding treatment A. $\lambda_0(t)$ is the baseline hazard function and β_1 represents the long-term effect. To account for potential different effects between clinical trial and the observational follow-up period, we also consider the

following model with a change point at time τ_1 (Liang et al., 1990; Pons et al., 2003)

$$\lambda(t|\mathbf{X};\beta_0,\boldsymbol{\beta}_1,\theta)=\lambda_0(t)\exp[(\beta_1+\theta I_{t>\tau_1})A+\boldsymbol{\beta}_0'^{T}\mathbf{X}_{-A}]. \tag{2.2}$$

β_1 now represents the effect in the clinical trial, while θ represents the difference of the effect between observational follow-ups and clinical trial. $\beta_1+\theta$ now represents the long-term effect. When $\theta=0$, model (2.2) reduces to (2.1). Without loss of generality, we will use the parameter notation from model (2.1) in the following. Following the notation in Lin and Wei (1989), for the i-th individual, let $\lambda_i(t)=\lambda(t|\mathbf{X}_i)$ be the true hazard function, $N_i(t)=I(\widetilde{T}_i\leqslant t,\Delta_i=1)$ and $Y_i(t)=I(\widetilde{T}_i\geqslant t)$. For $k=0,1,2$, define

$$\mathbf{S}_n^{(k)}(t)=\frac{1}{n}\sum_{i=1}^{n}Y_i(t)\lambda_i(t)\mathbf{X}_i^{\otimes k},\quad \mathbf{s}^{(k)}(t)=\mathbb{E}[\mathbf{S}_n^{(k)}(t)]$$
$$\mathbf{S}_n^{(k)}(\beta,t)=\frac{1}{n}\sum_{i=1}^{n}Y_i(t)\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\mathbf{X}_i^{\otimes k},\quad \mathbf{s}^{(k)}(\beta,t)=\mathbb{E}[\mathbf{S}_n^{(k)}(\boldsymbol{\beta},t)]$$

where for a column vector $\mathbf{a}$, denote $\mathbf{a}^{\otimes 2}=\mathbf{a}\mathbf{a}^T$, $\mathbf{a}^{\otimes 1}=\mathbf{a}$ and $\mathbf{a}^{\otimes 0}$ refers to the scalar 1. With Cox model as a working model, the parameter of interest is $\boldsymbol{\beta}_0^*$ that solves the following equations (Andersen and Gill, 1982; Lin and Wei, 1989)

$$\mathbf{U}_0(\boldsymbol{\beta})=\mathbb{E}\left[\Delta\left(\mathbf{X}-\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},\widetilde{T})}{\mathbf{s}^{(0)}(\boldsymbol{\beta},\widetilde{T})}\right)\right]=\int_0^{\tau_2}\mathbf{s}^{(1)}(t)dt-\int_0^{\tau_2}\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\mathbf{s}^{(0)}(t)dt. \tag{2.3}$$

When Cox model is correctly specified with true parameter $\boldsymbol{\beta}_0$, we have $\boldsymbol{\beta}_0^*=\boldsymbol{\beta}_0$. When Cox model is misspecified, parameter $\boldsymbol{\beta}_0^*$ that solves equation (2.3) is still well-defined. For example, if the true model follows equation 2.2 and we use 2.1 as our model, then the long-term effect β_1 can be interpreted as the average effect over the entire follow-up period.

2.2.2 The linkage problem and assumption

Now we consider the more realistic setup that not every participant is linked to the observational follow-up dataset. We use L to denote the linkage status. $L=1$ means that the participant is linked to the observational follow-up dataset and $L=0$ means unlinked. As

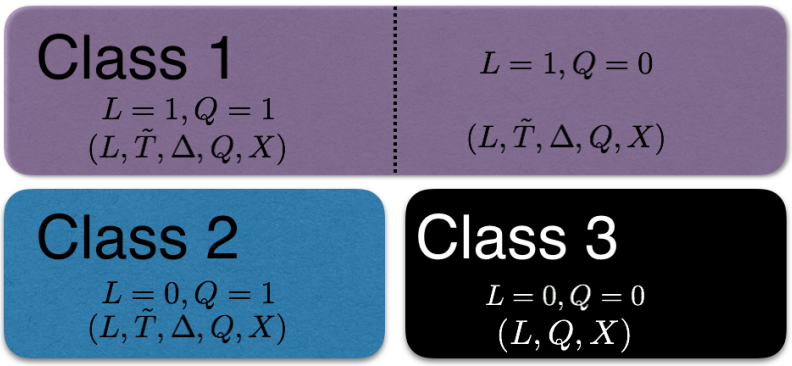

Figure 2.1: A diagram illustrating the three classes of participants and how it is defined via the linkage indicator L and in-trial censoring indicator Q.

C_2 is unobserved when $L = 0$, C is missing in this case. We classify participants into three classes based on their linkage status L and in-trial censoring indicator Q.

Class 1: $L = 1$**.** This class contains all participants linked to the observational follow-up dataset, where we have full observations: $(L, \widetilde{T}, \Delta, Q, \mathbf{X})$, but possibly censored.

Class 2: $L = 0, Q = 1$**.** This class represents participants diagnosed with the event of interest within the clinical trial but unlinked to observational follow-up. Thus, we still have complete information $(L, \widetilde{T}, \Delta, Q, \mathbf{X})$.

Class 3: $L = 0, Q = 0$**.** This class includes participants who did not experience the event of interest during the clinical trial, and were not linked to observational follow-up dataset. Both $\widetilde{T}$ and Δ are missing and we only observe $(L, Q, \mathbf{X})$.

Figure 2.1 provides a summary of the three classes defined above. We have completely observed data $(L, \widetilde{T}, \Delta, Q, \mathbf{X})$ in class 1 and 2, while two important variables $\widetilde{T}, \Delta$ are missing in class 3. To deal with the missing $\widetilde{T}$ and Δ, we propose the following conditional linking at random assumption (CLAR):

$$\textbf{(A1)}\ P(L = 1|\widetilde{T}, \Delta, Q = 0, \mathbf{X}) = P(L = 1|Q = 0, \mathbf{X}).$$

More specifically, (**A1**) states that for a participant who is censored in the clinical trial,

his/her linkage status is independent of the survival outcomes given his baseline covariates, clinical factors and treatment assignment. For example, clinical trial participants with higher social economic status might be more sensitive to personal privacy and not willing to share personal information that are important for data linkage.

We compare our CLAR assumption (A1) with the classical MAR type assumption (Rubin, 1976; Little and Rubin, 2019), which can be written as follows

$$P(L = 1|\widetilde{T}, \Delta, Q, \mathbf{X}) = P(L = 1|Q, \mathbf{X}). \tag{2.4}$$

CLAR is actually implied by MAR (2.4). However, CLAR is restricting the conditional independence to the subpopulation with $Q = 0$, while MAR (2.4) is assuming the conditional independence for the whole population. To see why this is important, note that MAR (2.4) implies that

$$P(L = 1|\widetilde{T}, \Delta, Q = 1, \mathbf{X}) = P(L = 1|Q = 1, \mathbf{X}) \tag{2.5}$$

and when $Q = 1$, both $\widetilde{T}$ and Δ are always observed, meaning that (2.5) might in fact contradict the data. In contrast, with no assumptions for participants with $Q = 1$, CLAR is non-parametrically identifiable, i.e., they will never contradict the data (Robins et al., 2000). Further discussions on potential linkage assumptions are given in Appendix A.5.

2.2.3 Alternative approaches and a motivating example

We now consider three alternative approaches that practitioners may use. We show that they all give inconsistent estimates for $\boldsymbol{\beta}_0^*$ with a simulated example when Cox model is misspecified and CLAR assumption (A1) holds. These three approaches are

- Complete-case (CC) analysis that only includes participants that are linked. These corresponds to participants in Class 1 with $L = 1$.

- Complete-case analysis plus (CC+) that includes not only participants that are linked, but also participants with $Q = 1$. This corresponds to participants in Class 1 and 2 in Figure 2.1.These are the participants with $L + Q > 0$.

- Non-linked-as-censored (NLAC) that treats participants from Class 3 ($L = 0, Q = 0$) as censored and sets their censoring time as $C = C_1$. These are the participants that are unlinked and censored in the clinical trial. Then we can fit the Cox regression with all participants from clinical trial.

We simulate data according to a Cox model with hazard function specified by covariates X_1, X_2, X_3^2 and a Cox model with covariates X_1, X_2, X_3 is fitted. More details are given in Appendix A.8. Figure 2.2 presents the 95% confidence intervals for one of the parameters. Oracle method refers to the approach that all participants in the clinical trial are linked. Among all four approaches, CC+ gives the most biased estimates. CC gives less biased estimates than CC+. NLAC also gives biased estimates compared to the oracle method. Although not shown here, the coverage of 95% confidence intervals for CC, CC+, NLAC all decrease as n increases.

We now discuss these three alternative approaches. One sufficient condition for CC to be consistent is linking completely at random (LCAR), i.e., $L \perp (\widetilde{T}, \Delta)$. This is similar to the missing completely at random (MCAR) (Little and Rubin, 2019) condition. The LCAR condition is a strong condition and may contradict the data[2] whereas the CLAR condition will not. Thus the CLAR condition (A1) is a preferred condition in a context similar to our setting. For CC+, the missingness of survival outcomes now also depends on survival outcomes itself as a participant would be included if $L + Q > 0$. Thus, the missingness can be viewed as MNAR and CC+ will always lead to biased estimates. Finally, for participants from Class 3, NLAC seems like a natural idea that simply uses their censoring time in clinical trial C_1 as the censoring time for the entire follow-up period. However, when Cox model is misspecified, NLAC in fact always give biased estimates of $\boldsymbol{\beta}_0^*$. More discussions of NLAC are deferred to section 2.6. As these three approaches are inconsistent, we have to propose a new approach to consistently estimate $\boldsymbol{\beta}_0^*$. Our proposed approach is different from NLAC as for participants in Class 3, we treat their survival outcomes as missing.

[2]we can easily test this condition

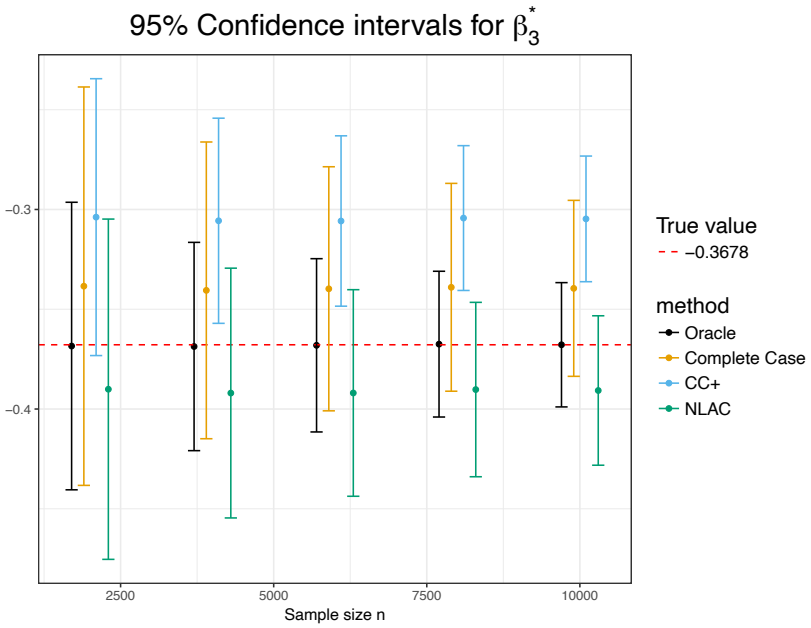

Figure 2.2: CC+, CC and NLAC all give inconsistent estimates of β_3^*.

2.3 Methods

2.3.1 IPLW Partial likelihood estimator

Due to the missingness of $\widetilde{T}$ and Δ, we can not use the classical partial likelihood for Cox model to estimate the parameters. We first illustrate our approach with time-independent covariates only. We start by writing the regular log-likelihood for Cox model as follows:

$$l_n(\boldsymbol{\beta}, \lambda_0) = \frac{1}{n}\sum_{i=1}^{n} l(\widetilde{T}_i, \Delta_i, \mathbf{X}_i; \boldsymbol{\beta}, \lambda_0) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\lambda(\widetilde{T}_i|\mathbf{X}_i)^{\Delta_i} S(\widetilde{T}_i|\mathbf{X}_i)\right)$$

where $S(t|\mathbf{x})$ is the conditional survival function for the failure time T. However, $l_n(\boldsymbol{\beta}, \lambda_0)$ is unidentifiable since we do not observe $(\widetilde{T}_i, \Delta_i)$ for participants in Class 3 of Figure 2.1. To resolve the identifiability issue, consider the expected log-likelihood

$$\mathbb{E}(l(\boldsymbol{\beta}, \lambda_0)) = \mathbb{E}\left[\log\left(\lambda(\widetilde{T}|\mathbf{X})^{\Delta} S(\widetilde{T}|\mathbf{X})\right)\right]$$

where $\mathbb{E}$ is the expectation with respect to random variable $(\widetilde{T}, \Delta, \mathbf{X})$ and

$$l(\boldsymbol{\beta}, \lambda_0) = l(\widetilde{T}, \Delta, \mathbf{X}; \boldsymbol{\beta}, \lambda_0).$$

By the law of large number, we have $l_n(\boldsymbol{\beta}, \lambda_0) \to_p \mathbb{E}(l(\boldsymbol{\beta}, \lambda_0))$.

Proposition 2.3.1. *Under assumption (A1), we have*

$$\mathbb{E}(l(\boldsymbol{\beta}, \lambda_0)) = \mathbb{E}\left[\frac{I(L+Q>0)l(\boldsymbol{\beta}, \lambda_0)}{Q + (1-Q)P(L=1|\mathbf{X}, Q=0)}\right] \tag{2.6}$$

Proposition 2.3.1 shows that $\mathbb{E}(l(\boldsymbol{\beta}, \lambda_0))$ can be expressed in the IPLW form and the proof can be found in Appendix A.3. We assume a logistic regression model for the linkage probability $P(L=1|\mathbf{X}, Q=0)$ for simplicity such that

$$P(L=1|\mathbf{X}, Q=0; \boldsymbol{\gamma}_0) = \frac{\exp(\boldsymbol{\gamma}_0^T \widetilde{\mathbf{X}})}{1+\exp(\boldsymbol{\gamma}_0^T \widetilde{\mathbf{X}})} = \pi_{\gamma_0}(\mathbf{X})$$

with $\boldsymbol{\gamma}_0 \in \mathbb{R}^{p+1}$ and $\widetilde{\mathbf{X}} = (1, \mathbf{X}^T)^T$. In particular, $\boldsymbol{\gamma}_0$ can be estimated by the maximum likelihood estimator $\widehat{\boldsymbol{\gamma}}_n$. Using result (2.6), an IPLW estimator of $\mathbb{E}(l(\boldsymbol{\beta}, \lambda_0))$ is

$$l_n(\boldsymbol{\beta}, \lambda_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(L_i + Q_i > 0)}{Q_i + (1-Q_i)\pi_{\widehat{\gamma}_n}(\mathbf{X}_i)}[\Delta_i \log \lambda(\widetilde{T}_i|\mathbf{X}_i) + \log S(\widetilde{T}_i|\mathbf{X}_i)]$$

with the log-likelihood being weighted by $\widehat{w}_i = I(L_i + Q_i > 0)/[Q_i + (1-Q_i)\pi_{\widehat{\gamma}_n}(\mathbf{X}_i)]$. IPLW partial score can then be derived as

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \widehat{w}_i \left(\mathbf{X}_i - \frac{\sum_{j=1}^{n} \widehat{w}_j I(\widetilde{T}_j \geqslant \widetilde{T}_i)\exp(\mathbf{X}_j^T \boldsymbol{\beta})\mathbf{X}_j}{\sum_{j=1}^{n} \widehat{w}_j I(\widetilde{T}_j \geqslant \widetilde{T}_i)\exp(\mathbf{X}_j^T \boldsymbol{\beta})}\right), \tag{2.7}$$

which is a sample analog of equation (2.3) and $\widehat{\boldsymbol{\beta}}_n$ can be obtained by solving equation (2.7) using standard statistical software. Detailed derivations of (2.7) can be found in Appendix A.1. In summary, our method for estimating the regression parameters of Cox model consists of two steps:

- **Step 1.** We estimate the linkage probability $P(L=1|Q=0, \mathbf{X}; \boldsymbol{\gamma}_0) = \pi_{\gamma_0}(\mathbf{X})$ with logistic regression. $\widehat{\boldsymbol{\gamma}}_n$ can be obtained by the maximum likelihood estimation.

- **Step 2.** An individual with $\Delta_i = 1$ is weighted with weight $\widehat{w}_i$ using the estimated linkage probability $\pi_{\widehat{\boldsymbol{\gamma}}_n}(\mathbf{X}_i)$. More specifically, for participants with $Q_i = 1$, the weight is 1; for participants with $Q_i = 0$ and $L_i = 1$, the weight is $1/\pi_{\widehat{\boldsymbol{\gamma}}_n}(\mathbf{X}_i)$. $\widehat{\boldsymbol{\beta}}_n$ is then obtained by solving (2.7).

2.3.2 Time-dependent covariates

In practice, it is common for Cox regression to include time-dependent covariates and we now extend our IPLW method to incorporate the time-dependent covariates. We build on the work in Lin and Wei (1989) to extend our theoretical results to include time-dependent covariates with the IPLW partial likelihood estimator. Let $\mathbf{X}_i(t) = (\mathbf{Z}_{1i}^T, \mathbf{Z}_{2i}(t)^T)^T \in \mathbb{R}^p$ denotes the covariates vector, where $\mathbf{Z}_{1i} \in \mathbb{R}^{d_1}$ corresponds to the baseline (time-independent) covariates and $\mathbf{Z}_{2i}(t) \in \mathbb{R}^{d_2}$ corresponds to the time-dependent covariates for $i = 1, \ldots, n$ at time t. We have $d_1 + d_2 = p$. $\mathbf{Z}_{2i}(t)$ can represent covariates that are continuously monitored during the clinical trial and observational follow-up datasets. For Cox model with a change point (2.2), $Z_{2i}(t) = I(t > \tau_1)A$. Let $\bar{\mathbf{X}}(t) = \{\mathbf{X}(s) : s \in [0, t]\}$ denote the history of covariate vector $\mathbf{X}(s)$, up to time t. To incorporate time-dependent covariates into the IPLW partial likelihood, we modify the CLAR assumption (A1) as following:

Assumptions.

(D1) The linkage status satisfies that

$$P(L = 1|\widetilde{T}, \Delta, Q = 0, \bar{\mathbf{X}}(\tau_M)) = P(L = 1|Q = 0, \mathbf{Z}_1).$$

The distribution of $\mathbf{Z}_1$ is not concentrated on a $(d_1 - 1)$ dimensional affine subspace of $\mathbb{R}^{d_1}$.

(D1) assumes that linkage only depends on time-independent covariates $\mathbf{Z}_1$ and also ensures the identifiability of $\boldsymbol{\gamma}_0$ (see, e.g., Example 5.40 of Van Der Vaart (2000)). We can further relax this assumption such that linkage also depends on $\mathbf{Z}_2(C_1)$, the value of time-dependent covariates at the censoring time in clinical trial. For simplicity, we assume that linkage

only depends on the baseline (time-independent) covariates. Based on assumption (D1), we modify the weights as follows:

$$w_i = \frac{I(L_i + Q_i > 0)}{Q_i + (1 - Q_i)\pi_{\gamma_0}(\mathbf{Z}_{1i})} \qquad \widehat{w}_i = \frac{I(L_i + Q_i > 0)}{Q_i + (1 - Q_i)\pi_{\widehat{\gamma}_n}(\mathbf{Z}_{1i})}$$

with $\boldsymbol{\gamma}_0 \in \mathbb{R}^{d_1+1}$. The IPLW partial score incorporating time-dependent covariates is now as follows:

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \widehat{w}_i \left\{ \mathbf{X}_i(\widetilde{T}_i) - \frac{\mathbf{S}^{(1)}_{n,w}(\boldsymbol{\beta}, \widetilde{T}_i)}{\mathbf{S}^{0}_{n,w}(\boldsymbol{\beta}, \widetilde{T}_i)} \right\} = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_i \int_0^{\tau_2} \mathbf{X}_i(t) dN_i(t) - \int_0^{\tau_2} \frac{\mathbf{S}^{(1)}_{n,w}(\boldsymbol{\beta}, t)}{\mathbf{S}^{(0)}_{n,w}(\boldsymbol{\beta}, t)} d\bar{N}(t) \tag{2.8}$$

with $\mathbf{S}^{(k)}_{n,w}(\boldsymbol{\beta}, t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_i Y_i(t) \exp(\boldsymbol{\beta}^T \mathbf{X}_i(t)) \mathbf{X}_i(t)^{\otimes k}$ and $\bar{N}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_i N_i(t)$.

We redefine

$$\mathbf{s}^{(k)}(\boldsymbol{\beta}, t) = \mathbb{E}[Y(t) \exp(\boldsymbol{\beta}^T \mathbf{X}(t)) \mathbf{X}(t)^{\otimes k}]$$

and

$$\mathbf{s}^{(k)}(t) = \mathbb{E}[Y(t) \lambda(t|\bar{\mathbf{X}}(t)) \mathbf{X}(t)^{\otimes k}]$$

where $\lambda(t|\bar{\mathbf{X}}(t))$ is the true hazard function for participants with covariates history $\bar{\mathbf{X}}(t)$. The estimated parameter $\widehat{\boldsymbol{\beta}}_n$ solves $\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \mathbf{0}$ and its population version $\boldsymbol{\beta}_0^*$ solves $\mathbf{U}_0(\boldsymbol{\beta}) = 0$ with

$$\mathbf{U}_0(\boldsymbol{\beta}) = \mathbb{E}\left[\Delta \left(\mathbf{X}(\widetilde{T}) - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, \widetilde{T})}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, \widetilde{T})} \right) \right] = \int_0^{\tau_2} \mathbf{s}^{(1)}(t) dt - \int_0^{\tau_2} \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)} \mathbf{s}^{(0)}(t) dt$$

In addition to (D1), we consider the following technical assumptions.

Assumptions.

(D2) The time-dependent covariates $\mathbf{X}_i(t)$ have bounded total variation such that $\|\mathbf{X}_i(0)\|_1 + \int_0^{\tau_2} \|d\mathbf{X}_i(t)\|_1 \leqslant B$ for a fixed constant $B > 0$.

(D3) $P(L = 1|Q = 0, \mathbf{Z}_1) = \pi_{\boldsymbol{\gamma}_0}(\mathbf{Z}_1) \geqslant \delta > 0$ for all possible values of $\mathbf{Z}_1$.

(D4) The failure time and censoring time satisfy

$$P(T \geqslant s|C_1, C_2, \bar{\mathbf{X}}(s)) = P(T \geqslant s|\bar{\mathbf{X}}(s))$$

for $s \in [0, \tau_2]$ and $P(\widetilde{T} \geqslant \tau_2) > 0$.

(D5) $\boldsymbol{\Sigma}_0 = \int_0^{\tau_2} \left\{ \frac{\mathbf{s}^{(2)}(\boldsymbol{\beta}_0^*,t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}_0^*,t)} - \left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}_0^*,t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}_0^*,t)} \right)^{\otimes 2} \right\} \mathbf{s}^{(0)}(t)dt$ is positive definite.

(D2) assumes that time-dependent covariates have bounded variation (Bilias et al., 1997). (D3) is a standard positivity assumption for IPW type approach. (D4) is an independent censoring assumption (Bilias et al., 1997) and basically requires that a positive fraction of participants are still at-risk after the end of the observational dataset. (D5) is a standard assumption for Cox models (Andersen and Gill, 1982; Lin and Wei, 1989) that ensures the uniqueness of $\boldsymbol{\beta}_0^*$. Now we present the consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$.

Theorem 2.3.2 (Asymptotic results of $\widehat{\boldsymbol{\beta}}_n$)**.** *Let $\widehat{\boldsymbol{\beta}}_n$ be the solution to the equation $\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = 0$. Under assumptions (D1) - (D5), we have $\widehat{\boldsymbol{\beta}}_n \to_p \boldsymbol{\beta}_0^*$ and*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0^*) \to_d N(0, \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_U\boldsymbol{\Sigma}_0^{-1})$$

and the form of $\boldsymbol{\Sigma}_U$ can be found in Theorem A.1.1 (supplementary material).

Our proof builds on the convergence results for the underlying IPLW emprical process and we give the relevant results for the IPLW process in Appendix A.2. In particular, we proved the Glivenko-Cantelli property of the IPLW empirical process and adopt the strategy in (Andersen and Gill, 1982) for the consistency proof. Next we follow Lin and Wei (1989) to derive the asymptotic linear form for our IPLW partial likelihood estimator and we further establish the weak convergence results of the IPLW empirical process to prove the asymptotic normality.

Remark 2.3.3. *We further consider the augmented inverse probability of linkage weighting (AIPLW) estimator in Appendix A.4. We give the augmented estimating equation and*

prove that AIPLW estimator is "doubly"-robust when either the linkage probability is consistently estimated or three regression functions are consistently estimated. The limitation of the AIPLW estimator is that we need to consistently estimate three regression functions. These three regression functions are themselves variational dependent and congenial parametric modeling can be difficult for all three functions. On the other hand, nonparametric estimation technique does not have the model congeniality problem, but suffers from the curse of dimensionality when there are a large number of covariates. For these reasons, we decide to not implement this "doubly"-robust estimator.

2.4 Simulation

We now compare the performances of our proposed IPLW method with several other methods, including complete-case analysis (CC), complete-case analysis plus (CC+), Non-linked as censored (NLAC) and the oracle method. The oracle method assumes that all participants in the clinical trial are linked to the observational follow-up dataset. We first revisit the motivating example in section 2.2.3. Figure 2.3 shows that our proposed IPLW method gives both consistent estimates and correct coverages for the 95% confidence intervals. On the other hand, CC+, CC and NLAC all give below nominal coverages and inconsistent estimates.

We next perform a more comprehensive set of simulations. We consider the following data generation settings. The hazard function is

$$\lambda(t|\mathbf{X}(t)) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3(t) \times X_1)$$

and $\boldsymbol{\beta}_0 = (\beta_1, \beta_2, \beta_3) = (-\ln(4), \ln(1.5), 0.5)$. X_1 follows a Bernoulli distribution with probability 0.5, X_2 follows a normal distribution with mean and variance both being 1 and $X_3 = I(t \geqslant \tau_1)$ where τ_1 is the end time of clinical trial. If we treat X_1 as the variable for treatment assignment, β_3 now represents the difference between the effect after and before τ_1. This is the Cox model with a change point at τ_1 (2.2). The baseline hazard function is $\lambda_0(t) = 0.06$. C_1 is exponentially distributed with rate $0.01X_1 + 0.03$ and the censoring time

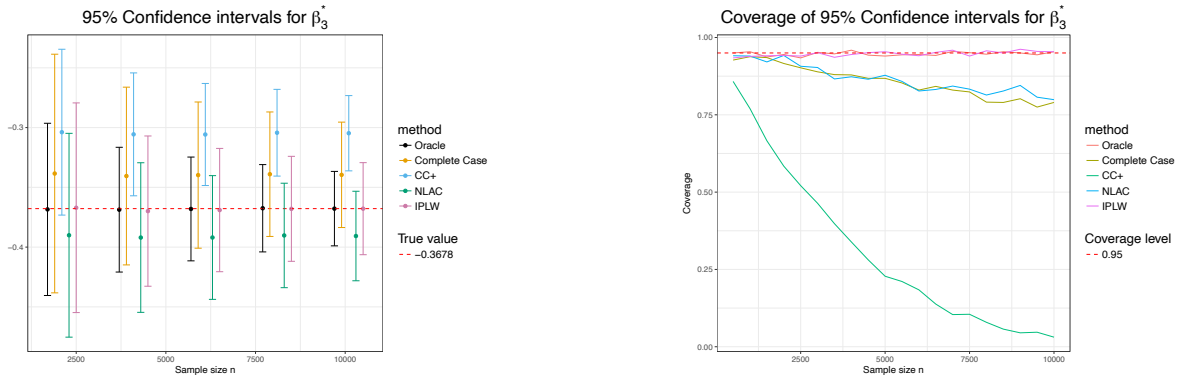

Figure 2.3: Revisiting the motivating example in section 2.2.3. Our approach and the oracle approach are the only methods leading to a valid confidence interval.

C_2 is set as C_1 plus an exponential random variable with rate $0.05X_1 + 0.03$. Further, we set $\tau_1 = 5$ and $\tau_2 = 16$. Three linkage mechanisms are considered as follows:

(1) $P(L = 1) = 0.5$ and L is independent of all other variables. This is the linking completely at random (LCAR) case.

(2)

$$\log\left\{P(L=1|\mathbf{X},Q=0,\widetilde{T},\Delta)/P(L=0|\mathbf{X},Q=0,\widetilde{T},\Delta)\right\} = -0.25 + 0.5X_1 + 0.5X_2$$

and $P(L = 1|Q = 1) = 0.5$. Thus the linkage satisfies CLAR assumption.

(3)

$$\begin{aligned}&\log\left\{P(L=1|\mathbf{X},Q=0,\widetilde{T},\Delta)/P(L=0|\mathbf{X},Q=0,\widetilde{T},\Delta)\right\}\\ &= -0.25 + 0.5X_1 + 0.5X_2 - 0.01\widetilde{T} - 0.01\Delta\end{aligned}$$

and $P(L = 1|Q = 1) = 0.5$. This is the linking not at random (LNAR($\widetilde{T}$)) case.

Both mechanisms (1) and (2) satisfy our CLAR assumption. Mechanism (2) shows that under the CLAR assumption, data linkage can still depend on the survival outcomes through the

in-trial censoring indicator Q. Mechanism (3) slightly violates the CLAR assumption and serves as a case for sensitivity analysis.

Table 2.1: Simulation results when Cox model is correctly specified when $n = 2000$.

		Bias (Mean SE)			Coverage of 95% CI		
Mechanism	Method	β_1	β_2	β_3	β_1	β_2	β_3
LCAR	Oracle	-0.00 (0.108)	0.00 (0.033)	-0.00 (0.140)	0.94	0.95	0.96
	CC	-0.00 (0.153)	0.00 (0.046)	-0.01 (0.199)	0.95	0.96	0.96
	CC+	0.13 (0.110)	-0.03 (0.038)	-0.13 (0.168)	0.77†	0.86†	0.89†
	NLAC	-0.00 (0.110)	0.00 (0.038)	-0.01 (0.168)	0.94	0.95	0.95
	IPLW	-0.00 (0.114)	0.00 (0.041)	-0.01 (0.170)	0.94	0.95	0.95
CLAR	Oracle	-0.00 (0.108)	0.00 (0.033)	-0.00 (0.140)	0.94	0.95	0.96
	CC	-0.18 (0.151)	-0.06 (0.045)	0.19 (0.188)	0.79†	0.75†	0.83†
	CC+	-0.07 (0.109)	-0.10 (0.038)	0.09 (0.157)	0.91	0.23†	0.91
	NLAC	-0.00 (0.109)	0.00 (0.037)	-0.00 (0.157)	0.94	0.95	0.95
	IPLW	-0.00 (0.110)	0.00 (0.040)	-0.00 (0.159)	0.94	0.95	0.96
LNAR ($\widetilde{T}$)	Oracle	-0.00 (0.108)	0.00 (0.033)	-0.00 (0.140)	0.94	0.95	0.96
	CC	-0.19 (0.151)	-0.07 (0.045)	0.20 (0.189)	0.79†	0.69†	0.84†
	CC+	0.07 (0.109)	-0.11 (0.038)	0.09 (0.158)	0.92	0.16†	0.92
	NLAC	-0.00 (0.109)	0.00 (0.037)	-0.01 (0.158)	0.94	0.95	0.95
	IPLW	-0.00 (0.111)	-0.00 (0.041)	-0.01 (0.160)	0.94	0.95	0.95
LNAR (C_2)	Oracle	-0.00 (0.108)	0.00 (0.033)	-0.00 (0.140)	0.94	0.95	0.96
	CC	-0.31 (0.154)	-0.16 (0.055)	0.33 (0.229)	0.47†	0.16†	0.70†
	CC+	-0.12 (0.111)	-0.22 (0.043)	0.14 (0.202)	0.83†	0.00†	0.88†
	NLAC	-0.00 (0.110)	0.00 (0.042)	-0.02 (0.202)	0.93	0.95	0.95
	IPLW	-0.00 (0.117)	0.01 (0.057)	-0.02 (0.211)	0.93	0.94	0.94

We use † to highlight settings with coverage below 90%.

Table 2.2: Simulation results for when Cox model is misspecified and $n = 2000$.

		Bias (Mean SE)		Coverage of 95% CI	
Mechanism	Method	β_1	β_2	β_1	β_2
LCAR	Oracle	-0.00 (0.068)	0.00 (0.033)	0.93	0.94
	CC	-0.00 (0.096)	0.00 (0.046)	0.93	0.96
	CC+	-0.00 (0.081)	-0.03 (0.038)	0.93	$0.84^{\dagger}$
	NLAC	-0.08 (0.081)	0.00 (0.038)	$0.82^{\dagger}$	0.95
	IPLW	-0.00 (0.087)	0.00 (0.042)	0.94	0.95
CLAR	Oracle	-0.00 (0.068)	0.00 (0.033)	0.93	0.94
	CC	-0.04 (0.087)	-0.06 (0.045)	0.92	$0.76^{\dagger}$
	CC+	-0.08 (0.075)	-0.10 (0.038)	$0.81^{\dagger}$	$0.25^{\dagger}$
	NLAC	-0.05 (0.076)	0.00 (0.037)	$0.89^{\dagger}$	0.95
	IPLW	-0.00 (0.079)	0.00 (0.040)	0.93	0.95
LNAR($\widetilde{T}$)	Oracle	-0.00 (0.068)	0.00 (0.033)	0.93	0.94
	CC	-0.05 (0.088)	-0.07 (0.046)	0.91	$0.70^{\dagger}$
	CC+	-0.09 (0.076)	-0.11 (0.038)	$0.78^{\dagger}$	$0.17^{\dagger}$
	NLAC	-0.06 (0.077)	-0.00 (0.038)	$0.88^{\dagger}$	0.94
	IPLW	-0.00 (0.080)	-0.00 (0.041)	0.93	0.95

We use † to highlight settings with coverage below 90%.

We consider sample sizes $n = 500, 1,000, \ldots, 10,000$ and we generate 1,000 samples for each simulation setting. We fit two Cox regressions. The first Cox regression is fitted with covariates $X_1, X_2, X_3(t) \times X_1$ and the second Cox regression is fitted with X_1 and X_2 only. Thus, Cox regression is correctly specified for the first regression and mis-specified for the second regression. For the misspecified case, $\boldsymbol{\beta}_0^*$ is estimated with the oracle method by

computing the averages of 1,000 parameter estimates with sample size $n = 10,000$. The mis-specified regression omits the time-dependent covariate $X_3(t) \times X_1$ and thus ignores the change-point at τ_1 for the effect. The corresponding parameter β_1^* for X_1 can be interpreted as an averaging effect for the entire follow-up period and is equal to -1.10, between the clinical trial effect $-\ln(4) \approx -1.39$ and the observational follow-up effect $-\ln(4) + 0.5 \approx -0.89$.

We use the robust variance estimate (Lin and Wei, 1989) when Cox model is misspecified for all methods other than the IPLW method. For the IPLW method, the variance estimate is automatically robust when Cox model is misspecified. When Cox model is correctly specified, one additional mechanism for linkage is considered as

(4)

$$\log\{P(L=1|\mathbf{X},Q=0,C_2,\Delta)/P(L=0|\mathbf{X},Q=0,C_2,\Delta)\} = \\ -0.25 + 0.5X_1 + 0.5X_2 - 0.1C_2 - 0.1\Delta$$

and $P(L=1|Q=1) = 0.5$. We call this linkage mechanism LNAR(C_2).

Mechanism (4) is a more serious violation of the CLAR assumption and linkage now depends on the unobserved censoring time C_2. As discussed in section 2.6, NLAC should still work under this linkage mechanism. The percentages of samples that are not linked and censored in the clinical trial are approximately 39%, 30%, 32%, 51% for these four mechanisms.

Simulation results are reported in Tables 2.1 - 2.2. In the table, bias is the difference of the average of 1,000 parameter estimates and the true parameter value. Mean standard error (SE) is the average of 1,000 SE estimates. CI stands for confidence interval. We first discuss the results when Cox model is correctly specified. When linkage satisfies LCAR, all methods give consistent estimates and correct coverages for the 95% confidence intervals except CC+. Oracle method gives the smallest variance estimates as each participant is linked. NLAC gives the second smallest variance estimates. CC can be viewed as an IPW method with known probability as the weights and it has the largest variance estimates among all methods. Our proposed IPLW method gives smaller variance estimates than CC for two reasons. First,

IPLW method uses more data than CC; second, IPLW method uses estimated weights, which is known to be more efficient than IPW method with known probability as weights.

When linkage satisfies CLAR but not LCAR, only oracle method, NLAC and our proposed IPLW method give consistent estimate. When CLAR is slightly violated, CC, CC+ all obtain severely biased estimates and confidence intervals with less than nominal coverages. NLAC and our proposed IPLW method still perform relatively well in this case. When linkage depends on the censoring time in observational follow-up C_2, NLAC gives consistent estimates and correct coverage as expected. For this particular simulation setting, our proposed IPLW method also works pretty well.

Next, we discuss the simulation results when Cox model is misspecified. When linkage satisfies LCAR, CC+ and NLAC give inconsistent estimates of the parameters and do not achieve nominal coverage for 95% confidence intervals. It is expected that CC would perform well in this case as discussed in section 2.2.3. Further, NLAC always obtains severely negatively biased estimate of β_1^*, the averaging effect. Our proposed IPLW method again obtains smaller variance estimates than CC as more data are fitted and estimated weights improve efficiency. When the linkage satisfies CLAR but not LCAR, only oracle method and IPLW method give consistent estimates and correct coverages. When CLAR is slightly violated, IPLW approach performs best among all methods other than the oracle method.

2.5 SWOG study

We apply the proposed IPLW method to the SWOG study that links medicare claims data to the PCPT data (Unger et al., 2018). The PCPT randomly assigned 18,880 eligible men from 1993 to 1997 to finasteride or placebo daily for seven years. PCPT clinical records are linked to participants' medicare claims data according to common social security number, sex and date of birth. Medicare claims are available from 1999 to 2011. The linkage enables PC to be identified by both clinical records and medicare claims. 14,176 (75.1%) participants

were linked to medicare claims (finasteride = 7069; placebo = 7107)[3]. The median time from treatment random assignment to the end of the linked trial medicare dataset was 16 years. We are interested in studying the effect of treatment finasteride on the time to diagnosis of PC. Death is treated as censoring.

Of the 14,176 participants with a link to the medicare, 2,037 have a gap between the end of SWOG trial and the start of medicare claims. The median length of the gap was 1.6 years. We exclude those participants with a gap. We fit Cox regression with covariates including the prostate-specific antigen (PSA) level at study entry, race, body mass index at study entry, first degree family history of prostate cancer, age at baseline. Additional covariates were included for logistic regression modeling linkage: participants' education level, marital status, employment status, type of jobs. We further remove participants with any missing covariates and we have 16,518 participants left in the study.

Following the studies in Unger et al. (2018), Cox regressions with two change points at 6.5 and 7.5 years are fitted to account for potential differing effects within critical periods of follow-up. We compared the results of CC, CC+, NLAC and our proposed IPLW method in table 2.3. Table 3.4 contains the parameter estimates and 95% confidence intervals for treatment Finasteride in different time periods[4]. Overall, the results do not differ much between all four methods based on the 95% confidence intervals. A key reason might be that the linkage rate was high for the original study (Unger et al., 2018) as 75% of the participants were linked. Further, they examined potential health care utilization differences by arm and other potential biases in Unger et al. (2018) and found no evidence of strong differences. This suggests that linkages might be following a LCAR mechanism. We also obtained robust variance estimates (Lin and Wei, 1989) and the corresponding confidence intervals. The results are very similar to the nonrobust ones. In summary, finasteride arm participants had a 30% decrease in the hazard ratio of prostate cancer (hazard ratio (HR) = 0.70, 95% confidence intervals (CI) = 0.61 - 0.80) during the first 6.5 years. The effect

[3]See Unger et al. (2018) for details on linkage criteria

[4]on the exponential level

of finasteride is strongest between 6.5 - 7.5 years (HR = 0.67, 95% CI = 0.60 - 0.75). The long-term effect of finasteride after the 7.5 years does not seem to increase the risk of PC (HR = 1.11, 95% CI = 0.95 - 1.30). It is worth noting that CC, CC+ and NLAC obtain more similar long-term effects estimates compared to our proposed IPLW methods. We further fit a Cox regression without any change points and the results are in table 2.4. The results again do not differ too much between all four methods, despite the fact that CC, CC+ and NLAC share more similar results compared to our IPLW method. In summary, the long-term effect of finasteride, now estimating the averaging effect over the entire follow-up period, is still beneficial (HR = 0.79, 95% CI = 0.73 - 0.85).

Table 2.3: SWOG study long-term effect estimation with two change points

Methods	Finasteride (0 - 6.5 years)	Finasteride (6.5 - 7.5 years)	Finasteride (7.5 years+)
IPLW	0.696 (0.607 - 0.797)	0.670 (0.599 - 0.749)	1.113 (0.951 - 1.303)
CC	0.683 (0.587 - 0.795)	0.662 (0.586 - 0.747)	1.087 (0.933 - 1.265)
CC+	0.699 (0.610 - 0.801)	0.663 (0.594 - 0.740)	1.086 (0.933 - 1.265)
NLAC	0.697 (0.608 - 0.798)	0.668 (0.600 - 0.744)	1.079 (0.927 - 1.257)

Table 2.4: SWOG study long-term effect estimation with no change points

Methods	Finasteride
IPLW	0.790 (0.732 - 0.853)
CC	0.767 (0.708 - 0.831)
CC+	0.758 (0.704 - 0.816)
NLAC	0.757 (0.703 - 0.814)

2.6 Comparison of NLAC and IPLW

Now we give a detailed comparison between NLAC and IPLW method. For notational simplicity, we only present the results with time-independent covariates. When Cox model is correctly specified, NLAC gives consistent estimates as long as censoring time C is independent of T given $\mathbf{X}$. Recall the censoring time is modified as

$$C_{\text{NLAC}} = L\max(C_1, C_2) + (1-L)C_1 = L[\max(C_1, C_2) - C_1] + C_1$$

with NLAC. The independent censoring assumption holds for C_{NLAC} if

$$\textbf{(N1)}\ L \perp\!\!\!\perp T|\mathbf{X}, C_1, C_2$$

holds since

$$\textbf{(N1)} + (C_1, C_2) \perp\!\!\!\perp T|\mathbf{X} \Rightarrow (L, C_1, C_2) \perp\!\!\!\perp T|\mathbf{X}.$$

Thus, $(L, C_1, C_2) \perp\!\!\!\perp T|\mathbf{X}$ implies that $C_{\text{NLAC}} \perp\!\!\!\perp T|\mathbf{X}$. Further, we study how NLAC works under the CLAR assumption. First, we can the CLAR assumption as

$$\textbf{(N2)}\ L \perp\!\!\!\perp (\widetilde{T}, \Delta)|\mathbf{X}, Q = 0, C_1$$

given C_1 is always observed when $Q = 0$ and one sufficient assumption for **(N2)** is

$$\textbf{(N3)}\ L \perp\!\!\!\perp (T, C_2)|(\mathbf{X}, Q = 0, C_1).$$

With a bit abuse of notation, we also call assumption **(N3)** the CLAR assumption (even though **(N3)** and **(N2)** are logically different from CLAR). Next, we have the following proposition.

Proposition 2.6.1. *When Cox model is correctly specified and* $(C_1, C_2) \perp\!\!\!\perp T|\mathbf{X}$*, if the following assumption*

$$\textbf{(N4)}\ L \perp\!\!\!\perp T|(\mathbf{X}, Q = 0, C_1, C_2)$$

holds, NLAC provides consistent estimates for $\boldsymbol{\beta}_0$*.*

The proof is given in Appendix A.5. By the weak union property of conditional independence, we have that

$$\textbf{(N3)} \Rightarrow \textbf{(N4)}$$

On the other hand, using the contraction property of conditional independence, we further have

$$\textbf{(N4)} + L \perp\!\!\!\perp C_2 | \mathbf{X}, Q = 0, C_1 \Rightarrow \textbf{(N3)}$$

Thus, to conclude, NLAC works under a slightly weaker assumption than CLAR **(N3)** in that the linkage can further depend on the potentially missing C_2. On the other hand, when Cox model is mis-specified, the parameter of interest $\boldsymbol{\beta}_0^*$ now depends on the actual distribution of the censoring time $C = \max(C_1, C_2)$ and NLAC always gives inconsistent parameter estimates of $\boldsymbol{\beta}_0^*$ since the distribution of the censoring time is modified. In contrast, our proposed IPLW method still gives consistent estimates under the CLAR assumption (A1).

2.7 Discussion

In this chapter, we consider the problem of long-term effect estimation by fitting a Cox model to a partially linked dataset. We propose a novel CLAR assumption that allows us to construct an elegant IPLW estimator that consistently estimates the underlying parameters as if all participants are linked. There have been a limited number of studies on incomplete linkages, other than Kim and Chambers (2012), but their focus is on linear regression with probabilistic record linkage. While Baldi et al. (2010) have discussed potential biases caused by incomplete linkages for Cox regression with simulation studies, no theoretical analysis has been conducted. In contrast, we consider the problem when data is linked by unique identifiers and thus we do not need account for incorrect linkages. This allows us to develop rigorous asymptotic theories for our proposed estimators and also compare with some other alternative methods. Here we point out some possible future directions.

- **Interval Censoring.** We have made the "no gap" assumption in this chapter to focus on the right censoring problem for simplicity. However, in practice, it is possible

that there might be gaps between the clinical trial and observational follow-up dataset. Thus, to fully deal with the problem, we need to extend our current procedure to the interval censoring case as mentioned in Appendix A.6. Saegusa and Wellner (2013) has studied the problem of two-phase sampling for Cox models under interval censoring. Generalizing their techniques to the current linkage problem remains an open question.

- **Beyond CLAR and sensitivity analysis.** CLAR may not hold in certain situations. For instance, if the data being linked is from another study, in which the time to event variable T may influence the chance that someone participates, then (A1) will no longer be true. In this case, we may need to model the linkage probability that depends on T, which could be seen as a sensitivity analysis (Little et al., 2012; Little and Rubin, 2019) on perturbing assumption (A1). How to analyze the data in this case is left as a future work.

- **Missing covariates.** Another direction that we will be exploring is the case of missing covariates (Tsiatis, 2007). Missing covariates is a common issue in medical research. When part of $\mathbf{X}$ is missing, CLAR will no longer be enough to identify the underlying parameter since the linkage probability cannot be computed for every individual. In this case, we have to impose additional assumptions on the missingness of $\mathbf{X}$. However, such assumption has to be carefully chosen so that it will not conflict with the assumption on the linkage.

Chapter 3

HANDLING NONMONOTONE MISSING DATA WITH AVAILABLE COMPLETE-CASE MISSING VALUE ASSUMPTIONS

3.1 Introduction

Missing data problems are very common in scientific research (Molenberghs et al., 2014; Little and Rubin, 2019). Based on the missing/response patterns, these problems can be categorized into monotone and nonmonotone missing data problems. For monotone missing data, variables subject to missing are ordered and if one variable is missing, all subsequent variables are missing. This occurs when individuals drop out of a study, which is common in longitudinal studies (Diggle et al., 2002). Nonmonotone missingness refers to the case when no such ordering exists (Molenberghs et al., 2014; Little and Rubin, 2019). For example, a participant might drop out and later return to a study. Nonmonotone missingness may also occur for regression analysis when outcomes and predictors are missing under arbitrary patterns.

Handling nonmonotone missing data is a very challenging task even if we assume missing-at-random (MAR) (Robins and Gill, 1997; Sun and Tchetgen Tchetgen, 2018). Inverse probability weighting (IPW) estimator for nonmonotone missing data may also be unstable under MAR (Sun and Tchetgen Tchetgen, 2018). Further, Robins and Gill (1997) and Vansteelandt et al. (2007) have argued that the MAR restriction should not be expected to hold in nonmonotone missing data.

In this chapter, we are interested in dealing with nonmonotone missing data that are missing-not-at-random (MNAR). Our study is motivated by an electronic health records (EHRs) dataset that contains longitudinal information of diabetes patients. For patients

with diabetes, one important variable is the glycated hemoglobin (HbA1c) measurement and a controlled HbA1c level ($\leqslant 7\%$) is known to reduce the risk of microvascular complications. However, EHR data also poses significant challenges. EHR data are incomplete as a patient's information is recorded only if and when they visit a clinic. This naturally leads to nonmonotone missing data when a patient reappeared after one or more missed visits. Another complication is that the missing patterns of HbA1c are associated with the underlying HbA1c levels. For example, sicker patients with higher HbA1c levels are likely to visit clinics often and thus have fewer missing values, while healthier patients are likely to miss visits and thus have more missing values. This suggests that the HbA1c missing mechanism is MNAR. Thus, we have nonmonotone and MNAR data for the HbA1c measurements.

The diabetes EHR dataset contains 8663 patients who were enrolled from 2003 to 2013, and who were followed up every 3 months until the 4th quarter of 2013. Thus, the longest follow up time is 11 years (44 quarters). For the purpose of this chapter, we will focus on first-year's data and define Y_i as the HbA1c measurement for the i-th quarter with $i = 0, 1, \ldots, 4$ and Y_0 as the baseline measurement. There are three main questions we would like to address:

- **Q1. Single variable of interest.** Given first-year's data $(Y_0, \cdots, Y_4)$, we are interested in estimating the mean HbA1c levels at the 4-th quarter, i.e., $\mathbb{E}[Y_4]$.

- **Q2. Multiple variables of interest: summary measures.** Given first-year's data, we want to estimate the probability that a patient successfully controls the HbA1c levels below 7% for the 3rd and 4th quarters, i.e., $P(Y_3 \leqslant 7\%, Y_4 \leqslant 7\%)$. Further, we are also interested in estimating the averages of the HbA1c levels for the last two quarters, i.e, $\mathbb{E}[(Y_3 + Y_4)/2]$.

- **Q3. Multiple variables of interest: marginal parametric model.** Given first-year's data, we want to study the linear relationship between Y_4 and Y_2, Y_3, i.e., we want to estimate the following linear regression model:

$$\mathbb{E}[Y_4|Y_2, Y_3] = \beta_0 + \beta_1 Y_2 + \beta_2 Y_3.$$

Addressing these questions is a non-trivial problem because we have nonmonotone missingness in the data and the missingness is MNAR. Several attemps have been made to handle nonmonotone missing data that is MNAR. One approach is to assume specific parametric models for both the study variables and the missing probability (Troxel et al., 1998a,b; Ibrahim et al., 2001). Another approach is the no self-censoring or itemwise conditionally independent nonresponse restriction (Shpitser, 2016; Sadinle and Reiter, 2017; Malinsky et al., 2021) and a variant of this idea is the causal graph approach (Nabi et al., 2020; Mohan and Pearl, 2021). Robins and Gill (1997) proposed the group permutation model and Zhou et al. (2010) proposed the block conditional MAR model. Little (1993) and Tchetgen et al. (2018) considered the complete-case missing value (CCMV) restriction. Tchetgen et al. (2018) used discrete-choice models to generate a class of MNAR assumptions. Linero (2017) introduced the transformed-observed-data restriction which requires specifying a transformation and it is also a partial identifying restriction. Chen (2022) introduced the idea of a pattern graph to generate further MNAR assumptions. However, all these existing work have limitations and cannot be applied to our problem. The no self-censoring restriction requires that no variable can be a direct cause of its own missingness status, which is unlikely to be true for the diabetes EHR data that we are investigating. Other methods such as the CCMV and pattern graph rely heavily on the size of the complete cases. However, for the first year's data $(Y_0, \cdots, Y_4)$, complete cases only account for 5% of the observations in our dataset.

In this chapter, we introduce a useful identifying assumption called available complete-case missing value (ACCMV) assumption for handling nonmonotone missing data that is MNAR. In practice we often have many variables at hand and only a few of them are of primary interest. We call them *primary variables*. For those *auxiliary variables* that are not of direct interest, they are often correlated with the primary variables and the missing mechanism of the primary variables. Thus they can be used to assist with the estimation for primary variables. For Q1 of the diabetes example, $(Y_0, Y_1, \ldots, Y_3)$ are auxiliary variables and Y_4 is the primary variable. We can use $(Y_0, Y_1, \ldots, Y_3)$ to help with the estimation of $\mathbb{E}[Y_4]$. In such a scenario, the conventional CCMV assumption will require all the variables

$(Y_0, \ldots, Y_4)$ to be fully observed for identification. However, requiring auxiliary variables to be fully observed is a strong condition to identify parameters that only involve the primary variable. Ideally, we should also use those observations with primary variables fully observed and auxiliary variables partially observed for identification.

On a high level, the principle of ACCMV imposes an assumption similar to the CCMV on the primary variables for identification and an assumption similar to the available-case missing value (ACMV) assumption (Molenberghs et al., 1998) on the auxiliary variables to improve the effective sample size. This allows a much larger set of observations to be used for identification. For the diabetes example, close to 48% of the patients have Y_4 observed, while only 5% of the patients have $Y_0, \ldots, Y_4$ fully observed. Thus, CCMV will only use 5% of the observation for identification and ACCMV instead will use 48% of the observations for identification. For this reason, ACCMV is particularly suitable for analyzing datasets with few complete cases if the assumption is plausible.

Outline. In Section 3.2, we introduce the relevant notations. In Section 3.3, we study the case with single primary variable. We show that ACCMV assumption leads to nonparametric identification of the distributions for the primary variable and develop an IPW estimator, regression adjustment estimator and a multiply-robust estimator. In Section 3.4, we extend our analysis to multiple primary variables and study the identification, estimation procedure and efficiency theory. We conduct a case study to investigate the scenario of marginal parametric models in Section 3.5. Section 3.6 studies the problem of sensitivity analysis of the ACCMV assumption. We conduct simulation studies in Section 3.7 and apply our approach to the diabetes dataset in Section 3.8.

3.2 Notations

In our analysis, we divide all the variables into two sets: a set of variables called *primary variables*, denoted as $L \in \mathbb{R}^d$, and another set of variables called *auxillary variables*, denoted as $X \in \mathbb{R}^p$. We are interested in structures involving the primary variables L. The auxillary variables X are not of primary interest and mainly help with the estimation involving the

primary variables. Namely, the parameter of interest is a statistical functional of L and does not involve X. We cannot ignore X in our analysis because X may be related to the missing data mechanism of L. We use $\|\cdot\|$ to denote the l_2 norm such that for a vector $x \in \mathbb{R}^d$, we have $\|x\| = \sqrt{\sum_{i=1}^d x_i^2}$. Further, we use $\|\cdot\|_{L_2(P)}$ to denote the $L_2(P)$ norm as $\|f\|_{L_2(P)} = \left(\int f(x)^2 dP(x)\right)^{1/2}$.

Both L and X are subject to missingness. We use the binary vector $A \in \{0,1\}^d$ to denote the response pattern of L, i.e., $A_j = 1$ if L_j is observed and $R \in \{0,1\}^p$ to denote the response pattern of X, i.e., $R_j = 1$ if X_j is observed. We use the notation $X_r = (X_j : r_j = 1)$ and $L_a = (L_j : a_j = 1)$ to denote the observed parts of X and L under pattern $R = r, A = a$. Let $1_p = (1, 1, \cdots, 1) \in \mathbb{R}^p$ and $1_d = (1, 1, \cdots, 1) \in \mathbb{R}^d$. We use the notation $\bar{r} = 1_p - r$ and $\bar{a} = 1_d - a$ to denote the vector after flipping 0 and 1 in r and a, respectively. The variable $X_{\bar{r}} = (X_j : r_j = 0)$ and $L_{\bar{a}} = (L_j : a_j = 0)$ will then refer to the missing variables under pattern $R = r$ and $A = a$. We further define $R \geqslant r$ if $R_i \geqslant r_i$ for $i = 1, 2, \ldots, p$. For instance, $1010 \geqslant 1000$ but 1010 cannot be compared with 0100.

Take the diabetes EHR data as an example, For question Q1 in Section 3.1, the primary variable is $L = Y_4 \in \mathbb{R}$ and the auxillary variable is $X = (Y_0, \cdots, Y_3) \in \mathbb{R}^4$. Suppose we only observe Y_0, Y_2, Y_4, then this individual would have response patterns $A = 1$ and $R = 1010$. For question Q2, our primary variables is $L = (Y_3, Y_4) \in \mathbb{R}^2$ and the auxillary variables $X = (Y_0, Y_1, Y_2) \in \mathbb{R}^3$. For the individual who we only observe Y_0, Y_2, Y_4, the response pattern is $A = 01$ and $R = 101$. In what follows, we will give concrete examples of what ACCMV assumption stands for in different contexts.

3.3 Single primary variable for ACCMV: estimation and inference

To start with, we consider a simple scenario where we only have one primary variable ($d = 1$), i.e., $L \in \mathbb{R}$ and $A \in \{0,1\}$, and we are interested in estimating the mean functional $\theta = \mathbb{E}(f(L))$ for some known function f. For the diabetes EHR data, this occurs when we are interested in estimating the average value of the HbA1c measurement at the end of the

first year, i.e., $L = Y_4$ and $\theta = \mathbb{E}(Y_4)$. A straightforward calculation shows that

$$\begin{aligned}\theta = \mathbb{E}(f(L)) &= \int f(\ell)p(\ell)d\ell \\ &= \underbrace{\int f(\ell)p(\ell, A=1)d\ell}_{\theta_1} + \sum_r \underbrace{\int f(\ell)p(\ell, x_r, R=r, A=0)dx_r d\ell}_{\theta_{0,r}} \\ &= \theta_1 + \sum_r \theta_{0,r}.\end{aligned}$$

Clearly, θ_1 is identifiable and can be estimated by a simple sample mean, i.e.,

$$\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} f(L_i)I(A_i = 1),$$

so we focus on identifying the second term $\theta_{0,r}$. We can show that

$$\theta_{0,r} = \int f(\ell)p(\ell|x_r, R=r, A=0)p(x_r, R=r, A=0)d\ell dx_r.$$

The quantity $p(x_r, R=r, A=0)$ is identifiable from the data. So the key is to identify the first component $p(\ell|x_r, R=r, A=0)$, which is also known as the extrapolation density.

The conventional CCMV assumption will impose the assumption

$$p(\ell|x_r, R=r, A=0) = p(\ell|x_r, R=1_p, A=1). \tag{3.1}$$

While equation (3.1) identifies the parameter θ, it has a limitation that all the information relies on the complete case $R = 1_p, A = 1$. For the diabetes dataset, only a very small fraction (5%) of the patients have $(Y_0, \ldots, Y_4)$ fully observed. So the CCMV might lead to an unreliable estimate.

The ACCMV is based on the insight that the complete case of L is enough for identifying the parameter of interest and we should be more flexible about the response patterns for the auxillary variables X. Formally, the ACCMV assumption imposes the following assumption:

$$p(\ell|x_r, R=r, A=0) = p(\ell|x_r, R \geqslant r, A=1). \tag{3.2}$$

Namely, to identify L under pattern $R = r, A = 0$, we use any patterns as long as the primary variable L is observed and the same set of auxillary variables X_r are also observed.

The assumption (3.2) allows the use of a much larger set of observations to infer the information in variable L. We can further prove that ACCMV assumption leads to nonparametric identification (Robins et al., 2000) of the marginal distribution $p(\ell, a)$ and this assumption will not conflict with the observed data.

Proposition 3.3.1. *Under the ACCMV assumption in equation* (3.2)*, $p(\ell, a)$ is nonparametrically identified.*

It is immediate from Proposition 3.3.1 that $p(\ell)$ is identifiable under the ACCMV assumption.

Example 3.3.2. *Consider the example where we have 4 auxiliary variables X_1, X_2, X_3, X_4 and we focus on the pattern $A = 0$ and $R = 1010$. The CCMV will assume that*

$$p(\ell|x_1, x_3, R = 1010, A = 0) = p(\ell|x_1, x_3, R = 1111, A = 1)$$

and the ACCMV will assume that

$$p(\ell|x_1, x_3, R = 1010, A = 0) = p(\ell|x_1, x_3, R \geqslant 1010, A = 1).$$

For CCMV, the extrapolation density is estimated by observations with $R_i = 1111, A_i = 1$ whereas in the ACCMV, the extrapolation density is estimated by observations with $R_i \in \{1010, 1110, 1011, 1111\}, A_i = 1$. Clearly, ACCMV allows us to estimate $p(\ell|x_1, x_3, R = 1010, A = 0)$ with a much larger set of observations, leading to a more reliable estimate if the ACCMV assumption holds.

3.3.1 The IPW estimation

Instead of directly estimating $p(\ell|x_r, R = r, A = 0)$, we now propose an IPW approach to estimate $\theta_{0,r}$.

Lemma 3.3.3. *The ACCMV assumption* (3.2) *can be equivalently written as follows.*

$$\frac{P(R = r, A = 0|X_r, L)}{P(R \geqslant r, A = 1|X_r, L)} = \underbrace{\frac{P(R = r, A = 0|X_r)}{P(R \geqslant r, A = 1|X_r)}}_{=O_r(X_r)}. \tag{3.3}$$

Lemma 3.3.3 suggests that the ACCMV can be expressed as requiring the odds $P(R = r, A = 0|X_r, L)/P(R \geqslant r, A = 1|X_r, L)$ to be independent of the variable L.

An important implication from Lemma 3.3.3 is that the quantity $O_r(X_r)$ is identifiable and we can estimate $O_r(x_r)$ by assuming a parametric model. For example, If we set $O_r(x_r; \alpha_r) = \exp(x_r^T \alpha_r)$, the odds can be estimated by simply fitting a logistic regression with covariates X_r that treats pattern $R = r, A = 0$ as class 1 and patterns $R \geqslant r, A = 1$ as class 0. Let $O_r(x_r; \widehat{\alpha}_r)$ be the estimated version of $O_r(x_r)$, where $\widehat{\alpha}_r$ is the estimated parameter.

Next, with equation (3.3), we can rewrite $\theta_{0,r}$ as an identifiable quantity as follows

$$\begin{aligned}
\theta_{0,r} &= \int f(\ell) p(\ell, x_r, R = r, A = 0) dx_r d\ell \\
&= \int f(\ell) \frac{p(\ell, x_r, R = r, A = 0)}{p(\ell, x_r, R \geqslant r, A = 1)} p(\ell, x_r, R \geqslant r, A = 1) dx_r d\ell \\
&= \int f(\ell) \frac{P(R = r, A = 0|\ell, x_r)}{P(R \geqslant r, A = 1|\ell, x_r)} p(\ell, x_r, R \geqslant r, A = 1) dx_r d\ell \\
&\overset{(3.3)}{=} \int f(\ell) O_r(x_r) p(\ell, x_r, R \geqslant r, A = 1) dx_r d\ell \\
&= \mathbb{E}\left(f(L) O_r(X_r) I(R \geqslant r, A = 1)\right).
\end{aligned} \tag{3.4}$$

This leads to the following IPW estimator:

$$\widehat{\theta}_{0,r,\mathsf{IPW}} = \frac{1}{n} \sum_{i=1}^{n} f(L_i) O_r(X_{i,r}; \widehat{\alpha}_r) I(R_i \geqslant r, A_i = 1).$$

Combining with the estimator $\widehat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} f(L_i) I(A_i = 1)$, our final estimator for θ will be

$$\widehat{\theta}_{\mathsf{IPW}} = \frac{1}{n} \sum_{i=1}^{n} f(L_i) I(A_i = 1) \left[1 + \sum_r O_r(X_{i,r}; \widehat{\alpha}_r) I(R_i \geqslant r)\right] \tag{3.5}$$

The expression in the last equality shows an elegant form–we can express IPW estimator as weighting the complete cases $A_i = 1$ with weight $1 + \sum_r O_r(X_{i,r}; \widehat{\alpha}_r) I(R_i \geqslant r)$ and we have the following asymptotic theory for $\widehat{\theta}_{\mathsf{IPW}}$.

Theorem 3.3.4. *Under the ACCMV assumption in equation* (3.3) *and the assumption that for every* r*,*

$$\sqrt{n}(\widehat{\alpha}_r - \alpha_r^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{r,\alpha_r^*}(X_{i,r}, R_i, A_i) + o_P(1)$$

for some function ψ_{r,α_r^*} *such that* $\mathbb{E}[\psi_{r,\alpha_r^*}(X_r, R, A)] = \vec{0}$, $\mathbb{E}\|\psi_{r,\alpha_r^*}\|^2 < \infty$ *and the true odds* $O_r(x_r) = O_r(x_r; \alpha_r^*)$. *We assume that* $O_r(X_r; \alpha_r)$ *is differentiable with respect to* α_r *and*

$$\mathbb{E}\|\nabla_{\alpha_r} O_r(X_r; \alpha_r) I(R \geqslant r) I(A = 1) f(L)\| < \infty$$
$$\mathbb{E}\|f(L) I(A = 1) O_r(X_r; \alpha_r) I(R \geqslant r)\|^2 < \infty$$

for $\alpha_r \in B(\alpha_r^*, \rho)$ *for some* $\rho > 0$. *Then*

$$\sqrt{n}(\widehat{\theta}_{\mathsf{IPW}} - \theta) \xrightarrow{d} N(0, \sigma^2_{\mathsf{IPW}})$$

for some $\sigma^2_{\mathsf{IPW}} > 0$.

We can compute the variance σ^2_{IPW} either through the influence function or using bootstrap. More specifically, we have

$$\sqrt{n}(\widehat{\theta}_{\mathsf{IPW}} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(X_i, L_i, R_i, A_i; \alpha^*) + o_P(1)$$

and we can estimate σ^2_{IPW} with

$$\widehat{\sigma}^2_{\mathsf{IPW}} = \frac{1}{n} \sum_{i=1}^{n} (\phi(X_i, L_i, R_i, A_i; \widehat{\alpha}) - \bar{\phi})^2$$

where $\bar{\phi} = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i, L_i, R_i, A_i; \widehat{\alpha})$. The form of the influence function ϕ can be found in B.1. In practice, we recommend bootstrap for its simplicity.

Assumptions in Theorem 3.3.4 are mild. The asymptotic linear form of $\widehat{\alpha}_r - \alpha_r^*$ is very common when we use a parametric model and estimate the parameter via the maximum likelihood estimation (MLE). The condition on the gradient of odds is also very mild. For conventional methods such as the logistic regression, this condition holds with covariates that have a bounded second order moment. The condition on the product of $f(L)$ and odds $O_r(X_r; \alpha_r)$ is also mild. This condition is required for $\widehat{\theta}_{\mathsf{IPW}}$ to have a bounded variance. Alternatively we may make the assumption that $O_r(x_r; \alpha_r)$ is bounded by a large constant for any x_r and $\alpha_r \in B(\alpha_r^*, \rho)$. This is very similar to the positivity assumption in the IPW literature.

3.3.2 The regression adjustment estimation

The ACCMV assumption in equation (3.2) leads to the following identification of $\theta_{0,r}$:

$$\begin{aligned}\theta_{0,r} &= \int f(\ell)p(\ell|x_r, R=r, A=0)p(x_r, R=r, A=0)d\ell dx_r \\ &\overset{(3.2)}{=} \int f(\ell)p(\ell|x_r, R\geqslant r, A=1)p(x_r, R=r, A=0)d\ell dx_r \\ &= \int m_{r,0}(x_r)p(x_r, R=r, A=0)dx_r \\ &= \mathbb{E}(m_{r,0}(X_r)I(R=r, A=0)),\end{aligned}$$

where

$$m_{r,0}(x_r) = \mathbb{E}(f(L)|X_r = x_r, R\geqslant r, A=1) \tag{3.6}$$

is the outcome regression model. Thus, we can estimate $\theta_{0,r}$ by imposing a model $m_{r,0}(x_r) = m_{r,0}(x_r;\beta_r)$ and estimate β_r via $\widehat{\beta}_r$ using observations with $R_i \geqslant r, A_i = 1$. For instance, we may regress the response $f(L)$ versus covariate X_r from observations with $R_i \geqslant r, A_i = 1$. Having estimated $\widehat{\beta}_r$, we then construct the estimator

$$\widehat{\theta}_{0,r,\mathsf{RA}} = \frac{1}{n}\sum_{i=1}^{n} m_{r,0}(X_{i,r};\widehat{\beta}_r)I(R_i = r, A_i = 0)$$

and the final estimator for θ will be

$$\widehat{\theta}_{\mathsf{RA}} = \frac{1}{n}\sum_{i=1}^{n}[f(L_i)A_i + m_{R_i,0}(X_{i,R_i};\widehat{\beta}_{R_i})(1-A_i)]. \tag{3.7}$$

and we have the following asymptotic theory for $\widehat{\theta}_{\mathsf{RA}}$.

Theorem 3.3.5. *Under the ACCMV assumption in equation* (3.2) *and assume that for every* r,

$$\sqrt{n}(\widehat{\beta}_r - \beta_r^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{r,\beta_r^*}(L_i, X_{i,r}, R_i, A_i) + o_P(1)$$

for some function ψ_{r,β_r^*} *such that* $\mathbb{E}[\psi_{r,\beta_r^*}] = \vec{0}$, $\mathbb{E}\|\psi_{r,\beta_r^*}\|^2 < \infty$ *and the true regression function is* $m_{r,0}(x_r) = m_{r,0}(x_r;\beta_r^*)$. *Also, we assume that* $m_{r,0}$ *is differentiable in* β_r *and*

$$\mathbb{E}\|\nabla_{\beta_r} m_{r,0}(X_r;\beta_r)I(R=r, A=0)\| < \infty$$

$$\mathbb{E}\|m_{r,0}(X_r;\beta_r)I(R=r, A=0)\|^2 < \infty$$

for $\beta_r \in B(\beta_r^*, \rho)$ *for some* $\rho > 0$. *Then*

$$\sqrt{n}(\widehat{\theta}_{\mathsf{RA}} - \theta) \xrightarrow{d} N(0, \sigma^2_{\mathsf{RA}})$$

for some $\sigma^2_{\mathsf{RA}} > 0$.

Assumptions in Theorem 3.3.5 takes a similar form as the ones in Theorem 3.3.4. These are mild modeling conditions. If we use a least square approach to fit the parameter β and the true parameter indeed solves the least square equation (occurs when the model is correct), then the asymptotic linear form exists. For linear regression, the gradient condition easily holds when the covariates X have bounded second moments.

3.3.3 Semi-parametric theory and multiply-robust estimation

It is known that the IPW and regression adjustment may not lead to an efficient estimator. In this section, we investigate the efficiency theory under the ACCMV assumption. Since $\theta = \theta_1 + \sum_r \theta_{0,r}$ and the first component is directly identifiable, we only need to study the efficiency theory of estimating $\theta_{0,r}$.

Theorem 3.3.6. *Under the ACCMV assumption, the efficient influence function of estimating* $\theta_{0,r}$ *is*

$$(f(L) - m_{r,0}(X_r))O_r(X_r)I(R \geqslant r, A = 1) + m_{r,0}(X_r)I(R = r, A = 0) - \theta_{0,r}.$$

Based on Theorem 3.3.6, the efficient estimator of $\theta_{0,r}$ is

$$\widehat{\theta}_{0,r,\mathsf{MR}} = \frac{1}{n}\sum_{i=1}^{n} \left(f(L_i) - \widehat{m}_{r,0}(X_{i,r})\right) \widehat{O}_r(X_{i,r})I(R_i \geqslant r, A_i = 1) +$$
$$\widehat{m}_{r,0}(X_{i,r})I(R_i = r, A_i = 0),$$

which leads to the estimator

$$\widehat{\theta}_{\mathsf{MR}} = \sum_r \widehat{\theta}_{0,r,\mathsf{MR}} + \widehat{\theta}_1 \tag{3.8}$$

where $\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^n f(L_i)I(A_i = 1)$. The two estimated functions $\widehat{m}_{r,0}$ and $\widehat{O}_r$ are the estimators of the regression function $m_{r,0}$ and odds O_r. We may use the same estimators as in Section 3.3.1 and 3.3.2. We make the following technical assumptions:

Assumptions:

(S1) For each r, $\widehat{O}_r$ is in a Donsker class $\mathcal{F}_r$ and $\widehat{m}_{r,0}$ is in another Donsker class $\mathcal{G}_r$. There exist functions $O_r^*(x_r)$ and $m_{r,0}^*(x_r)$ such that

$$\|\widehat{O}_r - O_r^*\|_{L_2(P)} = o_P(1) \quad \|\widehat{m}_{r,0} - m_{r,0}^*\|_{L_2(P)} = o_P(1)$$

(S2) $f(\ell), m_{r,0}(x_r), O_r(x_r), \widehat{m}_{r,0}(x_r), \widehat{O}_r(x_r)$ are uniformly bounded by a large constant $M > 0$ for all r, x_r, ℓ.

Assumption (S1) states that estimators $\widehat{m}_{r,0}$ and $\widehat{O}_r$ should converge to fixed functions. The Donsker condition is a common condition that controls the complexity of the estimators. Assumption (S2) is a technical condition and can be relaxed by stronger moment conditions on each function. $O_r(x_r)$ being bounded is related to the positivity assumption in the IPW literature and it is sensible to have $\widehat{O}_r$ also bounded as it estimates $O_r(x_r)$. Further, (S2) holds when these functions are smooth and X, L stay in compact sets. The estimator $\widehat{\theta}_{\mathsf{MR}}$ has the following multiply-robust property.

Theorem 3.3.7. *Under the ACCMV assumption, (S1), (S2) and appropriate assumptions for $\mathcal{F}_r$ and $\mathcal{G}_r$ that we define in B.2, the estimator $\widehat{\theta}_{\mathsf{MR}}$ in equation* (3.8) *satisfies the following properties:*

- **Consistency.** $\widehat{\theta}_{\mathsf{MR}} \xrightarrow{P} \theta$ *when*

$$\sum_r \|\widehat{m}_{r,0} - m_{r,0}\|_{L_2(P)} \|\widehat{O}_r - O_r\|_{L_2(P)} = o_P(1).$$

- **Asymptotic normality.** $\sqrt{n}(\widehat{\theta}_{\mathsf{MR}} - \theta) \xrightarrow{d} N(0, \sigma_{\mathsf{eff}}^2)$ *when*

$$\sqrt{n}\sum_r \|\widehat{m}_{r,0} - m_{r,0}\|_{L_2(P)} \|\widehat{O}_r - O_r\|_{L_2(P)} = o_P(1).$$

The quantity σ_{eff}^2 is the efficiency bound.

The first statement in Theorem 3.3.7 states that as long as for each pattern r, either the regression estimator $\widehat{m}_{r,0}$ or the odds estimator $\widehat{O}_r$ is consistent, the estimator $\widehat{\theta}_{\mathsf{MR}}$ will be consistent. This is known as the multiply-robust property. The second statement states that if both nuisance functions (m_r and O_r) are correctly specified and can be estimated sufficiently fast for all patterns r, the final estimator will be asymptotically normal and achieve the efficiency bound. The Donsker conditions might be relaxed if sample splitting is employed for estimation of $\widehat{O}_r$ and $\widehat{m}_{r,0}$.

Further, if we can assume that both $m_{r,0}$ and $O_r(x_r)$ are parametric functions, we are able to obtain asymptotic normality as long as either $m^*_{r,0}(x_r) = m_{r,0}(x_r)$ or $O^*_r(x_r) = O_r(x_r)$ for each r.

Corollary 3.3.8. *Under the ACCMV assumption and assuming that $m_{r,0}(x_r)$ and $O^*_r(x_r)$ are parametric functions for all r. We further assume that*

$$\|\widehat{m}_{r,0}(x_r;\widehat{\beta}_r) - m_{r,0}(x_r;\beta^*_r)\|_{L_2(P)} = o_P(1)$$
$$\|\widehat{O}_r(x_r;\widehat{\alpha}_r) - O_r(x_r;\alpha^*_r)\|_{L_2(P)} = o_P(1)$$

Then if $m_{r,0}(x_r;\beta^*_r) = m_{r,0}(x_r)$ or $O_r(x_r;\alpha^*_r) = O_r(x_r)$ for each r, we have

$$\sqrt{n}(\widehat{\theta}_{\mathsf{MR}} - \theta) \xrightarrow{d} N(0,\sigma^2)$$

When $m_{r,0}(x_r;\beta^*_r) = m_{r,0}(x_r)$ and $O_r(x_r;\alpha^*_r) = O_r(x_r)$ for each r, we have $\sigma^2 = \sigma^2_{\mathsf{eff}}$.

We can either estimate the variance σ^2 through the influence functions or using bootstrap. The form of the influence functions can be found in B.2. In practice, we recommend using bootstrap to compute the confidence intervals for its simplicity.

3.4 Multiple primary variables for ACCMV: estimation and inference

Now we consider the problem when $L \in \mathbb{R}^d$ is multivariate. As mentioned before, this occurs when we are interested in the last two HbA1c measurements for the first year. In this case, we have $L = (Y_3, Y_4)$ and $X = (Y_0, Y_1, Y_2)$. We assume that the parameter of interest

is $\theta = \mathbb{E}(f(L))$ for some known function f. Multiple primary variables also occur in the marginal parametric models, which we will have an in-depth discussion in Section 3.5.

When we have multiple primary variables, the complete-case that identifies the variable L will be $A = 1_d$. Thus, for $a \neq 1_d$, the ACCMV assumption in equation (3.2) will be revised as

$$p(\ell_{\bar{a}}|\ell_a, x_r, A = a, R = r) = p(\ell_{\bar{a}}|\ell_a, x_r, A = 1_d, R \geqslant r), \tag{3.9}$$

which is equivalent to

$$\frac{P(R = r, A = a|x_r, \ell)}{P(R \geqslant r, A = 1_d|x_r, \ell)} = \underbrace{\frac{P(R = r, A = a|x_r, \ell_a)}{P(R \geqslant r, A = 1_d|x_r, \ell_a)}}_{=O_{r,a}(x_r,\ell_a)}. \tag{3.10}$$

Equation (3.10) is the multivariate version of equation (3.3).

Proposition 3.4.1. *Under the ACCMV assumption in equation* (3.9)*, $p(\ell, a)$ is nonparametrically identified for any $a \neq 1_d$.*

Proposition 3.4.1 shows that the ACCMV assumption for multiple primary variables nonparametrically identifies the marginal density $p(\ell)$. So it is an assumption on the missing data without putting any constraints on the observed data. Our goal is to identify θ when our data is a collection of IID random elements $(R_i, A_i, X_{i,R_i}, L_{i,A_i})$ for $i = 1, 2, \cdots, n$.

3.4.1 The IPW estimation

For any function $f(\ell)$, we have

$$\theta = \mathbb{E}(f(L)) = \sum_{r,a} \mathbb{E}\left(f(L)I(A = a, R = r)\right) = \sum_{r,a} \theta_{r,a}$$

When $a = 1_d$, $\theta_{r,a} = \mathbb{E}\left(f(L)I(A = a, R = r)\right)$ is identifiable. When $a \neq 1_d$, through similar derivations as in (3.4), we have

$$\theta_{r,a} \overset{(3.10)}{=} \mathbb{E}(f(L)O_{r,a}(X_r, L_a)I(A = 1_d, R \geqslant r))$$

and the right hand side is clearly identifiable as long as we can estimate $O_{r,a}$.

Moreover, the following equality holds,

$$\begin{aligned}\sum_{r,a\neq 1_d} O_{r,a}(X_r, L_a) I(A = 1_d, R \geqslant r) &= \sum_r I(A = 1_d, R = r) \sum_{\tau \leqslant r, a\neq 1_d} O_{\tau,a}(X_\tau, L_a) \\ &= \sum_r Q_r(X_r, L) I(R = r, A = 1_d),\end{aligned}$$

with

$$Q_r(X_r, L) = \sum_{\tau \leqslant r, a\neq 1_d} O_{\tau,a}(X_\tau, L_a). \tag{3.11}$$

We can then rewrite the above equality as

$$\begin{aligned}\theta = \mathbb{E}(f(L)) &= \sum_{r,a\neq 1_d} \mathbb{E}(f(L) O_{r,a}(X_r, L_a) I(A = 1_d, R \geqslant r)) + \sum_r \mathbb{E}(f(L) I(A = 1_d, R = r)) \\ &= \mathbb{E}\left(f(L) \sum_r [1 + Q_r(X_r, L)] I(R = r, A = 1_d) \right).\end{aligned}$$

The quantity $1 + Q_r(X_r, L)$ behaves like the weight of observation with $A = 1_d, R = r$.

Based on the above analysis, our estimation procedure of θ will be the following three-step approach:

1. **Step 1: estimating individual odds $O_{r,a}$.** We first estimate $\widehat{O}_{r,a}(X_r, L_a)$ for $a \neq 1_d$. This can be done with a simple logistic regression, i.e., $\widehat{O}_{r,a}(X_r, L_a) = \exp(\widehat{\alpha}_{r,a}^T(X_r, L_a))$ where $\widehat{\alpha}_{r,a}$ is estimated by comparing pattern $(R = r, A = a)$ versus $(R \geqslant r, A = 1_d)$ using variables X_r, L_a.

2. **Step 2: computing total weights Q_r.** For each pattern $(R = r, A = 1_d)$, we compute

$$\widehat{Q}_r(X_r, L) = \sum_{\tau \leqslant r} \sum_{a\neq 1_d} \widehat{O}_{\tau,a}(X_\tau, L_a). \tag{3.12}$$

3. **Step 3: applying the IPW approach.** The final estimator is

$$\widehat{\theta}_{\mathsf{IPW}} = \frac{1}{n} \sum_{i=1}^n f(L_i) [\widehat{Q}_{R_i}(X_{i,R_i}, L_i) + 1] I(A_i = 1_d). \tag{3.13}$$

Theorem 3.4.2. *Under the assumption* (3.10) *and assume that for every* r *and* $a \neq 1_d$,

$$\sqrt{n}(\widehat{\alpha}_{r,a} - \alpha^*_{r,a}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{r,a}(X_{i,r}, L_{i,a}, R_i, A_i) + o_P(1)$$

for some function $\psi_{r,a}$ *such that* $\mathbb{E}[\psi_{r,a}] = \vec{0}$ *and* $\mathbb{E}\|\psi_{r,a}\|^2 < \infty$. *The true odds* $O_{r,a}(x_r, \ell_a) = O_{r,a}(x_r, \ell_a; \alpha^*_{r,a})$. *We assume that* $O_{r,a}(x_r, \ell_a; \alpha_{r,a})$ *is differentiable with respect to* $\alpha_{r,a}$ *and*

$$\mathbb{E}\|\nabla_{\alpha_{r,a}} O_{r,a}(X_r, L_a; \alpha_{r,a}) I(R \geqslant r, A = 1_d) f(L)\| < \infty$$
$$\mathbb{E}\|f(L) I(R \geqslant r, A = 1_d) O_{r,a}(X_r, L_a; \alpha^*_{r,a})\|^2 < \infty$$

for $\alpha_{r,a} \in B(\alpha^*_{r,a}, \rho)$ *for some* $\rho > 0$. *Then*

$$\sqrt{n}(\widehat{\theta}_{\mathsf{IPW}} - \theta) \xrightarrow{d} N(0, \sigma^2_{\mathsf{IPW}})$$

for some $\sigma^2_{\mathsf{IPW}} > 0$.

The conditions in Theorem 3.4.2 are very similar to the single primary variable case (Theorem 3.3.4). The difference is that here we have multiple response patterns of L that we need to consider. Again, assuming the logistic regression model (log odds is linear) is correct, then all these assumptions hold whenever X and L have bounded second moments. The proof can be found in B.3 and the variance can be estimated either through the influence function or bootstrap.

3.4.2 The regression adjustment estimation

Similar to the case of single primary variable scenario, we may apply a regression adjustment approach to estimate θ as well. The idea is based on the pattern mixture model formulation in equation (3.9) that links the extrapolation density to an observed density.

Specifically, for $a \neq 1_d$, Equation (3.9) implies that the parameter $\theta_{r,a} = \mathbb{E}(f(L)I(R = r, A = a))$ can be expressed via the following form:

$$\theta_{r,a} \overset{(3.9)}{=} \mathbb{E}(m_{r,a}(X_r, L_a) I(R = r, A = a))$$

where

$$m_{r,a}(X_r, L_a) = \mathbb{E}(f(L)|L_a, X_r, R \geqslant r, A = 1_d), \tag{3.14}$$

is the outcome regression model. As a result, the regression adjustment approach leads to the following two-stage estimator of θ:

1. **Step 1: estimating the outcome regression.** For each r, a with $a \neq 1_d$, we estimate $m_{r,a}(X_r, L_a)$ via an estimator $\widehat{m}_{r,a}(X_r, L_a)$ using observations with $R \geqslant r, A = 1_d$ and variables L, X_r. This can be done by placing a parametric model $m_{r,a}(X_r, L_a; \beta_{r,a})$ and estimate the underlying parameter $\widehat{\beta}_{r,a}$.

2. **Step 2: regression adjustment.** With the estimates from step 1, our final estimate will be
$$\widehat{\theta}_{\mathsf{RA}} = \frac{1}{n}\sum_{i=1}^{n}\left[f(L_i)I(A_i = 1_d) + \widehat{m}_{R_i,A_i}(X_{i,R_i}, L_{A_i})I(A_i \neq 1_d)\right]. \tag{3.15}$$

The regression adjustment estimator can be interpreted as follows. When we have a complete observation of the primary variable ($A_i = 1_d$), we observe $f(L_i)$. When any entries of L is missing, we find a proper model $\widehat{m}_{R,A}$ based on the response pattern in L, together with the response pattern in X, and compute the predicted value of $f(L)$.

Theorem 3.4.3. *Under the assumption of equation* (3.9) *and assume that for every* $r, a \neq 1_d$,

$$\sqrt{n}(\widehat{\beta}_{r,a} - \beta^*_{r,a}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{r,a}(X_{i,r}, L_{i,a}, R_i, A_i) + o_P(1)$$

for some function $\psi_{r,a}$ *such that* $\mathbb{E}[\psi_{r,a}] = 0$ *and* $\mathbb{E}\|\psi_{r,a}\|^2 < \infty$. *Further, assume that the true regression* $m_{r,a}(x_r, \ell_a) = m_{r,a}(x_r, \ell_a; \beta^*_{r,a})$, $m_{r,a}$ *is differentiable in* $\beta_{r,a}$ *and*

$$\mathbb{E}\|\nabla_{\beta_{r,a}} m_{r,a}(X_r, L_a; \beta_{r,a})I(R = r, A = a)\| < \infty$$
$$\mathbb{E}\|m_{r,a}(X_r, L_a; \beta_{r,a})I(R = r, A = a)\|^2 < \infty$$

for $\beta_{r,a} \in B(\beta^*_{r,a}, \rho)$ *for some* $\rho > 0$. *Then*

$$\sqrt{n}(\widehat{\theta}_{\mathsf{RA}} - \theta) \xrightarrow{d} N(0, \sigma^2_{\mathsf{RA}})$$

for some $\sigma^2_{\mathsf{RA}} > 0$.

Conditions in Theorem 3.4.3 is similar to the conditions in Theorem 3.3.5 except that L is multivariate. The modeling conditions are also mild; linear regression models will satisfy them when we have bounded second moments of both X and L. The proof can be found in B.3 and the variance can be computed either based on the influence functions or bootstrap.

3.4.3 Semi-parametric theory and multiply-robust estimation

Both IPW and regression adjustment are known to be inefficient. To improve the efficiency of the estimator, we first derive the efficient influence function of $\theta_{r,a}$.

Theorem 3.4.4. *Under the ACCMV assumption in equation* (3.10)*, the efficient influence function of estimating $\theta_{r,a}$ when $a \neq 1_d$ is*

$$[f(L) - m_{r,a}(X_r, L_a)]O_{r,a}(X_r, L_a)I(R \geqslant r, A = 1_d) + \\ m_{r,a}(X_r, L_a)I(R = r, A = a) - \theta_{r,a}.$$

The above theorem implies that we can construct an efficient estimator using the following approach:

$$\begin{aligned}\widehat{\theta}_{\mathsf{MR}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_{r,a\neq 1_d}\{[f(L_i) - \widehat{m}_{r,a}(X_{i,r})]\widehat{O}_{r,a}(X_{i,r}, L_{i,a})I(R_i \geqslant r, A_i = 1_d) \\ +\widehat{m}_{r,a}(X_{i,r}, L_{i,a})I(R_i = r, A_i = a)\} + f(L_i)I(A_i = 1_d)\Bigg],\end{aligned} \tag{3.16}$$

where $\widehat{O}_{r,a}$ and $\widehat{m}_{r,a}$ are estimators of the odds $O_{r,a}$ in equation (3.10) and the outcome regression $m_{r,a}$ in equation (3.14), respectively. We make the following technical assumptions:

Assumptions:

(M1) For each r and $a \neq 1_d$, $\widehat{O}_{r,a}$ is in a Donsker class $\mathcal{F}_{r,a}$ and $\widehat{m}_{r,a}$ is in a Donsker class $\mathcal{G}_{r,a}$. There exist functions $O^*_{r,a}(x_r, l_a)$ and $m^*_{r,a}(x_r, l_a)$ such that

$$\|\widehat{O}_{r,a} - O^*_{r,a}\|_{L_2(P)} = o_P(1) \quad \|\widehat{m}_{r,a} - m^*_{r,a}\|_{L_2(P)} = o_P(1)$$

(M2) $f(\ell), m_{r,a}(x_r, l_a), O_{r,a}(x_r, l_a), \widehat{m}_{r,a}(x_r, l_a), \widehat{O}_{r,a}(x_r, l_a)$ are uniformly bounded by a large constant $M > 0$ for all x_r, ℓ, r and $a \neq 1_d$.

Assumptions (M1) and (M2) are multivariate versions of (S1) and (S2).

Theorem 3.4.5. *Under the ACCMV assumption* (3.9)*, (M1), (M2) and appropriate assumptions for $\mathcal{F}_{r,a}$ and $\mathcal{G}_{r,a}$ that we define in B.3, the estimator $\widehat{\theta}_{\mathsf{MR}}$ has the following properties:*

- **Consistency.** $\widehat{\theta}_{\mathsf{MR}} \xrightarrow{P} \theta$ *when*

$$\sum_{r,a\neq 1_d} \|\widehat{m}_{r,a} - m_{r,a}\|_{L_2(P)} \|\widehat{O}_{r,a} - O_{r,a}\|_{L_2(P)} = o_P(1).$$

- **Asymptotic normality.** $\sqrt{n}(\widehat{\theta}_{\mathsf{MR}} - \theta) \xrightarrow{d} N(0, \sigma^2_{\mathsf{eff}})$ *when*

$$\sqrt{n} \sum_{r,a\neq 1_d} \|\widehat{m}_{r,a} - m_{r,a}\|_{L_2(P)} \|\widehat{O}_{r,a} - O_{r,a}\|_{L_2(P)} = o_P(1).$$

  The quantity σ^2_{eff} is the efficiency bound.

Theorem 3.4.5 implies that the estimator $\widehat{\theta}_{\mathsf{MR}}$ is multiply-robust in the sense that as long as we have either $m_{r,a}$ or $O_{r,a}$ being consistently estimated for all r and $a \neq 1_d$, the estimator $\widehat{\theta}_{\mathsf{MR}}$ will be consistent. Further it achieves the efficiency bound when the two sets of nuisance models are estimated sufficiently fast.

3.5 Multiple primary variables for ACCMV: marginal parametric model

In practice, we often impose a marginal parametric model over the primary variable L and use the data to estimate the underlying parameter. To start with, we consider two motivating examples.

Example 3.5.1. (Modeling the marginal distribution) *We assume that $L \sim p(\ell; \theta^*)$, where $p(\cdot; \theta)$ is a known parametric distribution such as a multivariate Gaussian, and the goal is to estimate the underlying parameter θ^*. A typical approach to estimate θ^* is the maximum likelihood estimator (MLE). Under usual regularity conditions, the true parameter solves the population score equation:*

$$\theta^* : 0 = \mathbb{E}(s(\theta^*|L)), \qquad s(\theta|\ell) = \nabla_\theta \log p(\ell; \theta).$$

When there is no missingness in L, the MLE is obtained from the following sample score equation:

$$\widehat{\theta}_{\mathsf{MLE}} : 0 = \frac{1}{n}\sum_{i=1}^{n} s(\widehat{\theta}_{\mathsf{MLE}}|L_i).$$

To give a concrete example of this, consider again the one-year diabetes data $Y_0, \cdots, Y_4$. Suppose that we are interested in the joint distribution of the last two visits, i.e., $L = (Y_3, Y_4)$, and we assume that it follows a bivariate Gaussian, i.e., $p(\ell;\theta) = p(y_3, y_4; \mu, \Sigma)$, where $\mu \in \mathbb{R}^2$ is the mean vector and $\Sigma \in \mathbb{R}^{2\times 2}$ is the covariance matrix. Then we can easily estimate μ and Σ using the MLE.

Example 3.5.2. (Modeling the marginal moment restricted model) *It is also very common that the parameter of interest may be a moment restricted model among variables in L. For instance, we may impose a linear model $\mathbb{E}(L_1|L_{-1}) = L_{-1}^T\theta^*$, where L_{-1} is all variables in L except the first one (for simplicity, we ignore the intercept). The parameter of interest is the regression coefficient θ^*. In this case, we often estimate the parameter θ^* by the least squares approach, i.e., at the population level, the parameter θ^* satisfies*

$$\theta^* = \mathsf{argmin}_\theta \mathbb{E}((L_1 - L_{-1}^T\theta^*)^2),$$

or equivalently, the parameter θ^* solves the following equation:

$$\vec{0} = \mathbb{E}(L_{-1}(L_1 - L_{-1}^T\theta^*)).$$

When there is no missingness in L, the least squares estimate solves the following estimating equation

$$\widehat{\theta}_{\mathsf{LS}} : \vec{0} = \frac{1}{n}\sum_{i=1}^{n} L_{i,-1}(L_{i,1} - L_{i,-1}^T\widehat{\theta}_{\mathsf{LS}}).$$

In the diabetes data, if we are interested in the linear relationship among Y_2, Y_3 and Y_4, we can use the model above and treat $Y_4 = L_1$ and $(Y_2, Y_3) = L_{-1}$.

In both examples, we see that the parameter of interest is now defined through a population estimating equation

$$0 = \mathbb{E}(s(\theta^*|L)). \tag{3.17}$$

So we will focus on the case of parameters defined through an estimating equation, and how to obtain a consistent estimate when there are missingness in L based on the ACCMV assumption (3.10).

3.5.1 IPW marginal parametric model

The IPW approach in Section 3.4.1 can be easily adapted to the marginal parametric model. Specifically, the population estimating equation of (3.17) can be written as

$$\begin{aligned}
0 &= \mathbb{E}(s(\theta^*|L)) \\
&\overset{(3.10)}{=} \sum_{r,a\neq 1_d} \mathbb{E}(s(\theta^*|L) O_{r,a}(X_r, L_a) I(A = 1_d, R \geqslant r)) + \sum_r \mathbb{E}(s(\theta^*|L) I(A = 1_d, R = r) \\
&= \mathbb{E}\left(s(\theta^*|L) \left[\sum_{r,a\neq 1_d} O_{r,a}(X_r, L_a) I(A = 1_d, R \geqslant r) + \sum_r I(A = 1_d, R = r) \right] \right) \\
&= \mathbb{E}\left(s(\theta^*|L) I(A = 1_d) \sum_r [Q_r(X_r, L) + 1] I(R = r) \right),
\end{aligned} \tag{3.18}$$

where the weight function Q_r is from equation (3.11).

As a result, the three-step procedure in Section 3.4.1 can be applied here with a mild modification:

1. **Step 1: estimating individual odds $O_{r,a}$.** For r and $a \neq 1_d$, we first estimate $\widehat{O}_{r,a}(X_r, L_a)$. This can be done by a simple logistic regression, i.e., $\widehat{O}_{r,a}(X_r, L_a) = O_{r,a}(X_r, L_a; \widehat{\alpha}_{r,a})$ where $\widehat{\alpha}_{r,a}$ is estimated by comparing pattern $(R = r, A = a)$ versus $(R \geqslant r, A = 1_d)$ using variables X_r, L_a.

2. **Step 2: computing total weights $Q_{r,a}$.** For each pattern $(R = r, A = 1_d)$, we compute its total weight

$$\widehat{Q}_r(X_r, L) = \sum_{\tau \leqslant r} \sum_{a\neq 1_d} \widehat{O}_{\tau,a}(X_\tau, L_a).$$

3. **Step 3: solving the weighted estimating equation.** The final estimator $\widehat{\theta}$ is from

$$\widehat{\theta} : 0 = \sum_{i=1}^{n} s(\theta|L_i)[\widehat{Q}_{R_i}(X_{i,R_i}, L_i) + 1]I(A_i = 1_d). \tag{3.19}$$

The first two steps are the same as Section 3.4.1. We only need to modify the last step by solving a weighted estimating equation. We have the following asymptotic results for $\widehat{\theta}$.

Theorem 3.5.3. *Under assumption* (3.10) *and assume that for every* r *and* $a \neq 1_d$,

$$\sqrt{n}(\widehat{\alpha}_{r,a} - \alpha^*_{r,a}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{r,a}(X_{i,r}, L_{i,a}, R_i, A_i) + o_P(1)$$

for some function $\psi_{r,a}$ *such that* $\mathbb{E}[\psi_{r,a}] = 0$ *and* $\mathbb{E}\|\psi_{r,a}\|^2 < \infty$. *The true odds*

$$O_{r,a}(x_r, \ell_a) = O_{r,a}(x_r, \ell_a; \alpha^*_{r,a}).$$

Next we assume that $O_{r,a}$ *is differentiable in* $\alpha_{r,a}$ *and*

$$\mathbb{E}\|s(\theta^*|L)\nabla_{\alpha_{r,a}} O_{r,a}(X_r, L_a; \alpha_{r,a})I(A = 1_d, R \geqslant r)\| < \infty$$

for $\alpha_{r,a} \in B(\alpha^*_{r,a}, \rho)$ *for some* $\rho > 0$. *Further we assume that*

$$I(\theta^*) = \mathbb{E}\left[\nabla_\theta s(\theta|L)|_{\theta=\theta_0}\left[\sum_{r,a\neq 1_d} O_{r,a}(X_r, L_a; \alpha^*_{r,a})I(A = 1_d, R \geqslant r) + I(A = 1_d)\right]\right]$$

exists and is invertible. Assuming that $\widehat{\theta} \to_p \theta^*$, *we have*

$$\sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow{d} N(0, \Sigma)$$

for some covariance matrix Σ.

3.5.2 Potential problems with regression adjustment

The marginal parametric model has one distinct property from the general case of multiple primary variables: the regression adjustment method and the multiply-robust approach may

be problematic. The main reason is: both regression-adjustment and multiply-robust approach will involve imposing a conditional model on one subset of L conditioned on another subset of L. This procedure implicitly places a model constraint on the distribution of L, which will conflict with the marginal parametric model when it is not designed well. The multiply-robust estimator also suffers from the same problem since it involves a model on the outcome regression. On the other hand, the odds in the IPW approach is a conditional model on the selection odds $P(R = r, A = a|x_r, \ell_a)/P(R \geqslant r, A = 1_d|x_r, \ell_a)$, so it is always compatible with the marginal parametric model. Hence, we recommend using the IPW approach in the case of marginal parametric model.

This phenomenon is similar to the model congeniality problem introduced in Meng (1994). The model congeniality problem refers to the case where the imputation model may not be compatible with the analysis model imposed on the imputed data. An imputation model can be viewed as a Monte Carlo approximation to the regression adjustment method and the marginal parametric model on L is the analysis model in Meng (1994). Thus, the model conflicting problem we encounter when using regression adjustment on a marginal parametric model can be viewed as another form of model congeniality problem.

3.6 Sensitivity analysis via exponential tilting

It is possible that the ACCMV assumption may not be correct or is only approximately correct. The sensitivity analysis (Little et al., 2012) is often conducted to study how the estimate changes when we slightly perturb the underlying missing data assumption.

Here we propose to perform sensitivity analysis of ACCMV via an exponential tilting approach (Kim and Yu, 2011; Shao and Wang, 2016; Zhao et al., 2017). For $a \neq 1_d$, recall that the ACCMV in equation (3.10) requires:

$$\frac{P(R = r, A = a|x_r, \ell)}{P(R \geqslant r, A = 1_d|x_r, \ell)} = \underbrace{\frac{P(R = r, A = a|x_r, \ell_a)}{P(R \geqslant r, A = 1_d|x_r, \ell_a)}}_{=O_{r,a}(x_r,\ell_a)}.$$

In reality, the odds on the left-hand-side of the above equality may depend on the unob-

served value of L. Using the concept of exponential tilting, we propose to perturb assumption (3.10) as follows:

$$\frac{P(R=r, A=a|x_r, \ell)}{P(R \geqslant r, A=1_d|x_r, \ell)} = O_{r,a}(x_r, \ell_a) \cdot \exp(\delta_{\bar{a}}^T \ell_{\bar{a}}), \tag{3.20}$$

where $\delta_{\bar{a}} \in \mathbb{R}^{|\ell_{\bar{a}}|}$ is a given vector that represents the amount of perturbation from the ACCMV assumption. Clearly, when $\delta_{\bar{a}}$ is a zero vector, equation (3.20) reduces to the usual ACCMV assumption.

In practice, we will choose a sensitivity parameter vector $\delta \in \mathbb{R}^d$ first, which implies $\delta_{\bar{a}}$ for every a. Then based on the perturbation (3.20), we compute the modified final estimate. For the IPW estimator, We only need to change

$$\widehat{Q}_r(X_r, L) = \sum_{\tau \leqslant r} \sum_{a \neq 1_d} \widehat{O}_{r,a}(X_\tau, L_a)$$

in equation (3.12) to

$$\widetilde{Q}_r(X_r, L; \delta) = \sum_{\tau \leqslant r} \sum_{a \neq 1_d} \widehat{O}_{r,a}(X_\tau, L_a) \exp(\delta_{\bar{a}}^T L_{\bar{a}})$$

and change the final estimator in equation (3.13) to

$$\widetilde{\theta}_{\mathsf{IPW},\delta} = \frac{1}{\widetilde{n}} \sum_{i=1}^{n} f(L_i)[\widetilde{Q}_{R_i}(X_{i,R_i}, L_i; \delta) + 1] I(A_i = 1_d). \tag{3.21}$$

with $\widetilde{n} = \sum_{i=1}^{n} [\widetilde{Q}_{R,i}(X_{i,R_i}, L_i; \delta) + 1] I(A_i = 1_d)$. Note that $\widehat{O}_{r,a}(X_\tau, L_a)$ is estimated under assumption (3.10).

Under a logistic regression model, the sensitivity parameter in the exponential tilting approach (3.20) has a nice interpretation. Recall that the logistic regression model will model

$$O_{r,a}(x_r, \ell_a) = \exp(\alpha_{r,a}^T(x_r, \ell_a)).$$

Thus, equation (3.20) will become

$$O_{r,a}(x_r, \ell) = \frac{P(R=r, A=a|x_r, \ell)}{P(R \geqslant r, A=1_d|x_r, \ell)} = \exp\left(\alpha_{r,a}^T(x_r, \ell_a) + \delta_{\bar{a}}^T \ell_{\bar{a}}\right).$$

Each δ_j and each element $\alpha_{r,a,j}$ have the same interpretation–they are the coefficient on linear model of the log odds. Consider a specific example that $L = (L_1, L_2) \in \mathbb{R}^2$, $a = 10$ (L_1 is observed) and the coefficient $\alpha_{r,a}$ on L_1 is 2. Then a sensitivity parameter $\delta_{01} = 1$ can be interpreted as the effect of the unobserved variable L_2 on the log odds is half of the estimated effect of the observed variable L_1. Thus, practitioners can use this as a way to think about a feasible range of the sensitivity parameter δ.

3.7 Simulation study

3.7.1 Single Primary Variable

We first consider the case when there is a single primary variable. We have $L = Y_3$ and $X = (Y_1, Y_2)$ and we are interested in estimating $\theta = \mathbb{E}[Y_3]$. Let $|r| = \sum_r r_i$ be the number of observed variables. Next, we generate data as follows:

1. $(L, X_r)|A = 1, R = r \sim N(\mu_{|r|+1}, \Sigma_{|r|+1})$

2. $X_r|A = 0, R = r \sim N(\mu_{|r|}, \Sigma_{|r|})$

with $\mu_1 = 1$, $\mu_2 = (1, -1)^T$, $\mu_3 = (0, -1, -1)^T$ and

$$\Sigma_1 = 1 \ , \ \Sigma_2 = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \text{ and } \Sigma_3 = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix}$$

Further, we assume that $P(A = j, R = r) = 1/8$ for $j = 0, 1$ and $r \in \{00, 01, 10, 11\}$. Note that under ACCMV assumption, $L|X_r, R = r, A = 0$ for $r \in \{00, 01, 10, 11\}$ are also specified given the data generations above.

We first consider estimation using the regression adjustment method. Under ACCMV, we can compute that $\theta = \mathbb{E}[Y_3] = \frac{89}{96}$ and the details are left in B.5. We fit linear regression models

$$m_{r,0}(x_r; \beta_r) = \mathbb{E}[Y_3|X_r = x_r, R \geqslant r, A = 1; \beta_r]$$

for $r \in \{00, 01, 10, 11\}$ and get $\widehat{\beta}_r$. The form of the linear regression models can be found in B.5. Then, we can get the estimates using regression adjustment as

$$\widehat{\theta}_{\mathsf{RA}} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_{i,3}I(A_i = 1) + \sum_r m_{r,0}(X_{i,r}; \widehat{\beta}_r)I(R_i = r, A_i = 0)\right]$$

Next, we consider the IPW estimates. We can compute the odds functions as follows.

$$\begin{aligned}
O_{00} &= \frac{P(A = 0, R = 00)}{P(A = 1, R \geqslant 00)} = \frac{1}{4} \\
O_{10}(y_1) &= \frac{P(A = 0, R = 10|y_1)}{P(A = 1, R \geqslant 10|y_1)} = \frac{1}{2}\exp(2y_1) \\
O_{01}(y_2) &= \frac{P(A = 0, R = 01|y_2)}{P(A = 1, R \geqslant 01|y_2)} = \frac{1}{2}\exp(2y_2) \\
O_{11}(y_1, y_2) &= \frac{P(A = 0, R = 11|y_1, y_2)}{P(A = 1, R = 11|y_1, y_2)} = \exp\left(\frac{8}{3}y_1 - \frac{4}{3}y_2 - \frac{4}{3}\right)
\end{aligned}$$

To get the estimate, we fit a logistic regression model

$$P(R = r, A = 0|X_r, \{R \geqslant r, A = 1\} \cup \{R = r, A = 0\}; \alpha_r) = \frac{O_r(x_r; \alpha_r)}{1 + O_r(x_r; \alpha_r)}$$

for each $r \in \{00, 01, 10, 11\}$ and get $\widehat{\alpha}_r$. Then we can get the estimate using IPW as

$$\widehat{\theta}_{\mathsf{IPW}} = \frac{1}{n}\sum_i Y_{3,i}I(A_i = 1)\left[1 + \sum_r O_r(X_{i,r}; \widehat{\alpha}_r)I(R_i \geqslant r)\right]$$

For the multiply-robust estimator, we can use the linear regression models and logistic regression models that we fitted before. Then for each r, we can get the estimate as

$$\begin{aligned}
\widehat{\theta}_{0,r,\mathsf{MR}} = \frac{1}{n}\sum_{i=1}^{n}&\left(f(L_i) - m_{r,0}(X_{i,r}; \widehat{\beta}_r)\right)O_r(X_{i,r}; \widehat{\alpha}_r)I(R_i \geqslant r, A_i = 1) \\
&+ m_{r,0}(X_{i,r}; \widehat{\beta}_r)I(R_i = r, A_i = 0),
\end{aligned}$$

Our final multiply-robust estimator is then

$$\widehat{\theta}_{\mathsf{MR}} = \sum_r (\widehat{\theta}_{0,r,\mathsf{MR}} + \widehat{\theta}_{1,r}) \tag{3.22}$$

where $\widehat{\theta}_{1,r} = \frac{1}{n}\sum_{i=1}^{n} f(L_i)I(A_i = 1, R_i = r)$.

For the multiply-robust estimator, we first consider the case when all the regression functions and odds functions are correctly specified. Next we consider the case when one of the regression function is misspecified. When $r = 11$, the correct regression model is $m_{11,0}(x_r; \beta_{11}) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2$ and we fit a linear regression model with Y_1 only. We also apply the same model misspecification to the regression adjustment estimator. We further consider the case when one of the odds function is mis-specified. When $r = 11$, the correct odds function is $O_{11}(x_r; \alpha_{11}) = \alpha_0 + \alpha_1 Y_1 + \alpha_2 Y_2$ and we fit a logistic regression with intercept only. Again we apply the same model misspecification to the IPW estimator. Finally we consider the case when both the regression function and the odds function is incorrect. When $r = 11$, we fit a linear regression model with Y_1 only and we fit a logistic regression model with intercept only for the odds function.

Table 3.1 contains the simulation results. We generate 1,000 samples with $n = 2000$. The bias is computed as the difference of the average of 1,000 parameter estimates and the true value of $\mathbb{E}[Y_3]$. Sample standard error (SE) is computed as the standard error of the 1,000 parameter estimates and the mean theoretical SE is computed as the mean of the 1,000 SE estimates. For all three methods, we estimate the SE for the estimators through their corresponding influence functions[1]. Note that bootstrap is an alternative approach to estimate the SE. The sample standard SE reflects the true SE of the estimator and the mean theoretical SE reflects the accuracy of the SE estimated through the influence functions. CI stands for Confidence Interval, RA stands for regression adjustment and MR stands for multiply-robust.

Based on the simulation results, we can see that 95% CI of IPW is undercovering when all the odds function are correctly specified. This is primarily due to the under-estimation of the SE for the IPW estimators in this setup. For this specific data generation setting, the difficulty is that we are estimating the variance of a very heavy-tailed distribution. We can see that the mean theoretical SE is much smaller than the sample SE, which suggests

[1]The actual form of the influen functions can be found in Appendix.

that the estimated SE is much smaller than the true SE. Next, IPW with misspecified odds function leads to much larger bias and even worse coverage for the 95 CI%. We also observe that the mean theoretical SE is much smaller than the sample SE. However, multiply-robust estimator with the same misspecified odds function obtains much better performance. We can see that the bias is very small, the coverage of the 95% CI is very close to the nominal coverage and the difference between mean theoretical SE and the sample SE is also much smaller now. This shows the robustness of the multiply-robust estimator.

Further, regression adjustment achieves nominal coverages and smallest SE estimates when all regression functions are correctly specified. RA with misspecified regression function also leads to large bias and bad coverage for the 95% CI. Again, multiply-robust estimator with the same model misspecification is able to reduce the bias and improve the coverages. We can see that in this case, all multiply-robust estimators underestimate their variances for the same reason as the IPW estimator. However, the coverages of the 95% CI are much better for multiply-robust estimators compared to the IPW estimator. When both the regression and odds function are misspecified, the multiply-robust estimator also obtained biased estimates and bad coverages for the 95% CI. Finally, we also include the results estimating $\mathbb{E}[Y_3]$ using complete-case analysis. It is clear from the simulations that under ACCMV assumption, the data is missing-not-at-random as complete-case analysis is severely biased.

3.7.2 Multiple primary variables

Next we consider the case when we have multiple primary random variables. We take $L = (Y_3, Y_4)$ and $X = (Y_1, Y_2)$. The parameter of interest is $\theta = \mathbb{E}[Y_3 Y_4]$ and this allows the estimation of the $\text{Cov}(Y_3, Y_4)$ given $\mathbb{E}[Y_3]$ and $\mathbb{E}[Y_4]$. For any a, let $|a| = \sum_i a_i$ be the number of observed primary variables. We generate the data as follows.

1. $(L, X_r)|A = 11, R = r \sim N(1_{2+|r|}, \Sigma_{2+|r|})$ for $r \in \{00, 01, 10, 11\}$.

2. $(L_a, X_r)|A = a, R = r \sim N(\mu_{1+|r|}, \Sigma_{1+|r|})$ for any $a \in \{01, 10\}$ and any $r \in \{00, 01, 10, 11\}$.

Table 3.1: Simulations results for estimating $\mathbb{E}[Y_3]$ when $n = 2000$

Methods	Bias	Sample SE	Mean theoretical SE	Coverage of 95% CI
IPW	-0.006	0.216	0.113	0.778
IPW (incorrect)	-0.084	0.174	0.090	0.536
RA	-0.001	0.044	0.043	0.955
RA (incorrect)	0.040	0.045	0.046	0.857
MR (correct)	-0.002	0.105	0.072	0.931
MR (IPW incorrect)	0.000	0.069	0.057	0.939
MR (RA incorrect)	-0.000	0.118	0.074	0.920
MR (Both incorrect)	0.041	0.069	0.058	0.870
Complete Case	-0.178	0.035	0.034	0.001

3. $X_r|A = 00, R = r \sim N(\mu_{|r|}, \Sigma_{|r|})$ for any $r \in \{01, 10, 11\}$.

where $1_d = (1, \ldots, 1)^T \in \mathbb{R}^d$, $\Sigma_d = 1/2I_d + 1/21_d1_d^T$ and

$$\mu_1 = 0.5 \quad \mu_2 = 1_2 \quad \mu_3 = 1_3.$$

Further, we also assume that $P(A = a, R = r) = 1/16$ for $a \in \{00, 01, 10, 11\}$ and $r \in \{00, 01, 10, 11\}$.

We first consider regression adjustment method. Under the ACCMV assumption, we can compute $\theta = \mathbb{E}[Y_3Y_4] = 175/128$ and the details can be found in B.5. Now to get the estimate, we need to fit the following regression models:

$$m_{r,a}(x_r; \beta_{r,a}) = \mathbb{E}[Y_3Y_4|X_r = x_r, R \geqslant r, A = 11; \beta_{r,a}]$$

for $a \neq 11$. Note that equivalently we can fit the following regression models

$$m'_{r,00}(x_r;\beta_{r,00}) = \mathbb{E}[Y_3Y_4|X_r = x_r, R \geqslant r, A = 11;\beta_{r,00}]$$
$$m'_{r,01}(x_r,y_4;\beta_{r,01}) = \mathbb{E}[Y_3|X_r = x_r, Y_4 = y_4, R \geqslant r, A = 11;\beta_{r,01}]$$
$$m'_{r,10}(x_r,y_3;\beta_{r,10}) = \mathbb{E}[Y_4|X_r = x_r, Y_3 = y_3, R \geqslant r, A = 11;\beta_{r,10}]$$

for $r \in \{00, 01, 10, 11\}$ and get $\widehat{\beta}_{r,a}$. The actual form of the regression models can be found in B.5 and we have

$$m_{r,a}(x_r,l_a;\beta_{r,a}) = \begin{cases} m'_{r,a}(x_r,l_a;\beta_{r,a}) & a = 00 \\ m'_{r,a}(x_r,l_a;\beta_{r,a})l_a & a = 01, 10 \end{cases}$$

Then we can get the estimate using regression adjustment as

$$\widehat{\theta}_{\mathsf{RA}} = \frac{1}{n}\sum_{i=1}^{n}[Y_{3,i}Y_{4,i}I(A_i = 11) + \sum_{r,a\neq 11} m_{r,a}(X_{r,i},L_{a,i};\widehat{\beta}_{r,a})I(R_i = r, A_i = a)]$$

Next, for IPW estimation, it is easy to get that for $a \neq 11$, $\log O_{r,a}(X_r, L_a)$ is a linear function of X_r and L_a. Then we can fit a logistic regression model as follows:

$$P(R = r, A = a|x_r, l_a, \{R \geqslant r, A = 11\} \cup \{R = r, A = a\};\alpha_{r,a}) = \frac{O_{r,a}(x_r,l_a;\alpha_{r,a})}{1 + O_{r,a}(X_r,L_a;\alpha_{r,a})}$$

for each r, $a \neq 11$ and get $\widehat{\alpha}_{r,a}$. We can get the estimate using IPW as

$$\widehat{\theta}_{\mathsf{IPW}} = \frac{1}{n}\sum_{i=1}^{n} Y_{3,i}Y_{4,i}I(A_i = 11)\left[\sum_{r,a\neq 11} O_{r,a}(X_{r,i},L_{a,i};\widehat{\alpha}_{r,a})I(R_i \geqslant r) + 1\right]$$

We now consider the multiply-robust estimation. We can get the estimate using the following multiply-robust estimator

$$\widehat{\theta}_{\mathsf{MR}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_{r,a\neq 11}\{(f(L_i) - m_{r,a}(X_{i,r},L_{i,a};\widehat{\beta}_{r,a}))O_{r,a}(X_{i,r},L_{i,a};\widehat{\alpha}_{r,a})I(R_i \geqslant r, A_i = 11)$$
$$+m_{r,a}(X_{i,r},L_{i,a};\widehat{\beta}_{r,a})I(R_i = r, A_i = a)\} + f(L_i)I(A_i = 11)\Bigg]$$

with the same estimators $m_{r,a}(X_r, L_a;\widehat{\beta}_{r,a})$ and $O_{r,a}(X_r, L_a;\widehat{\alpha}_{r,a})$ as before. Again, we consider the case when all the regression functions and odds functions are correctly specified.

Next, we consider the case when two of the regression estimators is misspecified. When $R = 00$ and $A = 01$, the true regression model is $\mathbb{E}[Y_3|Y_4] = \beta_0 + \beta_1 Y_4$ and we fit a linear regression with intercept only. When $R = 00$ and $A = 10$, the true regression model is $\mathbb{E}[Y_4|Y_3] = \beta_0 + \beta_1 Y_3$ and we also fit a linear regression model with intercept only. We apply the same model misspecification to the regression adjustment estimator. Further, we consider the case when two of the odds function are misspecified. When $R = 00$ and $A = 01$, the true odds function is $O_{r,a}(Y_4) = \exp(\alpha_0 + \alpha_1 Y_4)$ and we fit a logistic regression with intercept only. When $R = 00$ and $A = 10$, the true odds function is $O_{r,a}(Y_3) = \exp(\alpha_0 + \alpha_1 Y_3)$ and we also fit a logistic regression with intercept only. We also apply the same model misspecification to the IPW estimator. Finally, we consider the case when both the regression function and the odds function are misspecified. When $R = 00$ and $A = 10$, we fit a linear regression model with intercept only for the regression function and we also fit a logistic regression model with intercept only for the odds function. When $R = 00$ and $A = 01$, again we fit linear regression and logistic regression models with intercepts only.

Table 3.2: Simulations results for estimating $\mathbb{E}[Y_3 Y_4]$ when $n = 2000$

Methods	Bias	Sample SE	Mean theoretical SE	Coverage of 95% CI
IPW	-0.000	0.075	0.074	0.943
IPW (incorrect)	0.078	0.079	0.079	0.852
RA	-0.001	0.065	0.066	0.956
RA (incorrect)	-0.048	0.065	0.068	0.892
MR (correct)	-0.001	0.066	0.066	0.948
MR (IPW incorrect)	-0.001	0.065	0.066	0.949
MR (RA incorrect)	-0.001	0.067	0.067	0.952
MR (both incorrect)	0.014	0.068	0.065	0.943
Complete Case	0.131	0.092	0.092	0.723

From table 3.2, we can see that IPW, RA and multiply-robust estimators all obtained close to 0 bias and achieves nominal coverages for the 95% CIs when the regression or odds functions are correctly specified. In comparison, we also observe that IPW estimator obtained relatively large SE estimates, while multiply robust estimators and regression adjustment estimator obtained relatively similar SE estimates. As expected, both IPW and regression adjustment estimators fails when the regression or odds functions are misspecified. Multiply-robust estimators with the same model misspecification are able to achieve nominal coverages and small biases. Further, when both the regression and odds functions are misspecified, multiply-robust estimators obtain relatively large bias. Finally, the complete-case analysis again obtained severely biased estimates.

3.7.3 Marginal parametric model

In this section, we consider the following setup for the marginal parametric model. We have $L = (Y_2, Y_3)$ and $X = Y_1$. We want to estimate the following linear regression model

$$\mathbb{E}[Y_3|Y_2] = \beta_0 + \beta_1 Y_2$$

Further, we assume that

1. $(X, L) \sim_d N(\mu_0, \Sigma)$ with $\mu_0 = (1, 0, -1)^T$ and

$$\Sigma = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix}$$

2. $P(R = 0, A = a|X, L) = h(X, L) \exp\left(L^T 1_2\right)$ for $a \in \{00, 01, 10, 11\}$.

3. $P(R = 1, A = a|X, L) = h(X, L) \exp(L^T 1_2 + 0.5X)$ for $a \in \{00, 01, 10\}$.

4. $P(R = 1, A = 11|X, L) = h(X, L) \exp(L^T 1_2)$.

with $h(X, L) = 1/\left[5\exp(L^T 1_2) + 3\exp(L^T 1_2 + 0.5X)\right]$ being the normalization term. It can be verified that this data generation satisfies the ACCMV assumption. Given the data generation above, we have

$$\mathbb{E}[Y_3|Y_2] = -1 + \frac{1}{2}Y_2$$

As discussed in section 3.5, we will be using IPW to estimate the parameters for the linear regression model. We can estimate individual odds $O_{r,a}$ by fitting a logistic regression model as follows:

$$P(R = r, A = a|x_r, l_a; \{R \geqslant r, A = 11\} \cup \{R = r, A = a\}; \alpha_{r,a}) = \frac{O_{r,a}(x_r, l_a; \alpha_{r,a})}{1 + O_{r,a}(x_r, l_a; \alpha_{r,a})}$$

and get $\widehat{\alpha}_{r,a}$. Next, we compute the total weights for each r such that

$$\widehat{Q}_r(X_r, L) = \sum_{\tau \leqslant r} \sum_{a \neq 11} O_{\tau,a}(X_\tau, L_a; \widehat{\alpha}_{\tau,a})$$

Finally, we can get $\widehat{\beta}$ by solving the following weighted estimating equation:

$$\sum_{i=1}^{n} s(\widehat{\beta}|L_i) \left[\sum_r \widehat{Q}_r(X_{i,r}, L_i) I(R_i = r) + 1\right] I(A_i = 11) = 0$$

From table 3.3, we can see that IPW gives close to 0 bias and achieves nominal coverage for the 95% confidence intervals. On the other hand, complete-case analysis obtained biased estimates.

Table 3.3: Simulations results for the marginal parametric model when $n = 2000$.

	Bias (SE)		Coverage of 95% CI	
Methods	β_0	β_1	β_0	β_1
IPW	-0.001 (0.039)	-0.001 (0.046)	0.949	0.938
Complete Case	-0.061 (0.042)	-0.008 (0.043)	0.725	0.943

3.8 Applications to the Diabetes data

We apply the proposed estimation procedures to the diabetes EHR dataset, assuming that ACCMV holds. This dataset contains 8663 patients who were followed up every 3 months from 2003 to 2013.

3.8.1 Summary measures of the HbA1c levels

We focus on the HbA1c levels measured from the baseline and the first year, $(Y_0, Y_1, \ldots, Y_4)$. We now answer the first two questions raised in the introduction. We estimate

1. The mean HbA1c levels at the end of the first year, $\mathbb{E}[Y_4]$

2. The proportion of patients that have their HbA1c levels controlled, meaning that their HbA1c levels are below 7%, i.e., $P(Y_3 \leqslant 7, Y_4 \leqslant 7)$[2]

3. The averages of the HbA1c levels for the last two quarters, $\mathbb{E}[Y_3 + Y_4]/2$. All these are important summaries of the diabetes patients cohort.

For the estimation of $\mathbb{E}[Y_4]$, the primary variable is $L = Y_4$ and the auxillary variables are $X = (Y_0, \ldots, Y_3)$. For the estimations of $\mathbb{E}[Y_3 + Y_4]/2$ and $P(Y_3 \leqslant 7, Y_4 \leqslant 7)$, the primary variables are $L = (Y_3, Y_4)$ and the auxillary variables are $X = (Y_0, Y_1, Y_2)$.

We construct the 95% confidence intervals using bootstrap. The results are given in Table 3.4 and 3.5. We can see that IPW, regression adjustment and multiply robust estimators all obtain quite similar results. The only exception is with the estimation of $P(Y_3 \leqslant 7, Y_4 \leqslant 7)$, where IPW and multiply-robust estimators obtain non-overlap 95% confidence intervals. This could be due to the incorrect specifications of the odds functions. We can also see that results of complete-case analysis are different from the rest three methods. Complete-case analysis tends to over-estimate the HbA1c levels. This suggets that the missingness of HbA1c

[2]For convenience, we multiply the value of HbA1c levels by 100.

is relevant to the underlying values of HbA1c, which agrees with the intuition that healthier patients are more likely to miss their HbA1c measurements.

More specifically, our ACCMV approach shows that both $\mathbb{E}[Y_4]$ and $\mathbb{E}[Y_3 + Y_4]/2$ is less than 7 and more than 50% of the participants have their HbA1c levels controlled for the last two quarters; while complete-case analysis obtain completely opposite results. This highlights the importance of treating missing data in a real dataset.

Table 3.4: Summary statistics computed on diabetes dataset

Methods	$\mathbb{E}[Y_4]$	$P(Y_3 \leqslant 7, Y_4 \leqslant 7)$	$\mathbb{E}[Y_3 + Y_4]/2$
IPW	6.927 (6.897 - 6.956)	0.534 (0.515 - 0.556)	6.931 (6.900 - 6.958)
Regression Adjustment	6.936 (6.907 - 6.966)	0.569 (0.554 - 0.583)	6.955 (6.925 - 6.983)
Multiply Robust	6.931 (6.902 - 6.960)	0.575 (0.559 - 0.589)	6.949 (6.920 - 6.976)
Complete Case	7.011 (6.974 - 7.049)	0.454 (0.431 - 0.477)	7.197 (7.143 - 7.252)

We also perform sensitivity analysis by the exponential tilting approach proposed in section 3.6. We use the same sensitivity parameter for all patterns, i.e., every element of $\delta_{\bar{a}}$ in equation (3.21) is identical. From a practical perspective, we modify the exponential tilting as $\exp(\delta_{\bar{a}}^T(\ell_{\bar{a}} - 7))$. This allows the missingness of HbA1c measurements to follow the intuition that a healthier patient with controlled HbA1c levels are more likely to miss their HbA1c measurements, while a sick patients are less likely to miss their HbA1c measurement. For example, with a negative δ, $\delta(\ell_{\bar{a}} - 7)$ is more likely to be positive for healthier patients (because their HbA1c levels are more likely to be less than 7) and thus increase the missing probability. Similarly, a negative δ will decrease the missing probability for sicker patients. For completeness, we also displays the results with a positive δ, which will reverse the relationship between health status and HbA1c missing probability.

Figure 3.1 and 3.2 show the estimates and 95% confidence intervals for $\mathbb{E}[Y_4]$ and $P(Y_3 \leqslant$

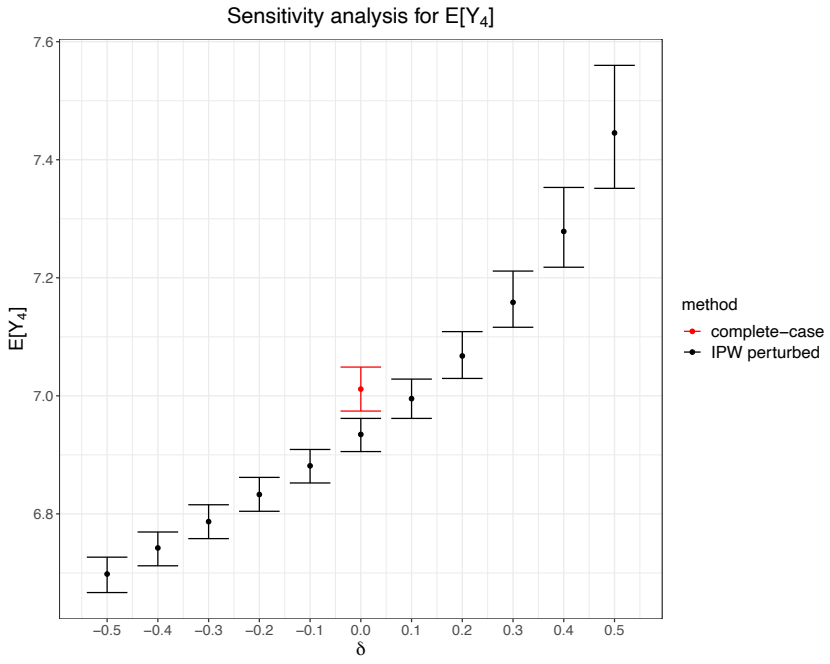

Figure 3.1: Sensitivity analysis of the ACCMV assumption by exponential tilting. We examine how the estimate $\mathbb{E}[Y_4]$ changes with respect to different values of the sensitivity parameter δ.

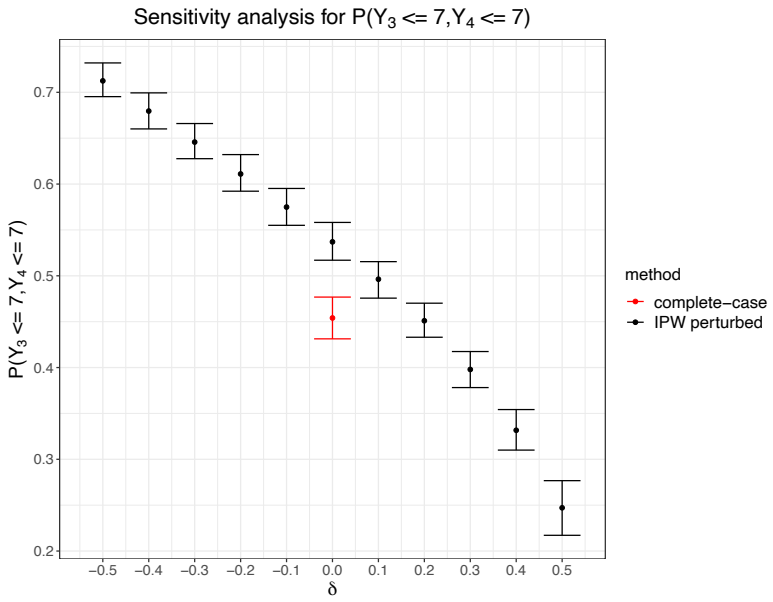

Figure 3.2: Sensitivity analysis of the ACCMV assumption by exponential tilting. We examine how the estimate $P(Y_3 \leqslant 7, Y_4 \leqslant 7)$ changes with respect to different values of the sensitivity parameter δ.

$7, Y_4 \leqslant 7$) as we vary the sensitivity parameter δ. We can see that the results highly agree with our intuition. When δ is negative and the magnitude of δ increases, the mean HbA1c levels decreases as we take into account of the fact that healthier patients with lower HbA1c values are more likely to be missing. For the same reason, the proportion of patients having their HbA1c levels controlled also increases. In the unrealistic scenario that δ is positive, we observe opposite results. The results for $\mathbb{E}[Y_3 + Y_4]/2$ are deferred to B.6.

3.8.2 Marginal parametric model

Table 3.5: Linear regression results for the diabetes dataset: $\mathbb{E}[Y_4|Y_2, Y_3] = \beta_0 + \beta_1 Y_2 + \beta_2 Y_3$.

Methods	β_0	β_1	β_2
IPW	1.183 (0.814 - 1.520)	0.138 (0.045 - 0.226)	0.692(0.586 - 0.798)
Complete Case	1.364 (1.127 - 1.601)	0.104 (0.048 - 0.160)	0.701(0.645 - 0.758)

Further, we also want to study the linear relationship between Y_2, Y_3 and Y_4. Our intuition is to predict Y_4, Y_3 should be more important compared to Y_2. We consider estimating the linear regression model as follows:

$$\mathbb{E}[Y_4|Y_2, Y_3] = \beta_0 + \beta_1 Y_2 + \beta_2 Y_3$$

For the linear regression model, the primary variable is $L = (Y_2, Y_3, Y_4)$ and the auxillary variables are $X = (Y_0, Y_1)$. Table 3.5 shows that indeed both Y_2 and Y_3 has a positive association with Y_4 and Y_3 has a stronger association than Y_2. Figure 3.3 shows the estimates and 95% confidence intervals for $\beta_0, \beta_1, \beta_2$ as δ varies. We can see that when δ is negative, the estimates are quite robust and do not change much.

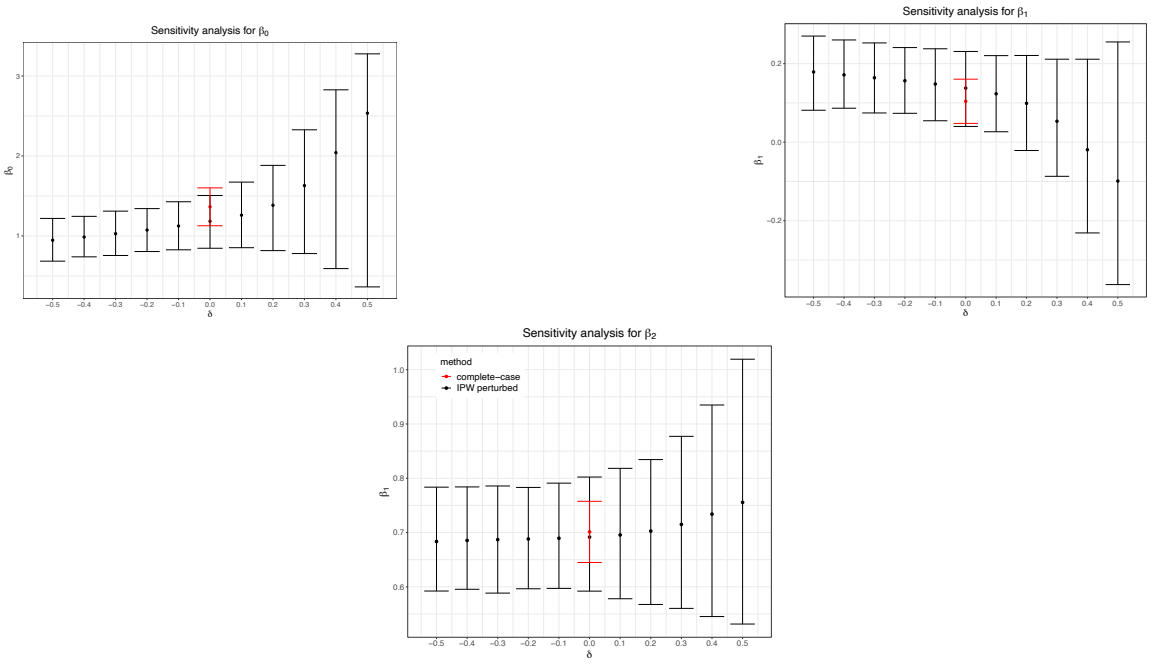

Figure 3.3: Sensitivity analysis of the ACCMV assumption by exponential tilting. We examine how the parameter estimates and confidence intervals changes with respect to different values of the sensitivity parameter δ.

3.9 Conclusion

In this chapter, we introduced the ACCMV assumption to handle nonmonotone and MNAR data. Our ACCMV assumption allows a much larger set of observations to be used for identification compared to the traditional CCMV assumption. Thus, ACCMV is particularly suitable for analyzing datasets with few complete cases. We further proposed IPW, regression adjustment and multiply-robust estimators. We also studied their asymptotic and efficiency theories. We then proposed a sensitivity analysis approach for the ACCMV model for the IPW estimator. Our simulation studies confirm the validity of the assumption. The real data results also highlight the effect of missing data on the final estimate and the importance of efficiently handling the missing data.

So far, we have focused on the first-year's data for the HbA1c measurements. However, the diabetes patients are followed up to 11 years and a patient could potentially have up to 44 measurements. Thus, it is helpful to also consider longer history of HbA1c measurements as this can provide more information of a patient. In particular, it will be of interest to recover the whole trajectory for a patient who has a bunch of missing values. This is a much more challenging task as the missing patterns increase exponentially when the number of measurements increases. We leave the trajectory recovery problem to future work.

Chapter 4

TRAJECTORY RECOVERY FOR NONMONOTONE MISSING NOT AT RANDOM DATA

4.1 Introduction

Missing data problems are very common in scientific research (Molenberghs et al., 2014). Based on the missing/response patterns, these problems can be categorized into monotone and nonmonotone missing data problems. For monotone missing data, variables subject to missing are ordered and if one variable is missing, all subsequent variables are missing. This occurs often as a result of dropout in a longitudinal study (Diggle et al., 2002). Nonmonotone missingness refers to the case when no such ordering exists (Little and Rubin, 2019). For example, a participant might drop out and later return to a study.

In this chapter, we are interested in recovering trajectories formed by repeated measurements collected from an individual. For example, an electronic health record (EHR) dataset contains quarterly measurements of glycated hemoglobin (HbA1c) from diabetes patients over 11 years. Thus, an individual would have a HbA1c trajectory with length 44 if no measurement is missing. There are three main challenges with recovering trajectories from a longitudinal study. First, trajectories are often subject to nonmonotone missing patterns and most existing methods in longitudinal data analysis can only handle monotone missing patterns. Second, measurements are often missing not at random (MNAR) which suggests that the missingness may depend on the unobserved outcome of interest so many classical missing data methodologies are not applicable. Third, the length of trajectory also creates additional difficulty for modeling and computation. For example, the number of possible missing patterns increase exponentially with the number of measurements.

Handling nonmonotone missing data is a very challenging task even if we assume that

data is missing-at-random (MAR) (Robins and Gill, 1997; Sun and Tchetgen Tchetgen, 2018). Robins and Vanstellandt Robins and Gill (1997); Vansteelandt et al. (2007) have argued that the MAR restriction should not be expected to hold in nonmonotone missing data. Several attemps have been made to handle non-monotone MNAR data (Troxel et al., 1998a; Robins and Gill, 1997; Troxel et al., 1998b; Ibrahim et al., 2001; Zhou et al., 2010; Sadinle and Reiter, 2017; Tchetgen et al., 2018; Nabi et al., 2020; Malinsky et al., 2021; Mohan and Pearl, 2021; Chen, 2022). However, all these existing work have limitations and are not suitable for recovering long trajectories.

In particular, Troxel et.al (Troxel et al., 1998a) employ a first-order Markov dependence structure for the study variables and allow the missing probability to depend on unobserved variables. However, their approaches suffer from computational difficulty and becomes intractable with more than three or four measurements. Multiple imputation by chained equations (MICE) (Van Buuren and Groothuis-Oudshoorn, 2011) is a very popular imputation method for missing data. MICE is able to recover long trajectories by imputing missing values through fitting a series of conditional distributions. However, MICE suffers from distribution incompatibility issues as a proper joint distribution might not exist for the corresponding conditional distributions (Raghunathan et al., 2001; Van Buuren, 2007). Further, MICE assumes that the missing data mechanism is MAR.

To deal with nonmonotone MNAR trajectories, we propose a block-Markov type assumption that decomposes trajectory into multiple missing blocks. This assumption leads to nonparametric identification of the joint distribution of the trajectory. It also greatly reduces the difficulty of identification and the complexity of modeling. Essentially, we just need to deal with the missing data for each smaller missing blocks, instead of working directly with the whole trajectory. For modeling purpose, we modify our assumption to a model-based version that assumes a multivariate normal distribution for each block. We propose to estimate the multivariate normal distribution with linear models or other flexible nonparametric/machine learning models. Further, we use multiple imputation to obtain multiple completed trajectories and estimate the uncertainty with bootstrap.

Outline. In section 4.2, we use an example to motivate and illustrate our block-Markov assumption. Then we formally define our missing data assumption and shows that it non-parametrically identifies the joint distribution for the trajectory. In section 4.3, we introduce the model-based assumption and propose to estimate the imputation distribution with linear and machine learning models. We further discuss using bootstrap to estimate the uncertainty of our estimator. We conduct simulation study in section 4.4 and apply our methods to recover HbA1c trajectories from a diabetes EHR dataset in section 4.5.

Notation. We use the vector $L \in \mathbb{R}^d$ to denote the trajectories that consist of longitudinal measured variables, where d denotes the total number of repeated measurements (total number of time points). We use $|\cdot|$ to denote the length of a vector and $|L| = d$. We use binary vector $A \in \{0,1\}^d$ to denote the missing pattern of L, i.e., $A_j = 1$ if L_j is observed. We use the notation $L_a = (L_j : a_j = 1)$ to denote the observed parts of L under pattern $A = a$. Let $1_d = (1, 1, \cdots, 1) \in \mathbb{R}^d$. We use the notation $\bar{a} = 1_d - a$ to denote the vector after flipping 0 and 1 in a. $L_{\bar{a}}$ refers to the missing variables under pattern $A = a$. We further define $A \geqslant a$ if $A_i \geqslant a_i$ for $i = 1, \ldots, d$. For instance, $101 \geqslant 100$ but 101 cannot be compared with 010. We use $I_d \in \mathbb{R}^{d \times d}$ to represent the identity matrix with dimension d.

4.2 Block-Markov assumption for trajectory recovery

Motivation and a simple example To motivate and illustrate our block-Markov assumption, we first start with a simple example. Suppose that the trajectory contains 5 measurements, i.e., $L \in \mathbb{R}^5$, and consider the missing pattern $A = 10010$. In this case, we only observe L_1, L_4 and L_2, L_3, L_5 are missing. We can then write the joint density as

$$p(\ell_1, \ell_2, \ell_3, \ell_4, \ell_5 | A = 10010) = p(\ell_1, \ell_4 | A = 10010) p(\ell_2, \ell_3, \ell_5 | \ell_1, \ell_4, A = 10010).$$

The quantity $p(\ell_1, \ell_4 | A = 10010)$ is identifiable from the data and we need to put missing data assumption on the second component $p(\ell_2, \ell_3, \ell_5 | \ell_1, \ell_4, A = 10010)$ to help identify the joint density. The second component is the density for the unobserved variables, which

is also known as the extrapolation density. The traditional complete case missing value (CCMV) assumption Little (1993); Tchetgen et al. (2018) identifies this density using only the complete cases:

$$p(\ell_2, \ell_3, \ell_5 | \ell_1, \ell_4, A = 10010) = p(\ell_2, \ell_3, \ell_5 | \ell_1, \ell_4, A = 11111) \tag{4.1}$$

However, in practice, it is possible that we might have very few individuals with fully observed trajectories. For example, for the diabetes EHR dataset, only 1.2% of the patients have their first 10 HbA1c fully observed and none of the patient has the whole 44 HbA1c fully observed. This leads to difficult and potentially unreliable identification of the extrapolation densities.

This motivates us to reduce the problem for the whole trajectory to each smaller missing block. We assume that each block of consecutive unobserved variables are conditionally independent of other variables given its two ends of observed variables. Concretely, this assumes that the extrapolation density can be decomposed as follows:

$$p(\ell_2, \ell_3, \ell_5 | \ell_1, \ell_4, A = 10010) = p(\ell_2, \ell_3 | \ell_1, \ell_4, A = 10010) p(\ell_5 | \ell_4, A = 10010). \tag{4.2}$$

Note that since L_5 is the last variable, the missing block L_5 only has one end of observed variable L_4. This shows the block-Markov property of our assumption. We do not impose any assumption on the observed variables and this assumption is only for the extrapolation density.

Under the above assumption, we decompose an extrapolation density involving 5 variables into two distributions involving 4 and 2 variables, respectively. For the two distributions on the right-hand side of equation (4.2), we then apply the idea of available case missing value (ACMV) assumption (Molenberghs et al., 1998) to identify them. For the first term $p(\ell_2, \ell_3 | \ell_1, \ell_4, A = 10010)$, this quantity can be identified as long as the first four variables are observed. So we assume that

$$p(\ell_2, \ell_3 | \ell_1, \ell_4, A = 10010) = p(\ell_2, \ell_3 | \ell_1, \ell_4, A \geqslant 11110). \tag{4.3}$$

This assumption identifies the extrapolation density for the first block using all observations with first block fully observed. In this scenario, available cases refer to all observations with

$A \geqslant 11110$, i.e., $A = 11110$ or $A = 11111$. Similarly for the second term, we assume that:

$$p(\ell_5|\ell_4, A = 10010) = p(\ell_5|\ell_4, A \geqslant 00011). \tag{4.4}$$

In this case, we only need the last two variables observed.

Thus, the idea of ACMV allows a potentially much larger set of observations being used for identification. Together, equations (4.2), (4.3) and (4.4) form our missing data assumption, termed block-Markov ACMV (BM-ACMV), for pattern $A = 10010$. Further, BM-ACMV also consists of similar assumptions for other missing patterns.

Formal Definition Now we formally introduce the definition of BM-ACMV. We first define some notations. For a missing pattern $A \in \{0,1\}^d$, recall $\bar{A} = 1_d - A$ represents the missing variables and let $J(\bar{A})$ denotes the number of missing blocks. We use $B_j \in \{0,1\}^d$ to represent the j-th block for $j = 1, \ldots, J(\bar{A})$ from left to right. For each block, B_j has elements

$$B_{j,k} = \begin{cases} 1 & \text{if } k \in [\underline{u}_j, \overline{u}_j] \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } k = 1, \ldots, d$$

where $\underline{u}_j$ is the starting position of the j-th missing block and $\overline{u}_j$ is the ending position of the j-th missing block with $\underline{u}_j \leqslant \overline{u}_j$. Then we can decompose $\bar{A}$ as

$$\bar{A} = B_1 + B_2 + \cdots + B_{J(\bar{A})}$$

We further define $B_j^\dagger \in \{0,1\}^d$ that adds two ends of observed variables to B_j as

$$B_{j,k}^\dagger = \begin{cases} 1 & \text{if } k \in [\underline{u}_j - 1, \overline{u}_j + 1] \cap \{1, 2, \ldots, d\} \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } k = 1, \ldots, d$$

For example, again consider pattern $A = 10010$ from the previous example. We have $\bar{A} = 01101$ and $J(\bar{A}) = 2$ as we have two missing blocks. We can decompose $\bar{A}$ as

$$\bar{A} = 01101 = \underbrace{01100}_{B_1} + \underbrace{00001}_{B_2}$$

with $\underline{u}_1 = 2$, $\overline{u}_1 = 3$ and $\underline{u}_2 = \overline{u}_2 = 5$. Taking the two ends of observed variables into account, we have

$$B_1^\dagger = 11110 \quad B_2^\dagger = 00011$$

and the difference $B_j^\dagger - B_j$ represents the two ends of the j-th block:

$$B_1^\dagger - B_1 = 11110 - 01100 = 10010 \quad B_2^\dagger - B_2 = 00011 - 00001 = 00010.$$

With the above notations, we can formally define our BM-ACMV assumption. The BM-ACMV assumption consists of the following two parts.

- **BM (block-Markov):** For any missing pattern $a \neq 1_d$, let b_j represents the block decomposition of $\bar{a}$ for $j = 1, \ldots, J(\bar{a})$. The extrapolation density can be factorized into product of block densities:

$$p(\ell_{\bar{a}}|\ell_a, A = a) = \prod_{j=1}^{J(\bar{a})} p(\ell_{b_j}|\ell_{b_j^\dagger - b_j}, A = a) \tag{4.5}$$

- **ACMV:** Each block density $p(\ell_{b_j}|\ell_{b_j^\dagger - b_j}, A = a)$ is identified via

$$p(\ell_{b_j}|\ell_{b_j^\dagger - b_j}, A = a) = p(\ell_{b_j}|\ell_{b_j^\dagger - b_j}, A \geqslant b_j^\dagger) \tag{4.6}$$

For the missing pattern $A = 10010$, equation (4.5) reduces to equation (4.2), equation (4.6) leads to equation (4.3) and (4.4).

Proposition 4.2.1. *Under the BM-ACMV assumption* (4.5) *and* (4.6)*, the marginal density* $p(\ell, A = a)$ *is nonparametrically identified.*

Proposition 4.2.1 shows that BM-ACMV is also a nonparametrically identifying assumption (Robins et al., 2000), i.e., the marginal distribution $p(\ell, a)$ is identifiable and this assumption will not conflict with the data. So any quantity involving the entire trajectory of L can be identified.

There are two major advantages of using the BM-ACMV assumption. First, the "complete case" needed for identifying the unobserved entries is relaxed. We only need to consider

complete cases of each block, rather than the entire trajectory. So this idea drastically increases the effective sample size for estimating the extrapolation density. Secondly, the model complexity is drastically reduced. The usual complete case missing value (CCMV) assumption requires models involving all variables in L. When the trajectory is long, it could be very challenging to place a model on it. For instance, d is 45 for the diabetes dataset. So a multivariate Gaussian model for L would lead to a mean vector of length 45 and a covariance matrix with $45 \times 46/2 = 1035$ parameters. On the other hand, the BM-ACMV only requires models on each block. So the bottleneck of model complexity is determined entirely by the longest block, which may be much smaller than the length of the entire trajectory.

In practice, we assume that the length of the longest block should not be too long. For example, we might set the longest number of consecutive unobserved variables to be three, which is practically meaningful. For a quarterly measurement, recovering the trajectory for a participant who was missing for more than a year is in general pretty difficult.

4.3 Modeling and imputation

4.3.1 Linear model approach

In practice, we need to model the block densities $p(\ell_{b_j}|\ell_{b_j^\dagger - b_j}, A \geqslant b_j^\dagger)$. To simplify the problem, we assume that L contains only continuous variables and consider fitting multivariate Gaussian distributions.

Model compatibility issue. While it is very tempting to directly model $p(\ell_{b_j}|\ell_{b_j^\dagger - b_j}, A \geqslant b_j^\dagger)$ as a Gaussian distribution, there is a subtle caveat for this modeling procedure. Consider the simple case when $d = 3$, let $b_j = 100$ and $b_j^\dagger = 110$, by definition of conditional probability,

$$P(\ell_1|\ell_2, A \geqslant 110) = \frac{P(\ell_1, \ell_2, A = 110) + P(\ell_1, \ell_2, A = 111)}{P(\ell_2, A = 110) + P(\ell_2, A = 111)}.$$

Similarly, for $b_j = 001$ and $b_j^\dagger = 011$, we have

$$P(\ell_3|\ell_2, A \geqslant 011) = \frac{P(\ell_3, \ell_2, A = 011) + P(\ell_2, \ell_3, A = 111)}{P(\ell_2, A = 011) + P(\ell_2, A = 111)}.$$

Thus, $p(\ell_1|\ell_2, A \geqslant 110)$ and $p(\ell_3|\ell_2, A \geqslant 011)$ are not variationally independent as they both depend on $p(\ell_2, A = 111)$. These variational dependences also occur for other missing patterns and might lead to model congeniality problems (Meng, 1994) without careful modeling. To resolve this issue, we choose to modify our BM-ACMV assumption and instead directly identify the extrapolation density with a model-based approach. As we directly model the extrapolation density, we no longer have the problem of model congeniality problems.

Model-based BM-ACMV We replace equation (4.6) by a parametric model

$$\ell_{b_j}|\ell_{b_j^\dagger - b_j}, A = a \sim N(\Gamma_{b_j}\ell_{b_j^\dagger - b_j} + \eta_{b_j}, \Sigma_{b_j}) \tag{4.7}$$

where $\Gamma_{b_j} \in \mathbb{R}^{|b_j|\times|b_j^\dagger - b_j|}$ is the matrix of coefficients (slopes), $\eta_{b_j} \in \mathbb{R}^{|b_j|}$ is the intercept vector and $\Sigma_{b_j} \in \mathbb{R}^{|b_j|\times|b_j|}$ is the covariance matrix. Further, we assume that the parameters $\Gamma_{b_j}, \eta_{b_j}, \Sigma_{b_j}$ are identified by fitting a conditional Gaussian distribution based on observations with $A \geqslant b_j^\dagger$. More specifically, we can fit linear models with $\ell_{b_j^\dagger - b_j}$ as predictors and ℓ_{b_j} as responses to estimate the conditional means. We can then use residuals from linear models to estimate the covarian matrix.

In more details, consider the example from (4.3), we have $\ell_{b_j^\dagger - b_j} = (\ell_1, \ell_4)^T$ and define $\ell_{1,4} = (1, \ell_1, \ell_4)^T$, model-based BM-ACMV assumes that

$$\ell_2, \ell_3|\ell_1, \ell_4, A = 10010 \sim N\left(\begin{pmatrix} \ell_{1,4}^T\beta_2^* \\ \ell_{1,4}^T\beta_3^* \end{pmatrix}, \Sigma_{b_j} = \begin{pmatrix} \sigma_2^2 & \sigma_{23} \\ \sigma_{23} & \sigma_3^2 \end{pmatrix}\right) \tag{4.8}$$

where $\beta_2^*, \beta_3^* \in \mathbb{R}^3$ and

$$\begin{aligned} \beta_2^* &= \underset{\beta_2}{\operatorname{argmin}}\, \mathbb{E}[(L_2 - L_{1,4}^T\beta_2)^2|A \geqslant 11110] \\ \beta_3^* &= \underset{\beta_3}{\operatorname{argmin}}\, \mathbb{E}[(L_3 - L_{1,4}^T\beta_3)^2|A \geqslant 11110] \\ \sigma_2^2 &= \mathbb{E}[(L_2 - L_{1,4}^T\beta_2^*)^2|A \geqslant 11110], \quad \sigma_3^2 = \mathbb{E}[(L_3 - L_{1,4}^T\beta_3^*)^2|A \geqslant 11110] \\ \sigma_{23} &= \mathbb{E}[(L_2 - L_{1,4}^T\beta_2^*)(L_3 - L_{1,4}^T\beta_3^*)|A \geqslant 11110] \end{aligned} \tag{4.9}$$

Equations (4.8) and (4.9) together form the linear model-based ACMV assumption. Γ_{b_j}, η_{b_j} can be recovered from β_2^*, β_3^*.

We can estimate β_2^* and β_3^* with least squares methods and get $\widehat{\beta}_2$ and $\widehat{\beta}_3$. Next, we can estimate the covariance matrix using residuals of the least squares fit, i.e.,

$$\begin{aligned}\widehat{\sigma}_2^2 &= \frac{1}{n_0}\sum_{i:A_i\geqslant 11110}(L_{i,2}-L_{i,1,4}^T\widehat{\beta}_2)^2, \quad \widehat{\sigma}_3^2 = \frac{1}{n_0}\sum_{i:A_i\geqslant 11110}(L_{i,3}-L_{i,1,4}^T\widehat{\beta}_3)^2 \\ \widehat{\sigma}_{23} &= \frac{1}{n_0}\sum_{i:A_i\geqslant 11110}(L_{i,2}-L_{i,1,4}^T\widehat{\beta}_2)(L_{i,3}-L_{i,1,4}^T\widehat{\beta}_3)\end{aligned} \tag{4.10}$$

where n_0 is the number of observations with $A_i \geqslant 11110$. After estimation, we can sample from the following distribution

$$\ell_2, \ell_3 | \ell_1, \ell_4, A = 10010 \sim N\left(\begin{pmatrix} \ell_{1,4}^T\widehat{\beta}_2 \\ \ell_{1,4}^T\widehat{\beta}_3 \end{pmatrix}, \begin{pmatrix} \widehat{\sigma}_2^2 & \widehat{\sigma}_{23} \\ \widehat{\sigma}_{23} & \widehat{\sigma}_3^2 \end{pmatrix}\right)$$

to impute the missing entries of ℓ_2, ℓ_3. Finally, we note that the model-based BM-ACMV also nonparametrically identifies $p(\ell, A = a)$.

Proposition 4.3.1. *Under the model-based BM-ACMV assumption, the marginal density $p(\ell, A = a)$ is nonparametrically identified.*

Imputation procedure We propose the following procedure to create multiple imputed trajectories from a dataset that consists of non-monotone missing entries:

1. **Step 1.** We first find out all possible response patterns $\mathcal{A} \subset \{0,1\}^d$. Namely, $\mathcal{A}$ is the smallest subset of $\{0,1\}^d$ such that $A_i \in \mathcal{A}$ for all i.

2. **Step 2.** For each $a \in \mathcal{A}$, we derive its blocks $\bar{a} = b_1 + b_2 + \cdots + b_{J(\bar{a})}$ and estimate the model parameters using (4.10) and derive the imputation density of each block.

3. **Step 3.** For $i = 1, \cdots, n$, suppose $A_i = a \neq 1_d$, we derive the blocks of $\bar{a} = b_1 + b_2 + \cdots + b_{J(\bar{a})}$. For block $j = 1, \cdots, J(\bar{a})$, we impute mising variables by sampling from

$$L_{i,b_j}^* | L_{i,b_j^\dagger - b_j} \sim N(\widehat{\Gamma}_{b_j} L_{i,b_j^\dagger - b_j} + \widehat{\eta}_{b_j}, \widehat{\Sigma}_{b_j})$$

where $\widehat{\Gamma}_{b_j}, \widehat{\eta}_{b_j}$ and $\widehat{\Sigma}_{b_j}$ can be constructed using estimates from the linear models. This procedure creates an imputed dataset $\mathcal{D}^* = \{L_1^*, \ldots, L_n^*\}$ such that $L_{i,j}^* = L_{i,j}$ if $L_{i,j}$ is observed and $L_{i,j}^*$ is imputed from the above density when it is missing.

4. **Step 4.** We repeat the above procedure for M times, leading to M complete datasets

$$\mathcal{D}^{*(1)}, \mathcal{D}^{*(2)}, \cdots, \mathcal{D}^{*(M)}$$

where

$$\mathcal{D}^{*(k)} = \{L_1^{*(k)}, \cdots, L_n^{*(k)}\}$$

is the k-th imputed dataset. We then use $\mathcal{D}^{*(1)}, \cdots, \mathcal{D}^{*(M)}$ to make inference about the trajectory.

4.3.2 Nonparametric and machine learning approaches

We can further extend our approach beyond the linear model assumptions. Instead of (4.7), we may assume that the block density satisfies the following assumption:

$$\ell_{b_j} | \ell_{b_j^\dagger - b_j}, A = a \sim N(\boldsymbol{f}_{b_j}(\ell_{b_j^\dagger - b_j}), \Sigma_{b_j}) \tag{4.11}$$

with $\boldsymbol{f}_{b_j}(\ell_{b_j^\dagger - b_j}) \in \mathbb{R}^{|b_j|}$ being a vector consisting of potentially nonlinear functions of $\ell_{b_j^\dagger - b_j}$. Further, we assume that $\boldsymbol{f}_{b_j}$ and Σ_{b_j} can be identified by fitting nonparametric or machine learning model with observations $A \geqslant b_j^\dagger$. For example, we can use k-nearest neighbors (kNN) or random forest to estimate $\boldsymbol{f}_{b_j}$.

Estimation of mean function We use kNN as an example to illustrate our idea. Consider again the example from (4.3) that $A = 10010$ and we focus on the first block (L_2, L_3 are missing while L_1 and L_4 are observed). The imputation distribution is

$$\ell_2, \ell_3 | \ell_1, \ell_4, A = 10010 \sim N\left(\widehat{\boldsymbol{f}}_{b_j}(\ell_{b_j^\dagger - b_j}) = \begin{pmatrix} \widehat{f}_{k,2}(\ell_1, \ell_4) \\ \widehat{f}_{k,3}(\ell_1, \ell_4) \end{pmatrix}, \widehat{\Sigma}_{b_j} = \begin{pmatrix} \widehat{\sigma}_{k,2}^2 & \widehat{\sigma}_{k,23} \\ \widehat{\sigma}_{k,23} & \widehat{\sigma}_{k,3}^2 \end{pmatrix}\right), \tag{4.12}$$

where $\widehat{f}_{k,2}(\ell_1, \ell_4)$ is the kNN estimate of the variable L_2 condition on $L_1 = \ell_1$ and $L_4 = \ell_4$ from observations with $A \geqslant 11110$, i.e., we fit the kNN regression to estimate the conditional mean of $L_2|L_1 = \ell_1, L_4 = \ell_4$ using observations with L_1, L_2, L_3, L_4 fully observed. $\widehat{f}_{k,3}(\ell_1, \ell_4)$ is estimated by a similar manner. Note that one can change the kNN estimator to any other nonparametric method or machine learning algorithm.

Estimation of the covariance matrix The estimation of the covariance matrix $\widehat{\Sigma}_{b_j}$ is more involved because the regression function is estimated nonparametrically. So naively using the residuals from the training sample to compute the covariance matrix will suffer from overfitting. Instead, we can use test error to estimate the covariance matrix as the test error can be decomposed into squared bias, variance of estimators $\widehat{f}$ and the irreducible error σ^2 (James et al., 2013). As sample size converge to infinity, with properly tuned parameters, both the squared bias and variance of estimators will converge to zero and thus the test error will eventually converge to the irreducible error σ^2, which also corresponds to the variance term here. Thus, we recommend using sample-splitting or cross-validation for estimating the covariance matrix. Here we describe the sample-splitting procedure.

First, we split the data into two parts: P_1 and P_2 such that $P_1 \cap P_2 = \varnothing$ and $P_1 \cup P_2 = \{i : A_i \geqslant 11110\}$. We use P_1 for training the kNN model and then use P_2 to compute the residuals for estimating the covariance matrix. Namely, P_1 is the training sample and P_2 is the validation sample. Specifically,

$$\widehat{\sigma}_2^2 = \frac{1}{n_2} \sum_{i:i \in P_2} (L_{i,2} - \widehat{f}_{k,2}(L_{i,1}, L_{i,4}))^2$$

where n_2 corresponds to the number of observations in part P_2 and $\widehat{f}_{k,2}$ is estimated using observations from P_1. $\widehat{\sigma}_3^2$ can be estimated similarly. The covariance is computed via a similar manner

$$\widehat{\sigma}_{23} = \frac{1}{n_2} \sum_{i:i \in P_2} (L_{i,2} - \widehat{f}_{k,2}(L_{i,1}, L_{i,4}))(L_{i,3} - \widehat{f}_{k,3}(L_{i,1}, L_{i,4}))$$

with $\widehat{f}_{k,2}$ and $\widehat{f}_{k,3}$ are both estimated using observations from P_1. Note that we can switch the position of P_1 and P_2 so that we train the machine learning model with data from P_2 and then estimate the covariance matrix using data from P_1. This leads to the cross-validation procedure and is also known as cross fitting in causal inference (Chernozhukov et al., 2018).

After estimating the imputation model, we impute the missing entries by

$$\ell_{b_j}|\ell_{b_j^\dagger - b_j}, A = a \sim N(\widehat{\boldsymbol{f}}_{b_j}(\ell_{b_j^\dagger - b_j}), \widehat{\Sigma}_{b_j})$$

and we apply the same multiple imputation procedure from the previous section.

Uncertainty estimation for the multiple imputation estimator So far, we have proposed to obtain multiple imputed datasets with our model-based BM-ACMV assumption. This allows us to obtain multiple imputed trajectories and we can then obtain estimates for parameters of interest using these completed datasets. However, to quantify the uncertainty for the estimates, we propose using bootstrap (Efron and Tibshirani, 1994) to estimate the standard errors and construct confidence intervals.

More specifically, the bootstrap procedure is as follows. We use $\mathcal{D}$ to denote the original dataset, n_B to denote the number of times for bootstrap and M for the number of times for multiple imputation. We want to estimate a parameter β.

1. **Step 1.** We sample with replacement to obtain a bootstrap sample $\mathcal{D}_{\mathsf{BT}}$. We apply the imputation procedure in section 4.3.1 to obtain multiple imputed datasets based on $\mathcal{D}_{\mathsf{BT}}$:

$$\mathcal{D}_{\mathsf{BT}}^{*(1)}, \mathcal{D}_{\mathsf{BT}}^{*(2)}, \cdots, \mathcal{D}_{\mathsf{BT}}^{*(M)}$$

2. **Step 2.** We obtain an estimate for each imputed dataset:

$$\widehat{\beta}^{*(1)}, \ldots, \widehat{\beta}^{*(M)}$$

and compute the average $\widehat{\beta} = \frac{1}{M}\sum_{i=1}^{M} \widehat{\beta}^{*(i)}$ as the bootstrap estimate.

3. **Step 3.** We repeat the above procedures (Step 1 and 2) for n_B times and get n_B bootstrap estimates $\widehat{\beta}_1, \ldots, \widehat{\beta}_{n_B}$.

We can then use the bootstrap estimates to either construct confidence intervals or estimate the standard errors.

4.4 Simulation Study

To illustrate the the effectiveness of our method, we design a simple simulation study with 10 time points, i.e., $L = (Y_1, \ldots, Y_{10}) \in \mathbb{R}^{10}$, and we assume that there are three blocks of variables subject to missing. The first block is $b_1 = (Y_2, Y_3)$, the second block is $b_2 = Y_5$ and the third block is $b_3 = (Y_7, Y_8, Y_9)$. The rest variables Y_1, Y_4, Y_6, Y_{10} are always observed. For convenience, we use $R \in \{0,1\}^3$ to denote the missing pattern of L, i.e., $R_j = 1$ if the j-th block is observed. We use L_r to represent the data observed, for example,

$$L_{100} = Y_1, \underbrace{Y_2, Y_3}_{b_1}, Y_4, Y_6, Y_{10} \quad L_{001} = Y_1, Y_4, Y_6, \underbrace{Y_7, Y_8, Y_9}_{b_3}, Y_{10}$$

Next, Let $|b_i|$ be the length of i-th block and define $|r| = \sum_r r_i |b_i| + 4$ as the length of the observed variables. We generate the data as follows:

$$L_r \sim N(\mathbb{1}_{|r|}, \Sigma_{|r|}) + \mu_r$$

where $\Sigma_{|r|} = 1/2 I_{|r|} + 1/2 1_{|r|} 1_{|r|}^T$ and μ_r can be viewed as a location shift for the mean of observed blocks. μ_r does not affect the mean of Y_1, Y_4, Y_6, Y_{10}. For this reason, we only specify μ_r for the corresponding observed blocks. We specify μ_r as follows:

$$\mu_{100} = (0,0) \quad \mu_{010} = (0.5) \quad \mu_{001} = (-0.5, -0.5, -0.5) \quad \mu_{110} = (-0.5, -0.5 \mid 0)$$
$$\mu_{101} = (0.5, 0.5 \mid 0, 0, 0) \quad \mu_{011} = (0.5 \mid 0.5, 0.5, 0.5) \quad \mu_{111} = (0, 0 \mid 0 \mid 0, 0, 0)$$

For example, when $r = 110$, $\mu_{110} = (-0.5, -0.5 \mid 0)$ and blocks b_1 and b_2 are observed. The mean for block 1 would then be $(0.5, 0.5)$ and the mean for block 2 would stay 1. Put

everything together, the mean for L_{110} would then be

$$(1, \underbrace{0.5, 0.5}_{b_1}, 1, \underbrace{\quad 1 \quad}_{b_2}, 1, 1)$$

Next, we assume that the extrapolation density follows the linear model-based BM-ACMV assumption. Under similar assumptions to (4.9), we can compute that for $r_1 \in \{000, 001, 010, 011\}$,

$$Y_2, Y_3 | Y_1, Y_4, R = r_1 \sim N\left(\begin{pmatrix} 1/3 + 1/3Y_1 + 1/3Y_4 \\ 1/3 + 1/3Y_1 + 1/3Y_4 \end{pmatrix}, \begin{pmatrix} 19/24 & 7/24 \\ 7/24 & 19/24 \end{pmatrix}\right)$$

Similarly, we can compute that for $r_2 \in \{000, 001, 100, 101\}$ and $r_3 \in \{000, 100, 010, 110\}$,

$$Y_5 | Y_4, Y_6, R = r_2 \sim N(1/3 + 1/3Y_4 + 1/3Y_6, 19/24)$$

$$Y_7, Y_8, Y_9 | Y_6, Y_{10}, R = r_3 \sim N\left(\begin{pmatrix} 1/3 + 1/3Y_6 + 1/3Y_{10} \\ 1/3 + 1/3Y_6 + 1/3Y_{10} \\ 1/3 + 1/3Y_6 + 1/3Y_{10} \end{pmatrix}, \begin{pmatrix} 19/24 & 7/24 & 7/24 \\ 7/24 & 19/24 & 7/24 \\ 7/24 & 7/24 & 19/24 \end{pmatrix}\right)$$

Finally, we assume that $P(R = r) = 1/8$ for all $r \in \{0, 1\}^3$. We are interested in estimating $\tau_1 = P(Y_1 \leqslant a, \cdots, Y_{10} \leqslant a) = P(\max_i Y_i \leqslant a)$ and the probability of trajectory crossing a, $\tau_2 = P(\max_i Y_i \geqslant a, \min_i Y_i \leqslant a)$ for a fixed constant a. We set $a = 1.8$. The true values $\tau_1 \approx 0.3141$ and $\tau_2 \approx 0.6797$ are computed by a Monte Carlo approximation with $100,000$ observations and 400 repetitions.

In the simulated data, we generate $n = 5,000$ observations. For our model-based approach, we perform multiple imputation 5 times with linear model, kNN and random forest. We also compared our methods with the popular MICE method and complete-case analysis (using only data without any missingness). We use $n_B = 1,000$ bootstrap replicates to estimate the uncertainty and the entire procedure is repeated $1,000$ times. For kNN, we use $k = 50$ and use cross-fitting to estimate the covariance matrix. For the random forest method, we use $n_{\mathsf{tree}} = 50$ trees and set the node size to be 350. We use out-of-bag errors to estimate the covariance matrix for random forest. For the MICE method, we use the Bayesian linear regression model for imputation.

Table 4.1: Simulation results when $n = 5000$ and $M = 5$

Method	Bias		TSE		Avg. Bootstrap SE		Coverage	
	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2
MICE	0.042	-0.045	0.0067	0.0067	0.0068	0.0068	0	0
Complete-case	0.052	-0.056	0.019	0.019	0.019	0.019	0.23	0.18
Linear	0.0003	-0.0003	0.0065	0.0065	0.0065	0.0065	0.96	0.96
kNN	-0.0046	0.0057	0.0067	0.0067	0.0069	0.0069	0.91	0.87
Random Forest	-0.0033	0.0046	0.0067	0.0067	0.0068	0.0068	0.96	0.96

Table 4.1 summarizes the result. Bias is computed as the difference of the average of 1,000 parameter estimates and the true value of τ_1 and τ_2. True standard errors (TSE) are computed by the 1000 repetitions of the experiment; it is an Monte Carlo approximation of the actual standard errors of the estimator. Average bootstrap standard error (Avg. Bootstrap SE) is the average bootstrap standard errors over the 1000 repetitions. Coverage stands for the coverage of 95% confidence intervals.

First, for all methods, bootstrap successfully capture the uncertainty in the sense that the Avg. Bootstrap SE is very close to the TSE. Both MICE and complete-case analysis lead to a biased estimate and a very poor confidence interval. The linear model-based BM-ACMV approach obtains close to zero bias and achieves nominal coverages for the 95% confidence intervals. This is because the true generating process is linear, so the linear model has a superior performance. Finally, both kNN and random forest obtain much better results than MICE and complete-case analysis. However, the confidence intervals of kNN do not achieve the nominal coverage because the biases cannot be ignored and is at a similar scale compared to the standard error estimates. For random forest, the coverages are much better, despite the biases appear to be on the same order as the kNN approach. We include some further simulation results in Appendix C.

4.5 Real data experiments

In this section, we apply our proposed model-based BM-ACMV assumption to the diabetes EHR dataset. The dataset contains 8663 individuals with the maximum number of possible HbA1c measurements being 45. Thus, the length of the HbA1c trajectory is 45. For simplicity, we focus on imputing the first 10 measurements from the trajectory. Even just for the first 10 measurements, we already observe many distinct missing patterns. For the first 10 measurements, the total number of possible missing patterns is $2^{10} = 1024$ and we observe 995 distinct missing patterns in our dataset. Further, there are only 108 individuals who have fully observed HbA1c measurements for the first 10 visits. For this reason, we choose to not restrict the maximum length of missing blocks and impute for all individuals in the diabetes dataset. In the meantime, we also have 154 individuals who do not have any HbA1c measurements.

Again, we are interested in estimating $\tau_1 = P(Y_1 \leqslant a, \cdots, Y_{10} \leqslant a)$ and $\tau_2 = P(\max_i Y_i \geqslant a, \min_i Y_i \leqslant a)$ for $a = 7\%$. Here 7% is an important threshold for diabetes patients as under 7% suggests that an individual has HbA1c successfully controlled. Thus, τ_1 is measuring the proportion of individuals who have their HbA1cs under control for the first 10 measurements. τ_2 is measuring the proportion of individuals who have their HbA1c measurements cross the important level of 7%.

The results are presented in table 4.2. For MICE, we again used Bayesian linear regression for the imputation. For kNN method, we set the number of neighbors to be 5. For the random forest, we set the number of trees to be 50. For the number of multiple imputations, we use $M = 20$. We can see that the results for all model-based BM-ACMV approach are very similar to each other, indicating the linear model fit might be enough in this case. Further, the complete-case analysis leads to a very wide confidence intervals for τ_1 and τ_2 because only 1.2% individuals have complete observations.

All our methods suggest a similar estimate in both parameter τ_1 and τ_2. The complete-case analysis suggests an estimate with a higher value of τ_1 and a lower value of τ_2 compared

Table 4.2: Real data results for HbA1c trajectories of length 10.

Method	Estimate (95% CI)	
	τ_1	τ_2
MICE	0.208 (0.199 - 0.219)	0.673 (0.662 - 0.685)
Complete-case	0.194 (0.127 - 0.272)	0.611 (0.521 - 0.701)
Linear	0.176 (0.163 - 0.191)	0.708 (0.688 - 0.722)
kNN	0.180 (0.171 - 0.197)	0.707 (0.689 - 0.717)
Random Forest	0.175 (0.170 - 0.197)	0.707 (0.682 - 0.712)

to our methods. The MICE leads to a result that τ_1 will be even higher and τ_2 is again smaller. Although there is no true label to tell us which method is the best, this experiment highlights the applicability of our method to an actual scientific problem.

4.6 Discussion

In this chapter, we proposed a block-Markov ACMV assumption for recovering trajectories of longitudinal measurements. For modeling purpose, we modify our assumption to a model-based BM-ACMV assumption that allows the estimation with linear models and flexible nonparametric / machine learning models. Our proposed assumption is nonparametrically identifiable and will not conflict with the real data. Currently we only consider modeling continuous data, but it is possible to further extend our approach to also consider binary, categorical and count data. It is also possible to allow each missing block to depend on more variables. The current procedure of using bootstrap to estimate the uncertainty is a valid, but computationally heavy approach. We might be able to use Rubin's combination rule to reduce the computation time and still obtain valid confidence intervals. We leave all this to future work.

BIBLIOGRAPHY

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. The annals of statistics, pages 1100–1120.

Athey, S., Chetty, R., and Imbens, G. (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. arXiv preprint arXiv:2006.09676.

Baldi, I., Ponti, A., Zanetti, R., Ciccone, G., Merletti, F., and Gregori, D. (2010). The impact of record-linkage bias in the Cox model. Journal of evaluation in clinical practice, 16(1):92–96.

Bilias, Y., Gu, M., and Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. The Annals of Statistics, 25(2):662–682.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. Biometrika, 79(1):139–147.

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I., and Brand, C. A. (2010). Data linkage: a powerful research tool with potential problems. BMC health services research, 10(1):1–7.

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. Scandinavian Journal of Statistics, 34(1):86–102.

Breslow, N. E. and Wellner, J. A. (2008). A z-theorem with estimated nuisance parameters and correction note for weighted likelihood for semiparametric models and two-phase

stratified samples, with application to cox regression. Scandinavian Journal of Statistics, 35(1):186–192.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. Biometrics, 59(3):570–579.

Casey, J. A., Schwartz, B. S., Stewart, W. F., and Adler, N. E. (2016). Using electronic health records for population health research: a review of methods and applications. Annual review of public health, 37.

Chen, Y.-C. (2022). Pattern graphs: a graphical approach to nonmonotone missing data. The Annals of Statistics, 50(1):129–146.

Cheng, G., Chen, Y.-C., Unger, J. M., Till, C., and Zhao, Y.-Q. (2022). Long-term effect estimation when combining clinical trial and observational follow-up datasets. arXiv preprint arXiv:2204.04309.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–202.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002). Analysis of longitudinal data. Oxford university press.

Efron, B. and Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

FDA (2018). Use of electronic health record data in clinical investigations guidance for industry.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328):1183–1210.

Fitzpatrick, T., Perrier, L., Shakik, S., Cairncross, Z., Tricco, A. C., Lix, L., Zwarenstein, M., Rosella, L., and Henry, D. (2018). Assessment of long-term follow-up of randomized trial participants by linkage to routinely collected data: a scoping review and analysis. JAMA network open, 1(8):e186019–e186019.

Ghosal, I. and Hooker, G. (2020). Boosting random forests to reduce bias; one-step boosted forest and its variance estimate. Journal of Computational and Graphical Statistics, 30(2):493–502.

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L.-C., Smith, P., Dibben, C., and Goldstein, H. (2018). Guild: Guidance for information about linking data sets. Journal of Public Health, 40(1):191–198.

Goldstein, H., Harron, K., and Wade, A. (2012). The analysis of record-linked data using multiple imputation with data value priors. Statistics in medicine, 31(28):3481–3493.

Han, Q. and Wellner, J. A. (2021). Complex sampling designs: Uniform limit theorems and applications. The Annals of Statistics, 49(1):459–485.

Han, Y. and Lahiri, P. (2019). Statistical analysis with linked data. International Statistical Review, 87:S139–S157.

Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B., and Goldstein, H. (2014). Evaluating bias due to data linkage error in electronic healthcare records. BMC medical research methodology, 14(1):1–10.

Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. Epidemiology, pages 561–570.

Hershman, D. L., Unger, J. M., Wright, J. D., Ramsey, S., Till, C., Tangen, C. M., Barlow, W. E., Blanke, C., Thompson, I. M., and Hussain, M. (2016). Adverse health events

following intermittent and continuous androgen deprivation in patients with metastatic prostate cancer. JAMA oncology, 2(4):453–461.

Hof, M. H., Ravelli, A. C., and Zwinderman, A. H. (2017). A probabilistic record linkage model for survival data. Journal of the American Statistical Association, 112(520):1504–1515.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association, 47(260):663–685.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. Biometrika, 88(2):551–564.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning, volume 112. Springer.

Kallus, N. and Mao, X. (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. arXiv preprint arXiv:2003.12408.

Kim, G. and Chambers, R. (2012). Regression analysis under incomplete linkage. Computational Statistics & Data Analysis, 56(9):2756–2770.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. Journal of the American Statistical Association, 106(493):157–165.

Kim, J. S. (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2):489–502.

Lahiri, P. and Larsen, M. D. (2005). Regression analysis with linked data. Journal of the American statistical association, 100(469):222–230.

Li, L., Shen, C., Li, X., and Robins, J. M. (2013). On weighting approaches for missing data. Statistical methods in medical research, 22(1):14–30.

Liang, K.-Y., Self, S. G., and Liu, X. (1990). The cox proportional hazards model with change point: An epidemiologic application. Biometrics, pages 783–793.

Lin, D. (2000). On fitting cox's proportional hazards models to survey data. Biometrika, 87(1):37–47.

Lin, D. Y. and Wei, L.-J. (1989). The robust inference for the cox proportional hazards model. Journal of the American statistical Association, 84(408):1074–1078.

Linero, A. R. (2017). Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness. Biometrika, 104(2):327–341.

Little, R. (1995). Modeling the drop-out mechanism in longitudinal studies. Journal of the American Statistical Association, 90(1):1.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association, 88(421):125–134.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., et al. (2012). The prevention and treatment of missing data in clinical trials. New England Journal of Medicine, 367(14):1355–1360.

Little, R. J. and Rubin, D. B. (1989). The analysis of social science data with missing values. Sociological Methods & Research, 18(2-3):292–326.

Little, R. J. and Rubin, D. B. (2019). Statistical analysis with missing data, volume 793. John Wiley & Sons.

Llewellyn-Bennett, R., Bowman, L., and Bulbulia, R. (2016). Post-trial follow-up methodology in large randomized controlled trials: a systematic review protocol. Systematic reviews, 5(1):1–7.

Malinsky, D., Shpitser, I., and Tchetgen Tchetgen, E. J. (2021). Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. Journal of the American Statistical Association, pages 1–9.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical Science, 9(4):538–558.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. Journal of the American Statistical Association, 116(534):1023–1037.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). Handbook of missing data methodology. CRC Press.

Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. J. (1998). Monotone missing data and pattern-mixture models. Statistica Neerlandica, 52(2):153–161.

Nabi, R., Bhattacharya, R., and Shpitser, I. (2020). Full law identification in graphical models of missing data: Completeness results. In International Conference on Machine Learning, pages 7153–7163. PMLR.

Neter, J., Maynes, E. S., and Ramanathan, R. (1965). The effect of mismatching on the measurement of response errors. Journal of the American Statistical Association, 60(312):1005–1027.

Padmanabhan, S., Carty, L., Cameron, E., Ghosh, R. E., Williams, R., and Strongman, H. (2019). Approach to record linkage of primary care data from clinical practice research datalink to other health-related patient data: overview and implications. European journal of epidemiology, 34(1):91–99.

Pons, O. et al. (2003). Estimation in a cox regression model with a change-point according to a threshold in a covariate. Annals of Statistics, 31(2):442–463.

Qi, L., Wang, C., and Prentice, R. L. (2005). Weighted estimators for proportional hazards regression with missing covariates. Journal of the American Statistical Association, 100(472):1250–1263.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., Solenberger, P., et al. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey methodology, 27(1):85–96.

Robins, J. M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In Proceedings of the Biopharmaceutical Section, American Statistical Association, volume 24, page 3. San Francisco CA.

Robins, J. M. and Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. Biometrics, 56(3):779–788.

Robins, J. M. and Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. Statistics in medicine, 16(1):39–56.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In AIDS epidemiology, pages 297–331. Springer.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Statistical models in epidemiology, the environment, and clinical trials, pages 1–94. Springer.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association, 89(427):846–866.

Rosenman, E., Basse, G., Owen, A., and Baiocchi, M. (2020). Combining observational and experimental datasets using shrinkage estimators. arXiv preprint arXiv:2002.06708.

Rosenman, E., Owen, A. B., Baiocchi, M., and Banack, H. (2018). Propensity score methods for merging observational and experimental datasets. arXiv preprint arXiv:1804.07863.

Roth, J. A., Etzioni, R., Waters, T. M., Pettinger, M., Rossouw, J. E., Anderson, G. L., Chlebowski, R. T., Manson, J. E., Hlatky, M., Johnson, K. C., et al. (2014). Economic return from the women's health initiative estrogen plus progestin clinical trial: a modeling study. Annals of internal medicine, 160(9):594–602.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3):581–592.

Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys, volume 81. John Wiley & Sons.

Sadinle, M. and Reiter, J. P. (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. Biometrika, 104(1):207–220.

Saegusa, T. and Wellner, J. A. (2013). Weighted likelihood estimation under two-phase sampling. Annals of statistics, 41(1):269.

Scheuren, F. and Winkler, W. E. (1993). Regression analysis of data files that are computer matched. Survey Methodology, 19(1):39–58.

Scheuren, F. and Winkler, W. E. (1997). Regression analysis of data files that are computer matched–part ii. Survey Methodology, 23(2):126–138.

Sellke, T. (1982). Large sample theory for sequential analysis of the proportional hazards model. Technical report, STANFORD UNIV CA DEPT OF STATISTICS.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. Biometrika, 70(2):315–326.

Shao, J. and Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. Biometrika, 103(1):175–187.

Shi, X., Li, X., and Cai, T. (2020). Spherical regression under mismatch corruption with application to automated knowledge translation. Journal of the American Statistical Association, pages 1–12.

Shpitser, I. (2016). Consistent estimation of functions of data missing non-monotonically and not at random. Advances in Neural Information Processing Systems, 29.

Sun, B. and Tchetgen Tchetgen, E. J. (2018). On inverse probability weighting for non-monotone missing at random data. Journal of the American Statistical Association, 113(521):369–379.

Tchetgen, E. J. T., Wang, L., and Sun, B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. Statistica Sinica, 28(4):2069.

Thompson, I. M., Goodman, P. J., Tangen, C. M., Lucia, M. S., Miller, G. J., Ford, L. G., Lieber, M. M., Cespedes, R. D., Atkins, J. N., Lippman, S. M., et al. (2003). The influence of finasteride on the development of prostate cancer. New England journal of medicine, 349(3):215–224.

Thompson, I. M., Tangen, C. M., Klein, E. A., and Lippman, S. M. (2005). Phase iii prostate cancer prevention trials: are the costs justified? Journal of clinical oncology, 23(32):8161–8164.

Troxel, A. B., Harrington, D. P., and Lipsitz, S. R. (1998a). Analysis of longitudinal data with non-ignorable non-monotone missing values. Journal of the Royal Statistical Society: Series C (Applied Statistics), 47(3):425–438.

Troxel, A. B., Lipsitz, S. R., and Harrington, D. P. (1998b). Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. Biometrika, 85(3):661–672.

Tsiatis, A. (2007). Semiparametric theory and missing data. Springer Science & Business Media.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society: Series B (Methodological), 38(3):290–295.

Unger, J. M., Hershman, D. L., Till, C., Tangen, C. M., Barlow, W. E., Ramsey, S. D., Goodman, P. J., and Thompson Jr, I. M. (2018). Using medicare claims to examine long-term prostate cancer risk of finasteride in the prostate cancer prevention trial. JNCI: Journal of the National Cancer Institute, 110(11):1208–1215.

Unger, J. M., Till, C., Thompson, I. M., Tangen, C. M., Goodman, P. J., Wright, J. D., Barlow, W. E., Ramsey, S. D., Minasian, L. M., and Hershman, D. L. (2016). Long-term consequences of finasteride vs placebo in the prostate cancer prevention trial. JNCI: Journal of the National Cancer Institute, 108(12).

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research, 16(3):219–242.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. Journal of statistical software, 45:1–67.

Van der Laan, M. J., Laan, M., and Robins, J. M. (2003). Unified methods for censored longitudinal data and causality. Springer Science & Business Media.

Van Der Vaart, A. and Wellner, J. A. (2000). Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In High dimensional probability II, pages 115–133. Springer.

Van Der Vaart, A. W. (2000). Asymptotic statistics, volume 3. Cambridge university press.

Van Der Vaart, A. W. (2002). Semiparametric statistics. Lecture Notes in Math., (1781).

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In Weak convergence and empirical processes, pages 16–28. Springer.

Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. Biometrika, 94(4):841–860.

Warren, J. L., Klabunde, C. N., Schrag, D., Bach, P. B., and Riley, G. F. (2002). Overview of the seer-medicare data: content, research applications, and generalizability to the united states elderly population. Medical care, pages IV3–IV18.

Zhao, P., Tang, N., Qu, A., and Jiang, D. (2017). Semiparametric estimating equations inference with nonignorable missing data. Statistica Sinica, pages 89–113.

Zhou, Y., Little, R. J., and Kalbfleisch, J. D. (2010). Block-conditional missing at random models for missing data. Statistical Science, 25(4):517–532.

Appendix A

APPENDIX OF CHAPTER 2

A.1 Derivation of IPLW partial score

Assume that there are n_{00} observations with $L = 0$ and $Q = 0$ (Class 3 of Figure 2.1) and define weight $\widehat{w}_i = \frac{I(L_i+Q_i>0)}{Q_i+(1-Q_i)\pi_{\widehat{\gamma}_n}(\mathbf{X}_i)}$. Now we are ready to derive the IPLW partial likelihood for estimating $\boldsymbol{\beta}_0^*$. Assume that $\widetilde{T}_{(1)} < \widetilde{T}_{(2)} < \cdots < \widetilde{T}_{(n-n_{00})}$ are the ordered $\widetilde{T}_i$'s, and $\mathbf{X}_{(i)}, L_{(i)}, Q_{(i)}, \Delta_{(i)}$ are the corresponding covariates, linkage indicator, in-trial censoring indicator and censoring indicator. Denote $\Lambda_0(t)$ as the cumulative hazard function. Let $h_i = d\Lambda_0(\widetilde{T}_{(i)}) = \Lambda_0(\widetilde{T}_{(i)}) - \Lambda_0(\widetilde{T}^-_{(i)})$ and $\Lambda_0(\widetilde{T}_{(i)}) = \sum_{j\leqslant i} h_j$. It is known(Van Der Vaart, 2000) that maximization with respect to λ_0 can be done by maximizing $\boldsymbol{h} = (h_1, \ldots, h_{n-n_{00}})$ and thus we only need to consider the case where $l_n(\boldsymbol{\beta}, \lambda_0) = l_n(\boldsymbol{\beta}, \mathbf{h})$. As a result, the IPLW log-likelihood can be rewritten as:

$$l_n(\boldsymbol{\beta}, \lambda_0) = l_n(\boldsymbol{\beta}, \mathbf{h}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_{(i)} \left[\Delta_{(i)} \log h_i + \Delta_{(i)} \mathbf{X}_{(i)}^T \boldsymbol{\beta} - \exp(\mathbf{X}_{(i)}^T \boldsymbol{\beta}) \sum_{j\leqslant i} h_j \right].$$

Thus, maximizing this with respect to $\boldsymbol{h} = (h_1, \ldots, h_{n-n_{00}})$ leads to

$$\widehat{h}_i = \frac{\Delta_{(i)}\widehat{w}_{(i)}}{\sum_{j\geqslant i} \widehat{w}_{(j)} \exp(\mathbf{X}_{(j)}^T \boldsymbol{\beta})}$$

Take $\widehat{\boldsymbol{h}} = (\widehat{h}_1, \cdots, \widehat{h}_{n-n_{00}})$ back to the empirical log-likelihood, we obtain the IPLW partial log-likelihood

$$\mathcal{L}_n(\boldsymbol{\beta}) = l_n(\boldsymbol{\beta}, \widehat{\boldsymbol{h}}) = \frac{1}{n}\sum_{i=1}^{n} \Delta_{(i)}\widehat{w}_{(i)} \left(\mathbf{X}_{(i)}^T \boldsymbol{\beta} - \log \left(\sum_{j\geqslant i} \widehat{w}_{(j)} \exp(\mathbf{X}_{(j)}^T \boldsymbol{\beta}) \right) \right)$$

which can be further simplified as

$$\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \widehat{w}_i \left(\mathbf{X}_i^T \boldsymbol{\beta} - \log \left(\sum_{j=1}^{n} I(\widetilde{T}_j \geqslant \widetilde{T}_i) \widehat{w}_j \exp(\mathbf{X}_j^T \boldsymbol{\beta}) \right) \right)$$

and then the partial score is

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \widehat{w}_i \left(\mathbf{X}_i - \frac{\sum_{j=1}^{n} \widehat{w}_j I(\widetilde{T}_j \geqslant \widetilde{T}_i) \exp(\mathbf{X}_j^T \boldsymbol{\beta}) \mathbf{X}_j}{\sum_{j=1}^{n} \widehat{w}_j I(\widetilde{T}_j \geqslant \widetilde{T}_i) \exp(\mathbf{X}_j^T \boldsymbol{\beta})} \right), \tag{A.1}$$

To derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n$, note that

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) - \widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0^*) = \left. \frac{\partial \widehat{\mathbf{U}}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0^*} (\boldsymbol{\beta} - \boldsymbol{\beta}_0^*) + o_P(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0^*\|)$$

Thus, choosing $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_n$ leads to

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0^*) \approx \left(- \left. \frac{\partial \widehat{\mathbf{U}}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0^*} \right)^{-1} n^{1/2} \widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0^*)$$

As a result, we just need to prove that $- \left. \frac{\partial \widehat{\mathbf{U}}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0^*} \to_p - \left. \frac{\partial \mathbf{U}_0(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0^*} = \boldsymbol{\Sigma}_0$ and $n^{-1/2}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0^*)$ converges to a normal distribution. We prove the second convergence by showing that $n^{1/2}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0^*)$ has a weighted asymptotically linear expansion: $n^{1/2}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0^*) = n^{-1/2}\sum_{i=1}^{n} \widehat{w}_i \mathbf{U}_i(\boldsymbol{\beta}_0^*) + o_p(1)$ with

$$\begin{aligned}
&\mathbf{U}_i(\boldsymbol{\beta}_0^*) \\
&= \int_0^{\tau_2} \left[\mathbf{X}_i(t) - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}_0^*, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}_0^*, t)} \right] dN_i(t) - \int_0^{\tau_2} \frac{Y_i(t) \exp(\boldsymbol{\beta}_0^{*T} \mathbf{X}_i(t))}{\mathbf{s}^{(0)}(\boldsymbol{\beta}_0^*, t)} \left[\mathbf{X}_i(t) - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}_0^*, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}_0^*, t)} \right] d\widetilde{N}(t)
\end{aligned}$$

and $\widetilde{N}(t) = \mathbb{E}[N(t)]$. Similar asymptotic linear expansions have appeared in Lin and Wei (1989) and Lin (2000). This weighted asymptotic linear expansion motivates the study of the IPLW empirical measure and processes.

We start by giving the weighted asymptotic linear expansion of the IPLW partial score $\widehat{\mathbf{U}}_n(\boldsymbol{\beta})$ and subsequently give its asymptotic distribution.

Theorem A.1.1 (Asymptotic linear expansion)**.** *Under assumptions (D1) - (D4), we have the following two results:*

1. *For each $\boldsymbol{\beta}$, $n^{1/2}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = n^{-1/2}\sum_{i=1}^{n} \widehat{w}_i \mathbf{U}_i(\boldsymbol{\beta}) + o_p(1)$ such that*

$$\begin{aligned}
&\mathbf{U}_i(\boldsymbol{\beta}) = \\
&\int_0^{\tau_2} \left[\mathbf{X}_i(t) - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)} \right] dN_i(t) - \int_0^{\tau_2} \frac{Y_i(t) \exp(\boldsymbol{\beta}^T \mathbf{X}_i(t))}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)} \left[\mathbf{X}_i(t) - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)} \right] d\widetilde{N}(t)
\end{aligned}$$

with $\widetilde{N}(t) = \mathbb{E}[N(t)]$.

2. $n^{-1/2}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0^*) \to_d N(0, \boldsymbol{\Sigma}_U)$ *with*

$$\boldsymbol{\Sigma}_U = \mathrm{Var}[\mathbf{U}_1(\boldsymbol{\beta}_0^*)] + \mathbb{E}\left[\mathbf{U}_1(\boldsymbol{\beta}_0^*)\mathbf{U}_1(\boldsymbol{\beta}_0^*)^T \frac{I(Q=0)[1-\pi_{\boldsymbol{\gamma}_0}(\mathbf{Z}_1)]}{\pi_{\boldsymbol{\gamma}_0}(\mathbf{Z}_1)}\right] - \mathbf{Q}_e(\mathbf{U}_1(\boldsymbol{\beta}_0^*))^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma}_0)\mathbf{Q}_e(\mathbf{U}_1(\boldsymbol{\beta}_0^*)),$$

where $\mathbf{Q}_e(\mathbf{U}_1(\boldsymbol{\beta}_0^*)) = \mathbb{P}_0[I(Q=0)(1-\pi_{\boldsymbol{\gamma}_0}(\mathbf{Z}_1))\widetilde{\mathbf{Z}}_1\mathbf{U}_1(\boldsymbol{\beta}_0^*)^T]$.

The proof of Theorem *A*.1.1 can be found in appendix A.3. We first prove that the IPLW partial score can be written in the weighted asymptotic linear expansion form (first assertion). Then we use the weak convergence result of the IPLW empirical process from Proposition *A*.2.2 to obtain the asymptotic distribution of $n^{1/2}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})$. Next recall $-\frac{\partial \widehat{\mathbf{U}}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{A}_n(\boldsymbol{\beta})$ and we can write $\mathbf{A}_n(\boldsymbol{\beta})$ equivalently as

$$\mathbf{A}_n(\boldsymbol{\beta}) = -\frac{\partial \widehat{\mathbf{U}}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \int_0^{\tau_2} \left\{ \frac{\mathbf{S}_{n,w}^{(2)}(\boldsymbol{\beta}, t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta}, t)} - \left(\frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta}, t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta}, t)}\right)^{\otimes 2} \right\} d\bar{N}(t) \tag{A.2}$$

and similarly define its population version $\mathbf{A}(\boldsymbol{\beta})$

$$\begin{aligned} \mathbf{A}(\boldsymbol{\beta}) = -\frac{\partial \mathbf{U}_0(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \int_0^{\tau_2} \left\{ \frac{\mathbf{s}^{(2)}(\boldsymbol{\beta}, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)} - \left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)}\right)^{\otimes 2} \right\} d\widetilde{N}(t) \\ &= \int_0^{\tau_2} \left\{ \frac{\mathbf{s}^{(2)}(\boldsymbol{\beta}, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)} - \left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)}\right)^{\otimes 2} \right\} \mathbf{s}^{(0)}(t) dt \end{aligned} \tag{A.3}$$

Under assumption (D5), $\mathbf{A}(\boldsymbol{\beta}_0^*) = \boldsymbol{\Sigma}_0$ is positive definite and we later prove that $\mathbf{A}_n(\boldsymbol{\beta}) \to_p \mathbf{A}(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ in a compact set $\mathbb{B}$ that contains $\boldsymbol{\beta}_0^*$. Together with the asymptotic normality of the IPLW partial score, we can derive the asymptotic normality of the estimator $\widehat{\boldsymbol{\beta}}_n$.

A.2 Empirical process theory for an IPLW process

To study the asymptotic properties of $\widehat{\boldsymbol{\beta}}_n$, we need to generalize empirical process theory into an IPW scenario. We first introduce an IPLW empirical measure

$$\mathbb{P}_n^{\pi} = \frac{1}{n}\sum_{i=1}^{n} \frac{I(L_i + Q_i > 0)}{Q_i + (1-Q_i)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}_i)} \delta_{\mathbf{X}_i, \widetilde{T}_i, \Delta_i} = \frac{1}{n}\sum_{i=1}^{n} w_i \delta_{\mathbf{X}_i, \widetilde{T}_i, \Delta_i}$$

where $\delta_{\mathbf{X}_i,\widetilde{T}_i,\Delta_i}$ is the Dirac measure placing unit mass on $(\mathbf{X}_i, \widetilde{T}_i, \Delta_i)$ and $w_i = \frac{I(L_i+Q_i>0)}{Q_i+(1-Q_i)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}_i)}$ such that $\mathbb{P}_n^{\pi} f = \frac{1}{n}\sum_{i=1}^n w_i f(\mathbf{X}_i, \widetilde{T}_i, \Delta_i)$ for a function $f = f(\mathbf{x}, \widetilde{t}, \delta)$. In practice, $\pi_{\boldsymbol{\gamma}_0}(X_i)$ is unknown, so we introduce the IPLW empirical measure with estimated weight

$$\mathbb{P}_n^{\pi,e} = \frac{1}{n}\sum_{i=1}^n \frac{I(L_i+Q_i>0)}{Q_i+(1-Q_i)\pi_{\widehat{\boldsymbol{\gamma}}_n}(\mathbf{X}_i)}\delta_{\mathbf{X}_i,\widetilde{T}_i,\Delta_i} = \frac{1}{n}\sum_{i=1}^n \widehat{w}_i\delta_{\mathbf{X}_i,\widetilde{T}_i,\Delta_i}$$

with $\boldsymbol{\gamma}_0$ replaced by $\widehat{\boldsymbol{\gamma}}_n$. Finally, we denote $\mathbb{P}_0$ as the probability measure corresponding to the true distribution such that $\mathbb{P}_0 f = \mathbb{E}[f(\mathbf{X}, \widetilde{T}, \Delta)]$. Note the usual empirical measure $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{X}_i,\widetilde{T}_i,\Delta_i}$ is unobserved due to the missingness in Class 3 of Figure 2.1. The IPLW empirical measure leads to the IPLW empirical processes $\mathbb{G}_n^{\pi} = \sqrt{n}(\mathbb{P}_n^{\pi} - \mathbb{P}_0)$ and $\mathbb{G}_n^{\pi,e} = \sqrt{n}(\mathbb{P}_n^{\pi,e} - \mathbb{P}_0)$. It turns out that our IPLW empirical measure and empirical process also enjoy similar asymptotic properties as the usual empirical measure and empirical processes. For any $\phi : \mathcal{F} \to \mathbb{R}$, we write $\|\phi(f)\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\phi(f)|$. We say that $\mathcal{F}$ is $\mathbb{P}$-Glivenko-Cantelli if and only if $\sup_{f\in\mathcal{F}} |(\mathbb{P}_n - \mathbb{P})f| = o_P(1)$. We first prove that the Glivenko-Cantelli property also holds for the IPLW empirical process.

Proposition A.2.1 (IPLW uniform convergence)**.** *Suppose that $\mathcal{F} = \{f(\mathbf{x}, \widetilde{t}, \delta)\}$ is $\mathbb{P}_0$-Glivenko-Cantelli with an integrable envelope function F such that $\mathbb{P}_0 F < \infty$. Under assumption (A1-3),*

$$\|\mathbb{P}_n^{\pi} - \mathbb{P}_0\|_{\mathcal{F}} \to_{P*} 0.$$

If $\widehat{\boldsymbol{\gamma}}_n \to_p \boldsymbol{\gamma}_0$, then $\|\mathbb{P}_n^{\pi,e} - \mathbb{P}_0\|_{\mathcal{F}} \to_{P*} 0$ also holds.

The proof can be found in Appendix A.3. Recall that $\widetilde{\mathbf{X}} = (1, \mathbf{X}^T)^T$ and $\widetilde{\mathbf{x}} = (1, \mathbf{x}^T)^T$. Now, to definite weak convergence, let X_n be a bounded process and X be a bounded process whose finite-dimensional laws correspond to the finite dimensional projections of a tight Borel law on $\ell^\infty(\mathcal{F})$. We say that $X_n \rightsquigarrow X$ in $\ell^\infty(\mathcal{F})$ if and only if $\mathbb{E}^* H(X_n) \to \mathbb{E}H(X)$ for all $H \in C_b(\ell^\infty(\mathcal{F}))$, where $C_b(\ell^\infty(\mathcal{F}))$ denotes all bounded continuous functions on $\ell^\infty(\mathcal{F})$ (Van Der Vaart and Wellner, 1996; Van Der Vaart, 2000). The next theorem states the weak convergence result for the IPLW empirical process.

Proposition A.2.2 (IPLW weak convergence)**.** *Under assumption (A1-3), suppose that* $\mathcal{F} = \{f(\mathbf{x}, \widetilde{t}, \delta\}$ *is* $\mathbb{P}_0$*-Donsker with an integrable envelope function* F *such that* $\mathbb{P}_0 F < \infty$*, then*

$$\mathbb{G}_n^{\pi} \rightsquigarrow \mathbb{G}(g_1 \cdot)$$
$$\mathbb{G}_n^{\pi,e} \rightsquigarrow \mathbb{G}^e = \mathbb{G}(g_1 \cdot - g_2 \mathbf{Q}_e(\cdot)^T \mathbf{g}_3)$$

in $l^\infty(\mathcal{F})$ *where* $g_1(l, q, \mathbf{x}) = \frac{I(l+q>0)}{q+(1-q)\pi_{\boldsymbol{\gamma}_0}(\mathbf{x})}$*,* $g_2(l, q, \mathbf{x}) = I(q = 0)[l - \pi_{\boldsymbol{\gamma}_0}(\mathbf{x})]$*,* $\mathbf{g}_3(\mathbf{x}) = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0}^{-1} \widetilde{\mathbf{x}}$ *and* $\mathbf{Q}_e(f) = \mathbb{E}[I(Q = 0)(1 - \pi_{\boldsymbol{\gamma}_0}(\mathbf{X}))f(\mathbf{X}, \widetilde{T}, \Delta)\widetilde{\mathbf{X}}]$*.* $\mathbb{G}$ *is the* $\mathbb{P}_0$*-Brownian bridge process, indexed by* $\mathcal{F}$*.*

A.3 Proofs

Proof of Proposition 2.3.1. For simplicity, denote $P(L = 1|Q = 0, \mathbf{X}) = \pi_0(\mathbf{X})$. First, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{I(L=1)l(\boldsymbol{\beta}, \lambda_0)}{\pi_0(\mathbf{X})}\middle| Q = 0\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{I(L=1)l(\boldsymbol{\beta}, \lambda_0)}{\pi_0(\mathbf{X})}\middle| Q = 0, \mathbf{X}, \widetilde{T}, \Delta\right]\middle| Q = 0\right] \\
&= \mathbb{E}\left[\frac{l(\boldsymbol{\beta}, \lambda_0)}{\pi_0(\mathbf{X})}\mathbb{E}\left[I(L=1)\middle| Q = 0, \mathbf{X}, \widetilde{T}, \Delta\right]\middle| Q = 0\right] \\
&= \mathbb{E}\left[l(\boldsymbol{\beta}, \lambda_0)|Q = 0\right]
\end{aligned}$$

The second to last equality holds as $l(\beta, \lambda_0)$ is a function of $\widetilde{T}, \mathbf{X}, \Delta$. The last equality holds by assumption (A1). To prove (2.6), we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{I(L+Q>0)}{Q+(1-Q)\pi_0(\mathbf{X})}l(\boldsymbol{\beta}, \lambda_0)\right] = \mathbb{E}\left[\frac{I(L=1)}{\pi_0(\mathbf{X})}l(\boldsymbol{\beta}, \lambda_0)\middle| Q = 0\right] P(Q = 0) \\
&+ \mathbb{E}\left[l(\boldsymbol{\beta}, \lambda_0)|Q = 1\right] P(Q = 1) \\
&= \mathbb{E}\left[l(\boldsymbol{\beta}, \lambda_0)|Q = 0\right] P(Q = 0) + \mathbb{E}\left[l(\boldsymbol{\beta}, \lambda_0)|Q = 1\right] P(Q = 1) = \mathbb{E}(l(\boldsymbol{\beta}, \lambda_0))
\end{aligned}$$

□

We first give the asymptotic distribution of $\widehat{\boldsymbol{\gamma}}_n$, the estimates of the logistic regression parameter $\boldsymbol{\gamma}_0$.

Lemma A.3.1. *Under assumptions (A1), (A2) and (A3), $\widehat{\boldsymbol{\gamma}}_n$ is consistent for $\boldsymbol{\gamma}_0$ and*

$$\sqrt{n}(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) \to_d N(0, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0}^{-1})$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0} = \mathbb{E}\left[I(Q=0)\frac{\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma}_0)\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T}{(1+\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma}_0))^2}\right] = \mathbb{E}[I(Q=0)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})(1-\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}))\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T]$.

Proof of Lemma *A*.3.1. First, the score for logistic regression converges as

$$\mathbf{S}_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} I(Q_i=0)\left[L_i\widetilde{\mathbf{X}}_i - \frac{\exp(\widetilde{\mathbf{X}}_i^T\boldsymbol{\gamma})\widetilde{\mathbf{X}}_i}{1+\exp(\widetilde{\mathbf{X}}_i^T\boldsymbol{\gamma})}\right] \to_p$$
$$\mathbb{E}\left[I(Q=0)\left(L\widetilde{\mathbf{X}} - \frac{\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma})\widetilde{\mathbf{X}}}{1+\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma})}\right)\right] = \mathbf{S}(\boldsymbol{\gamma})$$

and

$$\begin{aligned}\mathbf{S}(\boldsymbol{\gamma}_0) &= \mathbb{E}\left[I(Q=0)\left(L\widetilde{\mathbf{X}} - \frac{\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma}_0)\widetilde{\mathbf{X}}}{1+\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma}_0)}\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[I(Q=0)\left(L\widetilde{\mathbf{X}} - \mathbb{P}(L=1|Q=0;\mathbf{X})\widetilde{\mathbf{X}}\right)|Q,\mathbf{X}\right]\right] \\ &= \mathbb{E}\left(I(Q=0)\left[\mathbb{E}[L|Q,\mathbf{X}]\widetilde{\mathbf{X}} - \mathbb{P}(L=1|Q=0,\mathbf{X})\widetilde{\mathbf{X}}\right]\right) = 0\end{aligned}$$

since $\mathbb{E}[L|Q,\mathbf{X}] = I(Q=0)\mathbb{P}(L=1|Q=0;\mathbf{X}) + I(Q=1)\mathbb{P}(L=1|Q=1;\mathbf{X})$. Further, by the law of large numbers,

$$\nabla\mathbf{S}_n(\boldsymbol{\gamma}) = -\frac{1}{n}\sum_{i=1}^{n} I(Q_i=0)\frac{\exp(\widetilde{\mathbf{X}}_i^T\boldsymbol{\gamma})\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T}{\left[1+\exp(\widetilde{\mathbf{X}}_i^T\boldsymbol{\gamma})\right]^2}$$
$$\to_p -\mathbb{E}\left[I(Q=0)\frac{\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma})\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T}{(1+\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma}))^2}\right] = -\boldsymbol{\Sigma}(\boldsymbol{\gamma})$$

We assume that $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$ is positive-definite, thus $\boldsymbol{\gamma}_0$ is the unique solution for $\mathbf{S}(\boldsymbol{\gamma}) = \mathbf{0}$. Next we verify the uniform convergence condition:

$$\sup_{\boldsymbol{\gamma}} \|\mathbf{S}_n(\boldsymbol{\gamma}) - \mathbf{S}(\boldsymbol{\gamma})\| \to_p 0$$

Denote $\boldsymbol{\phi}_{\boldsymbol{\gamma}}(l,q,\mathbf{x}) = I(q=0)\left[l - \pi_{\boldsymbol{\gamma}}(\mathbf{x})\right]\widetilde{\mathbf{x}}$ and $\mathbf{S}_n(\boldsymbol{\gamma}) - \mathbf{S}(\boldsymbol{\gamma}) = (\mathbb{P}_n - \mathbb{P}_0)\boldsymbol{\phi}_{\boldsymbol{\gamma}}$. The function class $\{\boldsymbol{\phi}_{\boldsymbol{\gamma}}(l,q,\mathbf{x}) : \boldsymbol{\gamma}\}$ forms a VC-subgraph class by Lemma 2.6.15, 2.6.18 of Van Der Vaart and Wellner (1996). Thus, by Theorem 5.9 of Van Der Vaart (2000), we have $\widehat{\boldsymbol{\gamma}}_n \to_p \boldsymbol{\gamma}_0$.

For asymptotic normality of $\widehat{\boldsymbol{\gamma}}_n$, note that

$$\|\boldsymbol{\phi}_{\boldsymbol{\gamma}_1}(L,Q,\mathbf{X}) - \boldsymbol{\phi}_{\boldsymbol{\gamma}_2}(L,Q,\mathbf{X})\| \leqslant \left\| \left[\frac{\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma}_1)}{1+\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma}_1)} - \frac{\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma}_2)}{1+\exp(\widetilde{\mathbf{X}}^T\boldsymbol{\gamma}_2)} \right] \widetilde{\mathbf{X}} \right\|$$
$$\leqslant \frac{1}{4}\|\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\|\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|$$

Under the assumption that $\mathbf{X}$ is bounded, we have $\mathbb{E}\|\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\| < \infty$. $\mathbb{E}\boldsymbol{\phi}_{\boldsymbol{\gamma}}(L,Q,\mathbf{X})$ is differentiable at $\boldsymbol{\gamma}_0$ with derivative $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0}$. By Theorem 5.21 of Van Der Vaart (2000), we conclude that

$$\sqrt{n}(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\phi}_{\boldsymbol{\gamma}_0}(L_i, Q_i, \mathbf{X}_i) + o_P(1) \to_d N(0, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0}^{-1})$$

□

Proof of Proposition *A*.2.1. We start with bounding the difference between the IPLW empirical measure $\mathbb{P}_n^{\pi}$ and the usual empirical measure $\mathbb{P}_n$. By triangle inequality,

$$\|\mathbb{P}_n^{\pi} - \mathbb{P}_0\|_{\mathcal{F}} \leqslant \|\mathbb{P}_n - \mathbb{P}_0\|_{\mathcal{F}} + \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\frac{I(L_i + Q_i > 0)}{Q_i + (1-Q_i)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}_i)} - 1 \right) \delta_{\mathbf{X}_i, \widetilde{T}_i, \Delta_i} \right\|_{\mathcal{F}}$$

The first term is $o_{P^*}(1)$ since $\mathcal{F}$ is Glivenko-Cantelli. For the second term, define function $g(l,q,\mathbf{x}) = \frac{I(l+q>0)}{q+(1-q)\pi_{\boldsymbol{\gamma}_0}(\mathbf{x})} - 1$, then consider the function class $\mathcal{F}^* = \left\{ [g \cdot f](l,q,\mathbf{x},\delta,\widetilde{t}) : f \in \mathcal{F} \right\}$ where $f = f(\mathbf{x},\widetilde{t},\delta)$. By assumption (A3), $g(l,q,\mathbf{x})$ is bounded and $\mathcal{F}$ has an integrable envelope function F. These two together imply that $\mathcal{F}^*$ is $\mathbb{P}_0$-Glivenko-Cantelli by the Glivenko-Cantelli Preservation theorem (Van Der Vaart and Wellner, 2000). Then, for any $f \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{P}_0 g f &= \mathbb{E}\left[\mathbb{E}[g(L,Q,\mathbf{X}) f(\mathbf{X},\widetilde{T},\Delta)|Q] \right] \\
&= \mathbb{E}[g(L,Q,\mathbf{X}) f(\mathbf{X},\widetilde{T},\Delta)|Q=0]\mathbb{P}(Q=0) + \\
&\mathbb{E}[g(L,Q,\mathbf{X}) f(\mathbf{X},\widetilde{T},\Delta)|Q=1]\mathbb{P}(Q=1) \\
&= \mathbb{E}[0 * f(\mathbf{X},\widetilde{T},\Delta)|Q=1]\mathbb{P}(Q=1) + \\
&\mathbb{E}\left[\left(\frac{I(L=1)}{\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})} - 1 \right) f(\mathbf{X},\widetilde{T},\Delta) \middle| Q=0 \right] \mathbb{P}(Q=0) \\
&= \mathbb{E}\left[f(\mathbf{X},\widetilde{T},\Delta) \mathbb{E}\left(\frac{I(L=1)}{\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})} - 1 \middle| Q=0, \mathbf{X}, \widetilde{T}, \Delta \right) \middle| Q=0 \right] \mathbb{P}(Q=0) = 0
\end{aligned}$$

The last equality is due to the assumption (A1). Thus, the second term could be rewritten as

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{X}_i, \widetilde{T}_i, \Delta_i, L_i, Q_i} \right\|_{\mathcal{F}*} = \|\mathbb{P}_n - \mathbb{P}_0\|_{\mathcal{F}*}$$

which is again $o_{P*}(1)$.

Now consider $\mathbb{P}_n^{\pi,e}$. Since $\widehat{\boldsymbol{\gamma}}_n \to_p \boldsymbol{\gamma}_0$, it suffices to consider a small compact neighborhood $\mathbb{K} \subset \mathbb{R}^{p+1}$ of $\boldsymbol{\gamma}_0$. Let $\xi_{\boldsymbol{\gamma}}(\mathbf{x}, q) = \frac{q+(1-q)\pi_{\boldsymbol{\gamma}_0}(\mathbf{x})}{q+(1-q)\pi_{\boldsymbol{\gamma}}(\mathbf{x})}$. Since $\mathbf{X}$ is bounded and $\pi_{\boldsymbol{\gamma}}$ is continuous in $\mathbf{X}$, $\xi_{\boldsymbol{\gamma}}(\mathbf{X}, Q)$ is also bounded. Lemma 2.6.15 and 2.6.18 of Van Der Vaart and Wellner (1996) then imply that $\{\xi_{\boldsymbol{\gamma}}(\mathbf{x}, q) : \boldsymbol{\gamma} \in \mathbb{K}\}$ is a VC-subgraph class. Next the Glivenko-Cantelli Preservation theorem (Van Der Vaart and Wellner, 2000) implies that

$$\mathcal{G} = \left\{ h_{\boldsymbol{\gamma},f}(q, \mathbf{x}, \widetilde{t}, \delta) = \underbrace{\frac{q + (1-q)\pi_{\boldsymbol{\gamma}_0}(\mathbf{x})}{q + (1-q)\pi_{\boldsymbol{\gamma}}(\mathbf{x})}}_{\xi_{\boldsymbol{\gamma}}(\mathbf{x},q)} f(\mathbf{x}, \widetilde{t}, \delta) : f \in \mathcal{F}, \boldsymbol{\gamma} \in \mathbb{K} \right\}$$

is a $\mathbb{P}_0$-Glivenko-Cantelli class as $\mathcal{F}$ has an integrable envelope function and $\xi_{\boldsymbol{\gamma}}(x, q)$ is bounded. Then recognizing that

$$\mathbb{P}_n^{\pi,e} f = \frac{1}{n} \sum_{i=1}^{n} \frac{I(L_i + Q_i > 0)}{Q_i + (1-Q_i)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}_i)} \left\{ \frac{Q_i + (1-Q_i)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}_i)}{Q_i + (1-Q_i)\pi_{\widehat{\boldsymbol{\gamma}}_n}(\mathbf{X}_i)} \right\} f(\mathbf{X}_i, \widetilde{T}_i, \Delta_i) = \mathbb{P}_n^{\pi} \xi_{\widehat{\boldsymbol{\gamma}}_n} f$$

We have

$$\|\mathbb{P}_n^{\pi,e} - \mathbb{P}_0\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \|\mathbb{P}_n^{\pi} \xi_{\widehat{\boldsymbol{\gamma}}_n} f - \mathbb{P}_0 f\| \leqslant \sup_{f \in \mathcal{F}} \|\mathbb{P}_n^{\pi} \xi_{\widehat{\boldsymbol{\gamma}}_n} f - \mathbb{P}_0 \xi_{\widehat{\boldsymbol{\gamma}}_n} f\| + \sup_{f \in \mathcal{F}} \|\mathbb{P}_0 \xi_{\widehat{\boldsymbol{\gamma}}_n} f - \mathbb{P}_0 f\|$$

Notice the first term is $o_{P*}(1)$ since $\mathcal{G}$ is $\mathbb{P}_0$-Glivenko-Cantelli. For the second term, we have

$$\begin{aligned} &\frac{1}{Q + (1-Q)\pi_{\widehat{\boldsymbol{\gamma}}_n}(\mathbf{X})} - \frac{1}{Q + (1-Q)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})} \\ &= I(Q=0) \left(1 - \frac{1}{\pi_{\boldsymbol{\gamma}^*}(\mathbf{X})} \right) \widetilde{\mathbf{X}}^T (\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) \end{aligned} \tag{A.4}$$

where $\boldsymbol{\gamma}^*$ is some convex combinations of $\boldsymbol{\gamma}_0$ and $\widehat{\boldsymbol{\gamma}}_n$ and $\boldsymbol{\gamma}^* \to_p \boldsymbol{\gamma}_0$. This implies that

$$\xi_{\widehat{\boldsymbol{\gamma}}_n}(\mathbf{X}, Q) - 1 = \xi_{\widehat{\boldsymbol{\gamma}}_n}(\mathbf{X}, Q) - \xi_{\boldsymbol{\gamma}_0}(\mathbf{X}, Q) = I(Q=0)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}) \left(1 - \frac{1}{\pi_{\boldsymbol{\gamma}^*}(\mathbf{X})} \right) \widetilde{\mathbf{X}}^T (\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) \tag{A.5}$$

and the second term can be rewritten as following:

$$\sup_{f\in\mathcal{F}}\left\|\mathbb{E}\left\{f(\mathbf{X},\widetilde{T},\Delta)I(Q=0)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})\left(1-\frac{1}{\pi_{\boldsymbol{\gamma}^*}(\mathbf{X})}\right)\widetilde{\mathbf{X}}^T(\widehat{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma}_0)\right\}\right\|$$
$$=\sup_{f\in\mathcal{F}}\left\|\left\{f(\mathbf{X},\widetilde{T},\Delta)I(Q=0)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})\left(1-\frac{1}{\pi_{\boldsymbol{\gamma}^*}(\mathbf{X})}\right)\widetilde{\mathbf{X}}^T\right\}\right\|\|\widehat{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma}_0\|$$

Further, since $\pi_{\boldsymbol{\gamma}}(\mathbf{X})$ is bounded away from zero for $\boldsymbol{\gamma}\in\mathbb{K}$ and $\|\mathbb{P}_0\|_{\mathcal{F}}<\infty$, the second term is now determined by $\|\widehat{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma}_0\|$, which is $o_P^*(1)$. ▫

The proof of Proposition $A.2.2$ relies on the following lemma.

Lemma A.3.2. *Let* $\zeta_{\boldsymbol{\gamma}}(l,q,\mathbf{x})=\frac{I(l+q>0)}{q+(1-q)\pi_{\boldsymbol{\gamma}}(\mathbf{x})}$ *and* $\mathcal{F}=\{f(\mathbf{x},\widetilde{t},\delta\}$ *be a* $\mathbb{P}_0$*-Glivenko-Cantelli class with an integrable envelope function* F *such that* $\mathbb{P}_0F<\infty$*. Then*

$$\sup_{f\in\mathcal{F}}\left\|\sqrt{n}(\mathbb{P}_n-\mathbb{P}_0)\left(\zeta_{\widehat{\boldsymbol{\gamma}}_n}f-\zeta_{\boldsymbol{\gamma}_0}f\right)\right\|=o_P^*(1)$$

Proof. First, recall that $\xi_{\boldsymbol{\gamma}}(q,x)=\frac{q+(1-q)\pi_{\boldsymbol{\gamma}_0}(\mathbf{x})}{q+(1-q)\pi_{\boldsymbol{\gamma}}(\mathbf{x})}$, then we have that

$$\mathbb{P}_n\zeta_{\widehat{\boldsymbol{\gamma}}_n}f=\mathbb{P}_n^{\pi}\xi_{\widehat{\boldsymbol{\gamma}}_n}f;\quad \mathbb{P}_n\zeta_{\boldsymbol{\gamma}_0}f=\mathbb{P}_n^{\pi}\xi_{\boldsymbol{\gamma}_0}f$$

Further, since

$$\mathbb{P}_0\zeta_{\boldsymbol{\gamma}}=\mathbb{E}_0\left[\frac{I(L+Q>0)}{Q+(1-Q)\pi_{\boldsymbol{\gamma}}(\mathbf{X})}\right]$$
$$=\mathbb{E}_0\left[\mathbb{E}_0\left[\frac{I(L+Q>0)}{Q+(1-Q)\pi_{\boldsymbol{\gamma}}(\mathbf{X})}\middle|Q,\mathbf{X}\right]\right]=\mathbb{E}_0\left[\frac{Q+(1-Q)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})}{Q+(1-Q)\pi_{\boldsymbol{\gamma}}(\mathbf{X})}\right]=\mathbb{P}_0\xi_{\boldsymbol{\gamma}}$$

We further have that

$$\mathbb{P}_0\zeta_{\widehat{\boldsymbol{\gamma}}_n}f=\mathbb{P}_0\xi_{\widehat{\boldsymbol{\gamma}}_n}f;\quad \mathbb{P}_0\zeta_{\boldsymbol{\gamma}_0}f=\mathbb{P}_0\xi_{\boldsymbol{\gamma}_0}f$$

Given above results and equation (A.5), we have

$$\sup_{f\in\mathcal{F}}\left\|\sqrt{n}(\mathbb{P}_n-\mathbb{P}_0)(\zeta_{\widehat{\boldsymbol{\gamma}}_n}f-\zeta_{\boldsymbol{\gamma}_0}f)\right\|=\sup_{f\in\mathcal{F}}\left\|\sqrt{n}(\mathbb{P}_n^{\pi}-\mathbb{P}_0)(\xi_{\widehat{\boldsymbol{\gamma}}_n}f-\xi_{\boldsymbol{\gamma}_0}f)\right\|$$
$$\leqslant\sup_{f\in\mathcal{F}}\left\|(\mathbb{P}_n^{\pi}-\mathbb{P}_0)\left(I(Q=0)\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})\left(1-\frac{1}{\pi_{\boldsymbol{\gamma}^*}(\mathbf{X})}\right)f(\mathbf{X},\widetilde{T},\Delta)\widetilde{\mathbf{X}}^T\right)\right\|\left\|\sqrt{n}(\widehat{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma}_0)\right\|$$

where $\boldsymbol{\gamma}^*$ is a point lies between $\widehat{\boldsymbol{\gamma}}_n$ and $\boldsymbol{\gamma}_0$. Note that the consistency of $\widehat{\boldsymbol{\gamma}}_n$ implies that $\boldsymbol{\gamma}^* \xrightarrow{p} \boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}^*$ will fall into a compact small neighborhood $\mathbb{K}$ around $\boldsymbol{\gamma}_0$ with probability 1. As a result, $\mathcal{G} = \{I(q=0)\pi_{\boldsymbol{\gamma}_0}(\mathbf{x})\left(1-\frac{1}{\pi_{\boldsymbol{\gamma}}(\mathbf{x})}\right)\widetilde{\mathbf{x}}^T : \boldsymbol{\gamma} \in \mathbb{K}\}$ forms a VC-subgraph class by Lemma 2.6.15 and 2.6.18 of Van Der Vaart and Wellner (1996). Then by the Glivenko-Cantelli Preservation theorem (Van Der Vaart and Wellner, 2000), we have $\mathcal{F}_1 = \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$ is a $\mathbb{P}_0$-Glivenko-Cantelli class with an integrable envelope function. Then

$$\sup_{f\in\mathcal{F}} \left\|\sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)(\zeta_{\widehat{\boldsymbol{\gamma}}_n} f - \zeta_{\boldsymbol{\gamma}_0} f)\right\| \leqslant \|\mathbb{P}_n^{\pi} - \mathbb{P}_0\|_{\mathcal{F}_1}\sqrt{n}\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\|$$

By Proposition A.2.1 and Lemma A.3.1, $\sqrt{n}\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\| = O_P^*(1)$ and $\|\mathbb{P}_n^{\pi} - \mathbb{P}_0\|_{\mathcal{F}_1} = o_P^*(1)$. □

Proof of Proposition A.2.2. Let $\zeta_{\boldsymbol{\gamma}}(l,q,\mathbf{x}) = \frac{I(l+q>0)}{q+(1-q)\pi_{\boldsymbol{\gamma}}(\mathbf{x})}$. This implies that $\zeta_{\boldsymbol{\gamma}_0}(L,Q,\mathbf{X}) = g_1(L,Q,\mathbf{X})$, where g_1 is defined in Proposition A.2.2. Moreover, under assumption (A3) and $\|\mathbb{P}_0\|_{\mathcal{F}} < \infty$, $\mathcal{F}' = \{g_1 \cdot f : f \in \mathcal{F}\}$ is a Donsker Class by Example 2.10.10 of Van Der Vaart and Wellner (1996).

By the definition that $\mathbb{G}_n^{\pi} = \sqrt{n}(\mathbb{P}_n^{\pi} - \mathbb{P}_0)$, we have

$$\mathbb{G}_n^{\pi} f = \sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)\zeta_{\boldsymbol{\gamma}_0} f = \sqrt{n}(\mathbb{P}_n - \mathbb{P}_0) g_1 f.$$

Thus, the usual Donsker theorem (see, e.g., Section 19.2 of Van Der Vaart 2000) implies that $\mathbb{G}_n^{\pi} \rightsquigarrow \mathbb{G}(g_1\cdot)$ in $l^{\infty}(\mathcal{F})$ and

$$\mathrm{Var}(\mathbb{G}(g_1 f)) = \mathrm{Var}(f(\mathbf{X},\widetilde{T},\Delta)) + \mathbb{E}\left[f(\mathbf{X},\widetilde{T},\Delta)^2 \frac{I(Q=0)[1-\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})]}{\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})}\right]$$

Next,

$$\begin{aligned} \mathbb{G}_n^{\pi,e} f - \mathbb{G}_n^{\pi} f &= \sqrt{n}(\mathbb{P}_n^{\pi,e} - \mathbb{P}_0) f - \sqrt{n}(\mathbb{P}_n^{\pi} - \mathbb{P}_0) f \\ &= \mathbb{G}_n\left(\zeta_{\widehat{\boldsymbol{\gamma}}_n} - \zeta_{\boldsymbol{\gamma}_0}\right) f + \sqrt{n}\mathbb{P}_0\left(\zeta_{\widehat{\boldsymbol{\gamma}}_n} - \zeta_{\boldsymbol{\gamma}_0}\right) f \end{aligned} \tag{A.6}$$

By Lemma A.3.2, $\mathbb{G}_n\left(\zeta_{\widehat{\boldsymbol{\gamma}}_n} - \zeta_{\boldsymbol{\gamma}_0}\right) f$ is $o_P^*(1)$, which bounds the first term. For the second term, note that

$$\zeta_{\widehat{\boldsymbol{\gamma}}_n} - \zeta_{\boldsymbol{\gamma}_0} = -I(L=1)I(Q=0)\left(\frac{1}{\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})} - 1\right)\widetilde{\mathbf{X}}^T(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) + o(\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\|^2)$$

and $\mathbb{E}[I(L=1)|Q=0,\mathbf{X}] = \pi_{\boldsymbol{\gamma}_0}(\mathbf{X})$, we can show that

$$\sqrt{n}\mathbb{P}_0\left(\zeta_{\hat{\boldsymbol{\gamma}}_n} - \zeta_{\boldsymbol{\gamma}_0}\right) f = -\left\{\mathbb{E}\left[I(Q=0)(1-\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}))f(\mathbf{X},\widetilde{T},\Delta)\widetilde{\mathbf{X}}^T\right]\right\}\sqrt{n}(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) + o_P^*(1).$$

Together, equation (A.6) and Lemma A.3.1 implies that

$$\begin{aligned}
\mathbb{G}_n^{\pi,e} f &= \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[g_1(L_i,Q_i,\mathbf{X}_i)f(\mathbf{X}_i,\widetilde{T}_i,\Delta_i) - \mathbb{P}_0 f\right]}_{\mathbb{G}_n^{\pi} f} \\
&\underbrace{-\frac{1}{\sqrt{n}}\mathbb{E}\left[I(Q=0)(1-\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}))f(\mathbf{X},\widetilde{T},\Delta)\widetilde{\mathbf{X}}^T\right]\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0}^{-1}\sum_{i=1}^{n} I(Q_i=0)[L_i - \pi_{\boldsymbol{\gamma}_0}(\mathbf{X}_i)]\mathbf{X}_i + o_P^*(1)}_{=\sqrt{n}\mathbb{P}_0\left(\zeta_{\hat{\boldsymbol{\gamma}}_n}-\zeta_{\boldsymbol{\gamma}_0}\right)f} \\
&= \mathbb{G}_n\left[g_1\cdot f - g_2\mathbf{Q}_e(f)^T g_3\right] + o_P^*(1)
\end{aligned}$$

where $g_1(l,q,\mathbf{x}) = \frac{I(l+q>0)}{q+(1-q)\pi_{\boldsymbol{\gamma}_0}(\mathbf{x})}$, $g_2(l,q,\mathbf{x}) = I(q=0)[l-\pi_{\boldsymbol{\gamma}_0}(\mathbf{x})]$, $\mathbf{g}_3(\mathbf{x}) = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0}^{-1}\widetilde{\mathbf{x}}$ and $\mathbf{Q}_e(f) = \mathbb{E}[I(Q=0)(1-\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}))f(\mathbf{X},\widetilde{T},\Delta)\widetilde{\mathbf{X}}]$. This proves the finite dimensional convergence of $\mathbb{G}_n^{\pi,e}$. Next, we prove the asymptotic equicontinuity of $\mathbb{G}_n^{\pi,e}$. Define $\rho(f,g) = \mathbb{P}_0(f-g)^2$ and $\mathcal{F}_\delta = \{f-g : f,g\in\mathcal{F}, \rho(f-g)\leqslant\delta\}$. First, $\mathcal{F}$ is totally bounded by the metric ρ given (Van Der Vaart and Wellner, 1996, Problem 2.1.2) and $\|\mathbb{P}_0\|_{\mathcal{F}} < \infty$. Next,

$$\|\mathbb{G}_n^{\pi,e}\|_{\mathcal{F}_\delta} \leqslant \|\mathbb{G}_n^{\pi}\|_{\mathcal{F}_\delta} + \|\mathbb{G}_n^{\pi,e} - \mathbb{G}_n^{\pi}\|_{\mathcal{F}_\delta}$$

For the first term on the right-hand side, $\mathbb{G}_n^{\pi}$ is asymptotically equicontinuous with respect to ρ. For the second term,

$$\|\mathbb{G}_n^{\pi,e} - \mathbb{G}_n^{\pi}\|_{\mathcal{F}_\delta} \leqslant \|\mathbf{Q}_e^T\|_{\mathcal{F}_\delta}\sqrt{n}\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\| + o_P^*(1)$$

and

$$\|\mathbf{Q}_e^T\|_{\mathcal{F}_\delta} = \sup_{h\in\mathcal{F}_\delta}\left\|\mathbb{E}[I(Q=0)(1-\pi_{\boldsymbol{\gamma}_0}(\mathbf{X}))h(\mathbf{X},\widetilde{T},\Delta)\widetilde{\mathbf{X}}^T]\right\| \leqslant C\sup_{h\in\mathcal{F}_\delta}(\mathbb{P}(h^2))^{1/2} \leqslant C\delta$$

by the definition of h and some constant $C>0$. Thus, we have $\lim_{\delta\searrow 0}\|\mathbb{G}_n^{\pi,e}\|_{\mathcal{F}_\delta} = 0$. Thus, by Theorem 1.5.7 of Van Der Vaart and Wellner (1996), $\mathbb{G}_n^{\pi,e} \rightsquigarrow \mathbb{G}(g_1\cdot - g_2\mathbf{Q}_e(\cdot)^T\mathbf{g}_3)$ in

$l^\infty(\mathcal{F})$ and

$$\begin{aligned}&\mathrm{Var}(\mathbb{G}(g_1 \cdot f - g_2 \mathbf{Q}_e(f)^T \mathbf{g}_3)) = \mathrm{Var}(f(\mathbf{X}, \widetilde{T}, \Delta)) \\ &+ \mathbb{E}\left[f(\mathbf{X}, \widetilde{T}, \Delta)^2 \frac{I(Q=0)[1-\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})]}{\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})} \right] - \mathbf{Q}_e(f)^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0}^{-1} \mathbf{Q}_e(f),\end{aligned}$$

which completes the proof. □

Before proving Theorem A.1.1, we first introduce a useful lemma.

Lemma A.3.3. *Under assumption (D2-3), for a small compact set* $\mathbb{B}$ *that contains* $\boldsymbol{\beta}_0^*$,

$$\sup_{t \in [0,\tau_2], \boldsymbol{\beta} \in \mathbb{B}} \|\mathbf{S}_{n,w}^{(k)}(\boldsymbol{\beta}, t) - \mathbf{s}^{(k)}(\boldsymbol{\beta}, t)\| = o_P(1)$$

for $k = 0, 1, 2$.

Proof of Lemma A.3.3. By assumption (D2), $\mathbf{X}(t)$ can be written as the difference of two nondecreasing processes of $t \in [0, \tau_2]$, then by Example 2.11.16 of Van Der Vaart and Wellner (1996) and the fact that $\mathbf{X}(t)$ is bounded, $\mathbb{G}_n \mathbf{X} \rightsquigarrow \mathbb{G}_0$ in $l^\infty([0, \tau_2])$. Now we can view $\mathcal{G} = \{\mathbf{x}(t) : t \in [0, \tau_2]\}$ as a function class $\{f_t(\mathbf{x}) = \mathbf{x}(t); t \in [0, \tau_2]\}$. $\mathcal{G}$ is a $\mathbb{P}_0$-Donsker class. Next, $\{\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{B}\}$ is trivially a $\mathbb{P}_0$-Donsker class and $\{\mathbf{x}(t)^T \boldsymbol{\beta} : t \in [0, \tau_2], \boldsymbol{\beta} \in \mathbb{B}\}$ is also a Donsker class by theorem 2.10.6 of Van Der Vaart and Wellner (1996). Similarly, we can prove that $\{y(t) \exp(\boldsymbol{\beta}^T \mathbf{x}(t)) \mathbf{x}(t)^{\otimes k} : \boldsymbol{\beta} \in \mathbb{B}, t \in [0, \tau_2]\}$ is also a Donsker class for $k = 0, 1, 2$ as $\{y(t) : t \in [0, \tau_2]\}$ is also a $\mathbb{P}_0$-Donsker class by Theorem 2.11.16 of Van Der Vaart and Wellner (1996). Then the result can be proved by Proposition A.2.1. □

Proof of Theorem A.1.1. The proof of the asymptotic linear expansion is inspired by Lin and Wei (1989). By Lemma A.3.3, we have

$$\sup_{\boldsymbol{\beta} \in \mathbb{B}, t \in [0,\tau_2]} \|\mathbf{S}_{n,w}^{(k)}(\boldsymbol{\beta}, t) - \mathbf{s}^{(k)}(\boldsymbol{\beta}, t)\| = o_P(1) \qquad k = 0, 1, 2$$

for a compact set $\mathbb{B}$ that contains $\boldsymbol{\beta}_0^*$. Then we can decompose the partial score in equation (2.8) as follows:

$$\begin{aligned}
\sqrt{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Delta_i\widehat{w}_i\left\{\mathbf{X}_i(\widetilde{T}_i) - \frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},\widetilde{T}_i)}{\mathbf{S}_{n,w}^{0}(\boldsymbol{\beta},\widetilde{T}_i)}\right\} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{w}_i\int_0^{\tau_2}\mathbf{X}_i(t)dN_i(t) - \sqrt{n}\int_0^{\tau_2}\frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)}d\bar{N}(t) \\
&= \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{w}_i\int_0^{\tau_2}\mathbf{X}_i(t)dN_i(t)}_{(I)} - \underbrace{\sqrt{n}\int_0^{\tau_2}\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}d[\bar{N}(t)-\widetilde{N}(t)]}_{(II)} \\
&\underbrace{-\sqrt{n}\int_0^{\tau_2}\frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)}d\widetilde{N}(t)}_{(III)} - \underbrace{\sqrt{n}\int_0^{\tau_2}\left[\frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)} - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\right]d[\bar{N}(t)-\widetilde{N}(t)]}_{(IV)}
\end{aligned} \tag{A.7}$$

Term (I) and (II) are already linear expansions so we do not need to conduct any further derivation. In what follows, we will first show that term (IV) is $o_P(1)$ and then argue that term (III) has an asymptotic linear expansion.

Term (IV). By assumption (D4), for large enough n, both $\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)$ and $\mathbf{s}^{(0)}(\boldsymbol{\beta},t)$ are bounded away from 0, so Lemma $A.3.3$ implies that

$$\sup_{\boldsymbol{\beta}\in\mathbb{B},t\in[0,\tau_2]}\left\|\frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)} - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}\right\| = o_P(1)$$

Moreover, by Proposition $A.2.2$, $n^{1/2}(\bar{N}(\tau_2)-\widetilde{N}(\tau_2))$ converges to a mean zero normal random variable. Together, this implies that (IV) is $o_P(1)$.

Term (III). Next, we have that

$$\begin{aligned}
\frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)} &= \frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)\left[1+\frac{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)-\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\right]} \\
&= \frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\left[1-\frac{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)-\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}+o_P(1)\right] \\
&= \frac{1}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\left[\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t) - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)-\mathbf{s}^{(0)}(\boldsymbol{\beta},t)\}\right]+o_P(1)
\end{aligned}$$

Thus,

$$n^{1/2}\int_0^{\tau_2}\frac{\mathbf{S}^{(1)}_{n,w}(\boldsymbol{\beta},t)}{\mathbf{S}^{(0)}_{n,w}(\boldsymbol{\beta},t)}d\widetilde{N}(t)$$
$$=n^{1/2}\int_0^{\tau_2}\frac{1}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\left[\mathbf{S}^{(1)}_{n,w}(\boldsymbol{\beta},t)-\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\{\mathbf{S}^{(0)}_{n,w}(\boldsymbol{\beta},t)-\mathbf{s}^{(0)}(\boldsymbol{\beta},t)\}d\widetilde{N}(t)\right]+o_P(1)$$

Taking above equation back to equation (A.7), we get the desired asymptotic linear expansion of $\widehat{\mathbf{U}}_n(\boldsymbol{\beta})$: $\sqrt{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})=n^{-1/2}\sum_{i=1}^n\widehat{w}_i\mathbf{U}_i(\boldsymbol{\beta})+o_p(1)$ with

$$\mathbf{U}_i(\boldsymbol{\beta})=$$
$$\int_0^{\tau_2}\left[\mathbf{X}_i(t)-\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\right]dN_i(t)-\int_0^{\tau_2}\frac{Y_i(t)\exp(\boldsymbol{\beta}^T\mathbf{X}_i(t))}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\left[\mathbf{X}_i(t)-\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\right]d\widetilde{N}(t)$$

Finally, to apply Proposition *A*.2.2, note that $\pi_{\boldsymbol{\gamma}_0}(\mathbf{X})$ needs to be replaced by $\pi_{\boldsymbol{\gamma}_0}(\mathbf{Z}_1)$ and $\mathbf{U}_i(\boldsymbol{\beta})$ can be viewed as a function $\eta_{\boldsymbol{\beta}}(\bar{\mathbf{X}}(\widetilde{T}),\widetilde{T},\Delta)$. As a result, we have $\sqrt{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})=n^{-1/2}\sum_{i=1}^n\widehat{w}_i\mathbf{U}_i(\boldsymbol{\beta})+o_p(1)=\mathbb{G}_n^{\pi,e}\eta_{\boldsymbol{\beta}}+o_P(1)$. Proposition *A*.2.2 implies the asymptotic normality of $\sqrt{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0^*)$[1], which completes the proof. □

Proof of Theorem 2.3.2. Based on the fact that $\widehat{\mathbf{U}}_n(\widehat{\boldsymbol{\beta}}_n)=0$ and $\mathbf{U}_0(\boldsymbol{\beta}_0^*)=0$, we have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0^*)=\mathbf{A}_n(\boldsymbol{\beta}_0^*)^{-1}n^{1/2}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0^*)+o_P(\sqrt{n}\|\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0^*\|),$$

where $\mathbf{A}_n$ is defined in equation (A.2). To prove that $\widehat{\boldsymbol{\beta}}_n\to_p\boldsymbol{\beta}_0^*$, we adopt the same strategy as Lemma 3.1 in Andersen and Gill (1982). From Theorem *A*.1.1, we obtained that $\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^n\widehat{w}_i\mathbf{U}_i(\boldsymbol{\beta})+o_P(n^{-1/2})$. Then by Proposition *A*.2.1,

$$\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})\to_p\mathbb{E}\left\{\int_0^{\tau_2}[\mathbf{X}(t)-\right.$$
$$\left.\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\right]dN(t)-\int_0^{\tau_2}\left[\mathbf{X}(t)-\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}\right]\frac{Y(t)\exp(\boldsymbol{\beta}^T\mathbf{X}(t))}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)}d\widetilde{N}(t)\Bigg\}$$

It is not hard to prove that above expectation equals to $\mathbf{U}_0(\boldsymbol{\beta})=\mathbb{E}\left[\Delta\left(\mathbf{X}(\widetilde{T})-\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},\widetilde{T})}{\mathbf{s}^{(0)}(\boldsymbol{\beta},\widetilde{T})}\right)\right]$. By assumption (D5), $\mathbf{A}(\boldsymbol{\beta}_0^*)=-\frac{\partial\mathbf{U}_0(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0^*}$ is positive definite and by a similar argument

[1] we only need the pointwise convergence result at $\boldsymbol{\beta}=\boldsymbol{\beta}_0^*$

as Lemma 3.1 in Andersen and Gill (1982), we have $\widehat{\boldsymbol{\beta}}_n \to_p \boldsymbol{\beta}_0^*$. We now prove the uniform convergence of $\mathbf{A}_n(\boldsymbol{\beta})$ to $\mathbf{A}(\boldsymbol{\beta})$ in a small compact subset $\mathbb{B}$ that contains $\boldsymbol{\beta}_0^*$.

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}} \|\mathbf{A}_n(\boldsymbol{\beta}) - \mathbf{A}(\boldsymbol{\beta})\| \leqslant \int_0^{\tau_2} \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left\| \frac{\mathbf{S}_{n,w}^{(2)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)} - \left(\frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)} \right)^{\otimes 2} \right\| |d(\bar{N}(t) - \widetilde{N}(t))|$$
$$+ \int_0^{\tau_2} \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left\| \left\{ \frac{\mathbf{S}_{n,w}^{(2)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)} - \left(\frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)} \right)^{\otimes 2} \right\} - \left\{ \frac{\mathbf{s}^{(2)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)} - \left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)} \right)^{\otimes 2} \right\} \right\| d\widetilde{N}(t) \tag{A.8}$$

For the first term of (A.8), let

$$h_n(t) = \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left\| \frac{\mathbf{S}_{n,w}^{(2)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)} - \left(\frac{\mathbf{S}_{n,w}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{S}_{n,w}^{(0)}(\boldsymbol{\beta},t)} \right)^{\otimes 2} \right\|$$

and

$$h(t) = \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left\| \frac{\mathbf{s}^{(2)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)} - \left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta},t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta},t)} \right)^{\otimes 2} \right\|$$

. We can replace $h_n(t)$ by $h(t)$ such that

$$\int_0^{\tau_2} |h_n(t)||d(\bar{N}(t) - \widetilde{N}(t))| = \int_0^{\tau_2} |h(t)||d(\bar{N}(t) - \widetilde{N}(t))| + o_P(1)$$

by Lemma $A.3.3$. By Proposition $A.2.1$, $\sup_{t\in[0,\tau_L]} |\bar{N}(t) - \widetilde{N}(t)| = o_P(1)$. Thus the first term is $o_P(1)$. The second term is also $o_P(1)$ by Lemma $A.3.3$. By the uniform convergence of $\mathbf{A}_n(\boldsymbol{\beta})$ to $\mathbf{A}(\boldsymbol{\beta})$, we then have $\mathbf{A}_n(\boldsymbol{\beta}_0^*) \to_p \mathbf{A}(\boldsymbol{\beta}_0^*)$. Then, by slutsky's theorem, we obtained that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0^*) \to_d \mathbf{N}(0, \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1})$$

□

A.4 Doubly Robust Estimation and its limitation

Now we propose an augmented inverse probability of linkage weighting (AIPLW) estimator for estimating the long-term effect. We first need to define several outcome regression

functions:

$$
\begin{aligned}
m_0(\mathbf{X}) &= \mathbb{E}\left[\Delta\left(\mathbf{X} - \frac{s^{(1)}(\widetilde{T},\boldsymbol{\beta})}{s^{(0)}(\widetilde{T},\boldsymbol{\beta})}\right)\middle|\mathbf{X}, Q=0, L=1\right] \\
&= \underbrace{\mathbb{E}\left[\Delta|\mathbf{X}, Q=0, L=1\right]\mathbf{X}}_{m_1(\mathbf{X})} - \underbrace{\mathbb{E}\left[\Delta\frac{s^{(1)}(\widetilde{T},\boldsymbol{\beta})}{s^{(0)}(\widetilde{T},\boldsymbol{\beta})}\middle|\mathbf{X}, Q=0, L=1\right]}_{m_2(\mathbf{X})} \\
m_3(\mathbf{X}) &= \mathbb{E}\Big[I(\widetilde{T}\geqslant t)\exp(\mathbf{X}^T\boldsymbol{\beta})|\mathbf{X}, Q=0, L=1\Big] \\
&= \mathbb{E}\Big[I(\widetilde{T}\geqslant t)|\mathbf{X}, Q=0, L=1\Big]\exp(\mathbf{X}^T\boldsymbol{\beta})
\end{aligned}
$$

The augmented IPW partial likelihood is then as following:

$$
\begin{aligned}
U_{n,AIPLW} = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{ & I(Q_i=1)\Delta_i\left[\mathbf{X}_i - \frac{S^{(1)}_{n,DR}(\boldsymbol{\beta},\widetilde{T}_i)}{S^{(0)}_{n,DR}(\boldsymbol{\beta},\widetilde{T}_i)}\right] + \\
& I(Q_i=0)\left[\frac{I(L_i=1)}{\pi_{\widehat{\gamma}_n}(\mathbf{X}_i)}\Delta_i\left[\mathbf{X}_i - \frac{S^{(1)}_{n,DR}(\boldsymbol{\beta},\widetilde{T}_i)}{S^{(0)}_{n,DR}(\boldsymbol{\beta},\widetilde{T}_i)}\right]\right. \\
& \left.+\left(1-\frac{I(L_i=1)}{\pi_{\widehat{\gamma}_n}(\mathbf{X}_i)}\right)\widehat{m}_0(\mathbf{X}_i)\right]\Bigg\}
\end{aligned}
\tag{A.9}
$$

with

$$
\begin{aligned}
S^{(k)}_{n,DR}(\boldsymbol{\beta},t) = \frac{1}{n}\sum_{i=1}^{n}\Big\{ & I(Q_i=1)I(\widetilde{T}_i\geqslant t)\exp(\mathbf{X}_i^T\boldsymbol{\beta}) + \\
& I(Q_i=0)\left[\frac{I(L_i=1)}{\pi_{\widehat{\gamma}_0}(\mathbf{X}_i)}I(\widetilde{T}_i\geqslant t)\exp(\mathbf{X}_i^T\boldsymbol{\beta})\right. \\
& \left.+\left(1-\frac{I(L_i=1)}{\pi_{\widehat{\gamma}_n}(\mathbf{X}_i)}\right)\widehat{m}_3(\mathbf{X}_i)\right]\Big\}\,\mathbf{X}_i^{\otimes k}
\end{aligned}
$$

where $\widehat{m}_k(\mathbf{x})$ are certain estimators of $m_k(\mathbf{x})$ for $k=1,2,3$ such that $\widehat{m}_0(\mathbf{x}) = \widehat{m}_1(\mathbf{x}) - \widehat{m}_2(\mathbf{x})$. Let $\widehat{\boldsymbol{\beta}}_{AIPW}$ be the solution to $U_{n,AIPW} = \mathbf{0}$. $\widehat{\boldsymbol{\beta}}_{AIPW}$ is a doubly robust estimator in the sense that either $\pi_{\widehat{\gamma}_n}(\mathbf{x})$ being consistent for $\pi_0(\mathbf{x}) = P(L=1|\mathbf{x}, Q=0)$ or $\widehat{m}_k(\mathbf{x})$ being consistent for $m_k(\mathbf{x})$ for $k=1,2,3$ will guarantee that $\widehat{\boldsymbol{\beta}}_{AIPW}$ is a consistent estimator for $\boldsymbol{\beta}_0^*$. Informally, to prove the doubly-robust property of $\widehat{\boldsymbol{\beta}}_{AIPW}$, we need to prove that the population version of the estimating equation (A.9) have the same root $\boldsymbol{\beta}_0^*$ as the IPLW estimation equation.

We first argue that $S^{(k)}_{n,DR}(\boldsymbol{\beta}, t)$ is a doubly-robust estimator of $s_{(k)}(\boldsymbol{\beta}, t)$ in the above sense. It is not hard to see that $S^{(k)}_{n,DR}(\boldsymbol{\beta}, t)$ converges to $s^{(k)}_{DR}(\boldsymbol{\beta}, t)$ with

$$\begin{aligned} s^{(k)}_{DR}(\boldsymbol{\beta}, t) = \mathbb{E}\Big\{ \Big(& I(Q=1) I(\widetilde{T} \geqslant t) \exp(\mathbf{X}^T \boldsymbol{\beta}) \\ & + I(Q=0) \left[\frac{I(L=1)}{\pi_{\gamma_0}(\mathbf{X})} I(\widetilde{T} \geqslant t) \exp(\mathbf{X}^T \boldsymbol{\beta}) \right. \\ & \left. + \left(1 - \frac{I(L=1)}{\pi_{\gamma_0}(\mathbf{X})} \right) m^*_3(\mathbf{X}) \right] \Big) \mathbf{X}^{\otimes k} \Big\} \end{aligned}$$

where $\widehat{m}_3(\mathbf{x}) \to_p m^*_3(\mathbf{x})$ and $\pi_{\widehat{\gamma}_0}(\mathbf{x}) \to \pi_{\gamma_0}(\mathbf{x})$. $m^*_3(\mathbf{x})$ is not necessarily $m_3(\mathbf{x})$ and similarly $\pi_{\gamma_0}(\mathbf{x})$ might not be $\pi_0(\mathbf{x})$. Then when $\pi_0(\mathbf{x}) = \pi_{\gamma_0}(\mathbf{x})$ and $m^*_3(\mathbf{x}) \neq m_3(\mathbf{x})$, we have

$$\begin{aligned} s^{(k)}_{DR}(\boldsymbol{\beta}, t) &= \mathbb{E}\Big\{ \Big[I(Q=1) I(\widetilde{T} \geqslant t) \exp(\mathbf{X}^T \boldsymbol{\beta}) \\ &\qquad + I(Q=0) \frac{I(L=1)}{\pi_0(\mathbf{X})} I(\widetilde{T} \geqslant t) \exp(\mathbf{X}^T \boldsymbol{\beta}) \Big] \mathbf{X}^{\otimes k} \Big\} \\ &= s^{(k)}(\boldsymbol{\beta}, t) \end{aligned}$$

according to Proposition 2.3.1 and assumption (A1). On the other hand, when $\pi_{\gamma_0}(\mathbf{x}) \neq \pi_0(\mathbf{x})$ and $m^*_3(\mathbf{x}) = m_3(\mathbf{x})$, we can rewrite $s^{(k)}_{DR}(\boldsymbol{\beta}, t)$ as

$$\mathbb{E}\Bigg\{ \Bigg(I(Q=1) I(\widetilde{T} \geqslant t) \exp(\mathbf{X}^T \boldsymbol{\beta}) + \underbrace{I(Q=0) \frac{I(L=1)}{\pi_{\gamma_0}(\mathbf{X})} \left(I(\widetilde{T} \geqslant t) \exp(\mathbf{X}^T \beta) - m_3(\mathbf{X}) \right)}_{\mathbf{I}} \Bigg) + I(Q=0) m_3(\mathbf{X}) \Bigg) \mathbf{X}^{\otimes k} \Big\}$$

and term $\mathbf{I}$ is 0 by law of total expectation. Thus, again we have $s^{(k)}_{DR}(\boldsymbol{\beta}, t) = s^{(k)}(\boldsymbol{\beta}, t)$. Similarly, it is not hard to prove that the population version of the above estimating equation (A.9)

$$\begin{aligned} U_{AIPW} = \mathbb{E}\Bigg\{ & I(Q=1) \Delta \left[\mathbf{X} - \frac{s^{(1)}_{DR}(\boldsymbol{\beta}, \widetilde{T})}{s^{(0)}_{DR}(\boldsymbol{\beta}, \widetilde{T})} \right] + \\ & I(Q=0) \left[\frac{I(L=1)}{\pi_{\gamma_0}(\mathbf{X}_i)} \Delta \left[\mathbf{X} - \frac{s^{(1)}_{DR}(\boldsymbol{\beta}, \widetilde{T})}{s^{(0)}_{DR}(\boldsymbol{\beta}, \widetilde{T})} \right] + \right. \\ & \left. \left(1 - \frac{I(L=1)}{\pi_{\gamma_0}(\mathbf{X})} \right) m^*_0(\mathbf{X}) \right] \Bigg\} \end{aligned}$$

where $\widehat{m}_0(\mathbf{x}) \to m_0^*(\mathbf{x})$. Then when we have $\pi_{\gamma_0}(\mathbf{x}) = \pi_0(\mathbf{x})$, U_{AIPW} becomes

$$\mathbb{E}\left\{\left[I(Q=1) + I(Q=0)\frac{I(L=1)}{\pi_{\gamma_0}(\mathbf{X})}\right]\Delta\left[\mathbf{X} - \frac{s^{(1)}(\boldsymbol{\beta}, \widetilde{T})}{s^{(0)}(\boldsymbol{\beta}, \widetilde{T})}\right]\right\}$$

which is the same as the IPLW estimating equation according to Proposition 2.3.1 and assumption (A1). On the other hand, if we have $\pi_{\gamma_0}(\mathbf{x}) \neq \pi_0(\mathbf{x})$ and $m_k^*(\mathbf{x}) = m_k(\mathbf{x})$ for $k = 1, 2, 3$, then using the same argument as above, we have that U_{AIPW} becomes

$$E\left\{\Delta\left[\mathbf{X} - \frac{s^{(1)}(\boldsymbol{\beta}, \widetilde{T})}{s^{(0)}(\boldsymbol{\beta}, \widetilde{T})}\right]\right\}$$

which is just the original partial likelihood estimating equation for Cox model. Thus, $\widehat{\boldsymbol{\beta}}_{AIPW}$ is doubly-robust.

A.4.1 Difficulty of Estimation of Doubly-Robust Estimator

Based on (A.9), we need to estimate $m_k(\mathbf{x})$ for $k = 1, 2, 3$ to solve for $\widehat{\boldsymbol{\beta}}_{AIPW}$. However, there are major difficulties with estimating these three regression functions. To estimate $m_k(\mathbf{x})$, we need to estimate

$$\begin{aligned}
m_1(\mathbf{X}) &= \mathbb{E}[\Delta|\mathbf{X}; Q=0, L=1]\mathbf{X} = P(T \leqslant C|X, Q=0, L=1)\mathbf{X} \\
m_2(\mathbf{X}) &= \mathbb{E}\left[\Delta\frac{s^{(1)}(\widetilde{T}, \boldsymbol{\beta})}{s^{(0)}(\widetilde{T}, \boldsymbol{\beta})}\middle|\mathbf{X}, Q=0, L=1\right] \\
m_3(\mathbf{X}) &= P(\widetilde{T} \geqslant t|\mathbf{X}; Q=0, L=1)\exp(\mathbf{X}^T\boldsymbol{\beta})
\end{aligned}$$

We have a couple modeling strategies. We use $m_1(\mathbf{x})$ as an example to illustrate the modeling details. For the first modeling strategy, we can try to estimate $m_k(\mathbf{x})$ through modeling the distribution of C_1, C_2, T. More specifically, for $P(T \leqslant C|\mathbf{X}, Q=0, L=1)$, it is not hard to get that

$$P(T \leqslant C|\mathbf{X}, Q=0; L=1) = \frac{P(T \leqslant C_2, T \geqslant C_1|\mathbf{X}, L=1)}{P(T \geqslant C_1|\mathbf{X}, L=1)}$$

For the numerator, we have

$$P(T \leqslant C_2, T \geqslant C_1 | \mathbf{X}, L = 1) = \int_0^{\tau_2} P(C_2 \geqslant s, C_1 \leqslant s | T = s, \mathbf{X}, L = 1) f_T(s | \mathbf{X}, L = 1) ds$$

Thus, we need to model the joint distribution of C_1 and C_2 given $T, \mathbf{X}$ and $L = 1$. Note here with the independent censoring assumption, we have $(C_1, C_2) \perp\!\!\!\perp T | \mathbf{X}$. Together with assumption (A1), it is not clear if we also have $(C_1, C_2) \perp\!\!\!\perp T | \mathbf{X}, L = 1$. Next, even if we make the further assumption that $(C_1, C_2) \perp\!\!\!\perp T | \mathbf{X}, L = 1$, it still requires us to model the joint distribution of C_1 and C_2 given $\mathbf{X}$ and $L = 1$. Another thing is that this also requires us to model the distribution of failure time T given $\mathbf{X}$ and $L = 1$, which need careful modeling to avoid model conflict with the Cox model for T given $\mathbf{X}$ alone. Similar modelings are required for estimating $m_2(\mathbf{X})$ and $m_3(\mathbf{X})$.

For a second modeling strategy, we can directly model $P(\Delta = 1 | \mathbf{X}, Q = 0, L = 1)$ with a logistic regression as Δ is binary variable. Similarly, we can estimate $m_3(\mathbf{X})$ by directly modeling the distribution of the observed time $\widetilde{T}$ given $\mathbf{X}$, $Q = 0$ and $L = 1$ through Cox regression. To estimate $m_2(\mathbf{X})$, we need to either model the distribution of $\widetilde{T}$ given $\Delta = 1$, $Q = 0$ and $L = 1$ or model the distribution of Δ given $T, \mathbf{X}, Q = 0$ and $L = 1$. Take all things into consideration, very careful modelings need to be carried out to ensure model congeniality and whether such models exist is not clear to us.

Finally, to avoid the potential model congeniality issue, non-parametric estimation technique might be applied. However, nonparametric estimation in general suffers from the curse of dimensionality issue, which might require a very large number of samples to get a good estimate.

A.5 Linkage assumption and NLAC method

Note that one sufficient condition for CLAR is

$$L \perp\!\!\!\perp (T, C) | Q = 0, \mathbf{X}.$$

We can also modify the CLAR assumption such that linkage also depends on the censoring time in clinical trial C_1:

$$P(L = 1|\widetilde{T}, \Delta, Q = 0, \mathbf{X}, C_1) = P(L = 1|Q = 0, \mathbf{X}, C_1)$$

as C_1 is always observed when $Q = 0$. One sufficient assumption for this modified CLAR assumption is

$$L \perp\!\!\!\perp (T, C_2)|Q = 0, \mathbf{X}, C_1.$$

We now discuss some other potential assumptions for linkage. For an alternative approach, we might assume that

$$L \perp\!\!\!\perp (T, C)|\mathbf{X}.$$

However, the IPW type method for this assumption suffers from the same issue as the complete-case analysis in that unlinked participants that are diagnosed with PC within the clinical trial will not be included in analysis.

As our main goal is to deal with the missing survival outcome T and C_2 and the missingness only happens when a participant is not linked and censored in the clinical trial, an alternative approach would be to directly model $P(L = 0, Q = 0|\mathbf{X}, T, C_1, C_2)$ and a MAR type assumption would be

$$P(L = 0, Q = 0|\mathbf{X}, T, C_1, C_2) = P(L = 0, Q = 0|\mathbf{X})$$

since only $\mathbf{X}$ is always observed. However, this MAR assumption would never hold as we always have $Q = 0$ when $T \geqslant C_1$. Thus, we choose to model linkage alone as the CLAR assumption. Next, we give the proof for Proposition 2.6.1.

Proof of Proposition 2.6.1. The NLAC method modifies the censoring time C compared to the oracle method. To prove the consistency of the estimator obtained by naive method, we only need to prove that the population version of the partial likelihood for NLAC method has a solution at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. The rest is the same as the consistency proof in Andersen and Gill (1982). For notational simplicity, we illustrate the proof with time-independent covariates only.

Recall for NLAC method, we have

$$\widetilde{T}_j = \begin{cases} \widetilde{T}_j & L_j + Q_j > 0 \\ C_{1j} & L_j + Q_j = 0 \end{cases}$$

and $\Delta_j = 0$ if $L_j + Q_j = 0$. Thus, the partial likelihood for NLAC method solves

$$\widehat{U}_{naive}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} I(L_i + Q_i > 0)\Delta_i \left[\mathbf{X}_i - \frac{S^{(1)}_{n,w}(\boldsymbol{\beta}, \widetilde{T}_i)}{S^{(0)}_{n,w}(\boldsymbol{\beta}, \widetilde{T}_i)}\right] = \mathbf{0}$$

with

$$\begin{aligned} & S^{(k)}_{n,w}(\boldsymbol{\beta}, t) \\ &= \sum_{j=1}^{n} \Big[I(L_j + Q_j > 0) I(\widetilde{T}_j \geqslant t) \exp(\mathbf{X}_j^T \boldsymbol{\beta}) + \\ & I(L_j + Q_j = 0) I(C_{1j} \geqslant t) \exp(\mathbf{X}_j^T \boldsymbol{\beta}) \Big] \mathbf{X}_j^{\otimes k} \end{aligned}$$

Then, by similar technique in Andersen and Gill (1982), we can prove that

$$\widehat{U}_{naive}(\boldsymbol{\beta}) \to_p U_0(\boldsymbol{\beta}) = \mathbb{E}\left[\Delta I(L + Q > 0)\left[\mathbf{X} - \frac{S^{(1)}(\boldsymbol{\beta}, \widetilde{T})}{S^{(0)}(\boldsymbol{\beta}, \widetilde{T})}\right]\right]$$

with

$$\begin{aligned} S^{(k)}(\boldsymbol{\beta}, t) = \mathbb{E}\Big[\Big\{ & I(L + Q > 0) I(\widetilde{T} \geqslant t) \exp(\mathbf{X}^T \boldsymbol{\beta}) + \\ & I(L = 0) I(Q = 0) I(C_1 \geqslant t) \exp(\mathbf{X}^T \boldsymbol{\beta}) \Big\} \mathbf{X}^{\otimes k} \Big]. \end{aligned}$$

Next, we prove that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ solves $U_0(\boldsymbol{\beta}) = \mathbf{0}$. We first have that

$$\mathbb{E}[\Delta I(L + Q > 0)\mathbf{X}] = \mathbb{E}[\Delta I(Q = 1)\mathbf{X}] + \mathbb{E}[\Delta I(L = 1) I(Q = 0)\mathbf{X}]$$

Further, we have

$$\begin{aligned} \mathbb{E}[\Delta I(Q = 1)\mathbf{X}] &= \mathbb{E}[I(T \leqslant C) I(T \leqslant C_1)\mathbf{X}] \\ &= E[I(T \leqslant C_1)\mathbf{X}] = \int_0^{\tau_1} \mathbb{E}[P(C_1 \geqslant t|\mathbf{X}) f_T(t|\mathbf{X})] dt \\ &= \int_0^{\tau_1} \mathbb{E}[\mathbf{X} I(C_1 \geqslant t) I(T \geqslant t) \exp(\mathbf{X}^T \boldsymbol{\beta}_0)] \lambda_0(t) dt \end{aligned}$$

since $\lambda_T(t|\mathbf{X}) = \lambda_0(t)\exp(\mathbf{X}^T\boldsymbol{\beta}_0)$ and $C_1 \perp\!\!\!\perp T|\mathbf{X}$. Next, we have

$$\begin{aligned}
\mathbb{E}[\Delta I(L=1)I(Q=0)\mathbf{X}] &= \mathbb{E}[I(T \leqslant C)I(T \geqslant C_1)I(L=1)\mathbf{X}] \\
&= \mathbb{E}[I(T \leqslant C_2)I(T \geqslant C_1)I(L=1)\mathbf{X}] \\
&= \mathbb{E}[\mathbf{X}\pi(\mathbf{X}, C_1, C_2)I(T \geqslant C_1)I(T \leqslant C_2)] \qquad \text{(A.10)} \\
&= \mathbb{E}[\mathbf{X}\mathbb{E}[\mathbb{E}[I(T \geqslant C_1)I(T \leqslant C_2)\pi(\mathbf{X}, C_1, C_2)|T, \mathbf{X}]|\mathbf{X}]]
\end{aligned}$$

where $\mathbb{E}[I(L=1)|T, Q=0, C_1, C_2, \mathbf{X}] = \pi(\mathbf{X}, C_1, C_2)$. Next, denote

$$g(T, \mathbf{X}) = \mathbb{E}[I(T \geqslant C_1)I(T \leqslant C_2)\pi(\mathbf{X}, C_1, C_2)|T, \mathbf{X}]$$

, we further have

$$\begin{aligned}
g(s, \mathbf{X}) &= \mathbb{E}[I(T \geqslant C_1)I(T \leqslant C_2)\pi(\mathbf{X}, C_1, C_2)|T = s, \mathbf{X}] \\
&= \mathbb{E}[I(C_1 \leqslant s)I(C_2 \geqslant s)\pi(\mathbf{X}, C_1, C_2)|\mathbf{X}]
\end{aligned}$$

since $(C_1, C_2) \perp\!\!\!\perp T|\mathbf{X}$. Further, we have

$$\begin{aligned}
&\mathbb{E}[\Delta I(L=1)I(Q=0)\mathbf{X}] = \mathbb{E}[\mathbf{X}\mathbb{E}[g(T, \mathbf{X})|\mathbf{X}]] = \mathbb{E}\left[\mathbf{X}\int_0^{\tau_2} g(s, \mathbf{X})f_T(s|\mathbf{X})ds\right] \\
&= \int_0^{\tau_2} \mathbb{E}\left[\mathbf{X}\mathbb{E}[I(C_1 \leqslant s)I(C_2 \geqslant s)\pi(\mathbf{X}, C_1, C_2)|\mathbf{X}]f_T(s|\mathbf{X})\right] ds \\
&= \int_0^{\tau_2} \mathbb{E}\left[\mathbf{X}I(C_1 \leqslant s)I(C_2 \geqslant s)\pi(\mathbf{X}, C_1, C_2)P(T \geqslant s|\mathbf{X})\exp(\mathbf{X}^T\boldsymbol{\beta}_0)\right] \lambda_0(s)ds \\
&= \int_0^{\tau_2} \mathbb{E}\left[\mathbf{X}I(C_1 \leqslant s)I(C_2 \geqslant s)\pi(\mathbf{X}, C_1, C_2)\mathbb{E}[I(T \geqslant s)|\mathbf{X}]\exp(\mathbf{X}^T\boldsymbol{\beta}_0)\right] \lambda_0(s)ds \\
&= \int_0^{\tau_2} \mathbb{E}\left[\mathbf{X}\mathbb{E}[I(C_1 \leqslant s)I(C_2 \geqslant s)I(T \geqslant s)\pi(\mathbf{X}, C_1, C_2)|\mathbf{X}]\exp(\mathbf{X}^T\boldsymbol{\beta}_0)\right] \lambda_0(s)ds \\
&= \int_0^{\tau_2} \mathbb{E}\left[\mathbf{X}I(C_1 \leqslant s)I(C_2 \geqslant s)I(T \geqslant s)\pi(\mathbf{X}, C_1, C_2)\exp(\mathbf{X}^T\boldsymbol{\beta}_0)\right] \lambda_0(s)ds
\end{aligned}$$

Thus, together, we have

$$\begin{aligned}
&\mathbb{E}[\Delta I(L+Q>0)\mathbf{X}] = \\
&\int_0^{\tau_1} \mathbb{E}[\mathbf{X}\{I(C_1 \geqslant s)I(T \geqslant s)+ \\
&\pi(\mathbf{X},C_1,C_2)I(C_1 \leqslant s)I(C_2 \geqslant s)I(T \geqslant s)\}\exp(\mathbf{X}^T\boldsymbol{\beta}_0)]\lambda_0(s)ds \\
&+\int_{\tau_1}^{\tau_2} \mathbb{E}[\mathbf{X}\pi(\mathbf{X},C_1,C_2)I(C_2 \geqslant s)I(T \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta}_0)]\lambda_0(s)ds
\end{aligned}$$

as $C_1 \in [0,\tau_1]$. Similarly, we have

$$\begin{aligned}
&\mathbb{E}\left[\Delta I(L+Q>0)\frac{S^{(1)}(\boldsymbol{\beta},T)}{S^{(0)}(\boldsymbol{\beta},T)}\right] = \\
&\int_0^{\tau_1} \mathbb{E}[\{I(C_1 \geqslant s)I(T \geqslant s)+ \\
&\pi(\mathbf{X},C_1,C_2)I(C_1 \leqslant s)I(C_2 \geqslant s)I(T \geqslant s)\}\exp(\mathbf{X}^T\boldsymbol{\beta}_0)]\lambda_0(s)\frac{S^{(1)}(\boldsymbol{\beta},s)}{S^{(0)}(\boldsymbol{\beta},s)}ds \\
&+\int_{\tau_1}^{\tau_2} \mathbb{E}[\pi(\mathbf{X},C_1,C_2)I(C_2 \geqslant s)I(T \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta}_0)]\lambda_0(s)\frac{S^{(1)}(\boldsymbol{\beta},s)}{S^{(0)}(\boldsymbol{\beta},s)}ds
\end{aligned}$$

Now as long as we can prove that

$$\mathbb{E}[\Delta I(L+Q>0)\mathbf{X}] - \mathbb{E}\left[\Delta I(L+Q>0)\frac{S^{(1)}(\boldsymbol{\beta}_0,T)}{S^{(0)}(\boldsymbol{\beta}_0,T)}\right] = \mathbf{0}$$

Then we are done. We prove this by proving the following two equalities:

$$\begin{aligned}
&\int_0^{\tau_1} \mathbb{E}[\mathbf{X}\{I(C_1 \geqslant s)I(T \geqslant s)+ \\
&\pi(\mathbf{X},C_1,C_2)I(C_1 \leqslant s)I(C_2 \geqslant s)I(T \geqslant s)\}\exp(\mathbf{X}^T\boldsymbol{\beta}_0)]\lambda_0(s)ds- \\
&\int_0^{\tau_1} \mathbb{E}[\{I(C_1 \geqslant s)I(T \geqslant s)+ \\
&\pi(\mathbf{X},C_1,C_2)I(C_1 \leqslant s)I(C_2 \geqslant s)I(T \geqslant s)\}\exp(\mathbf{X}^T\boldsymbol{\beta}_0)]\lambda_0(s)\frac{S^{(1)}(\boldsymbol{\beta},s)}{S^{(0)}(\boldsymbol{\beta},s)}ds = \mathbf{0}
\end{aligned} \tag{A.11}$$

and

$$\begin{aligned}
&\int_{\tau_1}^{\tau_2} \mathbb{E}[\mathbf{X}\pi(\mathbf{X},C_1,C_2)I(C_2 \geqslant s)I(T \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta}_0)]\lambda_0(s)ds- \\
&\int_{\tau_1}^{\tau_2} \mathbb{E}[\pi(\mathbf{X},C_1,C_2)I(C_2 \geqslant s)I(T \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta}_0)]\lambda_0(s)\frac{S^{(1)}(\boldsymbol{\beta},s)}{S^{(0)}(\boldsymbol{\beta},s)}ds = \mathbf{0}
\end{aligned} \tag{A.12}$$

We first prove equation (A.11). For $s \in [0, \tau_1]$, we have

$$\begin{aligned} S^{(0)}(\boldsymbol{\beta}, s) &= \mathbb{E}[I(\widetilde{T} \geqslant s) \exp(\mathbf{X}^T\boldsymbol{\beta})[I(Q=1) + I(L=1)I(Q=0)] \\ &\quad + I(C_1 \geqslant s)I(L=0)I(Q=0)\exp(\mathbf{X}^T\boldsymbol{\beta})] \\ &= \mathbb{E}[I(\widetilde{T} \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta})I(Q=1) + I(C_1 \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta})I(Q=0)] \\ &\quad + \mathbb{E}[I(L=1)I(Q=0)\exp(\mathbf{X}^T\boldsymbol{\beta})[I(\widetilde{T} \geqslant s) - I(C_1 \geqslant s)]] \end{aligned}$$

Further,

$$\begin{aligned} &\mathbb{E}[I(\widetilde{T} \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta})I(Q=1) + I(C_1 \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta})I(Q=0)] \\ &= \mathbb{E}[I(T \geqslant s)I(T \leqslant C_1)\exp(\mathbf{X}^T\boldsymbol{\beta})] + \mathbb{E}[I(C_1 \geqslant s)I(T \geqslant C_1)\exp(\mathbf{X}^T\boldsymbol{\beta})] \\ &= \mathbb{E}[I(T \geqslant s)I(C_1 \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta})] \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}[I(L=1)I(Q=0)\exp(\mathbf{X}^T\boldsymbol{\beta})[I(\widetilde{T} \geqslant s) - I(C_1 \geqslant s)]] \\ &= \mathbb{E}[I(L=1)I(Q=0)\exp(\mathbf{X}^T\boldsymbol{\beta})I(T \geqslant s)I(C_2 \geqslant s)I(C_1 \leqslant s)] \\ &= \mathbb{E}[\pi(\mathbf{X}, C_1, C_2)I(T \geqslant s)I(C_2 \geqslant s)I(C_1 \leqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta})] \end{aligned} \tag{A.13}$$

Similarly, we can prove that

$$\begin{aligned} &S^{(1)}(\boldsymbol{\beta}, s) \\ &= \mathbb{E}[\mathbf{X}\{I(C_1 \geqslant s)I(T \geqslant s) + \pi(\mathbf{X}, C_1, C_2)I(C_1 \leqslant s)I(C_2 \geqslant s)I(T \geqslant s)\}\exp(\mathbf{X}^T\boldsymbol{\beta})] \end{aligned}$$

All these results suggest that when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, equation (A.11) is $\mathbf{0}$. When $s \in [\tau_1, \tau_2]$, we have

$$\begin{aligned} S^{(0)}(\boldsymbol{\beta}, s) &= \mathbb{E}[I(\widetilde{T} \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta})I(L=1)I(Q=0)] \\ &= \mathbb{E}[I(T \geqslant s)I(C_2 \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta})I(T \geqslant C_1)\pi(\mathbf{X}, C_1, C_2)] \\ &= \mathbb{E}[\pi(\mathbf{X}, C_1, C_2)I(T \geqslant s)I(C_2 \geqslant s)\exp(\mathbf{X}^T\boldsymbol{\beta})] \end{aligned}$$

as $C_1 \in [0, \tau_1]$. By the same idea, we can prove that equation (A.12) is $\mathbf{0}$. Then we have finished the proof. □

Note that we only uses assumption **(N4)** in (A.10) and (A.13) and in fact, we can further assume that

$$\textbf{(N5)}\ L \perp\!\!\!\perp T|(\mathbf{X}, Q = 0, C_1, C_2, \Delta)$$

and NLAC method still gives consistent estimate in this scenario. Finally for the case of time-dependent covariates, we can get similar results under the assumption that

$$P(T \geqslant s|C_1, C_2, \bar{\mathbf{X}}(\tau_M)) = P(T \geqslant s|\bar{\mathbf{X}}(s))$$
$$P(C_k \geqslant s|T, \bar{\mathbf{X}}(\tau_M)) = P(C_k \geqslant s|\bar{\mathbf{X}}(s)), k = 1, 2.$$

The proof is overall very similar to the case when there are only time-independent covariates and we omit it.

A.6 Relaxation of the "no gap" assumption

So far we have made the "no gap" assumption to focus on the right censoring problem. Now we consider relaxations of this assumption as it is quite common that a participant might not be under observation for some time in practice. This allows for the possibility of interval censoring as a participant might be diagnosed with the event of interest during the gap when he is not under observation. Further, this creates a situation that we have both right censored and interval-censored data, which is also known as partly interval-censored data (Turnbull, 1976).

Partly interval-censored data for Cox regression has been studied in Kim (2003),Cai and Betensky (2003) and the estimation is more difficult than right-censored data. For simplicity, we do not deal with interval-censoring in the current Chapter and leave that to future work. Instead we consider an alternative approach that transforms the interval-censored data to right-censored data. This approach is in the same spirit as the NLAC approach. However, one has to be careful with the transformation. We first discuss an intuitive but problematic approach.

A.6.1 A problematic approach

For illustration, we consider the oracle setting such that each participant is linked to the observational follow-up datasets. For participants that are known to be interval-censored during the gap between clinical trial and observational follow-up, we treat such participants as being right censored at the last recorded date of clinical trial. Thus, we transform the partly interval-censored problem to a right-censored only problem. On the other hand, for participants with gaps, it is also possible that they might not be interval-censored. It is then tempting to use their survival information in the observational dataset, i.e, failure time T or the censoring time C_2. However, this approach is problematic as this would lead to biased estimates. To see the effect of bias with this approach empirically, we conducted a simulation study[2] with approximately 4.5% of the participants being interval-censored. The coverage of the 95% confidence interval for parameter β_1 is only about 65% with $n = 10,000$ and 1,000 repetitions.

To see why we cannot use the survival information in the observational follow-up dataset for a participant with gap and not interval-censored, we need to think about the corresponding censoring distribution. Considering participants with gaps, effectively the censoring time C is set as

$$C = \begin{cases} C_1 & \text{if } C_1 < T < C_1 + U \\ \max(C_1, C_2) & \text{if } C_1 + U \leqslant T \end{cases},$$

where U is a random variable for the length of the gap between the clinical trial and the start of observational follow-up dataset. Thus, it is clear that the censoring time C now depends on the failure time T and violates the independent censoring assumption.

A.6.2 A remedy

We now propose a remedy approach that properly transforms the partly interval-censored data to right-censored data and conventional statistical software can then be applied to

[2]The detailed simulation setting is provided in the appendix A.7.

estimate the parameter for Cox models. Again we consider the oracle setting that each participant is linked to the observational follow-up dataset. For participants with gaps and censored in the clinical trial, we simply view them as being right censored at the last recorded date of clinical trial. Thus, we set $C = C_1$ whenever there is a gap between a participant's last recorded date in clinical trial and the start time of observational follow-up. Let G denote whether gap exists for a participant. Equivalently, the censoring time C is set as

$$C = I(G = 1)C_1 + I(G = 0)\max(C_1, C_2).$$

Thus, similar to the linkage assumption for the NLAC method, above proposed method works if $G \perp\!\!\!\perp T|\mathbf{X}, C_1, C_2$, under the oracle setting that each participant is linked.

For the more practical setting with incomplete linkages, we can similarly apply the methods developed in the current chapter. To be more specific, for NLAC, the censoring time C can be written as

$$C = I(L = 1)[I(G = 1)C_1 + I(G = 0)\max(C_1, C_2)] + I(L = 0)C_1$$

Thus, NLAC again sets the censoring time as C_1 for participants that are not linked and censored in the clinical trial. One sufficient condition for NLAC to work is

$$L \perp\!\!\!\perp T|\mathbf{X}, C_1, C_2, G,$$

which is similar to assumption **(N1)**. For the IPLW method, we might modify the CLAR assumption as

$$P(L = 1|\mathbf{X}, Q = 0, \widetilde{T}, \Delta, G) = P(L = 1|\mathbf{X}, Q = 0, G).$$

We present relevant simulation results in appendix A.7 due to space limit. The limitation of this remedy approach is similar to NLAC: it only works when Cox model is correctly specified; when Cox model is mis-specified, our proposed approach will no longer work as we modify the censoring times.

A.7 More simulation results

We now present the simulation results when there are gaps between the clinical trial and the observational follow-up dataset. We consider the following simulation scenario. The hazard function is $\lambda(t|\mathbf{X};\boldsymbol{\beta}_0) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)$ and $\boldsymbol{\beta}_0 = (\beta_1,\beta_2,\beta_3)^T = (-\log(2),\log(2),0.2)^T$. X_1 is a Bernoulli variable that takes value 1 with probability 0.5. X_2 is a normal random variable with mean -1 and standard deviation 1. X_3 is a normal random variable with mean 1 and standard deviation 2. The baseline hazard function is $\lambda_0(t) = 0.15$. The censoring time in clinical trial C_1 is uniformly distributed between 0 and 3.5.

For each patient, the probability for a gap between clinical trial and the observational follow-up dataset to exist is 0.5 and the length of the gap U is set as a uniform random variable between 1 and 2. Thus, the starting time of the observational follow-up period for a participant is set as C_1 plus U. C_2 is set as start time of the observational follow-up time plus an exponential random variable with rate $0.8 * X_1 + 0.03$. We further set $\tau_1 = 3.5$ and $\tau_2 = 16$[3]. The percentage of interval-censored patient is approximately 4.5%. This is the scenario we used in section A.6.

I consider the same three mechanisms for linkage to the medicare data as the simulations with time-dependent covariates in Section 2.4. The only difference is that for LCAR, we have $P(L=1) = 0.4$. When Cox model is correctly specified, one additional mechanism for linkage is considered as

$$P(L=1|\mathbf{X},C_2,\Delta;Q=0) = \frac{\exp(-0.25+0.5*X_1+0.5*X_2-0.1*C_2-0.1*\Delta)}{1+\exp(-0.25+0.5*X_1+0.5*X_2-0.1*C_2-0.1*\Delta)}$$

$$P(L=1|Q=1) = 0.5.$$

This leads to a more serious violation of the CLAR assumption (A1) and linkage now depends on the censoring time in clinical trial C_2 conditional on $\mathbf{X}$. As expected, NLAC should still work under this linkage mechanism as linkage does not depend on the failure time T. We again consider sample sizes $n = 500, 1,000, 1,500, \ldots, 10,000$. For each simulation setting,

[3] C_1 and C_2 will be administratively censored by τ or τ_M

we generate 1,000 repetitions. The simulation results are given in Table A.1 to Table A.2. As the results are similar to the simulation studies in the main text, we omit the discussion here.

A.8 Simulation setting in section 2.2.3

Now we present the simulation setting for the motivating example in section 2.2.3. The hazard function is $\lambda(t|x) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3^2)$ and $\boldsymbol{\beta}_0 = (\beta_1, \beta_2, \beta_3)^T = (-\log(2), \log(2), 0.2)^T$. X_1 is a Bernoulli variable that takes value 1 with probability 0.5. X_2 is a normal random variable with mean -1 and standard deviation 1. X_3 is a normal random variable with mean X_2 and standard deviation 2. The baseline hazard function is $\lambda_0(t) = 0.05$. The censoring time in clinical trial C_1 is exponentially distributed with rate $0.1 * X_1 + 0.05$. The censoring time C_2 is defined as C_1 plus an exponential random variable with rate $0.8 * X_1 + 0.03$. Further, we set $\tau_1 = 3$ and $\tau_2 = 16$.

Table A.1: Simulation results for linkage mechanism (LCAR) and (CLAR).

			Bias (Mean SE)			Coverage of 95% CI		
Mechanism	Method	n	β_1	β_2	β_3	β_1	β_2	β_3
LCAR	Oracle	500	-0.01 (0.219)	0.01 (0.095)	0.00 (0.045)	0.96	0.94	0.95
		2000	-0.00 (0.108)	0.00 (0.047)	0.00 (0.022)	0.94	0.95	0.94
	CC	500	-0.03 (0.355)	0.02 (0.155)	0.01 (0.073)	0.97	0.94	0.96
		2000	-0.01 (0.172)	0.00 (0.075)	0.00 (0.036)	0.96	0.96	0.94
	CC+	500	0.11 (0.241)	-0.06 (0.111)	-0.02 (0.054)	0.92	0.91	0.934
		2000	0.12 (0.118)	-0.07 (0.054)	-0.02 (0.026)	$0.81^{\dagger}$	$0.72^{\dagger}$	$0.86^{\dagger}$
	NLAC	500	-0.01 (0.240)	0.01 (0.112)	0.00 (0.054)	0.96	0.95	0.95
		2000	-0.00 (0.118)	0.00 (0.055)	0.00 (0.026)	0.94	0.95	0.94
	IPLW	500	-0.02 (0.271)	0.02 (0.125)	0.01 (0.060)	0.95	0.92	0.93
		2000	-0.01 (0.134)	0.00 (0.063)	0.00 (0.030)	0.95	0.95	0.93
CLAR	Oracle	500	-0.01 (0.219)	0.01 (0.095)	0.00 (0.045)	0.96	0.94	0.95
		2000	-0.00 (0.108)	0.00 (0.047)	0.00 (0.022)	0.94	0.95	0.94
	CC	500	-0.18 (0.320)	-0.14 (0.153)	0.00 (0.070)	0.92	$0.83^{\dagger}$	0.95
		2000	-0.16 (0.156)	-0.15 (0.074)	-0.00 (0.034)	$0.80^{\dagger}$	$0.45^{\dagger}$	0.95
	CC+	500	-0.08 (0.239)	-0.22 (0.117)	-0.02 (0.054)	0.95	$0.51^{\dagger}$	0.94
		2000	-0.07 (0.118)	-0.23 (0.057)	-0.02 (0.027)	$0.89^{\dagger}$	$0.03^{\dagger}$	$0.88^{\dagger}$
	NLAC	500	-0.01 (0.238)	0.01 (0.114)	0.00 (0.054)	0.95	0.95	0.95
		2000	-0.00 (0.117)	0.00 (0.056)	0.00 (0.027)	0.95	0.95	0.94
	IPLW	500	-0.02 (0.263)	0.02 (0.134)	0.01 (0.063)	0.94	0.91	0.93
		2000	-0.00 (0.130)	0.01 (0.069)	0.00 (0.032)	0.95	0.95	0.94

We use † to highlight settings with coverage below 90%.

Table A.2: Simulation results for linkage mechanism (LNAR($\widetilde{T}$)) and (LNAR(C_2)).

			Bias (Mean SE)			Coverage of 95% CI		
Mechanism	Method	n	β_1	β_2	β_3	β_1	β_2	β_3
LNAR($\widetilde{T}$)	Oracle	500	-0.01 (0.219)	0.011 (0.095)	0.00 (0.045)	0.96	0.94	0.95
		2000	-0.00 (0.108)	0.00 (0.047)	0.00 (0.022)	0.94	0.95	0.94
	CC	500	-0.20 (0.320)	-0.16 (0.154)	0.00 (0.071)	0.92	0.80†	0.95
		2000	-0.19 (0.156)	-0.17 (0.074)	-0.01 (0.034)	0.77†	0.38†	0.94
	CC+	500	-0.10 (0.239)	-0.24 (0.118)	-0.02 (0.055)	0.94	0.47†	0.94
		2000	-0.1 (0.118)	-0.24 (0.058)	-0.02 (0.027)	0.87†	0.02†	0.86†
	NLAC	500	-0.01 (0.238)	0.01 (0.114)	0.00 (0.055)	0.95	0.95	0.95
		2000	-0.01 (0.118)	0.00 (0.056)	-0.00 (0.027)	0.94	0.95	0.94
	IPLW	500	-0.02 (0.265)	0.01 (0.137)	0.01 (0.064)	0.94	0.92	0.92
		2000	-0.01 (0.131)	-0.00 (0.070)	-0.00 (0.032)	0.94	0.95	0.94
LNAR(C_2)	Oracle	500	-0.01 (0.219)	0.01 (0.095)	0.00 (0.045)	0.96	0.94	0.95
		2000	-0.00 (0.108)	0.00 (0.047)	0.00 (0.022)	0.94	0.95	0.94
	CC	500	-0.63 (0.340)	-0.27 (0.176)	-0.02 (0.082)	0.55†	0.63†	0.95
		2000	-0.60 (0.165)	-0.28 (0.083)	-0.02 (0.039)	0.04†	0.09†	0.91
	CC+	500	-0.46 (0.248)	-0.35 (0.127)	-0.05 (0.060)	0.54†	0.24†	0.87†
		2000	-0.44 (0.122)	-0.36 (0.061)	-0.05 (0.029)	0.04†	0.00†	0.62†
	NLAC	500	-0.01 (0.246)	0.01 (0.123)	0.00 (0.059)	0.96	0.95	0.95
		2000	-0.01 (0.121)	0.00 (0.060)	0.00 (0.029)	0.94	0.95	0.95
	IPLW	500	-0.04 (0.316)	0.04 (0.180)	0.02 (0.088)	0.94	0.87†	0.89†
		2000	-0.01 (0.157)	0.01 (0.101)	0.01 (0.047)	0.94	0.91	0.91

We use † to highlight settings with coverage below 90%.

Appendix B

APPENDIX OF CHAPTER 3

B.1 Proof for single primary variable

Proof of Proposition 3.3.1. $p(\ell, A = 1)$ is clearly identifiable. We can write $p(\ell, A = 0)$ as

$$p(\ell, A = 0) = \sum_r p(\ell, R = r, A = 0)$$

and we further have

$$\begin{aligned} p(\ell, R = r, A = 0) &= \int p(\ell, x_r, R = r, A = 0) dx_r \\ &= \int p(\ell | x_r, R = r, A = 0) p(x_r, R = r, A = 0) dx_r \\ &= \int p(\ell | x_r, R \geqslant r, A = 1) p(x_r, R = r, A = 0) dx_r \end{aligned}$$

Thus, $p(\ell, R = r, A = 0)$ is identifiable under ACCMV, which implies that $p(\ell, A = 0)$ is identifiable. □

Proof Of Lemma 3.3.3. We have

$$\begin{aligned} & p(\ell | x_r, R = r, A = 0) = p(\ell | x_r, R \geqslant r, A = 1) \\ \Leftrightarrow & \frac{p(\ell, x_r, R = r, A = 0)}{p(x_r, R = r, A = 0)} = \frac{p(\ell, x_r, R \geqslant r, A = 1)}{p(x_r, R \geqslant r, A = 1)} \\ \Leftrightarrow & \frac{p(R = r, A = 0, \ell, x_r)}{p(R \geqslant r, A = 1, \ell, x_r)} = \frac{p(R = r, A = 0, x_r)}{p(R \geqslant r, A = 1, x_r)} \\ \Leftrightarrow & \frac{p(R = r, A = 0 | \ell, x_r)}{p(R \geqslant r, A = 1 | \ell, x_r)} = \frac{p(R = r, A = 0 | x_r)}{p(R \geqslant r, A = 1 | x_r)} \end{aligned}$$

□

We start by giving the asymptotic linear expansion of $\widehat{\alpha}_r$. Given

$$O_r(X_r) = \frac{P(R = r, A = 0|X_r)}{P(R \geqslant r, A = 1|X_r)} = O_r(X_r; \alpha_r)$$

we have $P(R = r, A = 0|X_r, \{R \geqslant r, A = 1\} \cup \{R = r, A = 0\}; \alpha_r) = \frac{O_r(X_r;\alpha_r)}{1+O_r(X_r;\alpha_r)}$. The log-likelihood has the following form:

$$l_n(\alpha_r) = \frac{1}{n}\sum_{i=1}^{n}[I(R_i = r, A_i = 0)\log O_r(X_{i,r}; \alpha_r) - \{I(R_i = r, A_i = 0)+$$
$$I(R_i \geqslant r, A_i = 1)\}\log(1 + O_r(X_{i,r}; \alpha_r))],$$

The score is

$$S_n(\alpha_r) = \frac{1}{n}\sum_{i=1}^{n}\Bigg[I(R_i = r, A_i = 0)\frac{\nabla_{\alpha_r} O_r(X_{i,r}; \alpha_r)}{O_r(X_{i,r}; \alpha_r)}$$
$$-\{I(R_i = r, A_i = 0) + I(R_i \geqslant r, A_i = 1)\}\frac{\nabla_{\alpha_r} O_r(X_{i,r}; \alpha_r)}{1 + O_r(X_{i,r}; \alpha_r)}\Bigg]$$

Further, consider $S(\alpha_r)$ such that $S_n(\alpha_r) \to_p S(\alpha_r)$ and assuming that α_r^* is the unique solutions for $S(\alpha_r) = 0$. Next, we assume that $\sup_{\alpha_r} \|S_n(\alpha_r) - S_n(\alpha_r)\| = o_P(1)$. Then by theorem 5.9 of Van Der Vaart (2000), we have that $\widehat{\alpha}_r \to_p \alpha_r^*$. Further, we have that

$$\nabla S_n(\alpha_r) = \frac{1}{n}\sum_{i=1}^{n}\Bigg[I(R_i = r, A_i = 0)\left(\frac{\nabla^2_{\alpha_r} O_r(X_{i,r}; \alpha_r)}{O_r(X_{i,r}; \alpha_r)} - \frac{\nabla_{\alpha_r} O_r(X_{i,r}; \alpha_r)^{\otimes 2}}{O_r^2(X_{i,r}; \alpha_r)}\right)$$
$$-\{I(R_i = r, A_i = 0) + I(R_i \geqslant r, A_i = 1)\}$$
$$\left(\frac{\nabla^2_{\alpha_r} O_r(X_{i,r}; \alpha_r)(1 + O_r(X_{i,r}; \alpha_r)) - \nabla_{\alpha_r} O_r(X_{i,r}; \alpha_r)^{\otimes 2}}{(1 + O_r(X_{i,r}; \alpha_r))^2}\right)\Bigg]$$

and that $\nabla S_n(\alpha_r) \to_p -\Sigma(\alpha_r)$ with

$$\Sigma(\alpha_r^*) = \mathbb{E}\Bigg[P(R \geqslant r, A = 1|X_r)\frac{\nabla^2_{\alpha_r} O_r(X_r; \alpha_r^*)(1 + O_r(X_r; \alpha_r^*)) - \nabla_{\alpha_r} O_r(X_r; \alpha_r^*)^{\otimes 2}}{1 + O_r(X_r; \alpha_r^*)}$$
$$-P(R \geqslant r, A = 1|X_r)\frac{\nabla^2_{\alpha_r} O_r(X_r; \alpha_r^*) O_r(X_r; \alpha_r^*) - \nabla_{\alpha_r} O_r(X_r; \alpha_r^*)^{\otimes 2}}{O_r(X_r; \alpha_r^*)}\Bigg]$$
$$= \mathbb{E}\left[P(R \geqslant r, A = 1|X_r)\nabla_{\alpha_r} O_r(X_r; \alpha_r)^{\otimes 2}\frac{1}{O_r(X_r; \alpha_r^*)(1 + O_r(X_r; \alpha_r^*))}\right]$$

Under appropriate assumptions (see theorem 5.21 of Van Der Vaart (2000)), we have that

$$\sqrt{n}(\widehat{\alpha}_r - \alpha_r^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{r,\alpha_r^*}(X_{i,r}, R_i, A_i) + o_p(1) \to_d N(0, \Sigma(\alpha_r^*)^{-1})$$

where

$$\psi_{r,\alpha_r}(X_r, R, A) = \Sigma(\alpha_r)^{-1} \left[I(R = r, A = 0) \frac{\nabla_{\alpha_r} O_r(X_r; \alpha_r)}{O_r(X_r; \alpha_r)} - \right.$$
$$\left. \{I(R = r, A = 0) + I(R \geqslant r, A = 1)\} \frac{\nabla_{\alpha_r} O_r(X_r; \alpha_r)}{1 + O_r(X_r; \alpha_r)} \right]$$

Proof of Theorem 3.3.4. We prove the asymptotic normality of $\widehat{\theta}_{\mathsf{IPW}}$ through its asymptotic linear form. For notational convenience, we denote $g(L, X, R, A; \alpha) = f(L) I(A = 1)[1 + \sum_r O_r(X_r; \alpha_r) I(R \geqslant r)]$. Then, we have that $\widehat{\theta}_{\mathsf{IPW}} = \frac{1}{n} \sum_{i=1}^{n} g(L_i, X_i, R_i, A_i; \widehat{\alpha})$. We can rewrite $\widehat{\theta}_{\mathsf{IPW}} - \theta_0$ as

$$\widehat{\theta}_{\mathsf{IPW}} - \theta_0 = \underbrace{\frac{1}{n} \sum_{i=1}^{n} g(L_i, X_i, R_i, A_i; \widehat{\alpha}) - \frac{1}{n} \sum_{i=1}^{n} g(L_i, X_i, R_i, A_i; \alpha^*)}_{\mathbf{I}}$$
$$+ \underbrace{\frac{1}{n} \sum_{i=1}^{n} g(L_i, X_i, R_i, A_i; \alpha^*) - \theta_0}_{\mathbf{II}}$$

Term **II** is already in the linear expansion form. For term **I**, we have that

$$\begin{aligned}
\sqrt{n}\mathbf{I} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(L_i) I(A_i = 1) \sum_r I(R_i \geqslant r)[O_r(X_{i,r}; \widehat{\alpha}_r) - O_r(X_{i,r}; \alpha_r^*)] \\
&= \frac{1}{n} \sum_{i=1}^{n} f(L_i) I(A_i = 1) \sum_r I(R_i \geqslant r) \nabla_{\alpha_r} O_r(X_{i,r}; \alpha_r^*)^T \sqrt{n}(\widehat{\alpha}_r - \alpha_r^*) + o_p(1) \\
&= \sum_r \frac{1}{n} \sum_{i=1}^{n} f(L_i) I(A_i = 1) I(R_i \geqslant r) \nabla_{\alpha_r} O_r(X_{i,r}; \alpha_r^*)^T \sqrt{n}(\widehat{\alpha}_r - \alpha_r^*) + o_p(1) \\
&= \sum_r \left(\mathbb{E}[f(L) I(A = 1) I(R \geqslant r) \nabla_{\alpha_r} O_r(X_r; \alpha_r^*)^T] + o_p(1) \right) \sqrt{n}(\widehat{\alpha}_r - \alpha_r^*) + o_p(1) \\
&= \sum_r \mathbb{E}[\nabla_{\alpha_r} g(L, X, R, A, \alpha^*)^T] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{r,\alpha_r^*}(X_{i,r}, R_i, A_i) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_r \mathbb{E}[\nabla_{\alpha_r} g(L, X, R, A, \alpha^*)^T] \psi_{r,\alpha_r^*}(X_{i,r}, R_i, A_i) + o_p(1)
\end{aligned}$$

Thus, combined with term **II**, we have

$$\sqrt{n}(\widehat{\theta}_{\mathsf{IPW}} - \theta_0) =$$
$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\sum_r \mathbb{E}[\nabla_{\alpha_r} g(L, X, R, A, \alpha^*)^T] \psi_{r,\alpha_r^*}(X_{i,r}, R_i, A_i) + g(L_i, X_i, R_i, A_i; \alpha^*) - \theta_0 \right]$$
$$+ o_P(1)$$

and denote

$$\phi(X, L, R, A; \alpha^*) =$$
$$\sum_r \mathbb{E}[\nabla_{\alpha_r} g(L, X, R, A, \alpha^*)^T] \psi_{r,\alpha_r^*}(X_r, R, A) + g(L, X, R, A; \alpha^*) - \theta_0.$$

Then we have $\sqrt{n}(\widehat{\theta}_{\mathsf{IPW}} - \theta_0) \to_d N(0, \sigma^2_{\mathsf{IPW}})$ with $\sigma^2_{\mathsf{IPW}} = \mathrm{Var}[\phi(X, L, R, A; \alpha^*)]$. □

Similarly, we first give the asymptotic linear expansion of $\widehat{\beta}_r$. The estimating equation for β_r now has the following form:

$$S_n(\beta_r) = \frac{1}{n} \sum_{i=1}^{n} I(R_i \geqslant r, A_i = 1)(m_{r,0}(X_{i,r}; \beta_r) - f(L_i)) \nabla_{\beta_r} m_{r,0}(X_{i,r}; \beta_r)$$

Again for parametric models, under appropriate assumptions (see theorem 5.9 of Van Der Vaart (2000)), we have that $\widehat{\beta}_r \to_p \beta_r^*$. Further, we have that

$$\nabla S_n(\beta_r) = \frac{1}{n} \sum_{i=1}^{n} I(R_i \geqslant r, A_i = 1)[\nabla^2_{\beta_r} m_{r,0}(X_{i,r}; \beta_r)(m_{r,0}(X_{i,r}; \beta_r) - f(L_i))$$
$$+ \nabla_{\beta_r} m_{r,0}(X_{i,r}; \beta_r)^{\otimes 2}]$$

and that $\nabla S_n(\beta_r) \to_p \nabla S(\beta_r)$ with

$$\nabla S(\beta_r^*) = \mathbb{E}\left[I(R \geqslant r, A = 1) \nabla_{\beta_r} m_{r,0}(X_r; \beta_r^*)^{\otimes 2} \right]$$

Next, we have that

$$\sqrt{n}(\widehat{\beta}_r - \beta_r^*) =$$
$$\frac{1}{\sqrt{n}} \nabla S(\beta_r^*)^{-1} \sum_{i=1}^{n} I(R_i \geqslant r, A_i = 1)(f(L_i) - m_{r,0}(X_{i,r}; \beta_r^*)) \nabla_{\beta_r} m_{r,0}(X_{i,r}; \beta_r^*) + o_p(1)$$

with

$$\psi_{r,\beta_r^*}(L, X_r, R, A) = \nabla S(\beta_r^*)^{-1} I(R \geqslant r, A = 1)(f(L) - m_{r,0}(X_r; \beta_r^*))\nabla_{\beta_r} m_{r,0}(X_r; \beta_r^*)$$

.

Proof of Theorem 3.3.5. Now we give the proof for the regression adjustment estimation. The proof is very similar to proof of Theorem 3.3.4. Similarly, denote $h(L, X, R, A; \beta) = f(L)A + \sum_r m_{r,0}(X_r; \beta_r)(1 - A)I(R = r)$, we have that

$$\widehat{\theta}_{\mathsf{RA}} - \theta = \underbrace{\frac{1}{n}\sum_{i=1}^{n} h(L_i, X_i, R_i, A_i; \widehat{\beta}) - \frac{1}{n}\sum_{i=1}^{n} h(L_i, X_i, R_i, A_i; \beta^*)}_{\mathbf{I}}$$
$$+ \underbrace{\frac{1}{n}\sum_{i=1}^{n} h(L_i, X_i, R_i, A_i; \beta^*) - \theta}_{\mathbf{II}}$$

For term **I**, we have

$$\begin{aligned}
\sqrt{n}\mathbf{I} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{r}\left[m_{r,0}(X_{i,r}; \widehat{\beta}_r) - m_{r,0}(X_{i,r}; \beta_r^*)\right] I(R_i = r, A_i = 0) \\
&= \sum_{r} \mathbb{E}[\nabla_{\beta_r} m_{r,0}(X_r; \beta_r^*)^T I(R = r, A = 0)]\sqrt{n}(\widehat{\beta}_r - \beta_r^*) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{r} \mathbb{E}[\nabla_{\beta_r} m_{r,0}(X_r; \beta_r^*)^T I(R = r, A = 0)]\psi_{r,\beta_r^*}(L_i, X_{i,r}, R_i, A_i) + o_p(1)
\end{aligned}$$

Thus, combined with term **II**, denote

$$\begin{aligned}
&\phi(L, X, R, A; \beta^*) = \\
&\sum_{r} \mathbb{E}[\nabla_{\beta_r} m_{r,0}(X_r; \beta_r^*)^T I(R = r, A = 0)]\psi_{r,\beta_r^*}(L, X_r, R, A) + h(L, X, R, A; \beta^*) - \theta
\end{aligned}$$

we have

$$\sqrt{n}(\widehat{\theta}_{\mathsf{RA}} - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi(L_i, X_i, R_i, A_i; \beta^*) + o_p(1) \to_d N(0, \sigma_{\mathsf{RA}}^2)$$

with $\sigma_{\mathsf{RA}}^2 = \mathrm{Var}[\phi(X, L, R, A)]$. □

B.2 Proof of multiple-robustness for single variable

PROOF OF THEOREM 3.3.6. Recall that the IPW formulation for $\theta_{0,r}$ is

$$\begin{aligned}\theta_{0,r} &= \mathbb{E}[f(L)I(A=1)I(R \geqslant r)O_r(X_r;\alpha_r^*)] \\ &= \int f(\ell)O_r(x_r)I(s \geqslant r)I(a=1)p_0(\ell,x_r,s,a)d\ell dx_r dsda.\end{aligned}$$

and $p_0(\ell, x_r, s, a)$ is the true model. We consider a pathwise perturbation $p_\epsilon(\ell, x_r, s, a) = p_0(\ell, x_r, s, a)(1 + \epsilon \cdot g(\ell, x_r, s, a))$ such that g satisfies

$$\int p_0(\ell, x_r, s, a)g(\ell, x_r, s, a)d\ell dx_r dsda = 0$$

Under p_ϵ, denote $\theta_{0,r}$ as $\theta_{0,r}^\epsilon$. We derive the EIF using the semi-parametric theory (see section 25.3 of Van Der Vaart (2000)), the EIF is a function $\mathsf{EIF}(\ell, x_r, s, a)$ such that $\mathbb{E}[\mathsf{EIF}(L, X_r, R, A)] = 0$ and

$$\lim_{\epsilon\to 0}\frac{\theta_{0,r}^\epsilon - \theta_{0,r}}{\epsilon} = \int \mathsf{EIF}(\ell, x_r, s, a)p_0(\ell, x_r, s, a)g(\ell, x_r, s, a)d\ell dx_r dsda$$

Under model p_ϵ, we also have perturbed odds $O_{r,\epsilon}(x_r)$. We denote $\Delta O_r(x_r) = O_{r,\epsilon}(x_r) - O_{r,0}(x_r)$. Then, a direct computation shows that

$$\begin{aligned}\theta_{0,r}^\epsilon &= \int f(\ell)O_{r,\epsilon}(x_r)I(s \geqslant r)I(a=1)p_\epsilon(\ell,x_r,s,a)d\ell dx_r dsda \\ &= \theta_{0,r} + \epsilon \underbrace{\int f(\ell)O_{r,0}(x_r)I(s \geqslant r)I(a=1)p_0(\ell,x_r,s,a)g(\ell,x_r,s,a)d\ell dx_r dsda}_{\mathbf{A}} \\ &+ \underbrace{\int f(\ell)\Delta O_r(x_r)I(s \geqslant r)I(a=1)p_0(\ell,x_r,s,a)d\ell dx_r dsda}_{\mathbf{B}} + O(\epsilon^2)\end{aligned}$$

Part **A** is already in the form of an EIF. For part **B**, we need to derive $\Delta O_r(x_r)$. Now we expand the difference $\Delta O_r(x_r)$ as following:

$$\begin{aligned}\Delta O_r(x_r) = O_{r,\epsilon}(x_r) - O_{r,0}(x_r) &= \frac{p_\epsilon(R=r, A=0, x_r)}{p_\epsilon(R \geqslant r, A=1, x_r)} - \frac{p_0(R=r, A=0, x_r)}{p_0(R \geqslant r, A=1, x_r)} \\ &= \frac{1}{p_0(R \geqslant r, A=1, x_r)}\left[\Delta p(R=r, A=0, x_r) - O_{r,0}(x_r)\Delta p(R \geqslant r, A=1, x_r)\right] \\ &+ O(\epsilon^2)\end{aligned}$$

with

$$\begin{aligned}\Delta p(R=r,A=0,x_r) &= p_\epsilon(R=r,A=0,x_r) - p_0(R=r,A=0,x_r)\\ &= \epsilon\int I(s=r)I(a=0)p_0(\ell,x_r,s,a)g(\ell,x_r,s,a)d\ell dsda\\ \Delta p(R\geqslant r,A=1,x_r) &= p_\epsilon(R\geqslant r,A=1,x_r) - p_0(R\geqslant r,A=1,x_r)\\ &= \epsilon\int I(s\geqslant r)I(a=1)p_0(\ell,x_r,s,a)g(\ell,x_r,s,a)d\ell dsda\end{aligned}$$

Thus, the difference $\Delta O_r(x_r)$ can be rewritten as

$$\begin{aligned}\Delta O_r(x_r) =& \frac{\epsilon}{p_0(R\geqslant r,A=1,x_r)}\int [I(s=r)I(a=0) - O_{r,0}(x_r)I(s\geqslant r)I(a=1)]\\ &\times p_0(\ell,x_r,s,a)g(\ell,x_r,s,a)d\ell dsda + O(\epsilon^2)\end{aligned}$$

Now part **B** can be rewritten as

$$\begin{aligned}\mathbf{B} &= \int \Delta O_r(x_r)f(\ell)I(s\geqslant r)I(a=1)p_0(\ell,x_r,s,a)d\ell dx_r dsda\\ &= \int \Delta O_r(x_r)p_0(R\geqslant r,A=1,x_r)f(\ell)p_0(\ell|R\geqslant r,A=1,x_r)d\ell dx_r\\ &= \int \Delta O_r(x_r)p_0(R\geqslant r,A=1,x_r)\underbrace{\left\{\int f(\ell)p_0(\ell|R\geqslant r,A=1,x_r)d\ell\right\}}_{\mathbb{E}_{[f(L)|R\geqslant r,A=1,X_r=x_r]}=m_{r,0}(X_r)}dx_r\\ &= \int \Delta O_r(x_r)p_0(R\geqslant r,A=1,x_r)m_{r,0}(x_r)dx_r\\ &= \epsilon\int [I(s=r)I(a=0) - O_{r,0}(x_r)I(s\geqslant r)I(a=1)]m_{r,0}(x_r)p_0(\ell,x_r,s,a)g(\ell,x_r,s,a)\\ &\quad d\ell dx_r dsda + O(\epsilon^2)\end{aligned}$$

Thus, combining part **A** and **B**, we conclude that

$$\begin{aligned}\mathsf{EIF}_{r,0}(\ell,x_r,s,a) &= f(\ell)O_{r,0}(x_r)I(s\geqslant r)I(a=1)\\ &\quad + [I(s=r)I(a=0) - O_{r,0}(x_r)I(s\geqslant r)I(a=1)]m_{r,0}(x_r) - \theta_{0,r}\\ &= [f(\ell) - m_{r,0}(x_r)]O_r(x_r)I(s\geqslant r,a=1) + I(s=r,a=0)m_{r,0}(x_r) -\\ &\quad \theta_{0,r}\end{aligned}$$

□

We first discuss assumptions for the Donsker classes $\mathcal{F}_r$ and $\mathcal{G}_r$ where we have assumed that $\widehat{O}_r \in \mathcal{F}_r$ and $\widehat{m}_{r,0} \in \mathcal{G}_r$. We denote the $L_2(Q)$ norm as

$$\|f\|_{Q,2} = \int f^2 dQ$$

for a probability measure Q. We assume that $\mathcal{F}_r$ satisifies the following uniform entropy condition (Van Der Vaart and Wellner, 1996):

$$\int_0^\infty \sup_Q \sqrt{\log N(\epsilon \|F_r\|_{Q,2}, \mathcal{F}_\nabla, L_2(Q))} d\epsilon < \infty$$

where the supremum of Q is taken over all finitely discrete probability measures on X. $N(\epsilon\|F_r\|_{Q,2}, \mathcal{F}_\nabla, L_2(Q))$ is the covering number of class $\mathcal{F}_r$ with respect to the $L_2(Q)$ norm (see Definition 2.1.5 of Van Der Vaart and Wellner (1996)) and F_r is an envelope function of $\mathcal{F}_r$ such that $\mathbb{P}F_r^2 < \infty$ with P being the probability measure for X. Similarly we can assume that $\mathcal{G}_r$ satisfies the same uniform entropy condition with envelop function G_r. We also assume that $\mathcal{F}_r$ and $\mathcal{G}_r$ are suitably measurable (see Definition 2.3.3 of Van Der Vaart and Wellner (1996)) and $\mathbb{P}F_r^2 G_r^2 < \infty$. Further, we let $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and $\mathbb{P}_0 f = \int f(x) dP(x)$.

Proof of Theorem 3.3.7. By assumption (M1), we have $\|\widehat{O}_r - O_r^*\|_{L_2(P)} = o_P(1)$ and $\|\widehat{m}_{r,0} - m_{r,0}^*\|_{L_2(P)} = o_p(1)$. The true functions are denoted as $O_r(x_r)$ and $m_{r,0}(x_r)$. When the models are correct, we have $O_r(x_r) = O_r^*(x_r)$ and $m_{r,0}(x_r) = m_{r,0}^*(x_r)$. The multiply-robust estimator has the following form:

$$\begin{aligned}\widehat{\theta}_{\mathsf{MR}} = \frac{1}{n}\sum_{i=1}^n [&f(L_i) I(A_i = 1) + \\ &+ \sum_r \Big(\{f(L_i) - \widehat{m}_{r,0}(X_{i,r})\} \widehat{O}_r(X_{i,r}) I(R_i \geqslant r) I(A_i = 1) \\ &+ \widehat{m}_{r,0}(X_{i,r}) I(R_i = r) I(A_i = 0))]\end{aligned}$$

We first consider another estimator that replaces the estimated functions by the true

functions in the multiply-robust estimator.

$$\begin{aligned}\widetilde{\theta}_{\mathsf{MR}} = \frac{1}{n}\sum_{i=1}^{n}[&f(L_i)I(A_i=1)+\\ &+\sum_r\left(\{f(L_i)-m_{r,0}(X_{i,r})\}O_r(X_{i,r})I(R_i\geqslant r)I(A_i=1)\right.\\ &\left.+m_{r,0}(X_{i,r})I(R_i=r)I(A_i=0)\right)]\end{aligned}$$

It is not hard to prove that $\widetilde{\theta}_{\mathsf{MR}} \to_p \theta$ and

$$\sqrt{n}(\widetilde{\theta}_{\mathsf{MR}} - \theta) \to_d N(0, \sigma^2_{\mathsf{eff}})$$

where σ^2_{eff} is the efficiency bound for estimating θ. Further we have that

$$\begin{aligned}\widehat{\theta}_{\mathsf{MR}} = \widetilde{\theta}_{\mathsf{MR}} + \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\underbrace{\sum_r I(R_i\geqslant r)I(A_i=1)(\widehat{O}_r(X_{i,r})-O_r(X_{i,r}))(f(L_i)-m_{r,0}(X_{i,r}))}_{(\mathbf{I})}\\ &+\underbrace{\sum_r(\widehat{m}_{r,0}(X_{i,r})-m_{r,0}(X_{i,r}))\left[I(R_i=r)I(A_i=0)-I(R_i\geqslant r)I(A_i=1)O_r(X_{i,r})\right]}_{(\mathbf{II})}\\ &\underbrace{-\sum_r I(R_i\geqslant r)I(A_i=1)\left[\widehat{m}_{r,0}(X_{i,r})-m_{r,0}(X_{r,i})\right]\left[\widehat{O}_r(X_{i,r})-O_r(X_{i,r})\right]}_{(\mathbf{III})}\Bigg\}\end{aligned} \tag{B.1}$$

We first prove the multiply-robust property when the odds are correctly specified and the regression functions are misspecified. Thus, $O_r(x_r) = O^*_r(x_r)$ and $m_{r,0}(x_r) \neq m^*_{r,0}(x_r)$. Since $\|\widehat{m}_{r,0} - m_{r,0}\|_{L_2(P)}\|\widehat{O}_r - O_r\|_{L_2(P)} = o_P(1)$, we have $\|\widehat{O}_r - O_r\|_{L_2(P)} = o_P(1)$. Denote $g_{r,n}(x_r, l, s, a) = I(s \geqslant r)I(a = 1)(\widehat{O}_r(x_r) - O_r(x_r))(f(l) - m_{r,0}(x_r))$ and similarly

$$g_{r,0}(x_r, l, s, a) = I(s \geqslant r)I(a = 1)(O^*_r(x_r) - O_r(x_r))(f(l) - m_{r,0}(x_r)) = 0.$$

Then we can write term **I** in (B.1) for pattern r as

$$\frac{1}{n}\sum_{i=1}^{n} I(R_i \geqslant r)I(A_i = 1)(\widehat{O}_r(X_{i,r}) - O_r(X_{i,r}))(f(L_i) - m_{r,0}(X_{i,r}))$$

$$= \mathbb{P}_n g_{r,n} = \mathbb{P}_n(g_{r,n} - g_{r,0})$$

Given that $\|\widehat{O}_r - O_r\|_{L_2(P)} = o_P(1)$, we have $\|g_{r,n} - g_{r,0}\|_{L_2(P)} = o_P(1)$ as $f(l), m_{r,0}(x_r)$ are uniformly bounded. Further, we have

$$\begin{aligned} g_{r,n}(x_r, l, s, a) &= I(s \geqslant r)I(a = 1)(\widehat{O}_r(x_r) - O_r(x_r))(f(l) - m_{r,0}(x_r)) \\ &= I(s \geqslant r)I(a = 1)(f(l) - m_{r,0}(x_r))\widehat{O}_r(x_r) - g_{1,r}(x_r, l, s, a) \end{aligned}$$

where $g_{1,r}(x_r, l, s, a)$ is a function that does not involve $\widehat{O}_r$. As $\widehat{O}_r(x_r)$ is in a Donsker class $\mathcal{F}_r$, we have $g_{r,n}$ is also in a Donsker class

$$\mathcal{F}^* = \{I(s \geqslant r)I(a = 1)(f(l) - m_{r,0}(x_r))h(x_r) - g_{1,r}(x_r, l, s, a) : h \in \mathcal{F}\}$$

by Example 2.10.7 and 2.10.23 of Van Der Vaart and Wellner (1996) and the assumptions we made before this proof. By Lemma 19.24 of Van Der Vaart (2000), we then have

$$(\mathbb{P}_n - \mathbb{P}_0)(g_{r,n} - g_{r,0}) = o_P(n^{-1/2}) \Rightarrow (\mathbb{P}_n - \mathbb{P}_0)g_{r,n} = o_P(n^{-1/2}) \Rightarrow \mathbb{P}_n g_{r,n} = o_P(n^{-1/2})$$

realizing that $\mathbb{P}_0 g_{r,n} = 0$. For term **II**, similarly define

$$g'_{r,n}(x_r, s, a) = (\widehat{m}_{r,0}(x_r) - m_{r,0}(x_r))[I(s = r)I(a = 0) - I(s \geqslant r)I(a = 1)O_r(x_r)]$$

and

$$g'_{r,0}(x_r, s, a) = (m^*_{r,0}(x_r) - m_{r,0}(x_r))[I(s = r)I(a = 0) - I(s \geqslant r)I(a = 1)O_r(x_r)]$$

Then we can rewrite term **II** in (B.1) for pattern r as

$$\begin{aligned} &\frac{1}{n}\sum_i (\widehat{m}_{r,0}(X_{i,r}) - m_{r,0}(X_{i,r}))[I(R_i = r)I(A_i = 0) - I(R_i \geqslant r)I(A_i = 1)O_r(X_{i,r})] \\ &= (\mathbb{P}_n - \mathbb{P}_0)(g'_{r,n} - g'_{r,0}) \\ &+ \frac{1}{n}\sum_{i=1}^{n}(m^*_{r,0}(X_{i,r}) - m_{r,0}(X_{i,r})) \\ &[I(R_i = r)I(A_i = 0) - I(R_i \geqslant r)I(A_i = 1)O_r(X_{i,r})] \end{aligned}$$

First note that

$$\frac{1}{n}\sum_{i=1}^{n}(m^*_{r,0}(X_{i,r}) - m_{r,0}(X_{i,r}))[I(R_i = r)I(A_i = 0) - I(R_i \geqslant r)I(A_i = 1)O_r(X_{i,r})]$$
$$\rightarrow_p 0$$

This implies that term **II** in (B.1) for pattern r can be written as

$$(\mathbb{P}_n - \mathbb{P}_0)(g'_{r,n} - g'_{r,0}) + o_P(1)$$

Then again given that $\widehat{m}_{r,0}(x_r)$ is in a Donsker class $\mathcal{G}_r$ and $\|\widehat{m}_{r,0} - m^*_{r,0}\|_{L_2(P)} = o_P(1)$, we have that $\|g'_{r,n} - g'_{r,0}\|_{L_2(P)} = o_P(1)$ and $g'_{r,n}$ is also in a Donsker class by a similar reasoning as before. By Lemma 19.24 of Van Der Vaart (2000), we have

$$(\mathbb{P}_n - \mathbb{P}_0)(g'_{r,n} - g'_{r,0}) = o_P(n^{-1/2}) \Rightarrow (\mathbb{P}_n - \mathbb{P}_0)g'_{r,n} = (\mathbb{P}_n - \mathbb{P}_0)g'_{r,0} + o_P(n^{-1/2})$$

This implies that

$$\mathbb{P}_n g'_{r,n} = o_P(1) + o_P(n^{-1/2}) + \mathbb{P}_0 g'_{r,n} = o_P(1) + o_P(n^{-1/2}) = o_P(1)$$

as $\mathbb{P}_0 g'_{r,n} = 0$. For term **III**, we can similarly define

$$h_n(x_r, s, a) = (\widehat{m}_{r,0}(x_r) - m_{r,0}(x_r))(\widehat{O}_r(x_r) - O_r(x_r))I(s \geqslant r)I(a = 1)$$

and $h_0(x_r, s, a) = 0$. Then $h_n(x_r, s, a)$ is in a Donsker Class by Example 2.10.23 of Van Der Vaart and Wellner (1996). Given that $\|\widehat{O}_r - O_r\|_{L_2(P)} = o_P(1)$ and $\widehat{m}_{r,0}, m_{r,0}$ are uniformly bounded, we have $\|h_n - h_0\|_{L_2(P)} = o_P(1)$. Then by Lemma 19.24 of Van Der Vaart (2000), we have

$$(\mathbb{P}_n - \mathbb{P}_0)(h_n - h_0) = o_P(n^{-1/2}) \Rightarrow (\mathbb{P}_n - \mathbb{P}_0)h_n = o_P(n^{-1/2})$$
$$\Rightarrow \mathbb{P}_n h_n = \mathbb{P}_0 h_n + o_P(n^{-1/2})$$

and

$$\mathbb{P}_0 h_n \leqslant \|\widehat{m}_{r,0} - m_{r,0}\|_{L_2(P)}\|\widehat{O}_r - O_r\|_{L_2(P)} = o_P(1)$$

given the assumption that

$$\sum_r \|\widehat{m}_{r,0} - m_{r,0}\|_{L_2(P)} \|\widehat{O}_r - O_r\|_{L_2(P)} = o_P(1)$$

Above results imply that $\widehat{\theta}_{\mathsf{MR}} = \widetilde{\theta}_{\mathsf{MR}} + o_P(1) \to_p \theta$. By similar reasoning, when we have odds function misspecified and regression function correctly specified, we also have $\widehat{\theta}_{\mathsf{MR}} \to_p \theta$.

When we have both models correctly specified, by a similar proof as above, term **I** has leading term on the order of $o_P(n^{-1/2})$ for each pattern r. Similarly, **II** is also $o_P(n^{-1/2})$. Thus, we have

$$\begin{aligned}
&\widehat{\theta}_{\mathsf{MR}} - \widetilde{\theta}_{\mathsf{MR}} = \\
&- \frac{1}{n} \sum_{i=1}^{n} \sum_r I(R_i \geqslant r) I(A_i = 1)(\widehat{m}_{r,0}(X_{i,r}) - m_{r,0}(X_{r,i}))(\widehat{O}_r(X_{i,r}) - O_r(X_{i,r})) \\
&+ o_p(n^{-1/2})
\end{aligned}$$

Finally, for term **III**, by the same proof, we have

$$\mathbb{P}_n h_n = \mathbb{P}_0 h_n + o_P(n^{-1/2}) \leqslant \|\widehat{m}_{r,0} - m_{r,0}\|_{L_2(P)} \|\widehat{O}_r - O_r\|_{L_2(P)} + o_P(n^{-1/2}) = o_P(n^{-1/2})$$

assuming that

$$\sqrt{n} \sum_r \|\widehat{m}_{r,0} - m_{r,0}\|_{L_2(P)} \|\widehat{O}_r - O_r\|_{L_2(P)} = o_P(1)$$

Together, we have proved that

$$\begin{aligned}
&\sqrt{n}(\widehat{\theta}_{\mathsf{MR}} - \theta) = \sqrt{n}(\widehat{\theta}_{\mathsf{MR}} - \widetilde{\theta}_{\mathsf{MR}}) + \sqrt{n}(\widetilde{\theta}_{\mathsf{MR}} - \theta) = o_P(1) + \sqrt{n}(\widetilde{\theta}_{\mathsf{MR}} - \theta) \\
&\to_d N(0, \sigma^2_{\mathsf{eff}})
\end{aligned}$$

□

In fact, when we use parametric estimators for both $\widehat{m}_{r,0}$ and $\widehat{O}_r$, as long as for each pattern r, either $m_{r,0}(x_r; \beta_r^*) = m_{r,0}(x_r)$ or $O_r(x_r; \alpha_r^*) = O_r(x_r)$, we have the following

asymptotic linear expansion for $\widehat{\theta}_{\mathsf{MR}}$ as

$$\sqrt{n}(\widehat{\theta}_{\mathsf{MR}} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(L_i, X_i, R_i, A_i; \beta^*, \alpha^*) + f(L_i) I(A_i = 1) +$$
$$\sum_r \big\{ \mathbb{E}[\nabla_{\beta_r} h(L, X, R, A; \beta^*)^T] \psi_{r,\beta_r^*}(L, X_r, R, A) +$$
$$\mathbb{E}[\nabla_{\alpha_r} h(L, X, R, A; \alpha^*)^T] \psi_{r,\alpha_r^*}(X_r, R, A) \big\} - \theta_0 \big] + o_P(1)$$

with

$$h(L, X, R, A; \beta^*, \alpha^*) = \sum_r \{ [f(L) - m_{r,0}(X_r; \beta_r^*)] O_r(X_r; \alpha_r^*) I(R \geqslant r) I(A = 1) +$$
$$m_{r,0}(X_r; \beta_r^*) I(R = r) I(A = 0) \}$$

This will be used when we run the simulation studies.

B.3 Proof for multiple primary variables

Now we present the proof for the results when there are multiple primary variables. The proof for Proposition 3.4.1 is omitted as it is very similar to the proof in the single variable case.

First, for the IPW estimator, we again give the influence function when we estimate $O_{r,a}$ with a parametric model. We have that

$$\sqrt{n}(\widehat{\alpha}_{r,a} - \alpha_{r,a}^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{r,a}(X_{i,r}, L_{i,a}, R_i, A_i) + o_p(1) \to_d N(0, \Sigma(\alpha_{r,a}^*)^{-1})$$

where

$$\psi_{r,a}(X_r, L_a, R, A) = \Sigma(\alpha_{r,a}^*)^{-1} \left[I(R = r, A = a) \frac{\nabla_{\alpha_{r,a}} O_{r,a}(X_r, L_a; \alpha_{r,a}^*)}{O_{r,a}(X_r, L_a; \alpha_{r,a}^*)} - \right.$$
$$\left. \{ I(R = r, A = a) + I(R \geqslant r, A = 1_d) \} \frac{\nabla_{\alpha_{r,a}} O_{r,a}(X_r, L_a; \alpha_{r,a}^*)}{1 + O_{r,a}(X_r, L_a; \alpha_{r,a}^*)} \right]$$

with

$$\Sigma(\alpha_{r,a}^*) = \mathbb{E}\big[P(R \geqslant r, A = 1_d | X_r, L_a) \nabla_{\alpha_{r,a}} O_{r,a}(X_r, L_a; \alpha_{r,a})^{\otimes 2}$$
$$\frac{1}{O_{r,a}(X_r, L_a; \alpha_{r,a}^*)(1 + O_{r,a}(X_r, L_a; \alpha_{r,a}^*))} \big]$$

PROOF OF THEOREM 3.4.2. The proof is similar to the proof of Theorem 3.3.4 and we directly give the results. Denote

$$\begin{aligned}\phi(X, L, R, A, \alpha^*) = \sum_{r,a\neq 1_d} \big(\mathbb{E}\left[f(L)I(R \geqslant r, A = 1_d)\nabla_{\alpha_{r,a}} O_{r,a}(X_r, L_a; \alpha^*_{r,a})\right] \\ \psi_{r,a}(X_r, L_a, R, A) \\ + f(L)I(R \geqslant r, A = 1_d)O_{r,a}(X_r, L_a)) + f(L)I(A = 1_d) - \theta_0\end{aligned}$$

and we have

$$\sqrt{n}(\widehat{\theta}_{\mathsf{IPW}} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(X_i, L_i, R_i, A_i, \alpha^*) + o_P(1) \to N(0, \sigma^2_{\mathsf{IPW}})$$

with $\sigma^2_{\mathsf{IPW}} = \mathrm{Var}[\phi(X, L, R, A; \alpha^*)]$. □

For the regression adjustment method, we have that

$$\sqrt{n}(\widehat{\beta}_{r,a} - \beta^*_{r,a}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{r,a}(X_{i,r}, L_{i,a}, R_i, A_i) + o_P(1) \to_d N(0, \Sigma(\beta^*_{r,a})^{-1})$$

where

$$\begin{aligned}\psi_{r,a}(X_r, L_a, R, A) = \nabla S(\beta^*_{r,a})^{-1} I(R \geqslant r, A = 1_d)[f(L) - m_{r,a}(X_r, L_a; \beta^*_{r,a})] \\ \nabla_{\beta_{r,a}} m_{r,a}(X_r, L_a, \beta^*_{r,a})\end{aligned}$$

with

$$\nabla S(\beta^*_{r,a}) = \mathbb{E}[I(R \geqslant r, A = 1)\nabla_{\beta_{r,a}} m^{\otimes 2}_{r,a}]$$

PROOF OF THEOREM 3.4.3. The proof is again very similar to the proof of Theorem 3.3.5 and we directly give the results. Now we have

$$\begin{aligned}\phi(X, L, R, A; \beta^*) = \sum_{r,a\neq 1_d} \big(\mathbb{E}[I(R = r, A = a)\nabla_{\beta_{r,a}} m_{r,a}(X_r, L_a; \beta^*_{r,a})]\psi_{r,a}(X_r, L_a, R, A) \\ + m_{r,a}(X_r, L_a; \beta^*_{r,a})I(R = r, A = a)\big) + f(L)I(A = 1_d)\end{aligned}$$

Then, we have

$$\sqrt{n}(\widehat{\theta}_{\mathsf{RA}} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(X_i, L_i, R_i, A_i; \beta^*) + o_P(1) \to_d N(0, \sigma^2_{\mathsf{RA}})$$

with $\sigma^2_{\mathsf{RA}} = \mathrm{Var}[\phi(X, L, R, A)]$. □

We assume that $\mathcal{F}_{r,a}$ and $\mathcal{G}_{r,a}$ satisfy the uniform entropy condition with envelop functions $F_{r,a}$ and $G_{r,a}$. We further assume that $\mathcal{F}_{r,a}$ and $\mathcal{G}_{r,a}$ are suitably measurable and $\mathbb{P}F^2_{r,a}G^2_{r,a} < \infty$. The proof of Theorems 3.4.4 and 3.4.5 is omitted as it is almost identical to the proof of Theorems 3.3.6 and 3.3.7.

B.4 Proof for marginal parametric model

Under mild regularity conditions, we can prove that $\widehat{\theta} \to_p \theta^*$ by theorem 5.9 of Van Der Vaart (2000).

Proof of theorem 3.5.3. The sample estimating equation for the marginal parametric model under the ACCMV assumption is as following:

$$\sum_{i=1}^{n} s(\widehat{\theta}|L_i) \left[\sum_{r,a \neq \mathbb{1}_d} O_{r,a}(X_{i,r}, L_{i,a}; \widehat{\alpha}_{r,a}) I(A_i = \mathbb{1}_d) I(R_i \geqslant r) + I(A_i = \mathbb{1}_d) \right] = 0$$

We can define

$$\psi_{\theta,\alpha}(L, X, R, A) = s(\theta|L) \left[\sum_{r,a \neq \mathbb{1}_d} O_{r,a}(X_r, L_a; \alpha^*_{r,a}) I(A = \mathbb{1}_d) I(R \geqslant r) + I(A = \mathbb{1}_d) \right]$$

then we have

$$\mathbb{P}_n \psi_{\widehat{\theta},\widehat{\alpha}} = 0 \Rightarrow \mathbb{P}_n \psi_{\widehat{\theta},\widehat{\alpha}} - \mathbb{P}_n \psi_{\widehat{\theta},\alpha^*} + \mathbb{P}_n \psi_{\widehat{\theta},\alpha^*} = 0$$

Define

$$\phi_{\theta,\alpha_{r,a}}(L, X, R, A) = s(\theta|L) \nabla_{\alpha_{r,a}} O_{r,a}(X_r, L_a; \alpha_{r,a}) I(A = \mathbb{1}_d) I(R \geqslant r)$$

Then we have that

$$\mathbb{P}_n \left[\psi_{\widehat{\theta},\widehat{\alpha}} - \psi_{\widehat{\theta},\alpha^*} \right] = \sum_{r,a \neq \mathbb{1}_d} \mathbb{P}_n \phi^T_{\widehat{\theta},\alpha^*_{r,a}} (\widehat{\alpha}_{r,a} - \alpha^*_{r,a}) + o_P(1) \|\widehat{\alpha} - \alpha\|$$

where we also have

$$\widehat{\alpha}_{r,a} - \alpha^*_{r,a} = \mathbb{P}_n \xi_{r,a} + o_P(1/\sqrt{n})$$

based on our assumption. Thus, put everything together and multiply by $\sqrt{n}$ on both sides of the equation, we have

$$\sum_{r,a \neq 1_d} \mathbb{P}_n \phi^T_{\widehat{\theta},\alpha^*_{r,a}} \sqrt{n} \mathbb{P}_n \xi_{r,a} + o_P(1) + \sqrt{n} \mathbb{P}_n \psi_{\widehat{\theta},\alpha^*} = 0$$

next, we have that

$$\mathbb{P}_n \phi_{\widehat{\theta},\alpha^*_{r,a}} = \mathbb{P}_n \phi_{\widehat{\theta},\alpha^*_{r,a}} - \mathbb{P}_0 \phi_{\theta_0,\alpha^*_{r,a}} + \mathbb{P}_0 \phi_{\theta_0,\alpha^*_{r,a}}$$

Further, we have

$$\mathbb{P}_n \phi_{\widehat{\theta},\alpha^*_{r,a}} - \mathbb{P}_0 \phi_{\theta_0,\alpha^*_{r,a}} =$$
$$\underbrace{(\mathbb{P}_n - \mathbb{P}_0)(\phi_{\widehat{\theta},\alpha^*_{r,a}} - \phi_{\theta_0,\alpha^*_{r,a}})}_{\text{I}} + \underbrace{\mathbb{P}_0(\phi_{\widehat{\theta},\alpha^*_{r,a}} - \phi_{\theta_0,\alpha^*_{r,a}})}_{\text{II}} + \underbrace{(\mathbb{P}_n - \mathbb{P}_0)\phi_{\theta_0,\alpha^*_{r,a}}}_{\text{III}}$$

Now for term (I), we may use Lemma 19.24 of Van Der Vaart (2000) to prove that term I is $o_P(n^{-1/2})$ under the condition that $\phi_{\theta,\alpha^*_{r,a}}$ lies in a Donsker class. For term III, it is simply $o_P(1)$ by weak law of large numbers. For term II, it is also $o_P(1)$ as $\widehat{\theta} \to_p \theta$. Thus, this implies that

$$\mathbb{P}_n \phi_{\widehat{\theta},\alpha^*_{r,a}} = o_P(1) + \mathbb{P}_0 \phi_{\theta_0,\alpha^*_{r,a}}$$

Thus, put everything together, we have that

$$\sqrt{n} \mathbb{P}_n \psi_{\widehat{\theta},\alpha^*} + \sum_{r,a \neq \mathbb{1}_d} \mathbb{P}_0 \phi_{\theta_0,\alpha^*_{r,a}} \sqrt{n} \mathbb{P}_n \xi_{r,a} + o_P(1) = 0$$

Next, we have

$$\begin{aligned}
\sqrt{n} \mathbb{P}_n \psi_{\widehat{\theta},\alpha^*} &= \sqrt{n}(\mathbb{P}_n \psi_{\widehat{\theta},\alpha^*} - \mathbb{P}_0 \psi_{\theta_0,\alpha^*}) \\
&= \sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)(\psi_{\widehat{\theta},\alpha^*} - \psi_{\theta_0,\alpha^*}) + \sqrt{n}\mathbb{P}_0(\psi_{\widehat{\theta},\alpha^*} - \psi_{\theta_0,\alpha^*}) \\
&+ \sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)\psi_{\theta_0,\alpha^*} \\
&= o_P(1) + \sqrt{n}\nabla_\theta \mathbb{P}_0 \psi_{\theta_0,\alpha^*}(\widehat{\theta} - \theta_0) + \sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)\psi_{\theta_0,\alpha^*}
\end{aligned}$$

Next, put everything together, we have

$$\sqrt{n}\nabla_\theta \mathbb{P}_0 \psi_{\theta_0,\alpha*}(\widehat{\theta}-\theta_0) = -\sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)\psi_{\theta_0,\alpha*} - \sum_{r,a\neq \mathbb{1}_d} \mathbb{P}_0 \phi_{\theta_0,\alpha^*_{r,a}} \sqrt{n}\mathbb{P}_n \xi_{r,a} + o_P(1)$$

which implies that

$$\sqrt{n}(\widehat{\theta}-\theta_0) = -(\nabla_\theta \mathbb{P}_0 \psi_{\theta_0,\alpha*})^{-1} \left[\sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)\psi_{\theta_0,\alpha*} + \sum_{r,a\neq \mathbb{1}_d} \mathbb{P}_0 \phi_{\theta_0,\alpha^*_{r,a}} \sqrt{n}\mathbb{P}_n \xi_{r,a} \right]$$
$$+ o_P(1)$$

Thus, we have the desired asymptotic normality for $\widehat{\theta}$. □

B.5 Derivations for simulation studies

We first derive $\mathbb{E}[Y_3]$ with regression adjustment for single variables case. We have that

$$p(y_3|A=1, R \geqslant 00) = \frac{P(y_3, A=1, R\geqslant 00)}{P(A=1, R\geqslant 00)} = \frac{3}{4}\phi_{1,1}(y_3) + \frac{1}{4}\phi_{0,1}(y_3)$$

Thus, we have that $E[Y_3|A=1, R\geqslant 00] = 3/4$. Under ACCMV assumption, we can then identify $P(y_3|A=0, R=00)$ as follows:

$$P(y_3|A=0, R=00) = p(y_3|A=1, R\geqslant 00) = \frac{3}{4}\phi_{1,1}(y_3) + \frac{1}{4}\phi_{0,1}(y_3)$$

Next, we have that

$$P(y_3|A=1, R\geqslant 01, y_2) = \frac{P(y_3, A=1, R\geqslant 01, y_2)}{P(A=1, R\geqslant 01, y_2)}$$

Under ACCMV assumption, we have that

$$P(y_3|A=1, R=01, y_2) = P(y_3|A=1, R\geqslant 01, y_2) = \frac{1}{2}\left(\frac{\phi_{\mu_2,\Sigma_2}(y_3,y_2)}{\phi_{-1,1}(y_2)} + \frac{\phi_{\mu_4,\Sigma_2}(y_3,y_2)}{\phi_{-1,1}(y_2)} \right)$$

with $\mu_4 = (0,-1)^T$. Then we can compute that

$$E[Y_3|A=1, R\geqslant 01, Y_2] = \frac{Y_2}{2} + 1$$

Similarly, we have that

$$E(Y_3|A=1, R \geqslant 10, Y_1) = \frac{Y_1}{2} + 1$$

Finally, we also have that

$$E(Y_3|A=1, R=11, Y_1, Y_2) = \frac{1}{3}(Y_1 + Y_2) + \frac{2}{3}$$

Thus, we could compute the parameter of interest $\mathbb{E}[Y_3]$ as

$$E[Y_3] = \mathbb{E}[Y_3 I(A=1)] + \sum_r \mathbb{E}[m_{r,0}(X_r) I(R=r, A=0)]$$

where $\mathbb{E}[Y_3 I(A=1)] = \mathbb{E}[Y_3 I(A=1, R \geqslant 00)] = 3/8$ and

$$\begin{aligned}
\mathbb{E}[m_{00,0}(X_{00}) I(R=00, A=0)] &= \frac{3}{32} \\
\mathbb{E}[m_{10,0}(X_{10}) I(R=10, A=0)] &= \frac{1}{2}\mathbb{E}[Y_1 I(A=0, R=01)] + \frac{1}{8} = \frac{3}{16} \\
\mathbb{E}[m_{01,0}(X_{01}) I(R=01, A=0)] &= \frac{1}{2}\mathbb{E}[Y_2 I(A=0, R=10)] + \frac{1}{8} = \frac{3}{16} \\
\mathbb{E}[m_{11,0}(X_{11}) I(R=11, A=0)] &= \frac{1}{3}\mathbb{E}[(Y_1 + Y_2) I(A=0, R=11)] + \frac{1}{12} = \frac{1}{12}
\end{aligned}$$

For IPW estimation of $\mathbb{E}[Y_3]$, we first have that

$$O_{00} = \frac{P(A=0, R=00)}{P(A=1, R \geqslant 00)} = \frac{1}{4}$$

Further, when $r = 10$, we have

$$O_{10}(y_1) = \frac{P(A=0, R=10|y_1)}{P(A=1, R \geqslant 10|y_1)} = \frac{P(y_1, A=0, R=10)}{P(y_1, A=1, R \geqslant 10)} = \frac{1}{2}\exp(2y_1)$$

Similarly, we have that

$$O_{01}(y_2) = \frac{P(A=0, R=01|y_2)}{P(A=1, R \geqslant 01|y_2)} = \frac{P(y_2, A=0, R=01)}{P(y_2, A=1, R \geqslant 01)} = \frac{1}{2}\exp(2y_2)$$

Finally, for $r = 11$, we have that

$$\begin{aligned}
O_{11}(y_1, y_2) &= \frac{P(A=0, R=11|y_1, y_2)}{P(A=1, R=11|y_1, y_2)} = \frac{P(y_1, y_2, A=0, R=11)}{P(y_1, y_2, A=1, R=11)} \\
&= \frac{\phi_{\mu_1, \Sigma_1}(y_1, y_2)}{\phi_{(-1,-1), \Sigma_1}(y_1, y_2)} = \exp\left(\frac{8}{3}y_1 - \frac{4}{3}y_2 - \frac{4}{3}\right)
\end{aligned}$$

Now we move to the case of multiple primary variables and derive $\mathbb{E}[Y_3 Y_4]$. We have that

$$P(y_3, y_4 | A = 00, R = 00) = P(y_3, y_4 | A = 11, R \geqslant 00) = \phi_{\mathbb{1}_2, \Sigma_2}(y_3, y_4)$$

Then we have

$$\mathbb{E}[Y_3 Y_4 | A = 11, R \geqslant 00] = 3/2$$

Next, we have

$$P(y_3, y_4 | A = 00, R = 01, y_2) = P(y_3, y_4 | A = 11, R \geqslant 01, y_2) = \frac{P(y_2, y_3, y_4, A = 11, R \geqslant 01)}{P(y_2, A = 11, R \geqslant 01)}$$

and

$$Y_3, Y_4 | A = 11, R \geqslant 01, Y_2 \sim N\left(\begin{pmatrix} \frac{1}{2} Y_2 + \frac{1}{2} \\ \frac{1}{2} Y_2 + \frac{1}{2} \end{pmatrix}, \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix} \right)$$

Then we have that

$$\mathbb{E}[Y_3 Y_4 | A = 11, R \geqslant 01, Y_2] = \frac{1}{4} + \left(\frac{1}{2} Y_2 + \frac{1}{2} \right)^2$$

Next, we have

$$\begin{aligned} &P(y_3, y_4 | A = 00, R = 10, y_1) = P(y_3, y_4 | A = 11, R \geqslant 10, y_1) \\ &= \frac{P(y_1, y_3, y_4, A = 11, R \geqslant 10)}{P(y_1, A = 11, R \geqslant 10)} \end{aligned}$$

Thus, we now have that

$$\mathbb{E}[Y_3 Y_4 | A = 11, R \geqslant 10, Y_1] = \frac{1}{4} + \left(\frac{1}{2} Y_1 + \frac{1}{2} \right)^2$$

Next, we have that

$$\begin{aligned} &P(y_3, y_4 | A = 00, R = 11, y_1, y_2) = P(y_3, y_4 | A = 11, R = 11, y_1, y_2) \\ &= \frac{\phi_{\mathbb{1}_4, \Sigma_4}(y_1, y_2, y_3, y_4)}{\phi_{\mathbb{1}_2, \Sigma_2}(y_1, y_2)} \end{aligned}$$

Thus, we have that

$$Y_3, Y_4 | A = 11, R = 11, Y_1, Y_2 \sim N \left(\begin{pmatrix} \frac{1}{3}(Y_1 + Y_2 + 1) \\ \frac{1}{3}(Y_1 + Y_2 + 1) \end{pmatrix}, \begin{pmatrix} 2/3 & 1/6 \\ 1/6 & 2/3 \end{pmatrix} \right)$$

and

$$\mathbb{E}[Y_3 Y_4 | A = 11, R = 11, Y_1, Y_2] = \frac{1}{6} + \frac{1}{9}(Y_1 + Y_2 + 1)^2$$

Next consider the case $A = 01$, we have

$$\begin{aligned} &P(y_3 | y_4, A = 01, R = 00) = p(y_3 | y_4, A = 11, R \geqslant 00) \\ &= \frac{p(y_3, y_4, A = 11, R \geqslant 00)}{p(y_4, A = 11, R \geqslant 00)} \end{aligned}$$

Thus, we have that

$$Y_3 | Y_4, A = 11, R \geqslant 00 \sim N \left(\frac{1}{2} Y_4 + \frac{1}{2}, \frac{3}{4} \right)$$

and then

$$\mathbb{E}[Y_3 Y_4 | Y_4, R \geqslant 00, A = 11] = Y_4 \mathbb{E}[Y_3 | Y_4, R \geqslant 00, A = 11] = \frac{1}{2} Y_4 (Y_4 + 1)$$

Next, we have

$$\begin{aligned} &P(y_3 | y_4, A = 01, R = 01, y_2) = P(y_3 | y_2, y_4, A = 11, R \geqslant 01) \\ &= \frac{P(y_2, y_3, y_4, A = 11, R \geqslant 01)}{P(y_2, y_4, A = 11, R \geqslant 01)} \end{aligned}$$

Then we have

$$Y_3 | Y_2, Y_4, A = 11, R \geqslant 01 \sim N \left(\frac{1}{3}(Y_2 + Y_4 + 1), \frac{2}{3} \right)$$

Thus, we have

$$\mathbb{E}[Y_3 Y_4 | Y_2, Y_4, A = 11, R \geqslant 01] = \frac{1}{3} Y_4 (Y_2 + Y_4 + 1)$$

Similarly, we have

$$P(y_3|y_4, A = 01, R = 10, y_1) = P(y_3|y_4, A = 11, R \geqslant 10, y_1)$$

and we can get that $Y_3|Y_1, Y_4, A = 11, R \geqslant 10 \sim N\left(\frac{1}{3}(Y_1 + Y_4 + 1), \frac{2}{3}\right)$. Thus, we have

$$\mathbb{E}[Y_3Y_4|Y_1, Y_4, A = 11, R \geqslant 10] = \frac{1}{3}Y_4(Y_1 + Y_4 + 1)$$

Next, we have

$$\begin{aligned} &P(y_3|y_4, A = 01, R = 11, y_1, y_2) = P(y_3|y_4, A = 11, R = 11, y_1, y_2) \\ &= \frac{P(y_1, y_2, y_3, y_4, A = 11, R = 11)}{P(y_1, y_2, y_4, A = 11, R = 11)} \end{aligned}$$

and we can get that

$$Y_3|Y_4, A = 11, R = 11, Y_1, Y_2 \sim N\left(\frac{1}{4}(Y_1 + Y_2 + Y_4 + 1), \frac{5}{8}\right)$$

Thus, we have

$$\mathbb{E}[Y_3Y_4|Y_4, A = 11, R = 11, Y_1, Y_2] = \frac{1}{4}(Y_1 + Y_2 + Y_4 + 1)Y_4$$

Next consider the case $A = 10$, we have

$$P(y_4|y_3, A = 10, R = 00) = P(y_4|y_3, A = 11, R \geqslant 00)$$

By symmetry, we have that $Y_4|Y_3, A = 11, R \geqslant 00 \sim N(\frac{1}{2}Y_3 + \frac{1}{2}, \frac{3}{4})$. Thus, we have that

$$\mathbb{E}[Y_3Y_4|Y_3, A = 11, R \geqslant 00] = \frac{1}{2}Y_3(Y_3 + 1)$$

Next we have

$$P(y_4|y_3, A = 10, R = 01, y_2) = P(y_4|y_2, y_3, A = 11, R \geqslant 01)$$

Again similarly, we have that $Y_4|Y_2, Y_3, A = 11, R \geqslant 01 \sim N\left(\frac{1}{3}(Y_2 + Y_3 + 1), \frac{2}{3}\right)$. Thus, we have that

$$\mathbb{E}[Y_3Y_4|Y_2, Y_3, A = 11, R \geqslant 01] = \frac{1}{3}Y_3(Y_2 + Y_3 + 1)$$

Next, we have

$$P(y_4|y_3, A = 10, R = 10, y_1) = P(y_4|y_3, A = 11, R \geqslant 10, y_1)$$

and similarly we have $Y_4|Y_1, Y_3, A = 11, R \geqslant 10 \sim N\left(\frac{1}{3}(Y_1 + Y_3 + 1), \frac{2}{3}\right)$. Thus, we have that

$$\mathbb{E}[Y_3Y_4|Y_1, Y_3, A = 11, R \geqslant 10] = \frac{1}{3}Y_3(Y_1 + Y_3 + 1)$$

Next, we have that

$$P(y_4|y_3, A = 10, R = 11, y_1, y_2) = P(y_4|y_1, y_2, y_3, A = 11, R = 11)$$

and similarly we have $Y_4|Y_1, Y_2, Y_3, A = 11, R = 11 \sim N\left(\frac{1}{4}(Y_1 + Y_2 + Y_3 + 1), \frac{5}{8}\right)$ and we have that

$$\mathbb{E}[Y_3Y_4|Y_1, Y_2, Y_3, A = 11, R = 11] = \frac{1}{4}Y_3(Y_1 + Y_2 + Y_3 + 1)$$

Thus, we could now compute the parameter of interest $\mathbb{E}[Y_3Y_4]$ as

$$\mathbb{E}[Y_3Y_4] = \mathbb{E}[Y_3Y_4I(A = 11)] + \sum_{r,a\neq 11} \mathbb{E}[Y_3Y_4I(A = a, R = r)]$$

where

$$\mathbb{E}[Y_3Y_4I(A = 11)] = \mathbb{E}[\mathbb{E}[Y_3Y_4|A = 11]I(A = 11)] = \frac{3}{2} * P(A = 11) = \frac{3}{8}$$

Next, when $a = 00$, we have

$$\mathbb{E}[Y_3Y_4I(A = 00, R = 00)] = \mathbb{E}[m_{00,00}(X_{00}, L_{00})I(A = 00, R = 00)] = \frac{3}{2} \times \frac{1}{16} = \frac{3}{32}$$
$$\mathbb{E}[Y_3Y_4I(A = 00, R = 01)] = \mathbb{E}[m_{01,00}(X_{01}, L_{00})I(A = 00, R = 01)] = \frac{17}{256}$$
$$\mathbb{E}[Y_3Y_4I(A = 00, R = 10)] = \mathbb{E}[m_{10,00}(X_{10}, L_{00})I(A = 00, R = 10)] = \frac{17}{256}$$
$$\mathbb{E}[Y_3Y_4I(A = 00, R = 11)] = \mathbb{E}[m_{11,00}(X_{11}, L_{00})I(A = 00, R = 11)] = \frac{3}{32}$$

Next, when $a = 01$, we have

$$\mathbb{E}[Y_3Y_4I(A = 01, R = 00)] = \mathbb{E}[m_{00,01}(X_{00}, L_{01})I(A = 01, R = 00)] = \frac{7}{128}$$
$$\mathbb{E}[Y_3Y_4I(A = 01, R = 01)] = \mathbb{E}[m_{01,01}(X_{01}, L_{01})I(A = 01, R = 01)] = \frac{3}{32}$$
$$\mathbb{E}[Y_3Y_4I(A = 01, R = 10)] = \mathbb{E}[m_{10,01}(X_{10}, L_{01})I(A = 01, R = 10)] = \frac{3}{32}$$
$$\mathbb{E}[Y_3Y_4I(A = 01, R = 11)] = \mathbb{E}[m_{11,01}(X_{11}, L_{01})I(A = 01, R = 11)] = \frac{3}{32}$$

Finally, the results for $a = 10$ are identical to $a = 01$. Thus, collecting all the terms, we can get that $\mathbb{E}[Y_3Y_4] = \frac{175}{128}$.

Now we prove that the simulation setup for the marginal parametric model satisfies the ACCMV assumption. For $a \neq 11$, we have

$$P(R = 1, A = a|X, L) = \frac{\exp(0.5X)}{5 + 3\exp(0.5X)}$$

and we have

$$P(R = 1, A = 11|X, L) = \frac{1}{5 + 3\exp(0.5X)}$$

Thus for $a \neq 11$,

$$\frac{P(R = 1, A = a|X, L)}{P(R = 1, A = 11|X, L)} = \exp(0.5X)$$

which does not depend on L. Next, for $R = 0$, we have

$$P(R = 0, A = a, x, \ell) = P(R = 0, A = a|x, \ell)f_{X,L}(x, \ell) = \frac{1}{5 + 3\exp(0.5x)}f_{X,L}(x, \ell)$$

and $f_{X,L}(x, \ell)$ is the density function for X, L. Thus, we have

$$P(R = 0, A = a, \ell) = \int \frac{1}{5 + 3\exp(0.5x)}f_{X,L}(x, \ell)dx = \int \frac{1}{5 + 3\exp(0.5x)}f_{X|L}(x|\ell)dx f_L(\ell)$$
$$\Leftrightarrow P(R = 0, A = a|\ell) = \int \frac{1}{5 + 3\exp(0.5x)}f_{X|L}(x|\ell)dx$$

and this holds for all a. Similarly, we have

$$P(R = 1, A = 11|\ell) = \int \frac{1}{5 + 3\exp(0.5x)}f_{X|L}(x|\ell)dx$$

Thus, for any $a \neq 11$,

$$\frac{P(R = 0, A = a|\ell)}{P(R \geqslant 0, A = 11|\ell)} = \frac{1}{2}$$

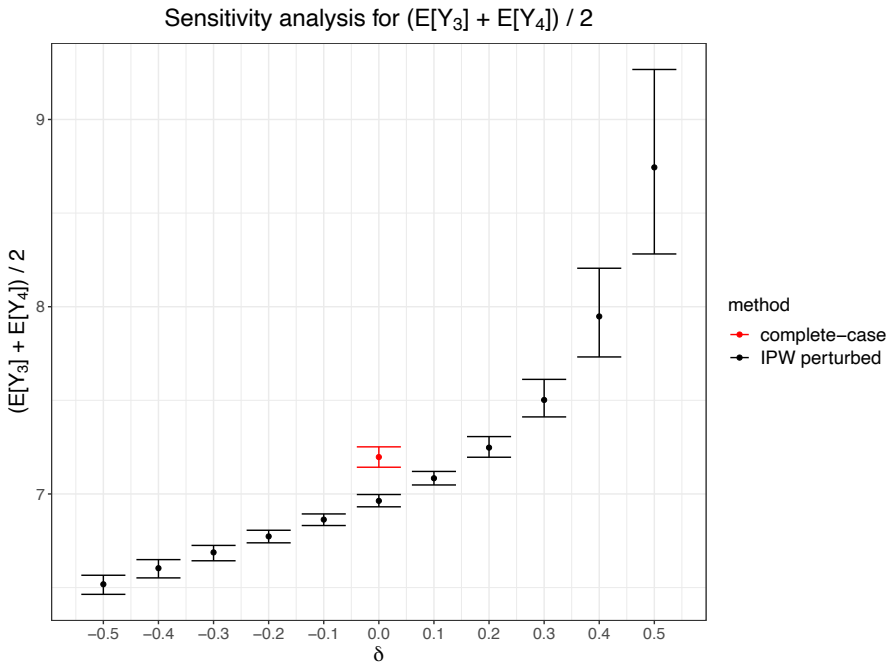

Figure B.1: Sensitivity analysis of the ACCMV assumption by exponential tilting. We examine how the estimate $\mathbb{E}[Y_3 + Y_4]/2$ changes with respect to different values of the sensitivity parameter δ.

B.6 Further sensitivity analysis

Appendix C

APPENDIX OF CHAPTER 4

C.1 Hyperparameters tuning

In the simulation study of Chapter 4, we set $k = 50$ for kNN. For random forest, we set $n_{\mathsf{tree}} = 50$ and set node size to be 350. Here, we present the results for different k, n_{tree} and node size. We set number of bootstrap samples $n_B = 200$ and number of multiple imputations $M = 5$. We generate 1,000 repetitions for each simulation setup and we use bootstrap percentile intervals for constructing 95% confidence intervals (CI).

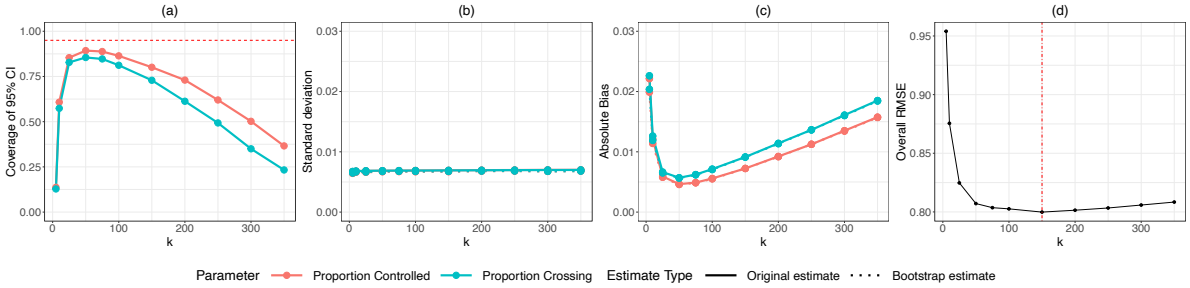

Figure C.1: We vary the number of neighbors and compare the coverages for 95% CI, standard deviation estimates and absolute bias estimates with sample size $n = 5,000$. We also use cross validation to compare the overall RMSE for different k.

From Figure C.1(a), we can see that the highest coverage is achieved with $k = 50$ and it is below the nominal level of 95%. We also compute the biases and standard deviations (SD) for our estimates. Recall for each repetition, we obtain a parameter estimate, a bootstrap parameter estimate by computing the averages of bootstrap estimates and a bootstrap standard error by computing the standard error of bootstrap estimates. For SD, the original SD

estimates are computed based on the standard errors of 1,000 parameter estimates and these reflect the true variabilities of our parameter estimates. The bootstrap SD estimates are computed based on the average of 1,000 bootstrap standard errors. We can see from Figure C.1(b) that bootstrap SD and Original SD estimates are essentially the same. We also compute the original bias and bootstrap bias over different k. The original bias is computed by the difference of the average of the 1,000 parameter estimates and the true parameter value. For bootstrap bias, we compute the difference between the averages of bootstrap parameter estimates and the true parameter value. Figure C.1(c) plots the bias in the absolute values and we can see that there are not much differences between the original bias and the bootstrap bias. Further, the absolute bias follows a U-shape as k increases and $k = 50$ minimizes the bias.

We also use cross validation to compare the overall root mean squared error (RMSE) for different k. We have in total 6 variables to impute and we compute the average RMSEs for all six variables as the overall RMSE. For cross validation, we can see from Figure C.1(d) that for the range of k we considered, the best k is 150 for minimizing the RMSE, which is different from k for optimizing coverages. This suggests that we cannot use cross validation to select the parameters in terms of coverages.

For random forest, we first set $n_{\mathsf{tree}} = 50$ and vary the node size. From Figure C.2(a), we can see that when the node size is between 250 and 400, the coverages are very close to the nominal level 0.95. Figure C.2(b) shows that the bootstrap SD is very similar to the true SD and the SD estimates do not vary much as node size changes. Figure C.2(c) shows a more interesting result. The biases again follow a U-shape and it seems that bootstrap bias is clearly smaller than the original bias when the node size is between 100 and 450. In particular, the bootstrap bias is much smaller than the original bias between 250 and 400, which explains the good coverages of the 95% confidence intervals as we use percentile intervals. The bootstrap bias now is much smaller than the bootstrap standard deviation. This also shows one scenario where bootstrap seems to "debias" the random forest. We further use out-of-bag error to compare the overall RMSE. Figure C.2(d) shows that the

best node size is 250 for minimizing RMSE, which is within the range of the node sizes that achieve the nominal coverages for the 95% confidence intervals.

Next, we set the node size to be 350 and vary the number of trees. From Figure C.3(a), we can see that when the number of trees is between 50 and 500, the coverages are all very close to the nominal level 0.95. Thus, it appears that the coverages are not sensitive to the choice of number of trees. Figure C.3(b) also shows that the standard deviations do not change much. Further, Figure C.3(c) again shows that the bootstrap bias is much smaller than the original bias. Figure C.3(d) shows that the optimal number of trees for minimizing the RMSE is 500.

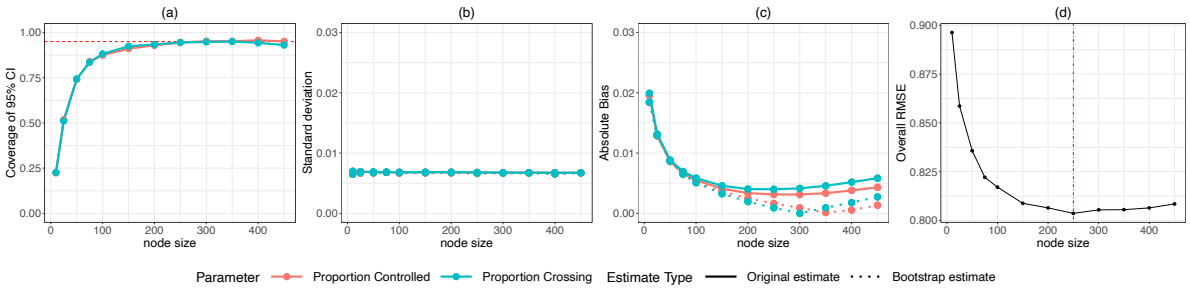

Figure C.2: We vary the node size and compare the coverages for 95% CI, SD estimates and absolute bias estimates with sample size $n = 5,000$ and $n_{\text{tree}} = 50$. We also use out-of-bag errors to compare different node size.

C.2 Additional simulation results with a nonlinear model

In Chapter 4, we consider the case when the extrapolation density follows the linear model-based BM-ACMV assumption. Now we present additional simulation results for the case when the extrapolation density follows a nonlinear model-based BM-ACMV assumption.

Data generation We modify the data generation process as follows. For $r \neq \{001, 100, 010\}$, the data is generated the same as the linear case. When $r = 001$, we have

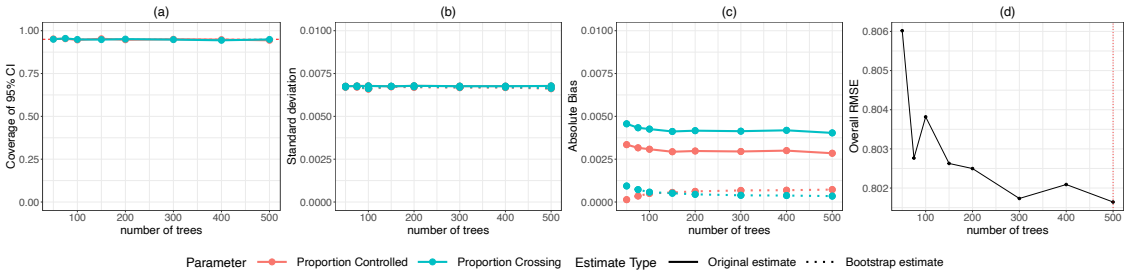

Figure C.3: We vary the number of trees and compare the coverages for 95% CI, SD estimates and absolute bias estimates with sample size $n = 5,000$ and node size being 350. We also use out-of-bag errors to compare different number of trees.

$Y_1, Y_4, Y_6, Y_{10} \sim N(1_4, \Sigma_4)$ with $\Sigma_4 = 1/2I_4 + 1/2 1_4 1_4^T$ and

$$Y_7, Y_8, Y_9 | Y_6, Y_{10}, R = 001 \sim N\left(f_1(Y_6, Y_{10}) \times 1_3, 0.5\Sigma_3\right)$$

where $f_1(y_6, y_{10}) = (y_6 - 1)^2 + (y_{10} - 1)^2$. For $r = 010$, we have $Y_1, Y_4, Y_6, Y_{10} \sim N(1_4, \Sigma_4)$ and

$$Y_5 | Y_4, Y_6, R = 010 \sim N(f_2(Y_4, Y_6), 0.7)$$

with $f_2(y_4, y_6) = (y_4 - 1)^2 + (y_6 - 1)^2$. For $r = 100$, again we have $Y_1, Y_4, Y_6, Y_{10} \sim N(1_4, \Sigma_4)$ and

$$Y_2, Y_3 | Y_1, Y_4, R = 100 \sim N(f_3(Y_1, Y_4) \times 1_2, 0.5\Sigma_2)$$

with $f_3(y_1, y_4) = (y_1 - 1)^2 + (y_4 - 1)^2$. Next, we assume that the extrapolation densities follow the model-based ACMV assumption. More specifically, we assume that for $r_1 \in \{000, 001, 010, 011\}$, we have

$$Y_2, Y_3 | Y_1, Y_4, R = r_1 \sim N(f_1^*(Y_1, Y_4) \times 1_2, \Sigma)$$

and we assume that f_1^* has the following form

$$f_1^*(y_1, y_4) = \beta_{10} + \beta_{11} y_1 + \beta_{11} y_4 + \beta_{13} y_1^2 + \beta_{14} y_4^2$$

Similarly, for $r_2 \in \{000, 001, 100, 101\}$ and $r_3 \in \{000, 100, 010, 110\}$, we have

$$Y_5|Y_4, Y_6, R = r_2 \sim N(f_2^*(Y_4, Y_6), \Sigma)$$
$$Y_7, Y_8, Y_9|Y_6, Y_{10}, R = r_3 \sim N(f_3^*(Y_6, Y_{10}) \times 1_3, \Sigma)$$

with f_2^* and f_3^* have the following form:

$$f_2^*(y_4, y_6) = \beta_{20} + \beta_{21}y_4 + \beta_{22}y_6 + \beta_{23}y_4^2 + \beta_{24}y_6^2$$
$$f_3^*(y_6, y_{10}) = \beta_{30} + \beta_{31}y_6 + \beta_{32}y_{10} + \beta_{33}y_6^2 + \beta_{34}y_{10}^2$$

We can estimate all the parameters through Monte carlo simulation. Again we are interested in estimating $\tau_1 = P(Y_1 \leqslant a, \cdots, Y_{10} \leqslant a) = P(\max Y_i \leqslant a)$, the probability of Y_i being controlled and the probability of trajectory crossing a, $\tau_2 = P(\max_i Y_i \geqslant a, \min_i Y_i \leqslant a)$ for a fixed constant $a = 1.8$. For this data generation, $\tau_1 \approx 0.2076$ and $\tau_2 \approx 0.7842$ are computed by a Monte Carlo approximation with $100,000$ observations and $1,000$ repetitions.

Simulation results We repeat the simulation for 1,000 times with sample size $n = 5,000$. For our model-based approach, we perform multiple imputations 5 times with linear model, kNN and random forest. We use $n_B = 500$ bootstrap replicates to estimate the uncertainty. For kNN, we set $k = 50$ or $k = 150$. For random forest, we use $n_{\mathsf{tree}} = 100$ and set the node size to be 25 or 100. For the MICE method, we use the Bayesian linear regression model for imputation.

Table C.1 summarizes the results. Again, complete-case analysis obtain biased estimates and zero coverages. Now linear model-based ACMV approach also obtains biased estimates and rather poor coverages. MICE obtains relatively small bias for τ_1, but the coverage for τ_2 is below 50%. Both kNN and random forest obtain much better results than MICE, complete-case analysis and linear model-based ACMV approach. However, kNN does not achieve nominal coverages with its confidence intervals. Random forest obtains nominal coverages for the 95% CIs with properly selected node size. It is worth noting that the biases obtained for random forest are rather large, compared to the standard deviations. The

Table C.1: Simulation results for nonlinear case when $n = 5000$ and $M = 5$

	Bias		TSE		Avg. Bootstrap SE		Coverage	
Method	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2
MICE	0.0084	-0.015	0.0074	0.0077	0.0076	0.0074	0.78	0.48
Complete-case	0.158	-0.160	0.019	0.019	0.019	0.019	0	0
Linear	-0.023	0.023	0.0060	0.0062	0.0062	0.0063	0.045	0.052
kNN (50)	-0.0081	0.0040	0.0059	0.0060	0.0059	0.0060	0.76	0.92
kNN (150)	-0.0065	0.0050	0.0058	0.0058	0.0059	0.0059	0.82	0.89
RF (25)	-0.015	0.010	0.0057	0.0059	0.0061	0.0062	0.95	0.85
RF (100)	-0.012	0.0069	0.0059	0.0060	0.0061	0.0062	0.86	0.95

coverages of 95% CIs are are reaching the nominal levels as we are using bootstrap percentile confidence intervals. From Figure C.5, C.6 and C.7, we can see that the bootstrap biases are very close to 0 and much smaller than the standard deviations. Thus, the bootstrap percentile intervals obtain correct coverages.

Hyperparameters tuning We also present the simulation results tuning the number of neighbors for kNN. From Figure C.4(a), we can see that the optimal k for coverages in fact differs for estimating τ_1 and τ_2. For estimating τ_1, the probability of being controlled, the optimal k is 150 and the optimal coverage is 81.7%, which is below the nominal coverages. For estimating τ_2, the probability of crossing the trajectory, the optimal k is 50 and the optimal coverage is 91.5%, which is closer to the nominal coverages. Figure C.4(b) shows that the standard deviations do not change much as k increases. Further, Figure C.4(c) shows that the biases roughly follow a U-shape and the bootstrap bias is slightly smaller than the original bias. We can further see that the optimal number of neighbors for bias are the same as the optimal number of neighbors for coverages. For cross validation, from figure C.4(d), we can see that the optimal k for RMSE is 25, which is different from the optimal k

for coverages.

For random forest, we first set $n_{\mathsf{tree}} = 100$ and then tune the node size. From figure C.5(a), we can see that the optimal node sizes for coverage are in fact different for estimating τ_1 and τ_2. For τ_1, the optimal node size is 25; for τ_2, the optimal node size is 100. The standard deviations do not vary much. From figure C.5(c), again we can observe that the bootstrap biases are much smaller than the original biases for random forest and the optimal node sizes for biases are the same as the optimal node sizes for coverages. From figure C.5(d), we can see that the best node sizes for optimizing the RMSE is 100, which is larger than the optimal node size for coverages of τ_1 and the same as the optimal node size for coverages of τ_2.

Next, we set the node size to be 25 or 100 and vary the number of trees. Again, the coverages are not very sensitive to the number of trees based on Figure C.6 and C.7. When node size is 100, the coverages for τ_2 are all close to the nominal level for the whole range of number of trees. However, the coverages for τ_1 always fall below the nominal level and improve a bit with a higher number of trees. Figure C.6(b) and (c) show that the SDs and biases do not vary much as number of trees increase. The bootstrap bias is again much smaller than the original bias. The out-of-bag errors also suggest that the optimal number of tree is 400 with node size 100.

When node size is 25, the coverages for τ_1 are now all very close to the nominal level for the whole range of number of trees. Instead, the coverages for τ_2 now always fall below the nominal level and improve a bit with a smaller number of trees. The SDs and biases again do not vary much. The bootstrap biases for τ_2 now increases as number of trees increase. The bootstrap biases for τ_1 on the other hand decreases and are all very close to 0. Further, the out-of-bag errors suggest that the optimal number of tree is 500 with node size 25.

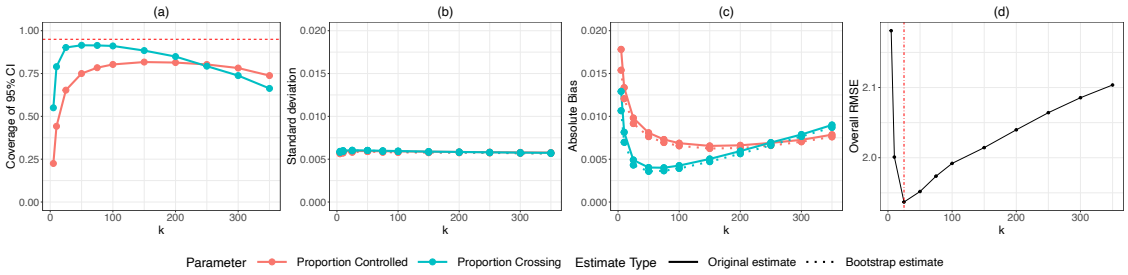

Figure C.4: We vary number of neighbors and compare the coverages for 95% CI, SD and biases with sample size $n = 5,000$. We also use cross validation to compare the overall RMSE.

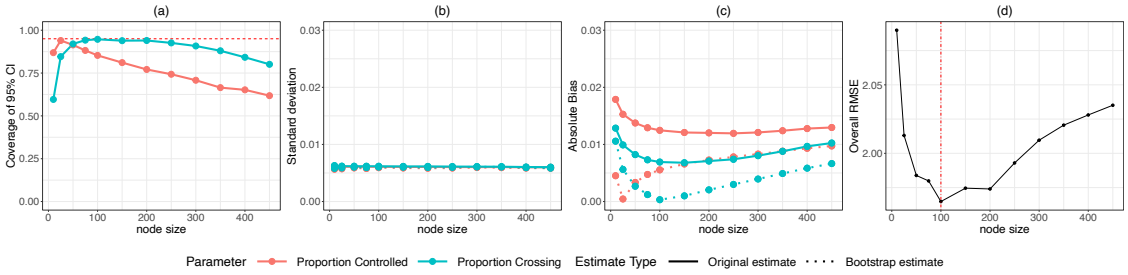

Figure C.5: We vary the node size and compare the coverages for 95% CI, SD and biases with sample size $n = 5,000$ and $n_{\mathsf{tree}} = 100$. We also use out-of-bag errors to compare the node size.

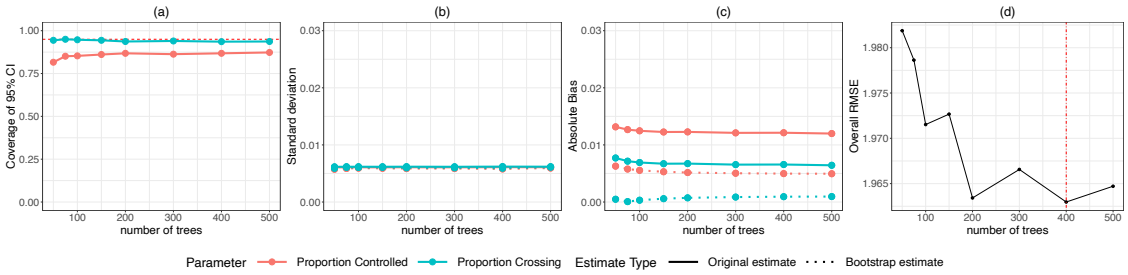

Figure C.6: We vary the number of trees and compare the coverages for 95% CI, SD and biases with sample size $n = 5,000$ and node size being 100. We also use out-of-bag errors to compare the number of trees.

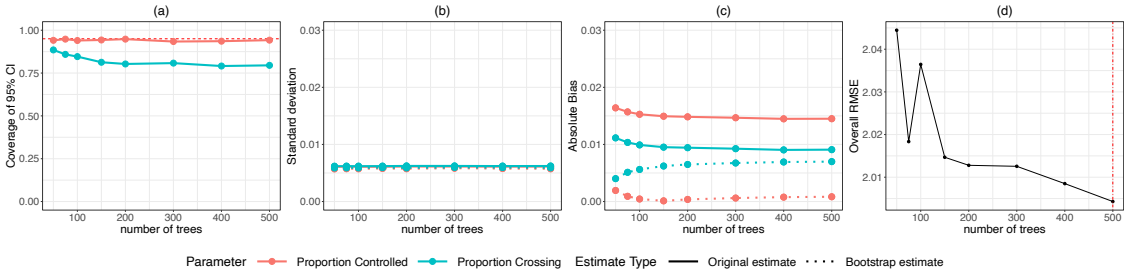

Figure C.7: We vary the number of trees and compare the coverages for 95% CI, SD and biases with sample size $n = 5,000$ and node size being 25. We also use out-of-bag errors to compare the number of trees.

C.3 Hyperparameter tuning for real data results

In the real data experiments of chapter 4, we set the number of neighbors to be 5 for kNN. For random forest, we set the number of trees to be 50 and the node size to be 93. Here, we present the results for different k, n_{tree} and node size. We set the number of bootstrap samples $n_B = 200$ and number of multiple imputations $M = 20$. We compute the overall root mean squared error (RMSE) by computing the sum of RMSE for all possible missing patterns.

From figure C.9, we can see that the optimal k for minimizing the overall RMSE is 15. From figure C.8, we can see that the 95% confidence intervals does not vary too much across different k. More specifically, when $k = 5$, the confidence interval is (0.173 - 0.199) for τ_1 and (0.685 - 0.714) for τ_2. When $k = 15$, the confidence interval for τ_1 is (0.163 - 0.190) and (0.693 - 0.722) for τ_2.

For random forest, we first set $n_{\mathsf{tree}} = 50$ and compare different node sizes. From figure C.11, we can see that the optimal node size is about 50, which is much smaller than the node size we use in the 4.5. From figure C.10, again the confidence intervals do not vary too much over different node sizes. When the node size is 50, the confidence interval is (0.169 - 0.198) for τ_1 and (0.680 - 0.713) for τ_2. Next, we set the node size to be 50 and compare different number of trees. From figure C.13, we can see that the optimal number of tree for minimizing the RMSE is 200. From figure C.12, the confidence intervals do not vary too much over different number of trees. When the number of tree is 50, the confidence interval is (0.170 - 0.197) for τ_1 and (0.682 - 0.712) for τ_2. When the number of tree is 200, the confidence interval is (0.169 - 0.197) for τ_1 and (0.684 - 0.709) for τ_2.

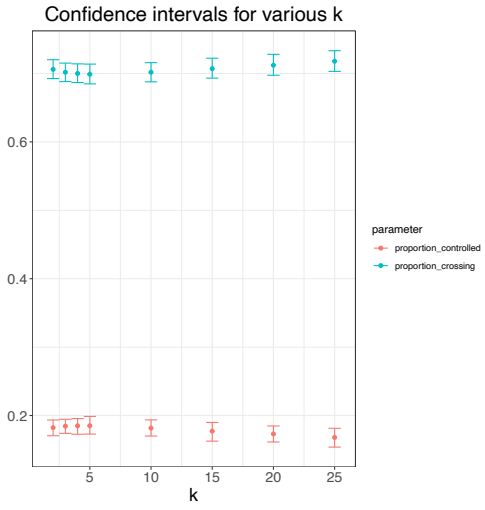

Figure C.8: We vary the number of neighbors and compare the 95% confidence intervals.

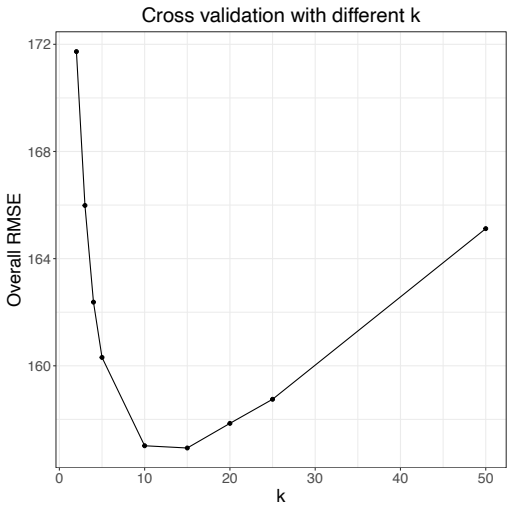

Figure C.9: We use cross validation to compare different k for the overall RMSE.

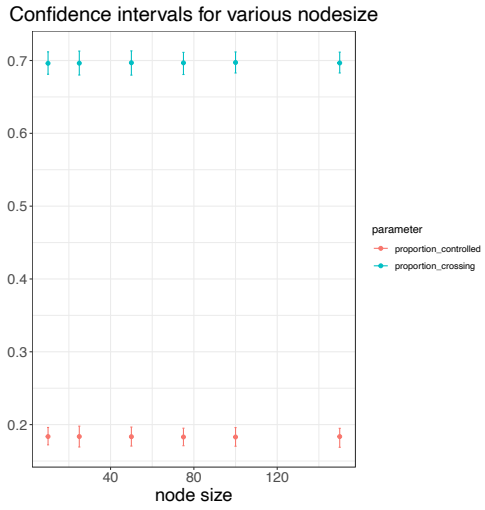

Figure C.10: We vary the node size and compare the 95% confidence intervals with $n_{\text{tree}} = 50$.

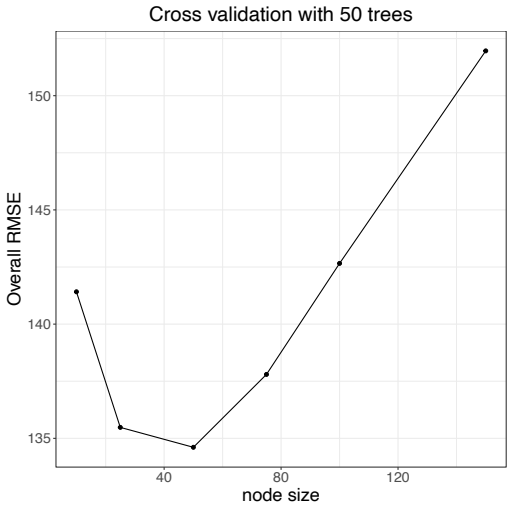

Figure C.11: We use cross validation to compare different node size for the overall RMSE with $n_{\text{tree}} = 50$.

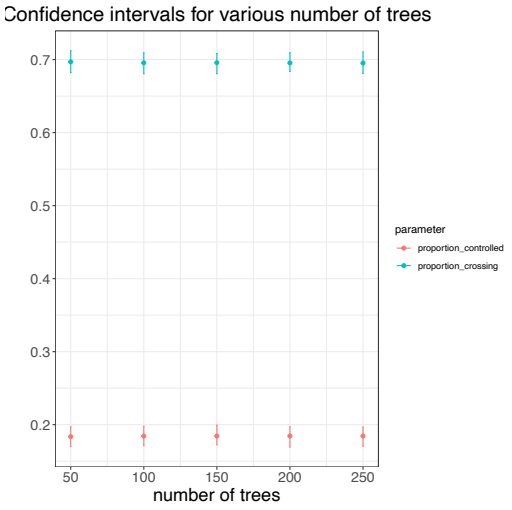

Figure C.12: We vary the number of trees and compare the 95% confidence intervals with node size being 50.

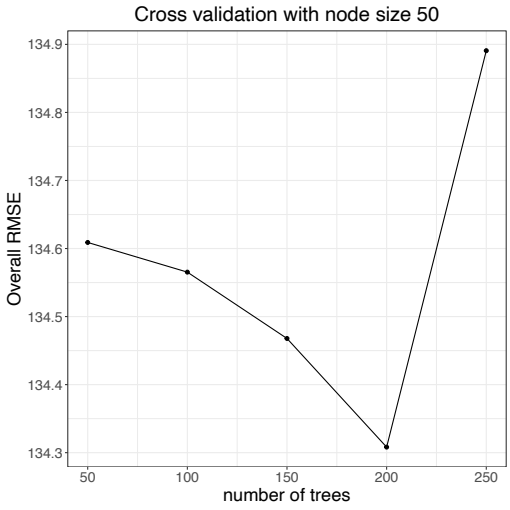

Figure C.13: We use cross validation to compare different number of trees for the overall RMSE with node size being 50.

C.4 Proof

In this section, we present the proof for Proposition 1 and 2. The proof of proposition 2 is very similar and thus omitted.

ProofProof for Proposition 1. Under our BM-ACMV assumption (5) and (6), we have

$$\begin{aligned} p(\ell, A=a) &= p(A=a)p(\ell|A=a) = p(A=a)p(\ell_a|A=a)p(\ell_{\bar{a}}|\ell_a, A=a) \\ &= p(A=a)p(\ell_a|A=a)\prod_{j=1}^{J(\bar{a})} p(\ell_{b_j}|\ell_{b_j^\dagger - b_j}, A=a) \\ &= p(A=a)p(\ell_a|A=a)\prod_{j=1}^{J(\bar{a})} p(\ell_{b_j}|\ell_{b_j^\dagger - b_j}, A \geqslant b_j^\dagger) \end{aligned}$$

Thus, $p(\ell, A=a)$ is nonparametrically identified. Note that this further implies that $p(\ell)$ is

identified as

$$p(\ell) = \sum_{a \in \{0,1\}^d} p(\ell, A = a).$$

□