Recovering Natural History: Modeling Cardiovascular Biomarkers in the Presence of Endogenous Medication Use

Andrew J. Spieker

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Robyn L. McClelland, Chair

Joseph A.C. Delaney

Patrick Heagerty

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Recovering Natural History: Modeling Cardiovascular Biomarkers in the Presence of Endogenous Medication Use

Andrew J. Spieker

Chair of the Supervisory Committee:
Professor Robyn L. McClelland
Biostatistics

In the modern era, cardiovascular biomarkers are often measured in the presence of medication use, whereby the observed value is different than the underlying untreated value for participants on medication. However, for certain problems, the natural history of the biomarker that would have occurred in the absence of medication use is of greater interest than the observed value. In observational data, medication use is nonrandom in that participants on medication tend to have higher underlying biomarker values than participants off medication. That is to say that medication use is endogenous. When faced with endogenous medication use, traditional methods such as adjustment for medication use in linear regression models are inappropriate. The goal of this dissertation is to develop methods to estimate associations between predictors of interest and biomarker outcomes in the presence of endogenous medication use.

First, we focus on methods for use in a cross-sectional setting. Heckman's treatment effects model, as suggested by its name, has historically been used to estimate the effect of medication use on a continuous outcome. In this research, we take a definitive departure from the historical use of the model, in that we utilize the Heckman framework in order to estimate associations between exposures and underlying (off-medication) outcomes, regarding the effect of medication on the biomarker as a nuisance rather than a parameter of interest. We show that the treatment effects model is fairly robust to departures from several of its main assumptions. One assumption to which the

treatment effects model is particularly sensitive, however, is the assumption of uniform treatment effects. In particular, the expected effect of medication use on the biomarker is presumed to be constant across participants (an assumption that is often thought to be unrealistic in practice). We extend the treatment effects model to allow effect modification, or "subgroup-specific" treatment effects.

The second major aim of this dissertation pertains to developing methodology to address endogenous medication use when repeated measures are available on subjects over time. Very little work has been done to address the challenges of endogenous medication use in longitudinal data. For certain types of probit analyses, existing methods invoke standard results on M-estimation theory to construct asymptotically valid estimates of marginal parameters. As cardiovascular biomarkers of interest show strong within-subject correlation over time, there is much efficiency to be gained by modeling that correlation. We seek to understand situations in which accounting for correlation can be advantageous (e.g., in the setting of deterministic covariates), and elucidate efficiency gains with specification of a working covariance.

These two objectives primarily target bias reduction for the challenge of addressing endogenous medication use in estimating biomarker associations. Improving estimation of these associations can help us better understand underlying biological mechanisms of disease and better motivate future clinical research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

BIC: Bayesian Information Criterion

CN: Censored Normal

CLOG-LOG: Complementary Log-Log

DAG: Directed Acyclic Graph

DBP: Diastolic Blood Pressure

FGP: Fasting Plasma Glucose

FRS: Framingham Risk Score

GEE: Generalized Estimating Equations

IPCW: Inverse Probability of Censoring Weights

IPTW: Inverse Probability of Treatment Weights

IPW: Inverse Probability Weighting

IV: Instrumental Variables

LCMM: Latent Class Mixture Modeling

LDL: Low-Density Lipoprotein

LEM: Longitudinal Endogeneity Model

MESA: Multi-Ethnic Study of Atherosclerosis

MSE: Mean Squared Error

OLS: Ordinary Least Squares

SBP: Systolic Blood Pressure

SSEM: Subgroup-Specific Effects Model

TEM: Treatment Effects Model

WLS: Weighted Least Squares

ACKNOWLEDGMENTS

I wish to express my sincerest appreciation to my dissertation committee chair, Robyn McClelland, for her outstanding mentorship and guidance. I would also like to thank my supervisory committee members Chris Delaney, Patrick Heagerty, and Andrew Zhou for their invaluable insights and encouragement throughout this research. I am deeply grateful for the guidance of Thomas Fleming, Jim Hughes, Lurdes Inoue, Barbara McKnight, Adam Szpiro, and Mary Lou Thompson who have all been instrumental in helping me develop as a teacher. I wish to also thank Scott Emerson, Gitana Garofalo, Ken Rice, and Pat Wahl who have all been highly influential mentors to me throughout my time here. Last, but not least, I wish to acknowledge my family and friends for their support.

DEDICATION

In loving memory of Anthony and Nancy Peluso.

Chapter 1
INTRODUCTION

Biomarkers can be useful tools for understanding subclinical cardiovascular disease. They can be used to inform physicians how to optimally treat their patients in order to prevent clinical cardiovascular disease, and/or to monitor disease progression over time. They can also be used to help quantify patients' risk of future adverse cardiovascular events such as stroke or myocardial infarction. High values of systolic or diastolic blood pressure (SBP and DBP, respectively), fasting plasma glucose (FPG), and low-density lipoprotein (LDL), are all indicative of underlying cardiovascular disease. Disease states are in fact defined by clinicians in terms of these biomarkers using specific cut-off points. For example, hypertension is typically defined as a SBP of $\geq$ 140 mmHg or a DBP of $\geq$ 90 mmHg (Chobanian et al., 2003).

Hypertension and hyperlipidemia are highly prevalent in the United States, at 31.7% and 29.1%, respectively (Nwankwo et al., 2013; Mozaffarian et al., 2015). Additionally, diabetes is also common in the United States, at a prevalence of 9.3% in 2012 (National Diabetes Statistics Report, 2014). A high observed value of any one of these biomarkers will often prompt some sort of intervention or management strategy in order to reduce the biomarker value, and in turn, risk of adverse cardiovascular events. Medication is primarily the strategy used, specifically targeting the biomarker. Thus, in modern observational studies, many participants are on medication to to reduce their biomarker value, such that the observed value is different than the underlying value that would have occurred in the absence of treatment. Medication is taken in a non-random fashion, in that participants with higher underlying biomarker values tend to be more likely to

be on medication than participants with lower underlying biomarker values. That is to say that medication use is endogenous, and therefore non-ignorable in problems that seek to use the untreated biomarker values.

For association studies involving certain types of exposures (e.g., single nucleotide polymorphisms, age, gender, and race category), the natural history of the biomarker that would have occurred in the absence of medication use is often of greater interest than the observed value. For example, we might be interested in estimating the difference in mean SBP values across race categories (Kramer et al., 2004), or the difference in mean LDL cholesterol values between those with a particular gene mutation and those without it (Chen et al., 2009). In studies of populations that are heavily treated (e.g., most modern cohort studies), medication use acts as a contaminant for estimating the association, such that simple approaches (e.g., ordinary least squares linear regression) do not characterize the association of interest. Despite this, these approaches are widely used in the literature. Estimators derived from these sorts of simple approaches estimate parameters that are heavily dependent on both medication use prevalence in the sample and the magnitude of the medication's effect on the biomarker. If our goal is to better understand underlying biological processes, it is of greater interest to study the association between these predictors and the biomarker outcomes that would be observed in the (typically counterfactual) setting in which no participants are on medication. That is to say that we are most interested in what we refer to as the "natural history association" between the predictor and the biomarker. This parameter is most useful to us in identifying underlying differences in biomarker values between groups, or across values of a predictor.

Conclusions regarding whether certain subgroups of patients tend to have riskier biomarker values can motivate subsequent clinical research, for example, to evaluate whether prophylactic medication use in certain sub-populations reduces the risk of ad-

verse cardiovasvular events. Patients with diabetes, for example, are often placed on hypolipidemic drugs precisely for this reason (Choesterol Treatment Trialists' Collaborators, 2008). Insights surrounding the effectiveness of prophylactic drug use can be elucidated exclusively with well conducted randomized controlled trials (RCTs). However, as RCTs are often motivated by results from observational studies, it is of interest to develop models that provide us with more accurate insights about underlying associations in order to reduce the number of misleading results from observational studies and, in turn, reduce futility rates from clinical research.

For the purposes of illustrating the importance of the natural history association, consider our motivating example in which we wish to quantify the difference in mean LDL between patients with and without a particular gene mutation, using cross-sectional observational data. Suppose we regress participants' observed LDL on their observed allele status (wild-type vs. mutant). Particularly in studies of older participants, many are likely to be on lipid-lowering drugs. If the allele type, a primordial exposure, is associated with the underlying LDL value, this simple regression approach will mischaracterize the influence of the allele on LDL. Hence, the estimand for this naïve analysis does not provide insight into whether the allele of interest is, in the most fundamental sense, associated with higher values of LDL. Instead, it instead estimates the expected difference in observed LDL between two randomly selected people from the population differing in their allele status, ignoring the fact that the on-medication participants would have tended to have a higher LDL value off medication. In order to understand how a predictor of interest is naturally associated with a biomarker, we need to somehow obtain information on the underlying, off-medication biomarker value. Of course, such an outcome measure is not observable for participants on medication, as it is masked by the effects of medication use. Figure 1.1 illustrates how endogenous medication use could impact estimation of the difference in means in this example.

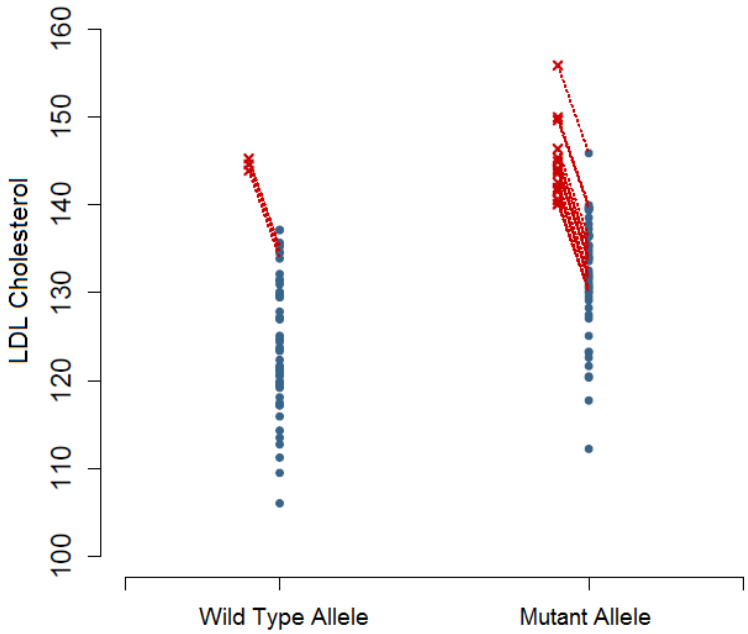

Figure 1.1: A simple illustration of endogenous medication use. In this scenario, any participant who would have had an underlying biomarker value exceeding 140 was on medication at the time of observation. The observed data points are shown in blue, and the (counterfactual) underlying biomarker is in red. The difference in mean observed biomarkers between groups is approximately 10 units, whereas the difference in the mean underlying biomarkers between groups is 20 units.

The purpose of this dissertation is to develop methodology in order to account for endogenous medication use when seeking to estimate the natural history association using observational data. First, we focus on the commonly encountered setting in which only cross-sectional data are available. Measurements on certain variables may be taken exclusively at baseline, for example, if data on those variables are time-consuming, costly, or inconvenient to obtain. Safety concerns may also preclude the possibility of ethically obtaining repeated measures. In addition, it may be difficult to obtain follow-up data on certain populations (e.g., the homeless), such that cross-sectional data are all that are available. Based on our methodology development in cross-sectional data, we then wish to extend our models to accommodate settings in which repeated measures are available on subjects over time.

Given limitations of the data we collect, we are faced with the challenge of making potentially untestable assumptions about the data generating mechanism. This is particularly the case in a cross-sectional setting. Features such as (i) the mechanism by which on-medication participants differ from off-medication participants, and (ii) the magnitude/distribution of the effect of the medication are essentially ignored by traditional approaches. Ordinary least squares (OLS) based approaches such as performing linear regression only on off-medication participants (excluding those on medication), or adjusting for medication use as though a confounder, rely on highly unrealistic assumptions about these aspects of medication use. Although the limitations of these simple approaches have been documented (Tobin et al., 2005), these methods are still widely used in biomarker association studies.

Modifications to these standard approaches have been proposed in order to handle medication use in observational data. Tobin et al. (2005) proposed a censored normal regression approach for cross-sectional data that relies on non-informative censoring, an assumption unlikely to be met in the scenario we have described. Wang and Fang (2011) propose an inverse probability-of-treatment weighting (IPTW) approach for use when two measurements are available on each participant and medication use occurs after baseline in at least a subset of the participants. This approach could naïively be modified for use in a cross-sectional approach. While the proposed IPTW approach does not presume having observed the exact underlying off-medication biomarker value, it relies on knowledge of a dichotomized version of the underlying biomarker ("high" or "low") which is ostensibly unobservable in cross-sectional data, unless, for example, external information could be obtained (e.g., from self reported status or prior medical records). Use of information from new medication users to improve propensity score models has been proposed to handle confounding by indication when estimating treatment effects (Jorgensen et al., 2013). While simple OLS approaches are known to result

in bias, these slightly more sophisticated approaches also rely on stringent assumptions about medication use that are likely violated.

Instrumental variable based approaches are also known to be effective in handling endogenous medication use in certain settings (Bowden and Turkington, 1984; Wooldridge, 2013), and have been applied/suggested for use in pharmacoepidemiologic studies (Brookhart et al., 2006). Often, such approaches are applied to estimate the effect of medication use on a biomarker, but these approaches also provide consistent estimates of the natural history association in the presence of a true instrument. The standard one-stage and two-stage instrumental variable approaches are appealing due to straightforward implementation, but they rely on the existence of an instrument (i.e., a single variable known to be associated with medication use, but not associated with the biomarker). In randomized trials, one may be able to find such a variable. For example, if one wants to estimate the effect of an intervention on a continuous outcome of interest in a setting where some study participants are non-adherent to the randomized group, one could analytically justify using randomized group as the instrument and believe that such a variable is not associated with the underlying biomarker. In the setting of observational data, finding a variable with such properties has been exceedingly difficult. If such a variable exists, consistent estimates of the natural history association can be obtained, although the average effect of medication use on the outcome has historically been the parameter of interest.

Heckman (1978), proposed what he referred to as the "hybrid model with structural shift," better known as the Treatment Effects Model (TEM). In short, the TEM jointly models the expected value of the biomarker as a linear function of predictors, together with an underlying probit model that describes the probability of medication use (conditional on observable covariates that may be distinct from those in the biomarker model). As suggested by the name, this model was proposed to estimate the

population-average effect of an intervention (in this case, medication use) on an outcome (in this case, a biomarker) when the intervention is endogenous. Just as with the instrumental variables approach, consistent estimates of the natural history association can be obtained from this method, despite the fact that this is not the historical use. In this sense, we regard the treatment effect as a nuisance rather than a parameter of interest. Hence, it is of interest to evaluate the extent to which Heckman's TEM serves as a useful alternative to existing approaches in order to achieve bias reduction for estimating associations between biomarkers and exposures in a cross-sectional, observational setting when endogenous medication use is present. As we are taking such a clear departure from what has been done in the literature, there is an unmet need to evaluate the sensitivity of the TEM to departures from its assumptions when estimating the natural history association. Moreover, the subsequent model extensions we develop will seek to address challenges that are motivated from this alternative use of the model.

Since the TEM was originally designed to estimate a population-average treatment effect, it relies heavily on the assumption that the effects of medication do not vary in expectation across covariates. In practice, the expected magnitude of a treatment effect typically varies with observable covariates such as genetic factors, demographic variables, medication class, or dose. If certain covariates are known to be associated with the effects of medication use, Heckman's TEM systematically over- or under-corrects the observed biomarker for certain subgroups of on-medication participants. We expect, of course, that this is most problematic when an effect modifier is an exposure of interest in the biomarker model. It is, however, unclear how this sort of effect measure modification impacts estimation of the natural history association when the source of heterogeneity is related only to the probability of medication use, or when the effect modifiers are only associated with the effects of medication use. There is hence an unmet need to

extend Heckman's TEM to accommodate subgroup-specific effects in order to provide better estimates of the natural history association using cross-sectional data.

Endogenous medication use also poses challenges when repeated measures are available on participants over time. In recent years, age-trend modeling of cardiovascular biomarkers has also received a great deal of attention in the literature in longitudinal data (Singh et al., 2012; Gurven et al., 2012; Carroll et al., 2005; Allen et al., 2014). However, results from these studies are based on simple approaches similar to those naïvely used in cross-sectional data. These standard approaches are not sufficient for estimating the natural history association, as they do not adequately address the challenges that arise from endogenous medication use. When repeated measures are available on study subjects, as in the case of studies seeking to evaluate biomarker age trends, there is no well-documented analogue of the TEM in order to account for endogeneity. The approach suggested in the literature for similar probit-response type models in order to accommodate correlation is to first fit the cross-sectional model to the entire data set (ignoring dependence altogether) to obtain a parameter estimate, and to then use a cluster-based robust variance-covariance estimator (Wooldridge, 2011). This is the approach implemented in modern software such as Stata (StataCorp, College Station, TX), and it is very similar in spirit to the approach of generalized estimating equations (GEE) with working independence (Liang and Zeger, 1986).

Although this approach provides a valid estimation procedure, there is efficiency to be gained from exploiting within-subject correlation in the biomarker; for cardiovascular biomarkers in particular, the intra-subject correlation over time is often very high. As will be made apparent, extending the TEM by fully specifying correlation in the underlying biomarker and in medication use over time is analytically intractable as it demands a computational approximation to high order integrals. We propose partial specification of the correlation structure for estimation of associations in longi-

tudinal data, bypassing the computational challenges associated with full specification of a covariance model. This approach provides a computationally tractable alternative to fully parametric likelihood approaches, while still accounting for correlation in the biomarker. Given the known limitations associated with using working correlation structures other than independence in marginal models (e.g., generalized estimating equations) when covariates are time-varying and non-deterministic, specifically focusing attention to settings in which it is appropriate to model correlation in the biomarker should be exploited is warranted.

The remainder of this dissertation is organized as follows. Chapters 2 through 5 focus specifically on understanding and developing methods for use in cross-sectional data. Chapters 2 through 4 specifically expand upon details surrounding use of the TEM for the purposes of estimating the natural history association (elaborating on details presented by Spieker et al. (2015)). Chapter 5 focuses on accounting for subgroup-specific effects; a substantive portion of the content is under review (Spieker et al., 2016). In Chapter 2, we provide background for models which have been previously used in the setting of nonrandom medication use, and examine analytically why they are inadequate in our setting. The failure of simple approaches in this setting has been documented to an extent, but we will explore bias in greater depth (Tobin et al., 2005). Instrumental variable approaches have been examined thoroughly and have well understood challenges, particularly in observational data. What has not been documented as well is the failure of slightly more sophisticated methods that have been previously presented as addressing nonrandom medication use such as inverse probability weighting and censored normal regression.

In Chapter 3, we specifically focus on providing background information on Heckman's TEM. Devoting an entire chapter to this one particular model is justified for two reasons: (i) doing so will set the stage for virtually all of the subsequent methods devel-

opment in this dissertation, and (ii) the seminal paper Heckman, 1978 focuses almost entirely on estimation of the population-average treatment effect, in contrast to out goal of estimating the natural history association. We present the modeling framework for this alternative purpose, and show how Heckman's TEM can accommodate dependence of medication use on the underlying biomarker. We also propose a sandwich based variance-covariance estimator. In Chapter 3, we also empirically compare Heckman's TEM to alternative approaches by simulation in order to elucidate the trade-off between bias reduction and efficiency cost in the setting of endogenous medication use. We will further compare Heckman's TEM to the classic instrumental variables estimator to demonstrate that Heckman's TEM does not demand that an instrumental variable exists in order to estimate the natural history association. Bias reduction will be of primary interest.

In Chapter 4, we aim to better understand the settings in which Heckman's TEM performs well, which includes a thorough evaluation of its robustness to departures from assumptions. We evaluate through several simulation studies how sensitive Heckman's TEM is to departures from its main assumptions, including non-normally distributed errors (distributions with skewness and heavier tails than the normal distributin), non-differential misclassification of the predictor of interest, probit-model misspecification, and non-uniform treatment effects. Our finding on sensitivity to departures from the assumption of uniform treatment effects motivates an extension, which will be the subject of Chapter 5.

In Chapter 5, we extend Heckman's TEM to allow for subgroup-specific treatment effects. Similar to Chapter 3, we are interested in comparing this model extension to alternatives. Parameters that describe effect modification are not the primary focus of this research, although we do present some results verifying that such parameters are estimable. Instead, our main goal is to understand which types of effect modifiers

(e.g., predictors associated with the underlying biomarker, with medication use, or with neither) are most helpful to accommodate to achieve bias reduction. We view this modification as a way to correct resulting bias when the assumption of uniform treatment effects is violated. We will also propose a Wald-based statistic to test for the presence of effect modification.

At this point, we transition out of the cross-sectional setting and into the setting where repeated measures are available over time. In Chapter 6, we provide a brief background of existing models for longitudinal data, including the "working independence" approach of Wooldridge (2011) for related probit models, in which an approach similar to GEE is used. We then demonstrate why full specification of a likelihood model results in a computationally intractable estimation problem, and proceed to devise a compromise between the working independence approach and fully parametric specification. We will proceed to prove a robustness property that allows consistent estimation of the natural history association even when the correlation structure is not correctly specified. A number of simulation studies will be conducted to further elucidate the advantages of modeling correlation where it exists.

In Chapter 7, we illustrate how our developed methods can be used in real data with a series of examples from the Multi-Ethnic Study of Atherosclerosis (MESA), specifically focusing on LDL cholesterol (Bild et al., 2002). In keeping with the original objectives of MESA, we will consider racial differences in subclinical cardiovascular disease as well as age and gender associations. First, we will utilize the Exam 1 (baseline) data to compare cross-sectional models as developed in Chapters 1-5. Then, we will present results from an analysis for estimating age trends in longitudinal data.

Finally, we provide a summary of our findings with in Chapter 8, and include a discussion of important conclusions and implications of our main results in both the cross-sectional and longitudinal models. We describe several potential ways these meth-

ods could be expanded and extended, and discuss directions for future research. We additionally attach the salient programming code for the main models derived, so that can be used by those interested in this type of research and/or those who wish to adopt these methods for application to real data.

Chapter 2

BACKGROUND: SIMPLE REGRESSION MODELS FOR CROSS-SECTIONAL DATA

The task of estimating the association between an exposure and a biomarker outcome can be rather simple in settings where no participants are on medication. Generally, in this setting, approaches such as OLS linear regression can be sufficient to estimate linear trends. When medication use is random (for instance, in the setting of a randomized controlled trial), simple modifications such as adjustment can be sufficient to address medication use. Among the simplest modifications to OLS that have been applied in observational data include: (i) excluding on-medication participants from analysis, and (ii) adjusting for medication use as though a confounder. These OLS approaches together make assumptions that are generally highly unreasonable if medication use is not random, as it would be in a placebo-controlled experiment. The adjustment method appears to be most widely used in practice for estimating associations between predictors of interest and biomarker outcomes (Brand et al., 2003; Matsubara et al., 2001; O'Donnell et al., 1998; Schunkert et al., 1998), although ignoring (Iwai et al., 2001; Sethi et al., 2003) and excluding on-medication participants (Rice et al., 2000) are also occasionally implemented in this setting.

Inverse probability weighting (IPW) can be implemented to address certain types of missing data and can achieve bias reduction when longitudinal data are available (Hernán et al., 2004; Robins et al., 2000). One might in fact think to apply an IPW-based technique to modify approaches like excluding on-medication participants or adjusting for medication use. We refer to an IPW approach for the former technique as

"inverse probability of censoring weights" (IPCW) and one for the latter approach as "inverse probability of treatment weights" (IPTW). The goal with these approaches is to attempt to "up-weight" participants who are less likely to be observed in the sample. Although IPW approaches were generally not designed for use in cross-sectional data, they do on occasion appear in the literature (Huynh et al., 2014). Indeed, advantages from IPW can also be seen in cross-sectional data when there is strong effect measure modification (Delaney et al., 2009).

Censored normal (CN) regression is another simple likelihood-based approach that has also been proposed to address medication use (Tobin et al., 2005). In this model, the outcome for participants on medication are presumed to be right-censored such that their true underlying biomarker values are modeled as lying somewhere above the observed value. The censored normal model is related to the Tobit model (Tobin, 1958; Amemiya, 1984).

Another method that has been devised to address endogenous medication use is the instrumental variables approach (Wooldridge, 2013; Bowden and Turkington, 1984; Brookhart et al., 2006), which relies on the existence of some variable that is associated with medication use, but not with the outcome of interest. These methods have been successful when one has access to an instrumental variable, but can be sensitive to departures from its unverifiable assumptions.

In this chapter, we outline the approaches described above, all of which have been previously applied in order to estimate associations between risk factors and biomarkers. We attempt to gain an intuition for the reasons they are inadequate. Specifically, we will illustrate analytically and heuristically their inappropriateness in the setting of endogenous medication use. In our discussion, we emphasize the point that bias from these approaches depends very heavily on characteristics of medication use that will ultimately need greater consideration in subsequent methods development: the

magnitude/distribution of the effect of the medication, and the prevalence of medication use. Other metrics of accuracy such as mean squared error (MSE) will be helpful and will be addressed empirically and in greater detail in subsequent chapters, particularly when conducting simulation studies. In turn, this chapter will set the stage for the rest of the dissertation, in which we will focus on developing methodology under one specific framework in order to address the unmet needs for estimating associations of interest.

2.1 Notation

Recall that our parameter of interest is the natural history association between some biomarker outcome and some predictor of interest–that is, the association between the two variables in the absence of medication use. Since many study participants may be on medication, we can think of this as a missing data problem, where the underlying off-medication biomarker value happens not to be observable in on-medication participants.

Throughout this dissertation, let $i = 1, \ldots, N$ index study participants, each with observed medication use status z_i, (1 if on medication, and 0 if off medication). Then, let $y_i(z_i)$ denote the biomarker of interest under medication use status z_i (a potential outcome). In this case, $y_i(0)$ denotes the underlying, off-medication biomarker value for subject i, and is the outcome of interest to us. Let $\boldsymbol{x}_i$ denote a $(p+1)$-vector of predictors of interest (including a 1 to allow for an intercept), and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ the corresponding vector of unknown regression coefficients, the parameter of interest in the following linear regression model: $y_i(0) = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$. One could fit this with ordinary least squares if $y(0)$ were observed for all $i = 1, \ldots, N$; however, $y_i(0)$ is not observable if $z_i = 1$. Let $\mathbf{X}$ denote a design matrix of predictors for all participants; let $\mathbf{Z}$ and $\mathbf{Y}$ denote the vectors for medication use indicators and observed biomarker values, respectively.

We do not at this point make distributional assumptions about the error terms: only that they are independent, of mean zero and finite variance. In this setting,

$y(0)$ is only observable if $z_i = 0$, so $\boldsymbol{\beta}$ may not be directly estimated. Instead, $y_i = y_i(0)(1-z_i)+y_i(1)z_i$ is observed. Let δ be an unknown real-valued parameter describing the effect of the medication on $y(0)$. We will make clear the assumptions made about medication use by each each method. At times in this chapter, we will make the assumption that $y_i = y_i(0) - \delta z_i$ for the purposes of demonstration, an assumption that will actually become necessary in Chapter 3 (but we will subsequently relax later in the dissertation).

2.2 Least Squares Linear Regression

We outline three basic approaches that one could take that make use of OLS linear regression based on the observed outcomes. The first is to simply fit the mean model with the observed biomarker as the outcome and the observed exposures of interest as the predictors: namely, the mean model $\mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{x}^T\boldsymbol{\beta}$. This approach completely ignores the medication use and its mechanism (henceforth referred to as the "Ignore" approach). In this case, $\hat{\boldsymbol{\beta}}_{\text{Ignore}}$ solves the estimating equations:

$$\mathbb{G}_N(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{N} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) = \mathbf{0}. \tag{2.1}$$

In the setting where $y_i = y_i(0) - \delta z_i$ (specifically, when the effect of the medication is uniform across participants on medication), the bias can be computed directly:

$$\begin{aligned}
\text{Bias}(\hat{\boldsymbol{\beta}}_{\text{Ignore}}) &= \mathbb{E}_{\mathbf{X},\mathbf{Z},\mathbf{Y}}[\hat{\boldsymbol{\beta}}_{\text{OLS}}] - \boldsymbol{\beta} \\
&= \mathbb{E}_{\mathbf{X},\mathbf{Z}}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}[\mathbf{Y}]] - \boldsymbol{\beta} \\
&= \mathbb{E}_{\mathbf{X},\mathbf{Z}}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \delta\mathbf{Z})] - \boldsymbol{\beta} \\
&= -\delta\mathbb{E}_{\mathbf{X},\mathbf{Z}}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}] \\
&= -\delta\mathbb{E}_{\mathbf{X}}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{Z}|\mathbf{X}]] \\
&= -\delta\mathbb{E}_{\mathbf{X}}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T P(\mathbf{Z}=1|\mathbf{X})].
\end{aligned} \tag{2.2}$$

This estimator is consistent for some parameter $\boldsymbol{\beta}_{\text{Ignore}}$ which is not, in general, the same as $\boldsymbol{\beta}$. Since medication use is generally not random, errors may not be of mean zero, conditional on $\boldsymbol{x}$. So, trivially sufficient conditions for unbiasedness include one or both of the following: (1) $\delta = 0$, and (2) $z_i = 0$ for all i. In general, this bias cannot be assumed to be zero. It is worth noting that the largest problem in this bias term is the fact that δ is not known. If information on δ were available, one could apply an approximate bias correction by predicting $P(z_i = 1|\boldsymbol{x}_i)$ via, say, logistic regression. However, if δ were thought to be known with great precision, one might simply impute $y_i(0)$ with the observed y_i plus the presumed value of δ for all on-medication participants. The major challenge with this approach, although it has been considered (Tobin et al., 2005), is that results from randomized trials in which estimates of δ are obtained may not generalize to the populations studied in cohort studies. Note that $\boldsymbol{\beta}_{\text{Ignore}}$ has a different interpretation than $\boldsymbol{\beta}$: specifically, $\beta_{k,\text{Ignore}}$ can be interpreted as the difference in expected *observed* biomarker values for two participants differing in x_k by one unit, but having the same value for all other predictors.

Another approach involving OLS linear regression is to exclude on-medication participants from analysis (we refer to this as the "Exclude" method), such that the fitted mean model is $\mathbb{E}[y|\boldsymbol{x}, z = 0] = \boldsymbol{x}^T\boldsymbol{\beta}$. Here, $\hat{\boldsymbol{\beta}}_{\text{Exclude}}$ solves the estimating equations:

$$\mathbb{G}_N(\boldsymbol{\beta}) \quad = \quad \sum_{z_i=0} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) = \mathbf{0}, \tag{2.3}$$

so that we restrict estimation to observations for which $z_i = 0$. Under the assumption that $\mathbb{E}[y(0)|\boldsymbol{x}_i] = \boldsymbol{x}_i^T\boldsymbol{\beta}$, the errors cannot be presumed to be of mean zero conditional on the predictors of interest if medication use depends upon the underlying biomarker: $\mathbb{E}[\epsilon_i|\boldsymbol{x}_i] \neq 0$. In short, this creates a problem of selection bias. This estimator described is consistent for some parameter $\boldsymbol{\beta}_{\text{Exclude}}$, which may be distinctly different from $\boldsymbol{\beta}$. Of note is that if medication use is unrelated to the underlying biomarker, residuals are still

of mean zero conditional on the exposures. The bias here cannot be computed analytically without more specific assumptions on the data generating mechanism. However, we note that the interpretation of the parameter $\boldsymbol{\beta}_{\text{Exclude}}$ is yet again different from the interpretation of $\boldsymbol{\beta}$: $\beta_{k,\text{Exclude}}$ can be interpreted as the difference in expected observed biomarker values for two randomly sampled participants not on medication, differing in x_k by one unit, but having the same value for all other predictors. The target parameter (namely the natural history association) is such that we do not want to condition on medication use status.

The second modification of OLS that we consider is to adjust for medication use as though a confounder (we will call this the "Adjust" approach). Obtaining $\hat{\boldsymbol{\beta}}_{\text{Adjust}}$ amounts to fitting the mean model $\mathbb{E}[y|\boldsymbol{x}, z] = \boldsymbol{x}^T\boldsymbol{\beta} + \delta z$, so that $\hat{\boldsymbol{\beta}}_{\text{Adjust}}$ solves the estimating equations

$$\mathbb{G}_N(\boldsymbol{\beta}) \quad = \quad \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix}^T \left(\mathbf{Y} - \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \delta \end{bmatrix} \right) = \mathbf{0}. \tag{2.4}$$

One intuitive way to understand why this model cannot be expected to perform well is by generalizing directly from the exclusion model. For the purposes of this explanation, consider the hypothetical estimator obtained from only including participants on medication, $\hat{\boldsymbol{\beta}}_{\text{Include}}$ (which is consistent for some parameter $\boldsymbol{\beta}_{\text{Include}}$). The estimating equations for this hypothetical estimator would be given by:

$$\mathbb{G}_N(\boldsymbol{\beta}) \quad = \quad \sum_{z_i=1} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) = \mathbf{0}. \tag{2.5}$$

This estimator solving these equation will tend to be biased for the same reason that $\hat{\boldsymbol{\beta}}_{\text{Exclude}}$ is: the errors are no longer of zero mean. The bias of $\hat{\boldsymbol{\beta}}_{\text{Adjust}}$ can be understood by the following decomposition:

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{Adjust}}] &= \mathbb{E}_z[\mathbb{E}_{\boldsymbol{x}|z}[\hat{\boldsymbol{\beta}}_{\text{Adjust}}]] \\
&= \mathbb{E}_z[\mathbb{E}_{\boldsymbol{x}|z}[\hat{\boldsymbol{\beta}}_{\text{Exclude}}\mathbf{1}(z=0) + \hat{\boldsymbol{\beta}}_{\text{Include}}\mathbf{1}(z=1)]] \\
&= \mathbb{E}_z[\boldsymbol{\beta}_{\text{Exclude}}\mathbf{1}(z=0) + \boldsymbol{\beta}_{\text{Include}}\mathbf{1}(z=1)] \\
&= P(z=0)\boldsymbol{\beta}_{\text{Exclude}} + P(z=1)\boldsymbol{\beta}_{\text{Include}}.
\end{aligned} \tag{2.6}$$

This quantity is, in general, not equal to $\boldsymbol{\beta}$ since both $\boldsymbol{\beta}_{\text{Exclude}}$ and $\boldsymbol{\beta}_{\text{Include}}$ are biased. Note that $\beta_{k,\text{Adjust}}$ can be interpreted as the difference in the expected observed outcome between two randomly sampled participants differing in x_k by one unit, but of the same medication use status and having equal values for all other predictors. If medication use is influenced by $y(0)$, then medication use does not serve as a confounder for the association of interest (that between $\boldsymbol{x}$ and $y(0)$) in the classical sense (Pearl, 2009). In particular, medication use is not a cause of the underlying biomarker. The OLS linear regression approaches described generally presume medication use not to be influenced by the underlying biomarker, an assumption highly unlikely to hold in observational data. As an illustration of the failure of naïve ordinary least squares approaches, see Figure 2.1, which is based on a single realization of a data generating mechanism in which medication use depends upon the underlying biomarker.

Figure 2.2 depicts a directed acyclic graph (DAG) summarizing a reasonable data generation mechanism for cross-sectional observational data with endogenous medication use. We acknowledge the existence of some variables associated with medication use, $\boldsymbol{w}$ (with association $\boldsymbol{\alpha}$, in a sense that will later be made more explicit, and allowing an intercept). Hence, let $\mathbf{W}$ denote a corresponding design matrix of these covariates. Additionally, we note that the variables $\boldsymbol{x}$ and $\boldsymbol{w}$ may share some common covariates. Some models (e.g., instrumental variables) do not permit complete overlap, which will be made apparent shortly. Table 2.1 summarizes the notation of covariates and parameters defined so far.

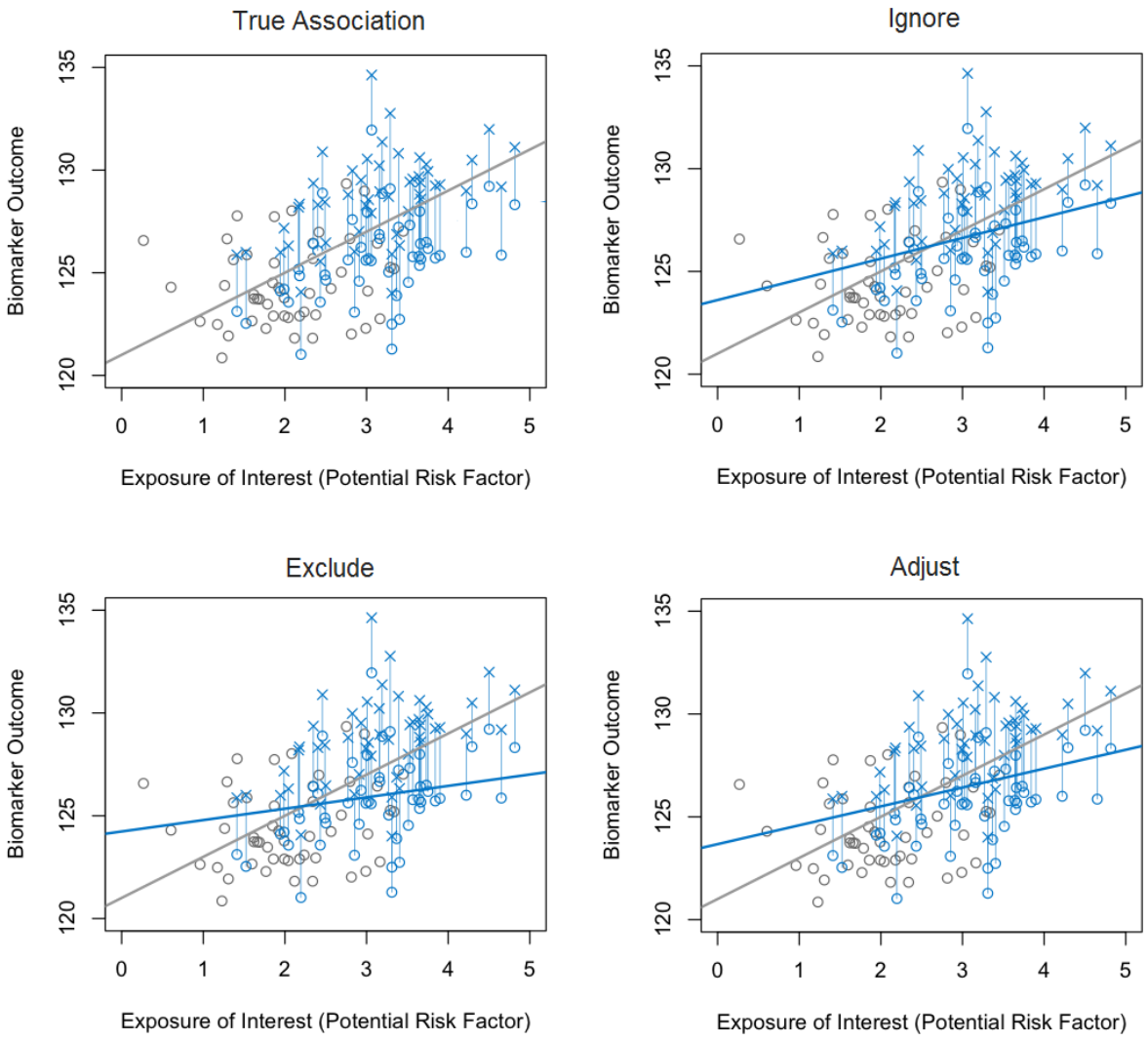

Figure 2.1: Illustration of the failure of naïve OLS linear regression approaches in the setting of endogenous medication use. Observed values for off-medication participants are shown as gray circles; on-medication participants are shown in blue. The observed ($\circ$) and the untreated ($\times$) values are connected with a line segment. The gray line signifies the true association ($\beta = 2$); the blue lines show results from the three naïve regression models: Ignore ($\hat{\beta} = 1.01$), Exclude ($\hat{\beta} = 0.55$), and Adjust ($\hat{\beta} = 0.92$).

2.3 Inverse Probability Weighting

IPW, along with other propensity score methods, can correct certain types of selection bias in a longitudinal setting by up-weighting data for observed participants who had a

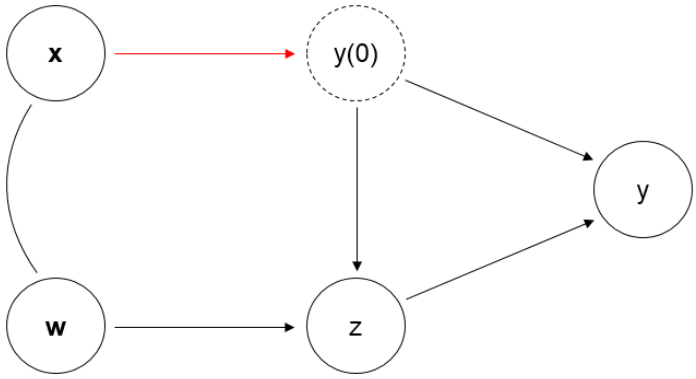

Figure 2.2: DAG illustrating relationship between covariates and outcomes. Solid indicates observed variables, dashed indicates partially observed variables. Note that the covariates of $\boldsymbol{x}$ and $\boldsymbol{w}$ need not be unique, as indicated by the curves lines connecting them. The arrow between $\boldsymbol{x}$ and $y(0)$ (red) corresponds to the association of interest.

Table 2.1: A summary and description of the notation we will be using throughout this dissertation.

	Description
Covariates	
$\boldsymbol{x}$	predictors of the underlying biomarker
$\boldsymbol{w}$	predictors of medication use status
z	medication use status (binary)
$y(0)$	underlying biomarker
y	vector of observed biomarker values
Parameters	
$\boldsymbol{\beta}$	natural history association between $\boldsymbol{x}$ and $y(0)$
$\boldsymbol{\alpha}$	association between $\boldsymbol{w}$ and z

low probability of being observed, thus allowing them to represent unobserved participants. Both on-medication participants with low risk covariates and the off-medication participants with high risk covariates are least likely to be observed. An approach was developed by Wang and Fang (2011) to estimate $\boldsymbol{\beta}$ when pre-treatment values were observable in study participants. This approach relies on medication use initiation taking place post-baseline, and also assumes that the underlying biomarker "category" (high

or low) is observable, even though the underlying biomarker value itself is not. This approach could naïvely be applied in a cross-sectional setting. In the cross-sectional setting, IPW can be described by the following two-stage procedure:

- Fit a logistic model using the binary indicator of medication use as the response, and known predictors of medication use as the predictors; obtain subject-specific probability-of-medication-use predictions (conditional on covariates).

- Fit a weighted least squares (WLS) linear regression model (choosing from either the exclude method or adjustment method) with weights given by the inverse of the predicted probablities of belonging to the observed medication group.

The Exclude approach with the weights as described above is referred to as "inverse probability of censoring weighting" (IPCW) if we think of $y_i(0)$ as being "censored," although the assumption that $y_i(0) - y_i \geq 0$ if $z_i = 1$ is not truly invoked in this setting. Re-weighting the Adjust method is referred to as "inverse probability of treatment weighting" (IPTW), which makes use of all observed data.

Stabilizing weights according to the unconditional probability of treatment can sometimes be of value in bias reduction, and so we choose to incorporate this feature (Cole et al., 2008). One additional important modification to this traditional IPW procedure is to only include variables that are associated with both medication use and the biomarker in Stage 1 to generate predicted probabilities. This is understood to produce better behavior than simply placing all covariates associated with medication use in the weighting model (Lefebvre et al., 2008). Let $\mathbf{W}_*$ denote the design matrix $\mathbf{W}$ restricting to covariates that also predict the underlying biomarker, and $\boldsymbol{\alpha}_*$ the corresponding parameter vector. Then, incorporating these two modifications, IPCW can be summarized by the following two-stage approach:

- Solve the following estimating equations for $\hat{\boldsymbol{\alpha}}_*$:

$$\mathbb{G}_N^1(\boldsymbol{\alpha}_*) \;=\; \mathbf{W}_*^T(\mathbf{Z} - \text{expit}(\mathbf{W}_*\boldsymbol{\alpha}_*)) = \mathbf{0}, \tag{2.7}$$

  and obtain predictions $\hat{\pi}_{0i}$ for the conditional probability of medication use. Then, solve the estimating equations $\mathbb{G}_N^0(\alpha_0) = \mathbf{1}^T(\mathbf{Z} - \text{expit}(\mathbf{1}\alpha_0)) = \mathbf{0}$ to obtain an estimate $\hat{\pi}_0$ of the unconditional probability of medication use (this is for stabilization).

- Solve the following (weighted) estimating equations to obtain $\hat{\boldsymbol{\beta}}_{\text{IPCW}}$:

$$\mathbb{G}_N(\boldsymbol{\beta}|\hat{\boldsymbol{\alpha}}_*, \hat{\alpha}_0) \;=\; \sum_{z_i=0} \frac{(1-\hat{\pi}_0)\boldsymbol{x}_i^T(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})}{1 - \hat{\pi}_{0i}(\boldsymbol{w}_i, z_i; \hat{\boldsymbol{\alpha}}_*, \hat{\alpha}_0)} = \mathbf{0}. \tag{2.8}$$

For IPTW, the first step of this two-stage model is the same; the estimating equations for the second stage incorporate information from the on-medication participants as well. They are given by:

$$\mathbb{G}_N(\boldsymbol{\beta}|\hat{\boldsymbol{\alpha}}_*, \hat{\alpha}_0) \;=\; \sum_{z_i=0} \frac{(1-\hat{\pi}_0)\boldsymbol{x}_i^T(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})}{1 - \hat{\pi}_{0i}(\boldsymbol{w}_i, z_i; \hat{\boldsymbol{\alpha}}_*, \hat{\alpha}_0)} + \sum_{z_i=1} \frac{\hat{\pi}_0\boldsymbol{x}_i^T(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})}{\hat{\pi}_{0i}(\boldsymbol{w}_i, z_i; \hat{\boldsymbol{\alpha}}_*, \hat{\alpha}_0)} = \mathbf{0}. \tag{2.9}$$

The greatest apparent challenge of this model as it applies to our estimand of interest is that the underlying biomarker is unobservable in on-medication participants, and hence not able to be incorporated in weight generation.

The ability of IPCW to impact bias not only depends on the mechanism of medication use, but also on violation of the linearity assumption after excluding participants on medication (that is, $\mathbb{E}[y|\boldsymbol{x}, z=0] = \boldsymbol{x}^T\boldsymbol{\beta}_{\text{Exclude}}$ for some $\boldsymbol{\beta}_{\text{Exclude}}$). If the conditional linearity assumption is satisfied, there will be no bias reduction. In IPTW, no bias reduction will be achieved if, in addition, $\mathbb{E}[y|\boldsymbol{x}, z=1] = \boldsymbol{x}^T\boldsymbol{\beta}_{\text{Include}}$ for some $\boldsymbol{\beta}_{\text{Include}}$, such that linearity holds after exclude all off-medication participants. To see this in the

IPCW case, let F denote the CDF of $\boldsymbol{x}$ and $y(0)$, conditional on $z = 0$. It will be useful to first realize $\hat{\boldsymbol{\beta}}_{\text{Exclude}}$ as a the unique solution to the unbiased estimating equations:

$$\sum_{z_i=0} \boldsymbol{x}_i^T (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) = \mathbf{0}, \tag{2.10}$$

which exists under mild regularity conditions even under a misspecified mean model. Here, we mean "unbiased" for the $\boldsymbol{\beta}_{\text{Exclude}}$, namely the functional $T(F)$ implicitly defined by the integral equation:

$$\int \boldsymbol{x}^T (y - \boldsymbol{x}^T \boldsymbol{\beta}) dF = \mathbf{0}. \tag{2.11}$$

Suppose that the conditional linearity assumption is met: $\mathbb{E}[y|\boldsymbol{x}, z = 0] = \boldsymbol{x}^T \boldsymbol{\beta}_{\text{Exclude}}$. Then trivially, the IPCW estimating equations (2.6) are unbiased estimating equations for $\boldsymbol{\beta}_{\text{Exclude}}$. This follows since $\hat{\pi}_{0i}$ is not a function of y_i, conditional on $z = 0$:

$$\begin{aligned} \mathbb{E}_{y_i|\boldsymbol{x}_i}[\boldsymbol{x}_i^T (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{\text{Exclude}})] &= \mathbb{E}_{y_i|\boldsymbol{x}_i}\left[\left(\frac{1-\hat{\pi}_0}{1-\hat{\pi}_{0i}}\right) \boldsymbol{x}_i^T (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{\text{Exclude}})\right] \\ &= \left(\frac{1-\hat{\pi}_0}{1-\hat{\pi}_{0i}}\right) \boldsymbol{x}_i^T \mathbb{E}_{y_i|\boldsymbol{x}_i}\left[y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{\text{Exclude}}\right] \\ &= \left(\frac{1-\hat{\pi}_0}{1-\hat{\pi}_{0i}}\right) \boldsymbol{x}_i^T \cdot 0 = \mathbf{0} \end{aligned} \tag{2.12}$$

Re-weightng may of course still alter efficiency in this setting, which we will explore empirically in Chapter 3. On the other hand, suppose instead that the conditional linearity assumption is not met, so that there is no $\boldsymbol{\beta}_{\text{Exclude}}$ such that $\mathbb{E}[y|\boldsymbol{x}, z = 0] = \boldsymbol{x}^T \boldsymbol{\beta}_{\text{Exclude}}$. Then, the estimating equations of (2.7) need not be unbiased for $\boldsymbol{\beta}_{\text{Exclude}}$. Instead, $\hat{\boldsymbol{\beta}}_{\text{IPCW}}$ is unbiased for the the functional $\tilde{T}(F)$ implicitly defined by

$$\int \left(\frac{1-\hat{\pi}_0}{1-\hat{\pi}_{0x}(\boldsymbol{x})}\right) \boldsymbol{x}^T (y - \boldsymbol{x}^T \boldsymbol{\beta}) dF = \mathbf{0}, \tag{2.13}$$

which may be distinct from that of $T(F)$.

Whether or not the conditional linearity assumption is met depends on the medication use mechanism. Take the following simple example for the purposes of illustration: suppose that $y_i(0) = x_i + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, 1)$, and that medication use is defined by one of the following rules:

$$\begin{aligned} \text{Rule 1: } z_i &= \mathbf{1}(y_i(0) < 2), \\ \text{Rule 2: } z_i &= \mathbf{1}(2x_i + y_i(0) < 4) \\ \text{Rule 3: } z_i &= \mathbf{1}(x_i + y_i(0)/20 < 2) \end{aligned}$$

In the first scenario, medication use depends only on the biomarker. In the second scenario, medication use depends both on the biomarker and the exposure. In the third scenario, medication use slightly depends on the biomarker, and the conditional linearity assumption is nearly met. Figure 2.3 shows the results empirically with a single realization ($N = 10,000$). In the first two scenarios, IPCW performs more poorly than the Exclude approach. In the third scenario, both estimates are very close to the truth.

This very simple example demonstrates the lack of reliability of inverse probability weighting to address the challenges of endogenous medication use, particularly when naïvely exended to cross-sectional data, in which this approach was not intended to be used. Firstly, weights cannot be generated using the unobservable off-medication biomarker value, and secondly, the impact of weighting on bias is highly dependent on some difficult to describe measure of mean-model misspecification induced by excluding participants. Intuitively, selection based on the outcome alters the solution to the estimating equations: in general, consistency will still hold, but the value for which $\hat{\boldsymbol{\beta}}_{\text{IPCW}}$ (and in turn, $\hat{\boldsymbol{\beta}}_{\text{IPTW}}$) will be consistent is altered depending on how badly linearity is violated after conditioning on medication use. IPCW up-weights the observations for which medication use is most likely, which in this example up-weights where the model

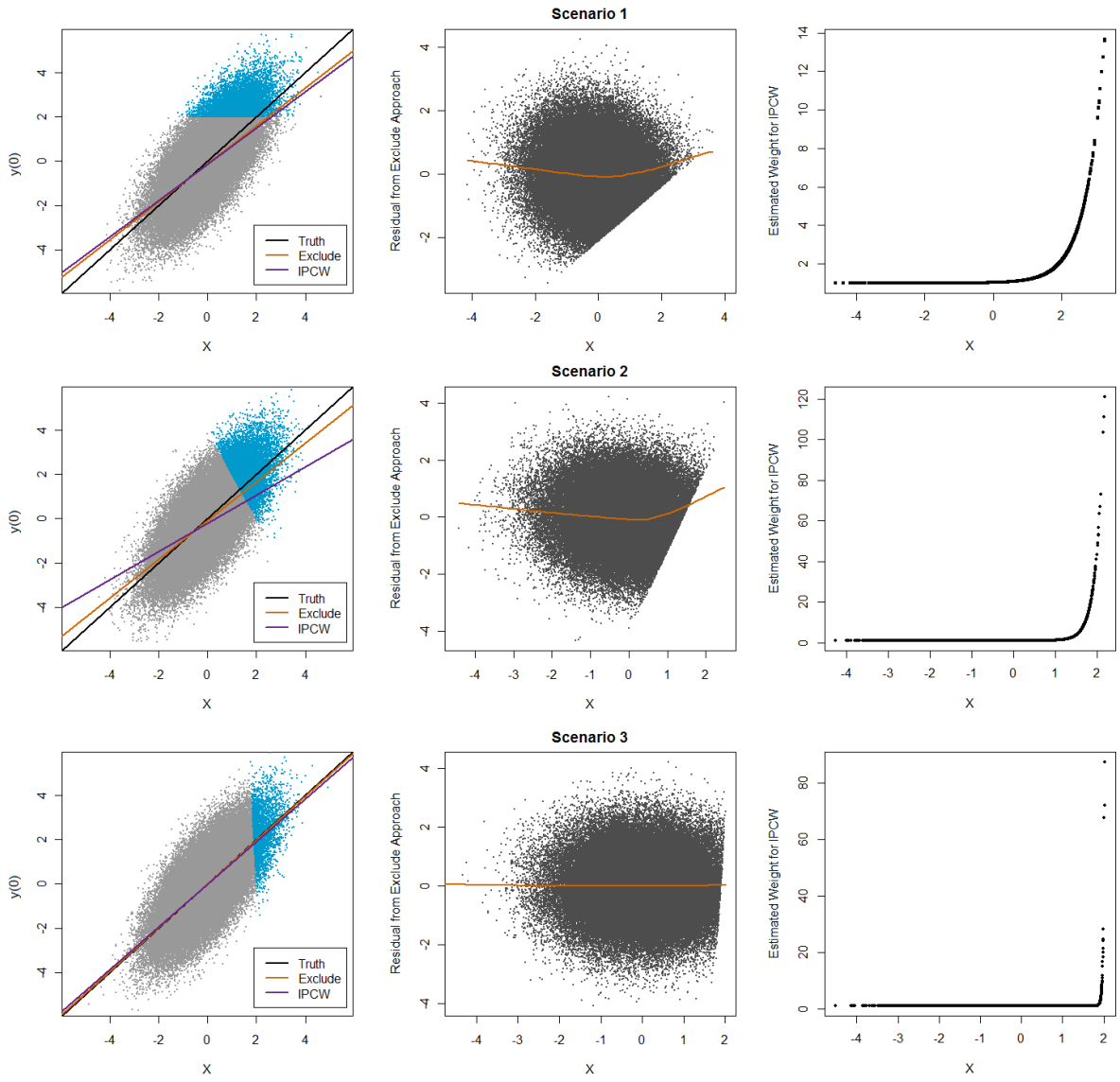

Figure 2.3: Empirical demonstration of the importance of conditional linearity in IPCW estimates. The left plots depict participants on medication in blue, and off-medication participants in gray. The center plots depict a plot of the residuals as a function of x with a LOESS smoother. There is a clear lack of mean-model linearity in the first two scenarios; the residuals for Scenario 3 are nearly of mean zero, conditional on x. The right plots depict the IPCW weights as a function of x.

is most violated. Hence, IPCW is not a particularly well defined procedure for addressing endogeneity in this setting, since it is not guaranteed to upweight observations for which the mean model is locally closer to that of the mean model of interest (that of

the underlying biomarker model).

The re-weighting approaches have other known limitations, including the requirement of positivity of weights. In the setting of cross-sectional data, we do not have the advantages of the approach described by Wang and Fang (2011), although their modeling assumptions, too, have limitations. In longitudinal data, participants are often on medication at baseline, and we also think of the underlying biomarker as being the corresponding off-medication biomarker value that would occur at the time of measurement.

2.4 Censored Normal Regression

Although normality is not a requirement of OLS linear regression, the estimating equations used to generate estimators arise from, and can be motivated from the score equations from the normal likelihood. The likelihood function and resulting score equations are given by:

$$\begin{aligned} \mathcal{L}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(\boldsymbol{\beta}, \sigma_y^2) &= \prod_{i=1}^{N} \frac{1}{\sigma_y} \phi\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma_y}\right) \\ \mathbb{G}_N(\boldsymbol{\beta}; \sigma_y) &= \frac{\partial \ell(\boldsymbol{\beta}; \sigma_y)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma_y^2} \sum_{i=1}^{N} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) = \mathbf{0}. \end{aligned} \tag{2.14}$$

where ϕ represents the standard normal density function and σ_y^2 the error variance. The choice of $\boldsymbol{\beta}$ that solves the estimating equations $\mathbb{G}_N(\boldsymbol{\beta}; \sigma_y) = \mathbf{0}$ is valid irrespective of σ_y^2, and so normality and homoscedasticity are not invoked.

In turn, recognizing that OLS approaches fail to properly account for medication use, Tobin et al. (2005) modified this normal likelihood model to account for the fact that $y_i(0)$ and y_i are not the same if $z = 1$. In particular, he assumes that $y_i(0) \geq y_i$ if $z = 1$, integrating the normal likelihood to from the observed value to y_i to infinity for participants on medication:

$$\begin{aligned}\mathcal{L}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(\boldsymbol{\beta},\sigma_y^2) &= \prod_{z_i=0}\frac{1}{\sigma_y}\phi\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma_y}\right)\prod_{z_i=1}\int_{y_i}^{\infty}\frac{1}{\sigma_y}\phi\left(\frac{t-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma_y}\right)dt \\ &= \prod_{z_i=0}\frac{1}{\sigma_y}\phi\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma_y}\right)\prod_{z_i=1}\left[1-\Phi\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma_y}\right)\right],\end{aligned} \tag{2.15}$$

where Φ represents the standard normal cumulative distribution function. The corresponding score equations for estimation of $\boldsymbol{\beta}$ still depend on the normal density and cumulative distribution function; the treatment effect is presumed to follow a truncated normal distribution beyond the observed biomarker. The censored normal model is appealing for a number of reasons: it is a single-stage approach that uses all participants; it also uses information in the observed biomarker, which can often be highly informative about $y(0)$. However, the censored normal model presumes that censoring is non-informative, an assumption that is almost certainly not satisfied in practice. If the mechanism that gives rise to medication use depends on the underlying biomarker or covariates in $\mathbf{X}$, non-informative censoring is violated. Tobin recognizes that the non-informative censoring assumption is likely not satisfied, but in his motivating example of blood pressure and genetic exposures, he believes this likelihood model to provide an adequate approximation to the truth.

To illustrate the impact of informative censoring on estimation, Figure 2.4 presents results from a data generation mechanism in which the non-informative censoring assumption is violated. Estimation of $\boldsymbol{\beta}$ shows substantial upward bias. The further the exposure values deviate from their mean, the higher their leverage. Here, those with high values of the exposure are most likely to be on medication. Correcting (integrating) the biomarker has a larger influence in estimating $\boldsymbol{\beta}$, hence the upward bias in this example. If the participants on medication tended to be those with the lower exposure values, the censored normal model would tend to provide downward-biased estimates.

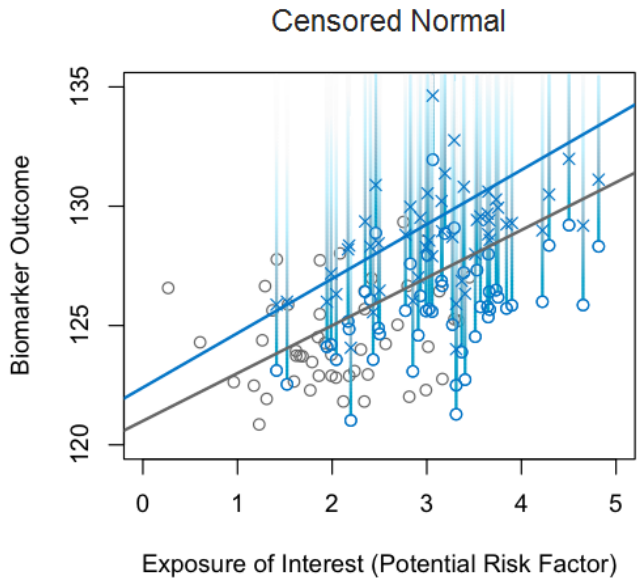

Figure 2.4: Demonstration of informative censoring. Gray circles depict off-medication participants, and on-medication participants are in blue, with ($\circ$) denoting the observed value and ($\times$) denoting the off-medication value. The true association ($\beta = 2$) is given in gray, and the blue line represents censored normal regression ($\hat{\beta} = 2.28$). The truncated normal distribution for each participant is denoted as a fading blue line.

If the effect of medication on the biomarker is believed to be somewhat small, integration to infinity may seem to overcompensate for uncertainty in the underlying biomarker. If we instead were to integrate to some unknown parameter τ which could in turn be estimated, and then optimize over $\boldsymbol{\beta}, \sigma_y$, and τ, we would get exactly the censored normal solution. This is the case because $\phi(\cdot) > 0$, and hence the integral $\int_{y_i}^{\tau} \sigma_y^{-1} \phi[(t - \boldsymbol{x}_i^T \boldsymbol{\beta})/\sigma_y] dt$ is monotone increasing as $\tau \nearrow \infty$. That is, $\tau = \infty$ would be the optimal choice for any value of $\boldsymbol{\beta}$ and σ_y.

In order to allow a parameter in the limit of integration and account for the potential overcompensation, one might think to utilize a regularization approach to penalize large values of τ, together with a cross-validation to determine the appropriate amount of shrinkage. However, it is not clear what criteria should be used to select an optimal value of τ, or how such a selection could be "guided" by the underlying biomarker value.

If we use, for example, mean squared error in prediction of the observed biomakrer, we are still faced with the challenge that $y_i(0)$ is unobservable in participants on medication. Moreover, regularization in this way may reduce the impact of overstating uncertainty, but fundamentally does not address the problem of informative censoring. It is unclear how the assumption of non-informative censoring could be relaxed without substantial modifications to the censored normal approach. The challenge of integration thresholding might be useful in other tangential settings in which censoring is not informative.

2.5 Instrumental Variables

Instrumental variable (IV) approaches are widely used in econometrics to address endogenous medication use and estimate causal effects. A standard one-stage approach, described by Wooldridge (2011) can be described as follows. Assume that the observed biomarker can be written as $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \delta z_i + \epsilon_i$, so that $\boldsymbol{x}_i$ are the (exogenous) predictors of interest, and z_i is potentially correlated with ϵ_i. A special variable (w_i), called an instrumental variable, is presumed to exist, such that the following properties are satisfied:

- The instrumental variable variable w_i is correlated with z_i (medication use).

- The variable w_i (the IV) is uncorrelated with ϵ_i (the error term); that is to say that $\mathbb{E}[w_i\epsilon_i] = 0$.

Note that in this case, it is presumed that there is no overlap between $\boldsymbol{x}$ and w. Let $\boldsymbol{z}_i' = (\boldsymbol{x}_i^T, z_i)^T$, so that the biomarker equation can be written more compactly as

$$y_i \;=\; (\boldsymbol{z}_i')^T \begin{bmatrix} \boldsymbol{\beta} \\ \delta \end{bmatrix} + \epsilon_i. \tag{2.16}$$

Further let $\boldsymbol{w}_i' = (\boldsymbol{x}_i^T, w_i)^T$. By the second assumption, together with the fact that $\boldsymbol{x}$ is a set of exogenous variables, $\mathbb{E}[\boldsymbol{w}_i'\epsilon_i] = \mathbf{0}$. Multiplying the biomarker equation through by $\boldsymbol{w}_i'$, we have that

$$\boldsymbol{w}_i' y_i = \boldsymbol{w}_i'(\boldsymbol{z}_i')^T\boldsymbol{\beta}' + \boldsymbol{w}_i'\epsilon_i, \tag{2.17}$$

where $\boldsymbol{\beta}' = (\boldsymbol{\beta}^T, \delta)^T$. Taking expectations, and invoking the second assumption, we have

$$\begin{aligned} \mathbb{E}[\boldsymbol{w}_i' y_i] &= \mathbb{E}[\boldsymbol{w}_i'(\boldsymbol{z}_i')^T\boldsymbol{\beta}'] + \mathbb{E}[\boldsymbol{w}_i'\epsilon_i] \\ &= \mathbb{E}[\boldsymbol{w}_i'(\boldsymbol{z}_i')^T\boldsymbol{\beta}'] + \mathbf{0} \\ &= \mathbb{E}[\boldsymbol{w}_i'(\boldsymbol{z}_i')^T]\boldsymbol{\beta}' \end{aligned} \tag{2.18}$$

By the first assumption, $\mathbb{E}[\boldsymbol{w}_i'(\boldsymbol{z}_i)^T]$ is of full rank, and hence we have that the true parameter vector $\boldsymbol{\beta}'$ can be written as $\left(\mathbb{E}[\boldsymbol{w}'(\boldsymbol{z}')^T]\right)^{-1}\mathbb{E}[\boldsymbol{w}'y]$. By the Weak Law of Large Numbers, we may estimate those expectations easily to obtain the instrumental variables estimator of $\boldsymbol{\beta}'$:

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}}' = \left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{w}_i'(\boldsymbol{z}_i')^T\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{w}_i' y_i\right]. \tag{2.19}$$

Note that the last component of $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}'$ contains an estimate of the treatment effect, δ. IV approaches can work very well when the conditions listed above are satisfied. In the setting of a randomized experiments, for example, one may wish to estimate the causal effect of an intervention on an outcome, where the intervention is not successfully given to the subjects. In this case, the "intent" to assign can act as the instrument, and by virtue of the fact that we are in the setting of a randomized trial, the assumptions listed above are reasonable.

In the setting of observational data, however, identification of such a variable w_i can be challenging since medication use cannot be assumed to be at random (or even close

to it in most settings). Unfortunately, the IV approach is understood to be sensitive to departures from the above assumptions (most importantly, the second assumption). The use of IV approaches in pharmacoepidemiologic studies has been criticized in the literature (Hernán et al., 2006).

2.6 Discussion

We have presented several easy-to-implement approaches that have previously been either inappropriately applied in the setting of endogenous medication use (OLS approaches) or have been presented as a way to address it (IPW, censored normal regression, and instrumental variables). Specifically, we have outlined the assumptions of each approach and discussed the major sources and mechanisms for bias.

The OLS approaches are most widely used in practice. Many researchers generally appear to be aware that medication use obscures estimation of the association between a predictor of interest and a biomarker outcome, so methods such as excluding and adjusting are more common than simply ignoring medication use altogether. However, these modifications address medication use in ways that are too simplistic and fail to appropriately account for endogeneity.

We have also presented two inverse-probability weighting approaches, which are inappropriate generalizations of the approach proposed by Wang and Fang (2011) in a longitudinal setting as a means to address endogeneity. Our choice to explore this method was initially done as an experimentation to evaluate IPW as a potential candidate for estimating the natural history association. However, upon exploration, the shortcomings revealed themselves quickly, prompting this brief exploration of the source of bias. The first clear limitation of IPW approaches in a cross-sectional setting is that the underlying biomarker, which we believe is highly correlated with medication use, is unobservable in participants on medication. However, also concerning is that the mechanism by which IPW approaches re-weighting participants in a cross-sectional set-

ting does not necessarily place greater weight on areas where the association between the predictor of interest and the observed biomarker outcome is closer to that of the association between the predictor and the underlying biomarker outcome (which is of interest). Hence, IPW approaches in this setting cannot be expected to solve the problems associated with endogenous medication use, and are more appropriately restricted to settings in which longitudinal data are available.

The censored normal model proposed by Tobin et al. (2005) relies on non-informative censoring, an assumption that is highly unlikely to hold in observational data precisely due to endogeneity. The assumption of a truncated normal treatment effect can also be of concern in settings where medication is not effective for every participant. Relaxing assumptions by expanding limits of integration would be challenging even with regularization approaches since there is no information in the data that could guide selection of the limits.

The instrumental variables approach can be of great use in estimating the natural history association if in fact one has an instrument. However, beyond the fact that the assumption that of identifying an instrument is unlikely to hold, this is inherently an unverifiable assumption since there is no estimator for the error term; moreover, the resulting estimator is extremely sensitive to departures from this assumption, particularly if the variable used as an instrument is only weakly associated with medication (Angrist and Krueger, 2001).

Ultimately, the unmet need this dissertation seeks to address going forward is to devise a set of approaches that place reasonable assumptions on: (1) the mechanism by which medication users differ from non-users, and (2) the nature of the effect of medication use on the underlying biomarker value (in terms of its magnitude and distribution, either marginally or conditional on observed covariates). When strong assumptions must be made, we will place emphasis on trying to evaluate sensitivity to departures

from these assumptions in order to have a richer understanding of the models considered, so that informed modeling choices can be made when working with cross-sectional data. Many of the approaches previously considered fall short in that they choose to condition on medication use. Instead, it is more appropriate to estimate the unconditional association between $\boldsymbol{x}$ and $y(0)$ that would occur in the absence of medication use. This will ultimately mean we must understand and model the mechanism by which medication users differ from non-users, and jointly model medication use with the underlying biomarker.

We acknowledge the existence of other approaches that could be implemented in an attempt to address medication use. For example, one could apply fixed or random addition (adding a constant to the observed biomarker for on-medication participants, with or without noise). The ability of these approaches to perform well depends very highly on having a prior understanding of treatment effect magnitude. If that knowledge is accurate, these approaches will perform well. If knowledge is lacking, they will not. In keeping with the goal of developing of models that can help us identify new epidemiologic predictors for biomarkers that may not be well understood, we restrict our attention to approaches that do not rely on the validity of external information beyond the constraints of the cross-sectional data available.

Bias has been of major interest to us in our exploration of these six approaches. Indeed, bias reduction will continue to be the major focus of this dissertation. Our motivating examples of cardiovascular biomarkers are all such that high values are undesirable and hence most likely to prompt medication use. Thus, all approaches except the censored normal model overwhelmingly provide downwardly biased estimates of associations of interest which can greatly understate their relevance quantitatively. However, even if fully understanding the magnitude of the association is not of interest, and one would like to simply test whether or not an association exists, this attenuation

can still be problematic. If the association is meaningfully attenuated, identification of important epidemiologic predictors can require a far larger sample size to compensate for the loss of power to detect an association.

While bias has been of most interest to us in this chapter, an exploration of efficiency will still be important (we leave this to later chapters). Challenges will be particularly apparent for the approaches which exclude subsets of the participants (Exclude and IPCW). If the prevalence of medication use is high in the cohort, a substantial loss of efficiency will be induced by using these approaches. In the settings of hypertension and hyperlipidemia, especially, efficiency challenges would be of great concern as medication use can exceed 50% in some populations.

Regarding estimation of standard errors, Huber-White "sandwich" based standard error estimates can be used for the least-squares and IPW approaches to accommodate heteroscedasticity and various forms of model misspecification (White, 1980). For the censored normal approach, it is straightforward to derive a robust variance estimator based on the observed information and an empirical "meat" matrix. We will do so more explicitly in our presentation of Heckman's TEM in Chapter 3.

Chapter 3

BACKGROUND: THE HECKMAN TREATMENT EFFECTS MODEL

Up until this point, we have focused primarily on analytically arguing the inadequacy of a number of common approaches to estimate the natural history association between some predictor and a biomarker when there is endogenous medication use. We focus in this chapter on providing a framework introduced by James Heckman (1978) to address challenges surrounding endogeneity. The general approach in this framework involves simultaneous equation modeling in order to estimate parameters of interest. In general, the parameter of most interest to Heckman and those working in econometrics in that time period was the population average treatment effect (δ); many of his motivating problems involve understanding salary disparities between genders, or the impact of policy implementation across demographic factors. There is a direct parallel to our problem in the sense that there are partially observed outcomes; in the example of salary disparities by gender, for example, the unobservable outcome would be the salary of a non-working individual. However, this context is more naturally identified with his sample selection model (Heckman, 1976). Instead, we will focus on what he referred to as the "hybrid model with structural shift," better known as the "treatment effects model" (TEM). As will be made apparent in this chapter, this framework can be used to consistently estimate the natural history association ($\boldsymbol{\beta}$) when applied to the setting of cardiovascular biomarkers. This is taking a clear departure from the historical context of these models, in which the effect of the intervention on the outcome of interest has been the parameter of interest. We instead treat this as a nuisance parameter.

In this chapter, we will show how Heckman's TEM provides a framework to explicitly account for and model the mechanism by which medication users differ from non-users. Such features are ignored by OLS-based approaches and censored normal regression, or are greatly oversimplified. As such, we believe Heckman's TEM corresponds to a data generation mechanism that we believe much more closely resembles a process that would give rise to true cross-sectional observational data as compared to many of the approaches of Chapter 2.

We will describe Heckman's TEM as it was originally proposed, including a proof of identifiability of parameters and a presentation of the maximum likelihood estimators. As we are seeking to utilize this framework as a foundation for estimating the natural history association, the subsequent modifications we make in this chapter (and in this dissertation in general) will be geared toward that purpose: we will show that the model can accommodate explicit dependence of medication use on the underlying biomarker, and allows for random treatment effects. We then proceed to compare the TEM to the simple approaches of Chapter 2 through a set of simulation studies. The primary goal of these studies is to evaluate (i) how large the treatment effect magnitude must be and (ii) what level of endogeneity must be present in order to observe meaningful gains from the TEM, which is admittedly more complicated than alternative approaches. We will characterize gains in terms of both bias and root mean squared error (rMSE) to confirm the inappropriateness of the alternative approaches in the setting of endogenous medication use. We will conclude this chapter with a comparison of the TEM to the IV approach when the presumed "instrument" is, in reality, weakly associated with the underlying biomarker (violating the major assumption of the IV approach). The goal of this study is to verify that the TEM does not share the same reliance on an IV that the IV approach does.

3.1 Heckman's TEM, The Principal Assumption, and Identifiability

Recall that $\mathbf{X}$ is composed of predictors of interest, and $\mathbf{W}$ is composed of predictors of medication use, and these design matrices may overlap. The TEM seeks to correct endogeneity bias by parametric specification of the mechanism by which medication users differ from non-users. In its most elementary form, a probit model $P(z_i = 1|\boldsymbol{w}_i, \boldsymbol{x}_i) = \Phi(\boldsymbol{w}_i^T\boldsymbol{\alpha})$ is used to describe this mechanism in addition to the biomarker model, $y_i(0) = \boldsymbol{x}_i^T\boldsymbol{\beta}+\epsilon_i$. The probit model can be rewritten as $z_i^* = \boldsymbol{w}_i^T\boldsymbol{\alpha}+\gamma_i$ for some latent, continuous variable z_i^*, where γ_i follows a normal distribution, and so that $z_i = \mathbf{1}(z_i^* > 0)$ is the observed medication use status. Presuming medication use reduces $y(0)$ by an unknown δ, we have that $\mathbb{E}[y_i|\boldsymbol{w}_i, \boldsymbol{x}_i, z_i] = \boldsymbol{x}_i^T\boldsymbol{\beta} - \delta z_i$. Thus, we have a system of simultaneous equations for the underlying biomarker and medication use. Figure 3.1 modifies the DAG in Figure 2.2 to incorporate the latent variable.

The simultaneous equations given above are a case of a general simultaneous system for $y_i(0)$ and z_i^*, which Heckman (1978) used to introduce and motivate his work in an attempt to encompass various forms of dependencies that could occur:

$$\begin{aligned} y_i(0) &= \boldsymbol{x}_i^T\boldsymbol{\beta} + \delta_2 z_i + \lambda_2 z_i^* + \epsilon_i \\ z_i^* &= \boldsymbol{w}_i^T\boldsymbol{\alpha} + \delta_1 z_i + \lambda_1 y_i(0) + \gamma_i. \end{aligned} \tag{3.1}$$

Here, $y_i(0)$ is partially observed (differing from the observed y_i by some real-valued, but unknown parameter δ_2 in participants on medication). Structural shift ($\delta_1, \delta_2 \neq 0$) is permitted subject to constraints we will soon discuss. The errors ϵ_i and γ_i are presumed to be i.i.d. and of zero mean; the two error terms are of unknown variances σ_y^2 and σ_γ^2, respectively, and have common correlation ρ. A semi-reduced form can be written as follows so that the observed biomarker and the latent medication use variables are written in terms of observable covariates (for the time being, assume covariates are allocated to either $\boldsymbol{x}$ or to $\boldsymbol{w}$, but not to both):

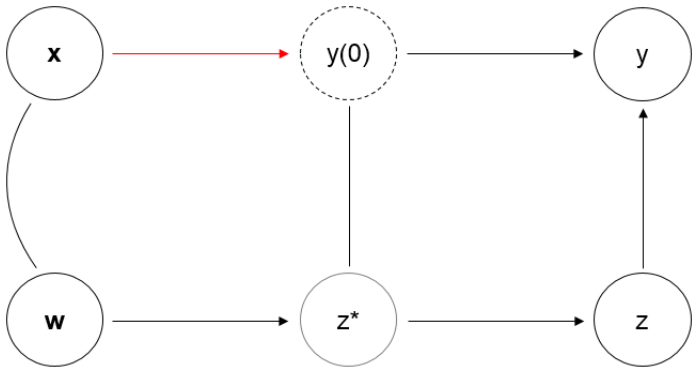

Figure 3.1: DAG illustrating relationship between covariates and outcomes. This extends the DAG of Figure 2.2 to accommodate the latent continuous variable z^*, shown in a gray circle, and its correlation with $y(0)$, represented by a straight line. The arrow between $\boldsymbol{x}$ and $y(0)$ (red) remains the association of interest.

$$\begin{aligned} y_i &= \boldsymbol{x}_i^T \boldsymbol{\pi}_{11} + \boldsymbol{w}_i^T \boldsymbol{\pi}_{12} + \pi_{13} z_i + \epsilon_i' \\ z_i^* &= \boldsymbol{x}_i^T \boldsymbol{\pi}_{21} + \boldsymbol{w}_i^T \boldsymbol{\pi}_{22} + \pi_{23} z_i + \gamma_i'. \end{aligned} \tag{3.2}$$

The total errors are given by, $\epsilon_i' = (\epsilon_i + \lambda_2 \gamma_i)/(1 - \lambda_1 \lambda_2)$, and $\gamma_i' = (\gamma_i + \lambda_1 \epsilon_i)/(1 - \lambda_1 \lambda_2)$, and the parameters given by:

$$\begin{aligned} &\boldsymbol{\pi}_{11} = \frac{\boldsymbol{\beta}}{1 - \lambda_1 \lambda_2}, \quad \boldsymbol{\pi}_{21} = \frac{\lambda_1 \boldsymbol{\beta}}{1 - \lambda_1 \lambda_2}, \quad \boldsymbol{\pi}_{12} = \frac{\lambda_2 \boldsymbol{\alpha}}{1 - \lambda_1 \lambda_2}, \\ &\boldsymbol{\pi}_{22} = \frac{\boldsymbol{\alpha}}{1 - \lambda_1 \lambda_2}, \quad \pi_{13} = \frac{\delta_2 + \lambda_2 \delta_1}{1 - \lambda_1 \lambda_2}, \quad \pi_{23} = \frac{\delta_1 + \lambda_1 \delta_2}{1 - \lambda_1 \lambda_2}. \end{aligned} \tag{3.3}$$

Heckman proceeds to prove that the condition $\pi_{23} = 0$ is necessary and sufficient in order for the parameters of the above model to exist. This condition, called the "principal assumption," is equivalent to writing $\lambda_1 \delta_2 + \delta_1 = 0$ from the original system (3.1). The fact that this principal assumption is a necessary condition for the parameters of the model to exist is not obvious, and the proof is provided in the seminal paper (1978). This assumption places a restriction on the structural shift and the endogeneity; by rewriting the partially reduced system, Heckman argues that the principal assumption intuitively allows one to uniquely define the probability of z_i being either zero or

one. Since z_i is defined to be $\mathbf{1}(z_i^* > 0)$ it is sensible to require that z_i^* may not depend on z_i, the binary variable it determines.

The classical version of the TEM as implemented in modern software presumes that $\lambda_1 = \lambda_2 = \delta_1 = 0$ so that $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \delta_2 z_i + \epsilon_i$ and $z_i^* = \boldsymbol{w}_i^T\boldsymbol{\alpha} + \gamma_i$. However, another less restrictive way to satisfy the principal assumption is to choose $\lambda_1 = \delta_1 = 0$ and let λ_2 be a real-valued, finite, free-varying parameter. Structural shift in the biomarker is still permitted, and medication use is still permitted to be correlated with the underlying biomarker value, but this allows systematic dependence of the probability of medication use on the underlying biomarker value (see Figure 3.2).

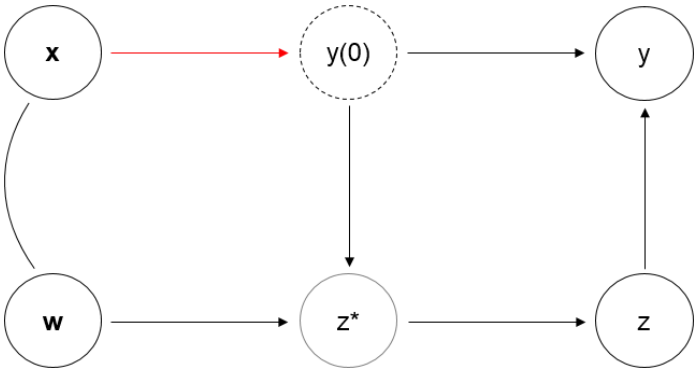

Figure 3.2: This DAG extends that of Figure 3.1 to accommodate systematic influence of $y(0)$ on z^*, represented by an arrow rather than a straight line. The arrow between $\boldsymbol{x}$ and $y(0)$ (red) remains the association of interest.

For the remainder of this dissertation, we invoke the assumptions that $\lambda_1 = \delta_1 = 0$ in order to satisfy the principal assumption, noting that (1) there are other ways to satisfy this restriction that give rise to models that may be of interest in other settings, and (2) this alternative formulation is indeed less restrictive than the models implemented in modern software in that it accommodates direct influence of $y_i(0)$ on z_i^*. Now, for simplicity in notation, let $-\delta$ replace δ_2 to denote the structural shift in the first equation, and let λ replace λ_1 to denote endogeneity; rewriting (3.1) to accommodate these imposed constraints, we have the following updated system:

$$\begin{aligned} y_i(0) &= \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i \\ z_i^* &= \boldsymbol{w}_i^T\boldsymbol{\alpha} + \lambda y_i(0) + \gamma_i \\ z_i &= \mathbf{1}(z_i^* > 0) \\ y_i &= y_i(0) - \delta z_i. \end{aligned} \tag{3.4}$$

With this updated parameterization, the principal assumption is satisfied, although the parameters of the second equation are not of interest in our problem (we will show shortly that they are not identifiable). It turns out that λ need not be estimated. Taking the second equation from (3.4), and decomposing the term $\lambda y_i(0)$ into its systematic and random components, we have:

$$\begin{aligned} z_i^* &= \boldsymbol{w}_i^T\boldsymbol{\alpha} + \lambda y_i(0) + \gamma_i \\ &= \boldsymbol{w}_i^T\boldsymbol{\alpha} + \lambda(\boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i) + \gamma_i \\ &= \boldsymbol{w}_i^T\boldsymbol{\alpha} + \lambda\boldsymbol{x}_i^T\boldsymbol{\beta} + \lambda\epsilon_i + \gamma_i \\ &\equiv \tilde{\boldsymbol{w}}_i^T\tilde{\boldsymbol{\alpha}} + \tilde{\gamma}_i. \end{aligned} \tag{3.5}$$

Here, $\tilde{\boldsymbol{w}}_i$ denotes the combined exposures in $\boldsymbol{x}$ and $\boldsymbol{w}$, and $\tilde{\boldsymbol{\alpha}}$ the corresponding parameter vector given by:

$$\tilde{\boldsymbol{\alpha}} = \begin{cases} \boldsymbol{\alpha} & \text{for covariates in } \boldsymbol{w} \text{ only} \\ \lambda\boldsymbol{\beta} & \text{for covariates in } \boldsymbol{x} \text{ only} \\ \boldsymbol{\alpha} + \lambda\boldsymbol{\beta} & \text{for covariates in } \boldsymbol{x} \text{ and } \boldsymbol{w} \end{cases}. \tag{3.6}$$

The total error is given by $\tilde{\gamma}_i = \lambda\epsilon_i + \gamma_i$. Thus, we have that the distribution of the error terms is given by:

$$\begin{bmatrix} \epsilon_i \\ \tilde{\gamma}_i \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \left(\mathbf{0}, \begin{bmatrix} \sigma_y^2 & \lambda\sigma_y^2 + \rho\sigma_y\sigma_\gamma \\ \lambda\sigma_y^2 + \rho\sigma_y\sigma_\gamma & \lambda^2\sigma_y^2 + \sigma_\gamma^2 \end{bmatrix} \right), \tag{3.7}$$

so that Heckman's TEM as modeled in (3.4) is indeed correctly specified if one simply places all covariates appearing in $\mathbf{X}$ into $\mathbf{W}$. Going forward, we will exploit this useful fact, but for simplicity of notation, we will use the symbol $\boldsymbol{\alpha}$ rather than $\tilde{\boldsymbol{\alpha}}$ to denote the parameter vector with covariates of $\boldsymbol{x}$ included; similarly, we will use the notation γ_i for the total error, with total variance given by $\sigma_z^2 = \lambda^2\sigma_y^2 + \sigma_\gamma^2$ as the total variance of z^*. Maddala (1983) outlined a procedure in order to determine which parameters in the model are identifiable; we utilize this framework to show that $\boldsymbol{\beta}$, δ, and σ_y are identifiable. This proof is trivial when $\boldsymbol{x}$ and $\boldsymbol{w}$ do not overlap, and so we only present the proof in case where they do.

Theorem: Suppose data are generated according to (3.5) such that the total errors are i.i.d. of mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. If we model the data according to (3.4) with all covariates of $\mathbf{X}$ included in $\mathbf{W}$, then $\boldsymbol{\beta}$, δ, and σ_y are identifiable parameters.

Proof: Let $\mathbf{A}$ denote a 2×2 nonsingular matrix. Let $\tilde{\boldsymbol{x}}$ denote a vector of length p containing all unique predictors in $\boldsymbol{x}$ and $\boldsymbol{w}$ (each having q_1 and q_2 predictors, respectively), and partition this complete covariate vector into three total classes: $\tilde{\boldsymbol{x}}_{(1)}$ denotes predictors in $\boldsymbol{x}$ only; $\tilde{\boldsymbol{x}}_{(2)}$ denotes predictors in $\boldsymbol{w}$ only, and $\tilde{\boldsymbol{x}}_{(3)}$ denotes predictors in both $\boldsymbol{x}$ and $\boldsymbol{w}$. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ can be partitioned accordingly as well. If $\boldsymbol{o} = (y, z^*)$ denotes the (partially observed) outcome vector, and $\tilde{\boldsymbol{x}}' = (1, \tilde{\boldsymbol{x}}^T, z)^T$ the covariate vector in the first structural equation, the simultaneous system may be written as $\mathbf{A}\boldsymbol{o} - \boldsymbol{\Gamma}\tilde{\boldsymbol{x}}' = (\epsilon, \gamma)^T$ for a $2 \times (p+2)$ dimensional matrix of coefficients, $\boldsymbol{\Gamma}$, given in this case by:

$$\boldsymbol{\Gamma} = \begin{bmatrix} \beta_0 & \boldsymbol{\beta}_{(1)}^T & \mathbf{0}^T & \boldsymbol{\beta}_{(2)}^T & \delta \\ \alpha_0 & \mathbf{0}^T & \boldsymbol{\alpha}_{(1)}^T & \boldsymbol{\alpha}_{(2)}^T & 0 \end{bmatrix} \tag{3.8}$$

The system can in turn be simplified to $\boldsymbol{o} = \boldsymbol{\Pi}\tilde{\boldsymbol{x}}' + (\epsilon', \gamma')^T$, where $\boldsymbol{\Pi} = \mathbf{A}^{-1}\boldsymbol{\Gamma}$, and the error vector has covariance matrix $\boldsymbol{\Omega} = \mathbf{A}^{-1}\boldsymbol{\Sigma}\mathbf{A}^{-T}$. Letting

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 \\ 0 & 1/\sigma_z \end{bmatrix} \tag{3.9}$$

where σ_z^2 is the total variance on z^*. Maddala (1983) shows that when the structural equations can be written in this form, parameters appearing in matrices $\boldsymbol{\Lambda\Pi}$ and $\boldsymbol{\Lambda\Omega\Lambda}^T$ are identifiable. The result that $\boldsymbol{\beta}$, δ, and σ_y are identifiable follows from setting $\mathbf{A}$ to be the 2×2 identity matrix. Q.E.D.

The parameters $\boldsymbol{\alpha}$ and ρ are identifiable up to a scale factor of σ_z. For this reason, it is computationally convenient to set $\sigma_z = 1$, as is standard in probit analysis (Freedman, 2010). The challenges associated with the fact that endogeneity cannot be parsed exactly between λ and ρ are bypassed by placing all covariates in $\mathbf{X}$ into $\mathbf{W}$ and not estimating λ. Identification of the natural history association relies on the stable unit treatment value and consistency assumptions. The former states that $(y_i(0), y_i(1))$ is independent of z_j for all $1 \leq i \neq j \leq N$ (such that the medication use status of one individual does not influence the medication use status of another individual). The latter states that the observed value y_i is equal to the potential outcome under the medication use status absolutely observed, $y_i(z_i)$.

3.2 The Two-Stage Approach and Maximum Likelihood

Heckman proposed a two-stage approach, or "indirect least squares" for estimation of parameters. We may standardize the error on z_i^* by σ_z^{-1} so that parameters are estimable. Bivariate normal theory can then be used to obtain the conditional distribution of $y_i|(z_i, \boldsymbol{x}_i, \boldsymbol{w}_i)$, for which the error term is normally distributed, and parameters of interest may be estimated. The specific details of the two-stage approach are discussed

by Heckman (1978), and in a later paper Heckman, 1979. This approach provides consistent estimates of the parameters $\boldsymbol{\beta}, \delta$, and σ_y, but it is not asymptotically efficient. Note that this approach for estimation is the first time we are invoking parametric assumptions on the errors; it was not necessary to invoke any assumptions about bivariate normality in order to show identifiability of parameters of interest.

Heckman makes note that although z_i^* is latent, the correlation, ρ, between z_i^* and $y_i(0)$ is still estimable in the presence of the dichotomous z_i. The assumption of normality provides a means of estimating the the point-biserial correlation, which estimates the correlation between continuous data and a dichotomous variable; this is discussed in greater detail by Tate (1954). Similar work has been done by Telser (1964).

Of greater interest to us is a likelihood-based approach in which the errors (ϵ_i, γ_i) are modeled as bivariate normal with unknown variance parameters σ_z^2 and σ_y^2, and correlation parameter ρ. In contrast to the two-stage approach, the likelihood approach is asymptotically efficient. Hence, we consider only the likelihood-based approach in this dissertation. Heckman (1978) points out that the two-stage estimator can be used to initialize likelihood estimation. Letting $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_y^2, \rho, \delta)$, the likelihood function for the original TEM can be derived as follows:

$$
\begin{aligned}
\mathcal{L}_{\mathbf{W},\mathbf{X},\mathbf{Y},\mathbf{Z}}(\boldsymbol{\theta}) &= \prod_{i=1}^{N} p(y_i|z_i)p(z_i) \\
&= \prod_{z_i=0} p(y_i|z_i^* \leq 0)p(z_i^* \leq 0) \prod_{z_i=1} p(y_i|z_i^* > 0)p(z_i^* > 0) \\
&= \prod_{z_i=0} \int_{-\infty}^{0} p(y_i) dF_{z_i^*|y_i}(z_i^*) \prod_{z_i=1} \int_{0}^{\infty} p(y_i) dF_{z_i^*|y_i}(z_i^*) \\
&= \prod_{i=1}^{N} p(y_i) \prod_{z_i=0} \int_{-\infty}^{0} p(z_i^*|y_i) dz_i^* \prod_{z_i=1} \int_{0}^{\infty} p(z_i^*|y_i) dz_i^* \\
&= \prod_{i=1}^{N} \frac{1}{\sigma_y} \phi\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} + \delta z_i}{\sigma_y}\right) \\
&\qquad \times \Phi\left((-1)^{1-z_i} \frac{\boldsymbol{w}_i^T\boldsymbol{\alpha} + \rho(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} + \delta z_i)/\sigma_y}{\sqrt{1-\rho^2}}\right)
\end{aligned}
\tag{3.10}
$$

Here, the expressions $\boldsymbol{w}_i^T\boldsymbol{\alpha} + \rho(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} + \delta z_i)/\sigma_y$ and $1 - \rho^2$ represent the conditional mean and variance of $z_i^*|y_i$, respectively. As we have noted, $\boldsymbol{\alpha}$ and ρ are weakly identifiable up to a scale factor of σ_z, and so we set $\sigma_z = 1$ without loss of generality.

All assumptions previously stated for identification of parameters still hold. To implement a maximum likelihood based approach, one requires a further assumption, as shown by Heckman (1978). That is, in order for the likelihood function above to possess an interior maximum, we need that

$$\min_{z_i=1}\{y_i\} < \max_{z_i=0}\{y_i\} \quad \text{and} \quad \max_{z_i=1}\{y_i\} > \min_{z_i=0}\{y_i\}. \tag{3.11}$$

In other words, the range of observed biomarker values between participants on medication and participants off medication must have a non-null intersection. If the groups are completely separated, the log-likelihood function fails to achieve a global maximum on the interior of the parameter space. This is fairly minor restriction; although there is always a nonzero probability that (3.11) will not be satisfied, the probability that it is satisfied converges almost surely to 1 as $N \uparrow \infty$ provided that $0 < p(z_i = 1|\boldsymbol{w}_i) < 1$.

We provide R code for the negative log-likelihood in Appendix A; this function can be optimized in order to fit Heckman's TEM. The function is written such that covariates of $\mathbf{X}$ should be manually included into covariates of $\mathbf{W}$.

3.3 Robust Covariance Estimation and Random Treatment Effects

Letting $\mathcal{U}(\boldsymbol{\theta}) \equiv \partial \log \mathcal{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ denote the score function for the likelihood of (3.9), and letting $\hat{\boldsymbol{\theta}}$ denote the solution to the score equations $\mathcal{U}(\boldsymbol{\theta}) = \mathbf{0}$, define $\mathcal{I}_N(\hat{\boldsymbol{\theta}})$ to be the expected Fisher information matrix at $\hat{\boldsymbol{\theta}}$. Under correct specification, we have a model-based variance estimator from standard asymptotic theory:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) \equiv \mathcal{I}_N^{-1}(\hat{\boldsymbol{\theta}}) \longrightarrow_p \mathrm{Var}(\hat{\boldsymbol{\theta}}). \tag{3.12}$$

In practice, though, it would be ideal not to invoke the assumption that the effect of medication use on the biomarker is a fixed constant; rather, we may assume that a subject specific effect is given by $\delta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\delta, \sigma_\delta^2)$. We have then, that $\delta_i = \delta + \varphi_i$, where φ_i is normally distributed with mean zero and variance σ_δ^2:

$$y_i \;=\; \boldsymbol{x}_i^T\boldsymbol{\beta} - (\delta + \varphi_i)z_i + \epsilon_i \sim \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_y^2 + \sigma_\delta^2 z_i). \tag{3.13}$$

In this setting, heteroscedasticity is introduced as the variance of y_i is now σ_δ^2 greater for on-medication subjects than that of off-medication subjects. Heteroscedasticity in the errors may also arise from other sources other than random treatment effects. The model-based variance estimator is invalid, and hence a robust, heteroscedasticity consistent estimator should be used in its place.

Let $\mathbf{B}_N(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^N \mathcal{U}_i(\hat{\boldsymbol{\theta}})\mathcal{U}_i(\hat{\boldsymbol{\theta}})^T$ to be the empirical "meat" matrix. We prefer to use the observed information $\mathcal{I}_N^{\text{obs}}(\hat{\boldsymbol{\theta}})$ for additional robustness to mean-model misspecification in simulation studies appearing in later chapters, although the mean-model is still correct under random treatment effects. We have that the following convergence result holds

$$\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}) = \left[\mathcal{I}_N^{\text{obs}}(\hat{\boldsymbol{\theta}})\right]^{-1} \mathbf{B}_N(\hat{\boldsymbol{\theta}}) \left[\mathcal{I}_N^{\text{obs}}(\hat{\boldsymbol{\theta}})\right]^{-T} \longrightarrow_p \text{Var}(\hat{\boldsymbol{\theta}}). \tag{3.14}$$

An analogous result holds for the censored normal regression model of Chapter 2.

3.4 Simulation: Gains under Correct Specification

We conduct a simulation study to evaluate the advantages of the TEM over alternative models when the TEM is correctly specified. We do not present the results from the IPCW approach in this study, as the relationship between its results and the Exclude approach is analogous to the relationship between IPTW and the Adjust method.

Let $N = 1000$ denote the number of study subjects, and suppose that the predictors are given by x_{1i}, x_{2i}, and x_{3i}, all i.i.d. $\mathcal{N}(0,1)$. Suppose that the data are generated as follows:

$$\begin{aligned} y_i(0) &= x_{1i} + x_{2i} + \epsilon_i \\ z_i^* &= x_{1i} + x_{3i} + 0.2y_i(0) + \gamma_i \\ &= 1.2x_{1i} + 0.2x_{2i} + x_{3i} + 0.2\epsilon_i + \gamma_i \end{aligned} \tag{3.15}$$

where ϵ_i and γ_i are all i.i.d. $\mathcal{N}(0,4)$. Then, using the results of Section 3.1, we have that the total errors are bivariate normally distributed with covariance matrix given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 0.8 \\ 0.8 & 4.16 \end{bmatrix}. \tag{3.16}$$

Further suppose that $\delta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(1, 1/6)$, and $y_i = y_i(0) - \delta_i z_i$ is the observed biomarker value for subject i. We simulate two-thousand replications from this data generation mechanism and estimate $\boldsymbol{\beta} = (0,1,1)^T$ from the Ignore, Exclude, Adjust, IPTW, CN, and TEM approaches. We also estimate δ from the the Adjust method, CN, and the TEM. Here, $\alpha_0 = 0$ so that there is an approximate 50% prevalence of medication use on average. To account for heteroscedasticity, we use the robust variance-covariance estimator of Section 3.3 for Heckman's TEM (and its analogue for the CN model) and the Huber-White sandwich estimator for the least squares and IPTW approaches. We focus our attention on β_1, β_2, and δ, as the intercept is rarely of interest in association studies. In this simulation, the data are generated such that the TEM is correctly specified for estimation of $\boldsymbol{\beta}$ and δ, whereas the alternative models are not. Hence, we expect the TEM to outperform other approaches; the goal of this study is to attempt to quantify bias reduction as compared to efficiency loss as compared to other approaches under a reasonable simulation setup. The simulation setup is illustrated in Figure 3.3, with the parameters noted by the corresponding associations (arrows) they represent.

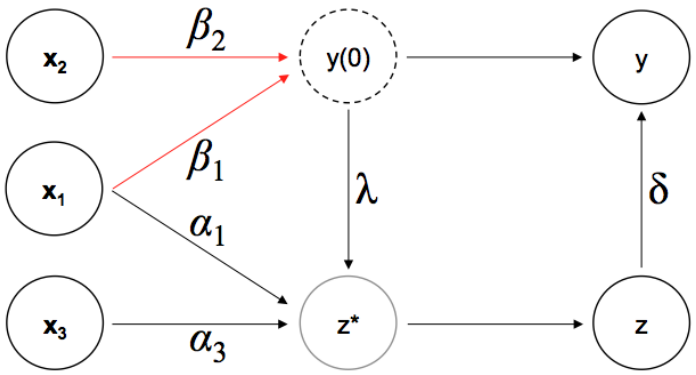

Figure 3.3: This DAG represents the simulation setup of Section 3.4. Note that $\alpha_2 = \lambda\beta_2$, so that the association of x_2 with z^* is indirect. Red arrows indicate the associations of primary interest.

Table 3.1 depicts results from this simulation study. As expected, the TEM shows substantially lower bias than alternative approaches. The root mean squared error (rMSE) is lowest for the TEM for $\boldsymbol{\beta}$ and δ; the gains we see in the rMSE are attributable to bias reduction; this is seen by the fact that the TEM tends to provide estimates with comparable or greater variance than the alternatives. The failure of CN is striking in this example, with bias and rMSE greatly exceeding those of the other approaches for both β_1 and β_2. The robust variance estimates appear to estimate the true variability of $\hat{\boldsymbol{\beta}}$ and $\hat{\delta}$ well, except perhaps for IPTW in estimation of Var($\hat{\boldsymbol{\beta}}$).

The two major sources that play a role in creating bias for approaches that are not equipped to address nonrandom medication use are (i) the strength of endogeneity, as measured by the total correlation on the biomarker/medication use model errors, and (ii) the magnitude of the treatment effect. It feels natural, then, to ask how strong endogeneity has to be for Heckman's TEM to show gains as strong as in Table 3.1, and to similarly ask how large the treatment effect magnitude has to be.

Consider a simulation mirroring the one described above, in which we vary the correlation through λ. So that we are varying the total correlation on the errors, but

Table 3.1: Results from a simulation study comparing six approaches when Heckman's TEM is correctly specified. We consider the bias and standard error as estimated from the simulations, as well as the average of the estimated robust standard errors, and the root mean squared errors.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$				
Ignore	-0.223	4.67	4.62	22.8
Exclude	-0.085	7.13	7.08	11.1
Adjust	-0.084	5.21	5.10	9.82
IPTW	-0.083	7.23	5.74	11.0
CN	0.407	6.17	6.10	41.2
Heckman's TEM	0.000	6.88	6.80	6.88
$\beta_2 = 1$				
Ignore	-0.039	4.62	4.64	6.04
Exclude	-0.015	6.28	6.34	6.46
Adjust	-0.015	4.56	4.57	4.80
IPTW	-0.016	5.00	5.59	5.25
CN	0.068	6.03	5.78	9.10
Heckman's TEM	-0.001	4.67	4.66	4.67
$\delta = 1$				
Adjust	-1.62	10.3	10.2	162
IPTW	-1.62	12.4	11.1	162
Heckman's TEM	-0.001	22.8	22.3	22.8

not their respective variances, set the total error $\sigma_\gamma^2 = 4.16 - 4\lambda^2$ for λ ranging between 0 and 0.8, so that the total error has variance $\sigma_z^2 = 4.16$ throughout. This choice is done to mirror the total variance of z_i^* of the previous simulation setup: $\text{Var}(\gamma_i) + \lambda^2 \text{Var}(y_i(0)) = 4.16$ as in (3.16). All other parameters from the main simulation study are carried over. Hence, the effective total error correlation ranges between 0 and approximately 0.8. Figure 3.4 displays the bias and rMSE as a function of this effective correlation, which we refer to as the "endogeneity strength." The bias and rMSE of Heckman's TEM does not appear to depend on endogeneity strength and consistently performs well; the other approaches are much more sensitive to departures from this assumption, although the Ignore method never does well. The Exclude, Adjust, and IPTW methods

all have acceptable performance when endogeneity is low or nonexistent. When the endogeneity strength is quite low, the adjustment method even outperforms Heckman's TEM in terms of MSE, although as endogeneity becomes nontrivial, the price of failing to adequately account for it becomes very high.

Figure 3.5 depicts the analogous plot for estimation of β_2; trends are quite similar to those of Figure 3.4, with two notable differences: the threshold for substantial worsening is a bit higher than for β_1, and the Ignore method is not biased for β_2 when no endogeneity is present. Recall that x_1 is associated with the underlying biomarker and the probability of medication use, whereas x_2 is only directly associated with the biomarker (and hence weakly associated with the probability of medication use through λ). Hence, the probability of medication use is not as dependent on x_2 as it is on x_1.

Now, consider a simulation which mirrors the main simulation, this time varying δ over a range of values from 0 to 2. All other parameters from the main simulation study are carried over. Figure 3.6 displays the bias and rMSE as a function of the expected treatment effect. Just in the previous case in which we varied endogeneity strength, the TEM performs well, and good performance does not appear to depend on the treatment effect magnitude. The Exclude, Adjust, and IPTW methods show bias that does not appear to vary meaningfully with δ. The performance of the Ignore method depends highly on δ, performing well when there $\delta = 0$ (the underlying biomarker, $y_i(0)$ is truly observed for all subjects in this case), and with bias and rMSE worsening as δ increases.

Although the CN method (not depicted in the figure) improves as δ increases, it is important to recognize that $\mathbb{E}[y_i(0)|z = 1] - \mathbb{E}[y_i(0)|z = 0] \approx 1.5$ under this simulation setup; a medication whose effect is greater than $\delta = 1.5$ reduces the observed biomarker of on-medication participants below of those who are not on medication, a setting that does not seem very realistic. Even at an unrealistically large treatment effect of $\delta = 2.5$, CN provides an estimate of $\hat{\beta}_1 = 1.14$ (Bias $= 0.14$).

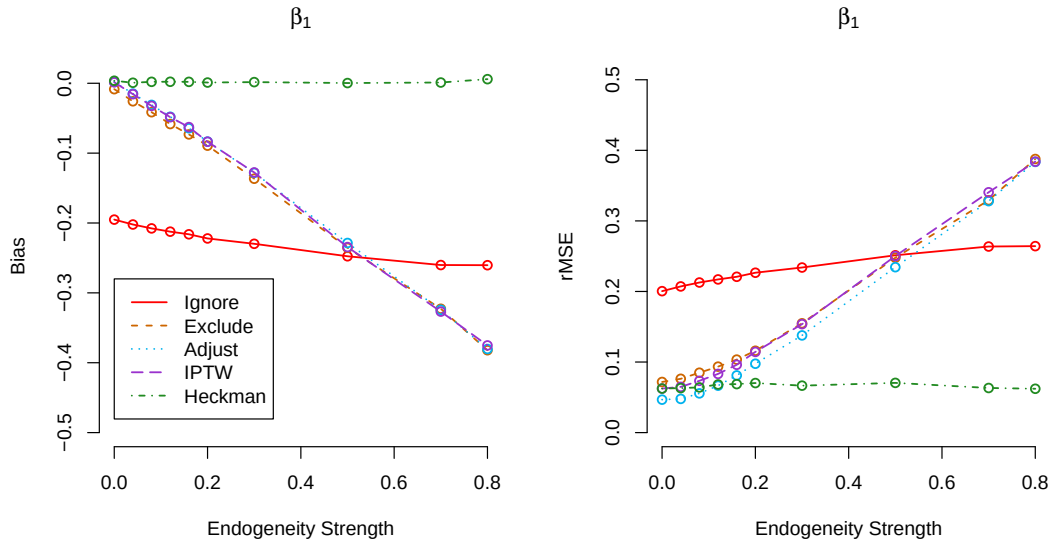

Figure 3.4: Bias and root mean squared error for estimation of β_1 while varying the endogeneity strength through the effective correlation. Results from the censored normal model are not included in the figure.

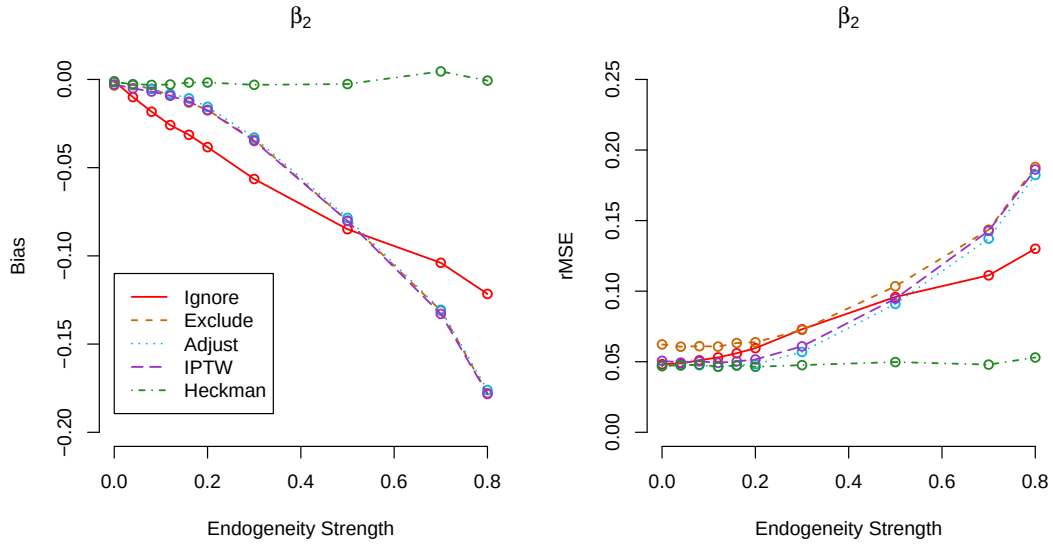

Figure 3.5: Bias and root mean squared error for estimation of β_2 while varying the endogeneity strength through the effective correlation. Results from the censored normal model are not included in the figure.

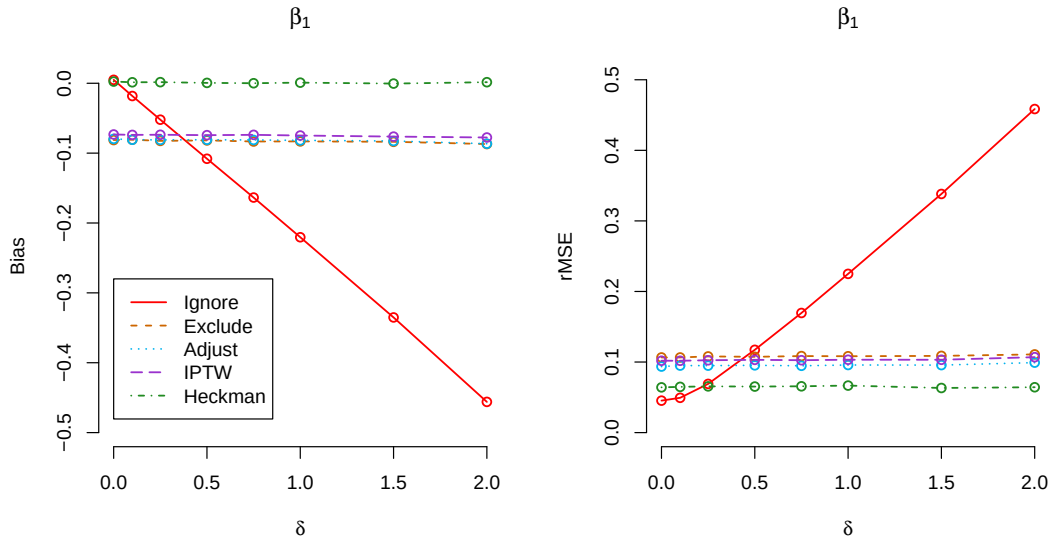

Figure 3.6: Bias and root mean squared error for estimation of β_1 while varying the expected treatment effect magnitude, δ. Results from the censored normal model are not included in the figure.

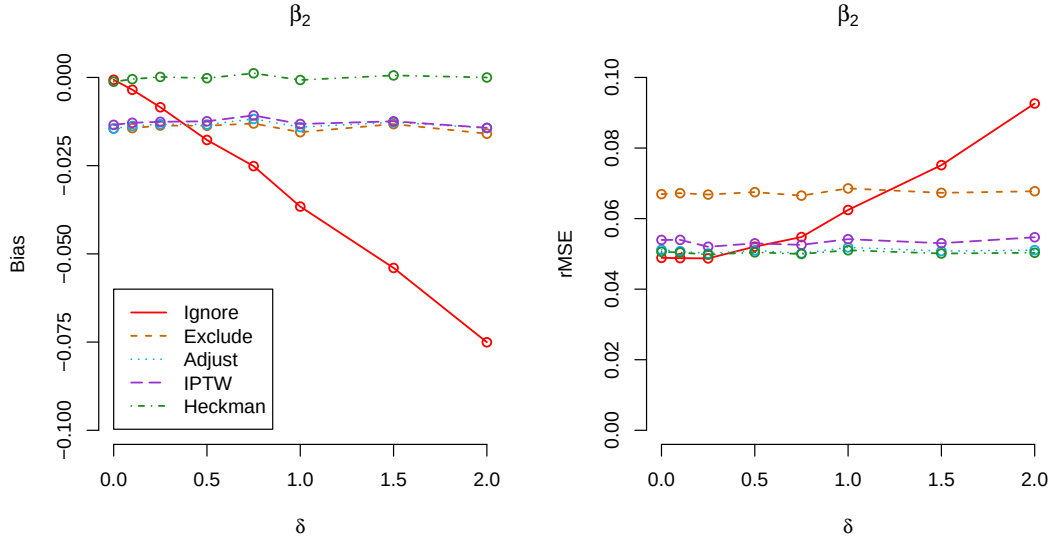

Figure 3.7: Bias and root mean squared error for estimation of β_2 while varying the expected treatment effect magnitude, δ. Results from the censored normal model are not included in the figure.

3.5 Absence of a True Instrumental Variable

In Chapter 2, we presented a classical one-stage IV approach when some measured w is associated with medication use but not the underlying biomarker (see x_3 in Figure 3.3). The IV approach does not permit inclusion of w in the biomarker model (it is not a true instrument if it is associated with the biomarker). However, if w is believed to be weakly associated with the biomarker, the TEM permits the inclusion of w in the biomarker model. We seek to evaluate whether Heckman's TEM provides valid estimates of the natural history association when the variable believed to be an instrument in the context of the IV approach is, in reality, weakly associated with the underlying biomarker. Suppose that the data generation mechanism of Section 3.4 holds with the following modification:

$$y_i(0) \quad = \quad x_{1i} + x_{2i} + \beta_3 x_{3i} + \epsilon_i, \tag{3.17}$$

where β_3 varies over from 0 to 0.5, so that it x_3 increases in its association with $y(0)$. We wish to compare the IV approach to the TEM for estimation of the natural history association. The IV estimator, as a reminder, is given by

$$\hat{\boldsymbol{\beta}}_{\text{IV}} \quad = \quad \left[(\mathbf{W}')^T\mathbf{Z}'\right]^{-1}(\mathbf{W}')^T\mathbf{Y}, \tag{3.18}$$

where $\mathbf{W} = \begin{bmatrix}\mathbf{1} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3\end{bmatrix}$ and $\mathbf{Z}' = \begin{bmatrix}\mathbf{1} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{Z}\end{bmatrix}$. Simplifying, the estimator is given by:

$$\hat{\boldsymbol{\beta}}_{\text{IV}} \quad = \quad \begin{bmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{X}_1 & \mathbf{1}^T\mathbf{X}_2 & \mathbf{1}^T\mathbf{Z} \\ \mathbf{X}_1^T\mathbf{1} & \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2 & \mathbf{X}_1^T\mathbf{Z} \\ \mathbf{X}_2^T\mathbf{1} & \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_2 & \mathbf{X}_2^T\mathbf{Z} \\ \mathbf{X}_3^T\mathbf{1} & \mathbf{X}_3^T\mathbf{X}_1 & \mathbf{X}_3^T\mathbf{X}_2 & \mathbf{X}_3^T\mathbf{Z} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}^T\mathbf{Y} \\ \mathbf{X}_1^T\mathbf{Y} \\ \mathbf{X}_2^T\mathbf{Y} \\ \mathbf{X}_3^T\mathbf{Y} \end{bmatrix} \tag{3.19}$$

If in reality x_3 is associated with $y(0)$, failure to specify it in the underlying biomarker model results in unmeasured confounding. Hence, we compare the above IV estimator to Heckman's TEM when x_3 included in the biomarker model. Unsurprisingly, naïve OLS approaches, along with the IPW approaches and CN were all found to perform poorly in this setting, for estimation of $\boldsymbol{\beta}$ and δ (where applicable). We do not report on these results in depth. Of greater interest is the comparison between Heckman's TEM and the IV approach.

Figure 3.8 depicts the results comparing the IV approach to the correctly specified TEM for estimation of of β_1 (both bias and rMSE are provided). As expected, the estimates of β_1 from the IV approach remain approximately unbiased when x_3 truly is an instrument (that is, when $\beta_3 = 0$). The simulated rMSE is greater for the TEM than that of the IV approach when β_3 is small (i.e., when x_3 is only weakly associated with $y(0)$); this is at least partially attributable to the greater variability associated with estimating more parameters in the TEM. However, the bias of the IV approach very clearly shows a steady downward trend as the association between x_3 and $y(0)$ grows, and the advantages in terms of rMSE vanish at around $\beta_3 = 0.08$. Relative to the other parameters ($\beta_1 = \beta_2 = 1$), this is not a very strong association–note that all three covariates are generated from a standard normal distribution. Importantly, Heckman's TEM does not show evidence of meaningful bias across β_3.

Figure 3.9 depicts the analogous results for β_2. A similar pattern is observed, although the bias in this setting is an order of magnitude lower for the IV approach than for estimation of β_1 (recall that x_2 is not as strongly associated with the medication use). Under this setup, the TEM and the IV approach show comparable levels of rMSE across a range of β_3 values. As β_3 increases, IV increases in both bias and rMSE.

Also of interest is to compare estimation of the average treatment effect, δ, between these two approaches. Figure 3.10 illustrates these results. Indeed, a similar pattern

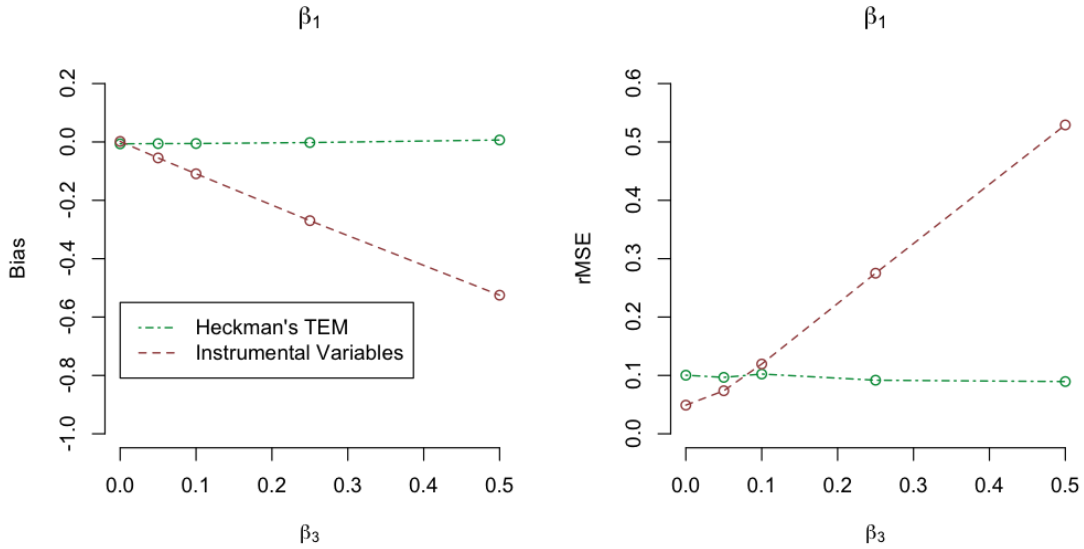

Figure 3.8: Bias and root mean squared error for estimation of β_1 when x_3 is weakly associated with the underlying biomarker (and accounted for in the biomarker model of Heckman's TEM).

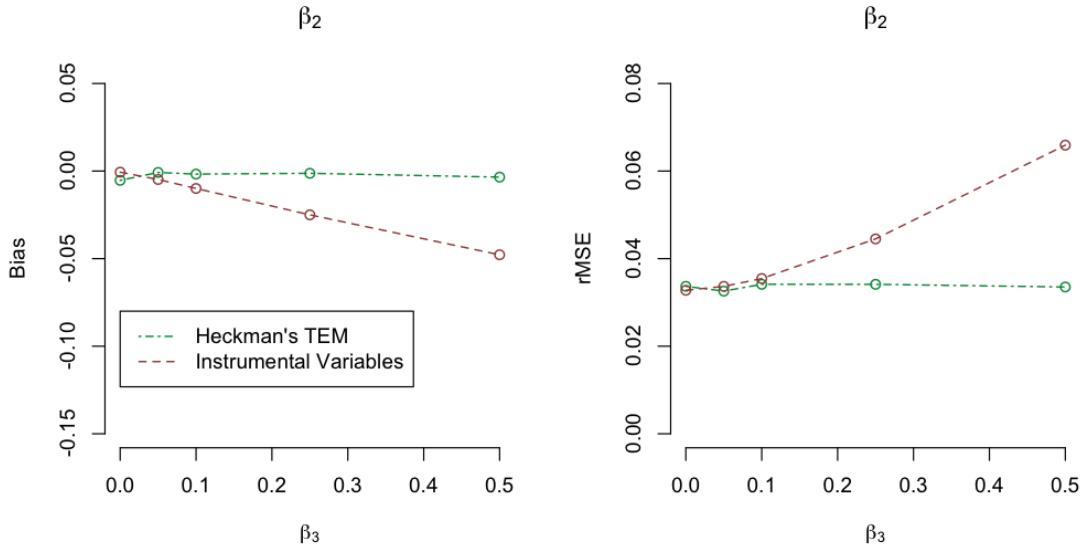

Figure 3.9: Bias and root mean squared error for estimation of β_2 when x_3 is weakly associated with the underlying biomarker (and accounted for in the biomarker model of Heckman's TEM).

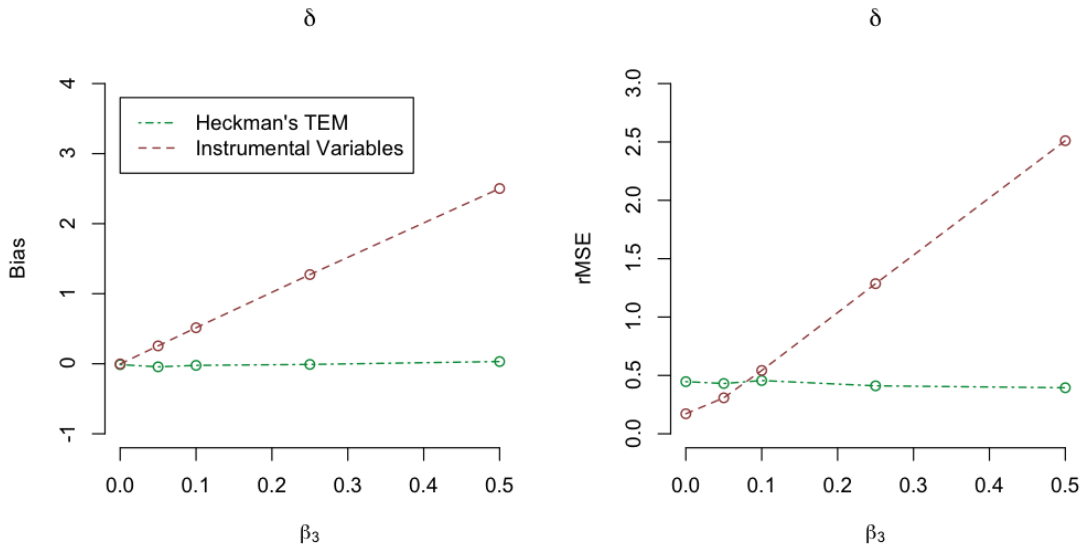

Figure 3.10: Bias and root mean squared error of for estimation of δ when x_3 is weakly associated with the underlying biomarker (and accounted for in the biomarker model of Heckman's TEM).

in bias is observed for the two approaches as with the natural history association. The estimate of the average treatment effect provided by Heckman's TEM shows larger variability, as seen by the fact that IV dominates the nearly unbiased Heckman's TEM over small values of β_3 in terms of rMSE. However, considering β_3 values beyond approximately 0.1, the TEM appears to dominate the IV approach for estimation of δ. If one believes that x_3 is truly an instrument (or close to it), and interested in estimating the population-average treatment effect, the IV approach would be the optimal choice on the basis of rMSE; however, in the setting of cross-sectional observational data, justifying this is difficult since the assumption of having an instrument is untestable, and the TEM performs well irrespective of this assumption.

Since $\boldsymbol{\beta}$ is our parameter of interest, and not δ, this finding supports use of Heckman's TEM in the setting where an instrument is not thought to exist. However, note that Heckman's TEM was correctly specified in this case, in that we chose to adjust for it in the biomarker model. The classical one-stage IV approach does not permit this

modification (otherwise the x_3 variable would not be an instrument). In that sense, it is crucial to adjust for x_3 in the biomarker model if it is believed to be even weakly associated with the biomarker to avoid confounding bias. Failure to specify x_3 in the biomarker model when weakly associated with medication use produces levels of bias that are similar to those observed from the IV approach in this study. In observational data, true instruments may be difficult to find (and the assumption of a variable being an instrument is not testable). Together with the fact that the TEM does not demand an instrument for estimates of $\boldsymbol{\beta}$ to outperform the IV estimates, this provides justification to be skeptical of the IV approach for estimating the natural history association in observational data. These findings are consistent with those of Marchenko et al. (2012), in which a minor modification to the Heckman sample selection model was evaluated under the setting of no true instrumental variable.

3.6 Discussion

3.6.1 Justification of Parameter Selection

Burton et al. (2006) point out that in the design and implementation of simulation studies to demonstrate results, careful attention should be paid to how closely the simulated data sets resemble data that would be observed in the real world. While we will take measures to do this more elaborately in Chapter 4 by means of perturbing model assumptions, selection of parameters plays a large role in how realistic the simulation setup is. Of course, variables can be scaled and shifted, so conditional on a (somewhat arbitrary) selection of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, σ_y, and σ_z, careful selection of the endogeneity parameter and treatment effect is important. As we have noted, the selection of $\lambda = 0.2$ yields an expected difference of approximately 1.5 in the underlying biomarker between participants on and off medication. Thus, selection of $\delta = 1$ suggests a medication that has an is moderately effective in restoring the biomarker values toward those of the

off-medication counterparts. The simultaneous selection of λ and δ in this way reveals that $\lambda = 0.2$ has the property of being the approximate "breaking point" at which the advantage of Heckman's TEM over alternatives becomes apparent.

3.6.2 Interpretation of Results

The primary purpose of the main simulation study was not to show that Heckman's TEM provides consistent estimates of $\boldsymbol{\beta}$ across various parameter values when correctly specified; this has already been established. Instead, this study was conducted to evaluate how severe the problem of endogenous medication use had to be in order for Heckman's TEM to provide gains that are scientifically relevant (as measured by both bias and mean squared error) over alternative approaches. The answer to this question was quite clearly that we see important advantages from Heckman's TEM even when dependence of medication use on the underlying biomarker and the treatment effect magnitude are somewhat small.

What we see in Table 3.1, as well as Figures 3.4 to 3.7 is that Heckman's TEM almost completely dominates the alternative approaches in terms of both bias and mean squared error across a range of λ and δ values. The Adjust method performs well when endogeneity is not present, and is sufficient. Since it estimates fewer parameters, it is not surprising that it performs slightly better than Heckman's TEM when endogeneity is low. In glancing at Table 3.1, we also note that most of the reduction in mean squared error comes from a bias reduction. While the variance of parameter estimates from Heckman's TEM are indeed slightly higher than those of the simpler approaches (Ignore, Adjust, and IPTW, in particular), the improvements in bias are substantial enough that we are willing to sacrifice this small amount of efficiency. This exemplifies the well know statistical phenomenon of the "bias-variance trade-off" and should be considered in model selection.

Importantly, we have also demonstrated how medication use is permitted to directly depend on the underlying biomarker. As the TEM is typically applied, this feature is generally not incorporated; however, allowing medication use to depend on the underlying biomarker other than through the error correlation is straightforward to implement: one simply places all covariates in the biomarker model into the medication use model. In Chapter 4, we will show that failure to account for the systematic components of the biomarker model in the medication use model can be problematic.

Econometricians tend to view these types of models in terms of structural equation modeling, particularly when solving the hybrid model equations via the two-stage approach. This may also be viewed as an instrumental variables approach because these exist variables associated with medication use but not the biomarker; however, since Heckman's TEM outperforms alternatives in the absence of an instrumental variable, it does not appear to suffer the challenges of classical "instrumental variables" approaches. We have demonstrated that the TEM does not require the existence of an IV in order to estimate the natural history association. Identification of the natural history association without the presence of an IV comes from the principal assumption and the distributional assumptions on the error terms. In similar models (e.g., the sample selection model), it has been shown that failure to include an IV in the medication use model results in an increase of variability of estimates, but not an introduction of bias (Marchenko et al., 2012). Our results confirmed this in the case of the TEM. Thus, when an IV is present, it should be accounted for in the medication use model. Additionally, we also require conditioning each outcome $y(0)$ and z^* on both $\boldsymbol{x}$ and $\boldsymbol{w}$ in order to estimate parameters. That is to say, in particular, that any variables in only $\mathbf{W}$ are presumed not to be associated with $y(0)$.

3.6.3 A Note on Missingness

When viewing Heckman's TEM through the single-stage likelihood framework, we can view it as a missing at random model with a simple likelihood-based imputation for the missing $y(0)$ in on-medication participants. The missing at random component comes from the fact that Heckman's model presumes that all predictors of medication use are known, measured, and included in the model. The imputation comes from correcting the observed biomarker for the effects of medication use with the likelihood-based δ (simultaneously estimated with the other parameters). In reality, we recognize that the true data contain no information about whether missingness is occurring at random or not at random. Hence, it is important to evaluate how sensitive Heckman's TEM is to misspecification of covariates which predict medication use (Chapter 4).

Framing Heckman's TEM as a likelihood-based imputation problem clearly motivates further study to evaluate the assumption of uniform treatment effects. This will be discussed in Chapter 4. In Chapter 5, we will further extend Heckman's TEM to allow effect modifiers in the model. Namely, we will expand the "imputation" component of the model and allow it to borrow information across subjects of nearby covariate values.

Additionally, framing Heckman's TEM in this fashion further helps distinguish it conceptually from inverse probability weighting approaches. Inverse probability weighting does not explicitly correct the off-medication biomarker. As we have seen, inverse probability weighting based on the exposures fails to change bias. In fact, these approaches were empirically observed to be less efficient in this setting.

Chapter 4

SENSITIVITY OF HECKMAN'S TREATMENT EFFECTS MODEL TO VIOLATIONS OF MODEL ASSUMPTIONS

In Chapter 2, we described why the assumptions of simple approaches are not reasonable when estimating biomarker associations in a cross-sectional, observational setting with endogenous medication use. In Chapter 3, we focused primarily on evaluating the advantages of Heckman's TEM over alternative approaches when correctly specified. A important consequence of modeling the underlying medication use model and treatment effect behavior is that we are forced to make assumptions that are challenging, if not impossible, to verify without external data or prior knowledge.

In this chapter, we focus on evaluating the sensitivity of Heckman's TEM to a variety of departures from its assumptions. We accomplish this by modifying, one by one, various aspects from the simulation study of Section 3.4 (in which Heckman's model was correctly specified) so that the assumptions of the TEM no longer hold. We evaluate how bias and variability are affected by these modifications. Since the Ignore approach and the CN method are very poorly behaved and inappropriate for the setting of nonrandom medication use, we do not give them a great deal of attention in this chapter. A researcher who is aware of the challenges of medication use is far more likely to turn to approaches that are more well known such as Excluding and Adjusting; we focus on comparing to these approaches.

This chapter is organized as follows. We consider first assumptions on the medication use model, including failure to specify important covariates, and departures from the underlying probit model. We then proceed to test sensitivity to non-normal er-

rors on the biomarker model, including right skewed and heavy-tailed error terms. We then investigate the impact of nondifferental measurement error on observed predictors. Finally, we consider departures from the assumption of uniform treatment effects by examining bias under proportionate treatment effects.

Certain forms of misspecification are absolute (e.g., omission of a variable from the medication use model). Other forms of misspecification have gradation; for instance, the existence of effect measure modification, or severity of probit misspecification. To that end, we approach many of these non-absolute categories of misspecification by examining behavior across a range of severity.

4.1 Failure to Specify Variables in the Medication Use Model

We have demonstrated in Chapter 3 that the TEM can perform well even when a true instrument does not exist, provided this variable is accounted for in the underlying biomarker model. However, it is of interest to understand how sensitive the TEM is when variables associated with medication use are omitted from the medication use model. Recall that for IPTW, both x_1 and x_2 appear in the logistic medication use model on the basis of the fact that they are associated with the biomarker (Lefebvre et al., 2008). In contrast, the medication use model in the TEM accommodates and demands variables associated with medication use regardless of their association with the biomarker; based on the data generating mechanism of our main simulation study, this suggests inclusion of x_1, x_2, and x_3 in the medication use model.

Suppose that we generate data according to the simulation setup of Section 3.4, but fail to include x_1 in the medication use model. We compare this modification for both IPTW and the TEM. Results are presented in Table 4.1. Note that this type of misspecification only applies to IPTW and the TEM among the models considered, as the other approaches do not model the characteristics of the treatment groups. Results for other approaches (Ignore, Exclude, and Adjust) are borrowed from Table 3.1 of

Section 3.4 for the purposes of comparison. The TEM shows sensitivity to this type of misspecification for estimation of β_1, showing substantially more bias and variability than when x_1 was properly included in the medication use model. This is unsurprising given that this is a problem of unmeasured confounding. In this case, the simpler approaches perform better in terms of both bias and rMSE. The TEM provides low-bias estimates of β_2 and δ, although with greater variability than when x_1 is modeled.

Similar patterns were observed when x_2 was not specified in the medication use model; the induced bias was not as severe. This is not surprising given that x_2 is only associated with z^* through λ in this setup. This is an important finding, since inclusion of x_2 (directly associated only with $y(0)$) is the distinguishing feature between the model that allows z^* to depend on $y(0)$ and the model that does not.

Table 4.1: Results from a simulation study comparing four approaches when x_1 is incorrectly omitted from medication use models. This type of misspecification only applies to IPTW and Heckman's TEM.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$				
Ignore†	-0.223	4.67	4.62	22.8
Exclude†	-0.085	7.13	7.08	11.1
Adjust†	-0.084	5.21	5.10	9.82
IPTW	-0.084	5.21	6.20	9.85
Heckman's TEM	-0.102	5.30	5.18	11.5
$\beta_2 = 1$				
Ignore†	-0.039	4.62	4.64	6.04
Exclude†	-0.015	6.28	6.34	6.46
Adjust†	-0.015	4.56	4.57	4.80
IPTW	-0.015	4.57	5.53	4.80
Heckman's TEM	-0.001	4.70	4.69	4.69
$\delta = 1$				
Adjust†	-1.62	10.3	10.2	162
IPTW	-0.622	10.3	12.4	162
Heckman's TEM	0.00	24.9	24.2	24.9

† - Results from Table 3.1, Section 3.4

Finally, suppose we fail to include x_3 in the medication use model in Heckman's TEM. This type of misspecification only applies to Heckman's TEM, as the naïve approaches do not model the characteristics of the treatment groups, and IPTW does not include variables that are only associated with medication use but not the biomarker. Results are presented in Table 4.2. Once again, we utilize estimates from Table 3.1 to compare results from the TEM to those of other approaches (Ignore, Exclude, and Adjust). It appears that bias in estimation of β_1 is not very substantial for this type of misspecification. However, the variability of all estimates from Heckman's TEM are much higher than when x_3 is not omitted. For estimation of β_1, the alternatives meaningfully outperform Heckman's TEM in terms of rMSE despite the fact that Heckman's TEM provides the lowest bias.

Table 4.2: Results from a simulation study comparing four approaches when x_3 is incorrectly omitted from the medication use model of Heckman's TEM. This type of misspecification only applies to Heckman's TEM.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$				
Ignore†	-0.223	4.67	4.62	22.8
Exclude†	-0.085	7.13	7.08	11.1
Adjust†	-0.084	5.21	5.10	9.82
IPTW†	-0.083	7.23	5.74	11.0
Heckman's TEM	-0.046	20.2	21.9	20.8
$\beta_2 = 1$				
Ignore†	-0.039	4.62	4.64	6.04
Exclude†	-0.015	6.28	6.34	6.46
Adjust†	-0.015	4.56	4.57	4.80
IPTW†	-0.016	5.00	5.59	5.25
Heckman's TEM	-0.009	5.74	6.45	5.81
$\delta = 1$				
Adjust†	-1.62	10.3	10.2	162
IPTW†	-1.62	12.4	11.1	162
Heckman's TEM	-0.21	88.2	94.5	90.6

† - Results from Table 3.1, Section 3.4

This confirms that instruments, when they exist, should be included in Heckman's TEM. The fact that the bias is lower may be a product of the fact that x_3 was generated as a normally distributed variable, and hence becomes absorbed into the error tern when not specified. In light of this realization, this simulation serves as further confirmation that true instrumental variables are not necessary for valid estimation of the natural history association. This study suggests that conditioning on variables known to be associated with medication use can greatly reduce variability for estimation of parameters of interest.

4.2 Misspecification of the Underlying Probit Model

As previously discussed, expressing the probit model in terms of a latent variable with a normally distributed error term is a convenient way to allow correlation between the biomarker and medication use errors, and in turn, derive a likelihood from the assumption of bivariate normality. The bivariate normal model is a convenient modeling assumption to obtain a tractable expression for the conditional distribution of $z_i^*|y_i(0)$. Since $y_i(0)$ and z_i^* are either completely or partially latent, it is not possible to estimate the underlying medication use curve based on observed data, and so the assumption of bivariate normality is not one that is directly testable. Hence, it is of interest to examine the behavior of Heckman's TEM when distributional assumptions are not met. First, we consider other link functions to describe the probability of medication use. We consider the logistic and complementary log-log (clog-log) links.

Specifically, we compare the TEM to the Exclude, Adjust, and IPTW approaches. Suppose first that the following logistic model is used to determine the probability of medication use for each observation, conditional on covariates:

$$P(z = 1|\boldsymbol{w}, y(0)) = \frac{\exp(x_1 + x_3 + 0.2y(0))}{1 + \exp(x_1 + x_3 + 0.2y(0))} \tag{4.1}$$

Then, suppose the clog-log link is used:

$$P(z = 1|\boldsymbol{w}, y(0)) \quad = \quad 1 - \exp(-\exp(x_1 + x_3 + 0.2y(0))) \tag{4.2}$$

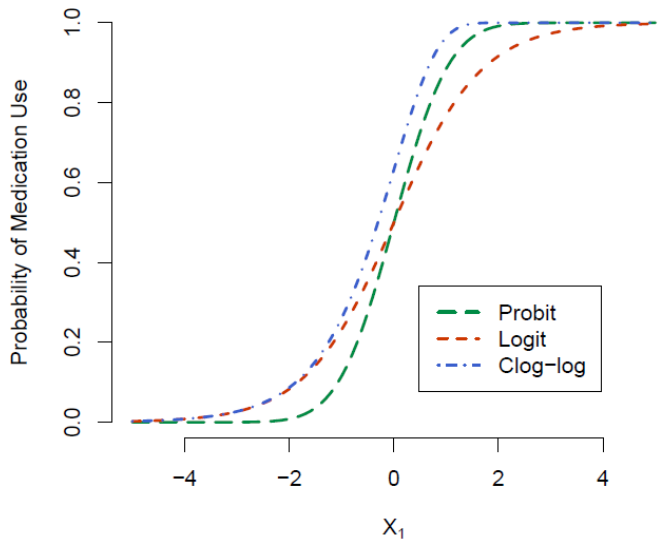

Figure 4.1: Link functions considered in Section 4.2, holding $x_2 = 0$ and $x_3 = 0$. For the logit model, the probability of medication use is given by $P(z = 1|x_1; x_2 = x_3 = 0) = \exp(1.2x_1)/(1 + \exp(1.2x_1))$, and for the clog-log link, the probability is given by $P(z = 1|x_1; x_2 = x_3 = 0) = 1 - \exp(-\exp(1.2x_1))$

Figure 4.1 illustrates how these three link functions differ from each other. There are some noted differences. For example, use of the logit link resembles the probit link in overall shape but has meaningfully heavier tails. The clog-log link maps to the reverse extreme-value distribution which is asymmetric with a heavier left tail.

Table 4.3 depicts results for when the logit link is used in the data generating mechanism in place of the probit link, and Table 4.4 depicts results for when the clog-log link is used. The patterns of bias appear to be quite similar to those observed under the probit model (i.e., correct specification of the TEM). Heckman's model outperforms the alternative approaches in terms of both bias and rMSE. The levels of bias seen in alternative approaches are mostly consistent with those of Table 3.1.

Table 4.3: Results from a simulation study comparing four approaches when the medication use model is based on a logit link rather than a probit link.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$				
Exclude	-0.068	6.90	6.95	9.65
Adjust	-0.067	4.97	5.01	8.38
IPTW	-0.068	6.58	5.62	9.48
Heckman's TEM	-0.001	7.14	6.90	7.14
$\beta_2 = 1$				
Exclude	-0.010	6.25	6.35	6.33
Adjust	-0.011	4.51	4.58	4.64
IPTW	-0.012	4.89	5.57	5.02
Heckman's TEM	0.000	4.61	4.66	4.61
$\delta = 1$				
Adjust	-1.67	9.93	10.0	167
IPTW	-1.67	11.4	11.1	167
Heckman's TEM	-0.007	26.0	25.0	26.0

Table 4.4: Results from a simulation study comparing four approaches when the medication use model is based on a complementary log-log link rather than a probit link.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$				
Exclude	-0.112	7.41	7.31	13.4
Adjust	-0.099	5.15	5.20	11.2
IPTW	-0.046	8.26	5.73	9.46
Heckman's TEM	0.001	6.66	6.70	6.66
$\beta_2 = 1$				
Exclude	-0.017	6.25	6.34	6.58
Adjust	-0.016	4.55	4.56	4.83
IPTW	-0.011	5.16	5.59	5.29
Heckman's TEM	0.001	4.63	4.64	4.63
$\delta = 1$				
Adjust	-1.58	10.4	10.4	159
IPTW	-1.66	14.5	11.1	167
Heckman's TEM	0.003	20.4	20.4	20.4

The results of these simulations suggest that Heckman's TEM is not overly sensitive to modest departures from the probit model. This result is important given that the assumption of bivariate normality/correct link function specification is not testable.

4.3 Right-Skewed Errors

In keeping with understanding the sensitivity of the TEM to departures from bivariate normality, we wish to understand behavior when the errors are right-skewed. Suppose that the errors for the biomarker model are sampled from an Exponential($\lambda = 3/5$), and the medication use errors are sampled from an Exponential($\lambda = 3$) distribution (both shifted to mean zero). Figure 4.2 illustrates how these distributions differ from the normal distributions of the same mean and variance. Results from this study are presented in Table 4.5. Patterns in bias and rMSE are similar to those observed in the main simulation; interestingly, the TEM provides biased estimates of the treatment effect, but the natural history association appears to be estimated with high accuracy. This suggests that Heckman's TEM is robust to right-skewed errors, as far as estimation of $\boldsymbol{\beta}$ is concerned.

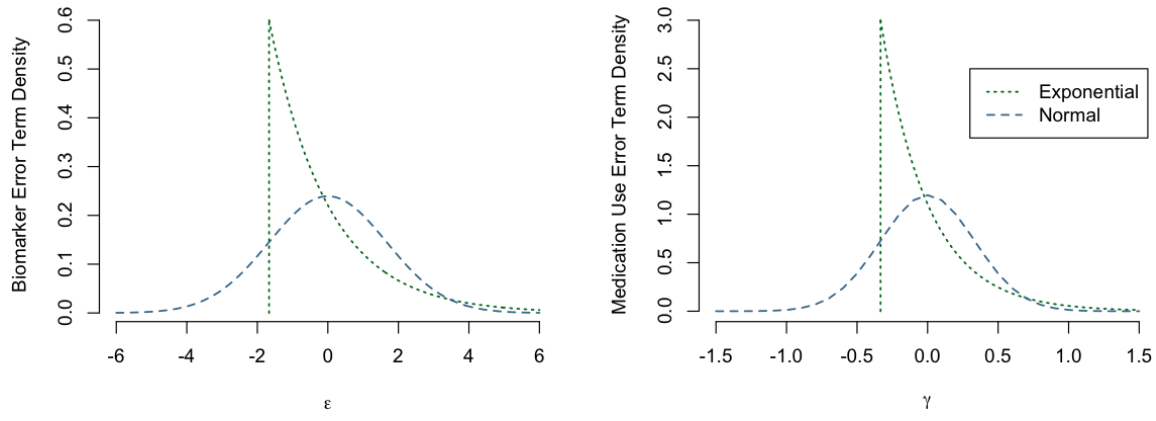

Figure 4.2: Density functions for a right-skewed shifted Exponential($\lambda = 3/5$) and Exponential($\lambda = 3$) distributions, centered to mean zero. The corresponding normal distributions of the same mean and variance is shown.

Table 4.5: Results from a simulation study comparing four approaches when the errors are right-skewed, generated from Exponential distributions and shifted to have mean zero (for the biomarker model, $\lambda = 3/5$; for the medication use model, $\lambda = 3$). The simulation setup otherwise mirrors that of Table 3.1.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$				
Exclude	-0.183	7.53	7.72	19.8
Adjust	-0.233	7.04	6.81	24.4
IPTW	-0.200	13.7	6.90	24.2
Heckman's TEM	0.000	7.47	7.18	7.48
$\beta_2 = 1$				
Exclude	-0.037	6.14	6.14	7.18
Adjust	-0.042	5.26	5.27	6.73
IPTW	-0.043	6.11	6.32	7.49
Heckman's TEM	-0.003	5.37	5.35	5.38
$\delta = 1$				
Adjust	-1.20	14.3	13.7	121
IPTW	-1.25	23.4	12.3	127
Heckman's TEM	-0.313	6.50	6.19	32.0

4.4 Heavy-Tailed Errors

Further exploring the sensitivity of the TEM to departures from bivariate normality, we wish to understand behavior when the errors on the underlying biomarker model are heavy-tailed. Suppose that the errors are generated from a bivariate t-distribution with parameters $\nu = 5$ and covariance parameter given by

$$\boldsymbol{\Sigma} = 2 \times \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}. \tag{4.3}$$

Results are shown in Table 4.6. As was the case in Section 4.3, the TEM proves low-bias estimates of $\boldsymbol{\beta}$, also outperforming other approaches in terms rMSE. However, there is noted bias in estimating the average treatment effect, δ (although the TEM

Table 4.6: Results from a simulation study comparing four approaches when the errors are heavy-tailed, generated from a bivariate t-distribution. The simulation setup otherwise mirrors that of Table 3.1.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$				
Exclude	-0.24	8.98	8.95	25.5
Adjust	-0.24	6.10	6.39	23.4
IPTW	-0.24	7.54	6.54	24.7
Heckman's TEM	0.024	10.4	10.5	10.6
$\beta_2 = 1$				
Exclude	-0.041	7.69	7.89	8.69
Adjust	-0.039	5.70	5.64	6.89
IPTW	-0.039	5.98	6.47	7.13
Heckman's TEM	0.0038	6.12	6.10	6.12
$\delta = 1$				
Adjust	-0.86	12.9	12.7	86.9
IPTW	-0.86	14.3	12.9	87.1
Heckman's TEM	-0.53	13.4	12.5	55.0

still outperforms the alternative approaches in bias and MSE). Sections 4.3 through 4.5 together suggest that the TEM is not particularly sensitive to moderate departures from the assumption of bivariate normally distributed error terms in the underlying biomarker and medication use models. This is particularly true for estimation of our parameter of interest, $\boldsymbol{\beta}$; importantly, bias in estimating δ can occur if distributional assumptions are violated.

4.5 Measurement Error Considerations

We now investigate the impact of measurement error in the exposures on bias and variability. Recall that, in the original simulation setup, x_1, x_2, and x_3 are each of unit variance. First, suppose that $(x_1, x_1^{\text{observed}})$ is generated from a bivariate normal distribution with zero mean and unit variance in each component, with correlation 0.95. Here x_1 denotes the true value of x_1 and x_1^{observed} is the observed value, which has some

modest (nondifferential) measurement error. Table 4.7 presents results on bias and rMSE for the naïve approaches, Ignore, Exclude, Adjust, as well as Heckman's TEM. Patterns mirror those observed Section 3.4. A modest amount of bias was induced by this minor measurement error on x_1; larger error (e.g., lower correlation between x_1 and x_1^{observed} in this example) resulted in further attenuation of estimates, consistent with prior results on nondifferential misclassification of the exposure (Lefebvre et al., 2008). Similar patterns were observed with measurement error on x_2.

Now suppose instead that x_3 is observed with the same type of nondifferential measurement error (the pair $(x_3, x_3^{\text{observed}})$ is generated from a bivariate normal distribution with zero mean and unit variance in each component, with correlation 0.95). Table 4.8 presents results on bias and rMSE for the same four approaches. No substantial changes in bias or variability were observed.

Table 4.7: Results from a simulation study comparing four approaches when there is modest measurement error on x_1.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$				
Ignore	-0.26	4.83	4.68	26.8
Exclude	-0.15	7.41	7.14	17.0
Adjust	-0.15	5.38	5.13	16.2
Heckman's TEM	-0.054	6.81	6.76	8.67
$\beta_2 = 1$				
Ignore	-0.038	4.74	4.68	6.06
Exclude	-0.018	6.47	6.45	6.72
Adjust	-0.018	4.70	4.64	5.03
Heckman's TEM	-0.001	4.82	4.74	4.82
$\delta = 1$				
Adjust	-1.52	10.6	10.3	152
Heckman's TEM	-0.014	22.7	22.9	22.7

Table 4.8: Results from a simulation study comparing four approaches when there is modest measurement error on x_3.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$				
Ignore	-0.22	4.60	4.61	22.8
Exclude	-0.086	7.15	7.08	11.2
Adjust	-0.084	5.15	5.09	9.84
Heckman's TEM	-0.001	7.17	6.99	7.17
$\beta_2 = 1$				
Ignore	-0.038	4.62	4.63	5.99
Exclude	-0.016	6.10	6.32	6.30
Adjust	-0.015	4.59	4.57	4.82
Heckman's TEM	-0.001	4.70	4.66	4.70
$\delta = 1$				
Adjust	-1.62	10.3	10.2	162
Heckman's TEM	-0.009	23.5	22.5	23.7

4.6 The Assumption of Uniform Treatment Effects: A Motivating Example

In Section 3.3, we allowed for a random treatment effect sampled from a normal distribution of unknown mean and variance rather than a constant unknown parameter. To account for heteroscedasticity induced, we presented a robust variance-covariance estimator. This generalization does not allow for treatment effects to systematically differ with covariates. Hence, it is of interest to examine the sensitivity of the TEM when effect measure modification is present.

One simple means of evaluating this sensitivity is to evaluate behavior when subgroup-specific treatment effects are taken to be proportionate to the expected underlying biomarker value. We conduct a simulation to elucidate this sensitivity. Suppose that we alter the simulation setup of Section 3.4 such that $\delta_i = \psi \mathbb{E}[y_i(0)|\boldsymbol{x}_i] = \psi \cdot \boldsymbol{x}_i^T \boldsymbol{\beta}$, and in turn, $y_i = y_i(0) - \psi z_i \cdot \boldsymbol{x}_i^T \boldsymbol{\beta}$. To ensure that the effects of medication are positive for virtually all subjects, set $\beta_0 = 10$; in turn, set $\sigma_y^2 = 9$ in order to increase the difference

between $\mathbb{E}[y(0)|z = 1]$ and $\mathbb{E}[y(0)|z = 0]$. Then, suppose that $\psi = 0.2$ so that the expected effect of medication use on the biomarker is approximately two units rather than one as in our main simulation of Section 3.4. Table 4.9 presents the results for Exclude, Adjust, IPTW, and Heckman's TEM. Heckman's TEM shows bias in estimating β_1 and β_2. In this setting, Heckman's TEM performs best in terms of rMSE for estimation of the biomarker associations, but the level of bias is still fairly substantial. Bias patterns were noted to increase as the proportionality constant ψ increased.

Table 4.9: Results from a simulation study comparing approaches when subject-specific treatment effects is proportionate to their expected underlying biomarker value.

	Bias	SE $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$			
Exclude	-0.35	14.4	38.3
Adjust	-0.45	10.3	46.4
IPTW	-0.45	11.1	46.6
Heckman's TEM	-0.10	13.9	17.1
$\beta_2 = 1$			
Exclude	-0.063	12.8	14.3
Adjust	-0.16	8.97	18.4
IPTW	-0.16	9.16	18.6
Heckman's TEM	-0.10	9.43	14.0
$\delta \approx 2$			
Adjust	-1.62	20.6	163.6
IPTW	-1.62	21.4	163.7
Heckman's TEM	-0.004	45.4	45.4

In practice, the effects of medication use on a biomarker may have both additive and proportionate characteristics. That is, medication may lower the biomarker by a fixed amount on average for all participants, and then further decrease the biomarker in a way that is proportionate the the biomarker. It is of interest to determine the behavior of these models as the we shift treatment effects from being more "uniform" in nature to being more "proportionate" in nature. In this simulation, let $\sigma_y^2 = 40$

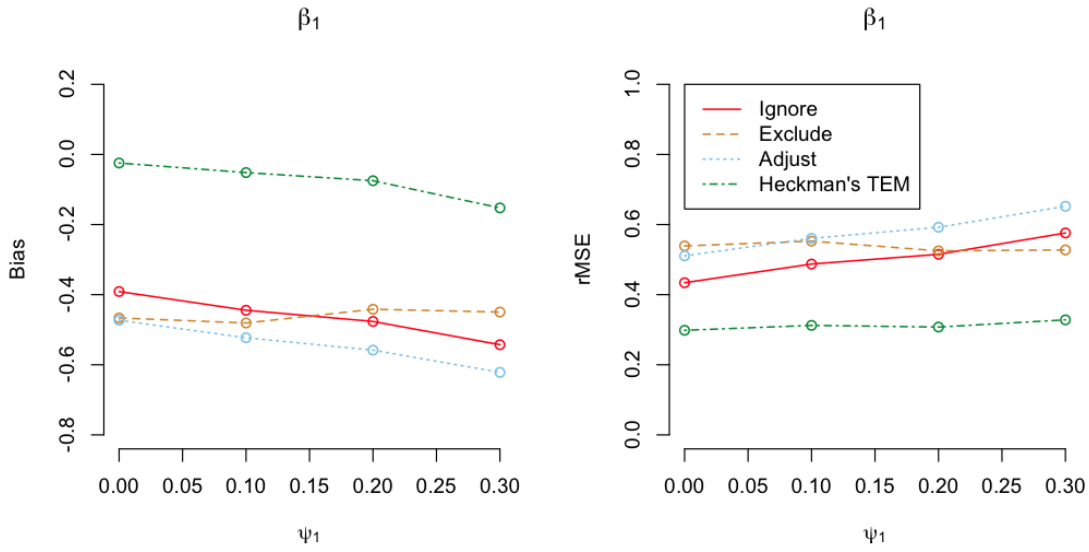

Figure 4.3: Bias and root mean squared error of for estimation of β_1 while varying the strength of proportionality of treatment effects.

and $\sigma_\gamma^2 = 10$ to increase the difference between $\mathbb{E}[y(0)|z = 1]$ and $\mathbb{E}[y(0)|z = 0]$ to approximately 4. Let $\delta_i = \psi_0 + \psi_1 \mathbb{E}[y_i(0)|\boldsymbol{x}_i]$, and we vary ψ_0 and ψ_1; we choose ψ_0 and ψ_1 so that the marginal treatment effect, $\mathbb{E}[\delta_i] = 3$. Specifically, we let ψ_0 range from 3 to 0, and $\psi_1 = 0.3 - 0.1\psi_0$ ranges linearly in ψ_0 from 0 to 0.3. $\psi_1 = 0$ indicates that the treatment effects are entirely additive, and $\psi_1 = 0.3$ indicates that the treatment effects are entirely proportionate the the expected underlying biomarker value. Figures 4.3-4.5 present the results for estimation of β_1, β_2, an δ from the Ignore, Exclude, Adjust approaches, and Heckman's TEM.

Is is apparent that Heckman's TEM performs very well compared to other approaches in terms of both bias and rMSE for estimation of β_1, although the bias is still moderately high when the treatment effects are entirely proportionate. For estimation of β_2, the level of observed bias is more substantial relative to that of the naïve approaches. The TEM generally outperforms the alternative approaches when compared to the other approaches in terms of rMSE, although the gap between the approaches

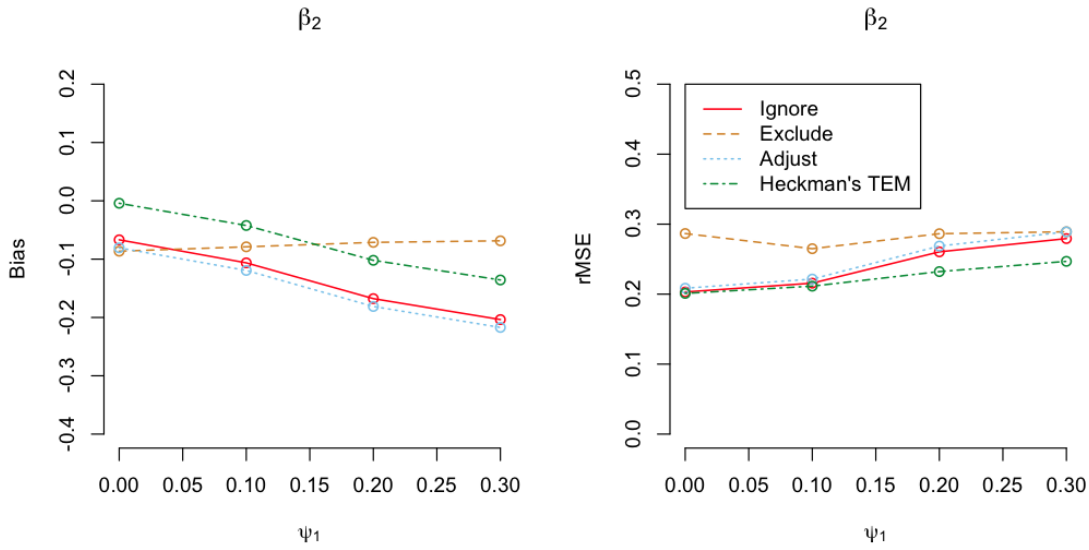

Figure 4.4: Bias and root mean squared error of for estimation of β_2 while varying the strength of proportionality of treatment effects.

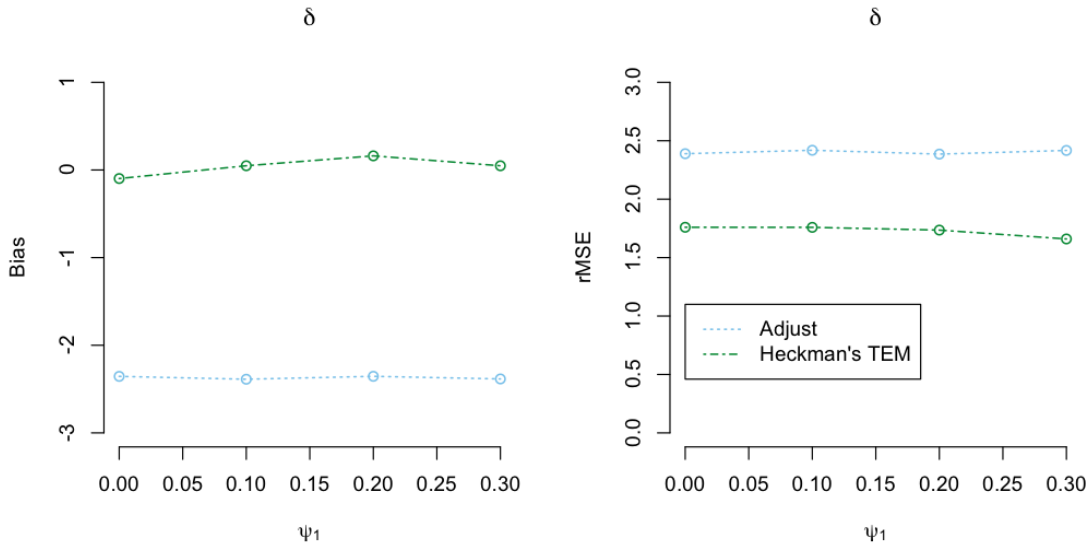

Figure 4.5: Bias and root mean squared error of for estimation of δ while varying the strength of proportionality of treatment effects.

appears to close as proportionality increases. For estimation of δ, the TEM outperforms the Adjust approach across levels of proportionality levels, with low bias and lower rMSE. This result is consistent with prior results (Table 3.1 and Table 4.9). This suggests that the presence of effect measure modification can have a substantial impact for estimating $\boldsymbol{\beta}$, although the marginal treatment effect, δ, may still be reasonably well estimated. This pattern appears to be the reverse of those seen in Section 4.3 and 4.4, which pertained to distributional assumptions on the errors.

4.7 Discussion

In this chapter, we have demonstrated that Heckman's TEM is fairly robust to departures from several of its main assumptions. In reality, if the errors terms in either the biomarker model or the medication use model are skewed or heavy-tailed, our estimates of the natural history association are not meaningfully affected. A similar result is true if the underlying treatment assignment model is not based on a probit link but rather from a logistic or complementary log-log link. This is to say that Heckman's TEM is fairly robust to departures from the assumption of bivariate normality. However, follow-up studies did reveal that bias can be induced if the biomarker errors become more heavily skewed, and challenges with convergence can occur when generating errors from a bivariate t-distribution with fewer than five degrees of freedom. Intuitively, the fact that there would be bias makes sense since the likelihood incorporates expressions for the the conditional mean and variance of $z_i^*|y_i(0)$, which under a skewed or heavy-tailed distribution is no longer guaranteed to be correct.

Although we have previously shown that Heckman's model performs well even when there is no true instrumental variable, they must be accounted for in the medication use model when they exist. Failure to specify important predictors of medication use can result in large increases in variability even if those variables are not associated with the underlying biomarker (that is, if they are instruments). This is an important

limitation of which it is important to be aware. Many times, predictors of medication use are either not well understood or they are not measured in the study. This makes it challenging to justify using Heckman's TEM in those settings. We note of course that challenges arising from failure to place predictors of interest (x_1 and x_2, in particular) in the medication use model can be addressed very easily by simply placing all variables in the biomarker model in the medication use model. The results of this chapter suggest that if z_i^* is systematically influenced by $y(0)$, making these adjustments can be very important to obtain consistent estimates. Since researchers may not necessarily think to collect data on variables associated with medication use only, misspecification of the x_3 class of variables is more concerning and less easy to address in practice. If predictors of medication use are thought to be well understood, Heckman's TEM should certainly be considered as a candidate for estimating biomarker associations.

While the Heckman model remains robust to certain assumptions, unverifiable misspecification of the medication use model can be difficult to overcome in settings where predictors of medication use are not well understood. However, if predictors of medication use are thought to be known and the effect of medication is to change the biomarker by a fixed amount on average, the Heckman model should be considered as a means of correcting bias. Special consideration of endogeneity and the associations of variables with medication use and biomarkers should inform model selection for estimating biomarker-to-exposure associations.

We also found that Heckman's TEM can be somewhat sensitive to departures from the assumption of uniform treatment effects. In our example simulation, effect modification was generated by allowing treatment effects to be proportional to the expected value of the biomarker conditional on covariates. Heckman's TEM does not accommodate this type of systematic dependence in the effects of medication use. Indeed, effect modification can arise not just from proportionality of treatment effects, but depen-

dence on other covariates separate from those in $\mathbf{X}$ or even those in $\mathbf{W}$. This type of effect modification will be explored in much greater depth in Chapter 5 when we extend Heckman's TEM to accommodate more general effect modifiers.

Chapters 3 and 4 together confirm the inappropriateness of the alternative approaches, as they are all sensitive to both the strength of endogeneity and the magnitude of treatment effects. Heckman's TEM still outperforms the alternative approaches even when there are modest assumption violations.

Chapter 5

EXTENSION TO ALLOW SUBGROUP SPECIFIC TREATMENT EFFECTS

Chapters 2 through 4 have primarily focused on identifying and evaluating a model to account for endogenous medication use when estimating the natural history association between a predictor of interest and a biomarker outcome. The TEM is appealing for this purpose since it provides us with an easy to implement model to characterize how medication users differ from non-users and allows medication use to be correlated with the underlying biomarker (that is, it incorporates endogeneity of medication use). In addition to demonstrating that several simple modifications to OLS linear regression are not adequate for removing bias when endogeneity is present, we have also evaluated the robustness of the TEM to departures from several of its main assumptions. We found that the gains from Heckman's TEM generally depend upon on the assumption of uniform treatment effects. In practice, we expect this assumption to be violated (for example, if a predictor of interest modifies the effect of medication use on the underlying biomarker). Differential efficacy may appear across race categories or genetic exposures, for example. Hence, there is an unmet need to address the assumption of uniform treatment effects and account for effect measure modification.

IV approaches have been presented and discussed for estimation of marginal treatment effects when there is heterogeneity is present (Heckman et al., 2006; Wang et al., 2015). Attention has not been given, however, to understanding how the existence of treatment effect modifiers impacts estimation of the natural history association, and no generalizations of the TEM have been made to accommodate this potential challenge.

This chapter aims to extend Heckman's TEM to allow for systematic differences in the expected treatment effect magnitude, based on observable covariates. We will first propose a model to explicitly accommodate effect modifiers and show identifiability of parameters. We will present a robust Wald-based procedure to test for the presence of effect modification across subsets of observable covariates. We will then follow up with a number of simulation studies in order to elucidate the advantages of accommodating effect modification when present.

5.1 Accounting for Covariate-Dependent Treatment Effects

As we have previously discussed, Heckman's TEM can be seen as a likelihood-based missing-at-random model for the partially missing outcome $y_i(0)$. If a predictor of interest is in fact an effect modifier, then using a single value δ as a correction for all study subjects serves as a systematic under- or over-correction to the observed value. In this case, we hypothesize that the resulting estimators of the natural history association from Heckman's TEM will not be consistent for the true value. Figure 5.1 depicts the updated corresponding DAG to allow for different classes of effect modifiers (those associated with the underlying biomarker, with the probability of medication use, with the effect of medication use, or any combination of these three).

Let $\boldsymbol{v}$ denote a subgroup-specific vector of covariates that predict the effect of medication use on the biomarker (allowing for an intercept), and let $\boldsymbol{\eta}$ denote an unknown parameter vector describing the effect modification. In turn, let $\mathbf{V}$ denote the $N\times(q_3+1)$ design matrix for covariates that influence the effect of medication use; these covariates may or may not be distinct from the covariates of $\mathbf{X}$ and $\mathbf{W}$. Figure 5.1 depicts the updated corresponding DAG to allow for effect modifiers. We may assume that

$$\boldsymbol{v}_i = (1 \quad \underbrace{0 \quad 0 \quad \cdots \quad 0}_{q_3}) \tag{5.1}$$

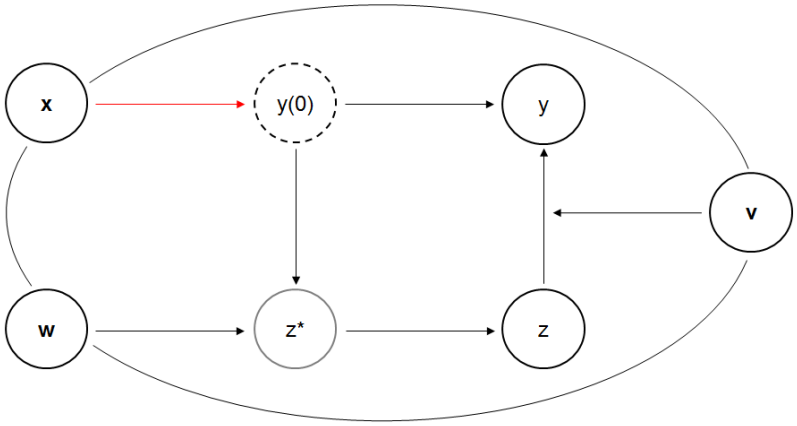

Figure 5.1: DAG illustrating relationship between covariates and outcomes when subgroup-specific effects are accommodated. Solid indicates observed variables, dashed indicates partially observed variables. Note that the covariates of $\boldsymbol{x}$, $\boldsymbol{w}$, and $\boldsymbol{v}$ need not be unique, indicated by the curves lines connecting them. The arrow between $\boldsymbol{x}$ and $y(0)$ (red) corresponds to the association of interest.

if $z_i = 0$, as $y_i(0)$ is observed in off-medication participants. This is done because some potential effect modifiers may only apply to on-medication participants (e.g., medication class or dose). Then, rather than modeling the expected treatment effect as a fixed, unknown scalar δ, we may model the expected treatment effect with a subject-specific δ_i given by $\delta_i = \boldsymbol{v}_i^T\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is an unknown $(q_3 \times 1)$-vector. By modeling $y_i(0) = y_i + \delta_i z_i$, we borrow information across covariate values for data to determine an appropriate correction to the observed biomarker at the subgroup-specific level. The likelihood model for Heckman's TEM is easily extended as follows:

$$\begin{aligned}\mathcal{L}_{\mathbf{V},\mathbf{W},\mathbf{X},\mathbf{Y},\mathbf{Z}}(\boldsymbol{\theta}) \;=\; & \prod_{i=1}^{N} \frac{1}{\sigma_y}\phi\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{v}_i^T\boldsymbol{\eta}z_i}{\sigma_y}\right) \\ & \times \Phi\left((-1)^{1-z_i}\frac{\boldsymbol{w}_i^T\boldsymbol{\alpha} + \rho(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{v}_i^T\boldsymbol{\eta}z_i)/\sigma_y}{\sqrt{1-\rho^2}}\right)\end{aligned} \tag{5.2}$$

The result that $y_i(0)$ is still permitted to influence z_i^* in the data generation mechanism still holds in this setting, and the result that λ need not be estimated also still holds.

A robust variance-covariance estimator analogous to the one described in Section 3.3 can be applied to account for random error terms in the treatment effect. We refer to this extension as the "subgroup-specific effects model" (SSEM).

5.2 Identifiability of Parameters of Interest

We prove that parameters of interest are identifiable. Let $\mathbf{\Sigma}$ denote the covariance matrix of the error terms.

Theorem: Suppose we have structural equations given by (1) $y = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{v}^T\boldsymbol{\eta}z_i + \epsilon_i$ and (2) $z^* = \boldsymbol{w}^T\boldsymbol{\alpha} + \gamma_i$, again including all covariates of $\mathbf{W}$ in $\mathbf{X}$. Then, $\boldsymbol{\beta}, \boldsymbol{\eta}$, and σ_y are identifiable parameters.

Proof: Note that by including all covariates of $\mathbf{X}$ in $\mathbf{W}$, we are accommodating dependence of z^* on $y(0)$, without having to estimate the additional term, λ. Let $\mathbf{A}$ denote a 2×2 nonsingular matrix. Let $\tilde{\boldsymbol{x}}$ denote a vector of length p containing all unique predictors in $\boldsymbol{x}$, $\boldsymbol{w}$, and $\boldsymbol{v}$ combined (each having q_1, q_2, and q_3 predictors, respectively). Partition this complete covariate vector into seven classes, based on which outcomes or combination of outcomes they predict (the underlying biomarker, the probability of medication use, or the expected treatment effect magnitude); perform this partition as in Figure 5.2. The parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\eta}$ can be partitioned analogously. Let $\boldsymbol{o} = (y, z^*)$ denote the partially observed outcome vector, and $\tilde{\boldsymbol{x}}' = (1, \tilde{\boldsymbol{x}}^T, \boldsymbol{v}^T \times z)^T$ the covariate vector for the first structural equation. Thus, the simultaneous equation system may be written as $\mathbf{A}\boldsymbol{o} - \mathbf{\Gamma}\tilde{\boldsymbol{x}}' = (\epsilon, \gamma)^T$ for a $2 \times (p + q_3 + 1)$ dimensional matrix of coefficients, $\mathbf{\Gamma}$, given in this case by:

$$\mathbf{\Gamma} = \begin{bmatrix} \beta_0 & \boldsymbol{\beta}_{(1)}^T & \boldsymbol{\beta}_{(2)}^T & \mathbf{0}^T & \boldsymbol{\beta}_{(3)}^T & \mathbf{0}^T & \mathbf{0}^T & \boldsymbol{\beta}_{(4)}^T & \boldsymbol{\eta}^T \\ \alpha_0 & \boldsymbol{\alpha}_{(1)}^T & \boldsymbol{\alpha}_{(2)}^T & \boldsymbol{\alpha}_{(3)}^T & \boldsymbol{\alpha}_{(4)}^T & \boldsymbol{\alpha}_{(5)}^T & \mathbf{0}^T & \boldsymbol{\alpha}_{(6)}^T & \mathbf{0}^T \end{bmatrix}. \tag{5.3}$$

The system can then be simplified to $\boldsymbol{o} = \boldsymbol{\Pi}\tilde{\boldsymbol{x}}' + (\epsilon', \gamma')^T$, where $\boldsymbol{\Pi} = \mathbf{A}^{-1}\boldsymbol{\Gamma}$, and the error vector has covariance matrix $\boldsymbol{\Omega} = \mathbf{A}^{-1}\boldsymbol{\Sigma}\mathbf{A}^{-T}$. Again letting

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 \\ 0 & 1/\sigma_z \end{bmatrix}, \tag{5.4}$$

where σ_z^2 is the total variance of z^*, we have that parameter appearing in the matrices $\boldsymbol{\Lambda\Pi}$ and $\boldsymbol{\Lambda\Omega\Lambda}^T$ are identifiable (Maddala, 1983). The result that $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and σ_y are identifiable follows from setting $\mathbf{A}$ to be the 2×2 identity matrix. Q.E.D.

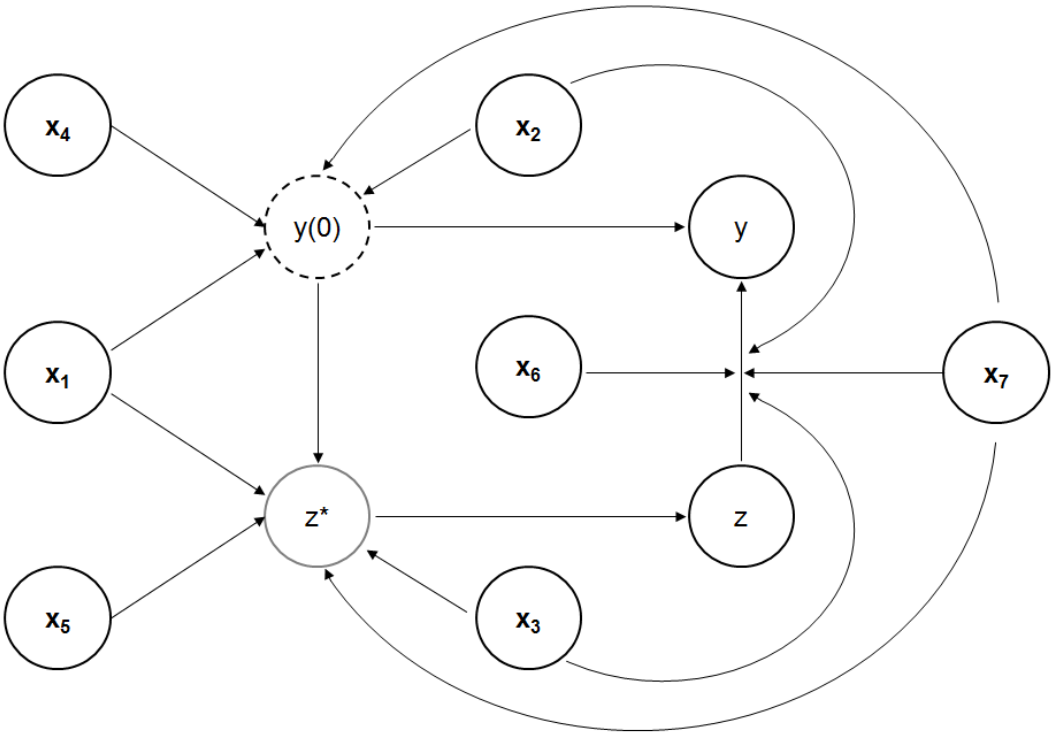

Figure 5.2: Partitioning of exposures into seven classes based on which of the three major outcomes they predict: (1) the underlying biomarker $y(0)$, (2) the latent medication use variable z^*, and (3) the treatment effect magnitude δ_i.

5.3 A Robust Wald-Based Procedure to Test for Effect Modification

In practice, one might be unsure of whether or not certain covariates serve as potential effect modifiers. As a tool to test for evidence of effect modification, a robust Wald test can be implemented to test the null hypothesis $H_0 : \boldsymbol{\eta} = (\eta_0, 0, \ldots, 0)^T$ against $H_1 : (\text{not } H_0)$, for any free-varying real-valued η_0. Suppose the parameter vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\eta}$ are of dimensions $(q_1 + 1)$, $(q_2 + 1)$, and $(q_3 + 1)$, respectively. Define the matrix $\mathbf{R}$ to identify the effect modifier parameters of interest:

$$\mathbf{R} = \left[\mathbf{0}_{q_3 \times (q_1+q_2+3)} \quad \mathbf{I}_{q_3 \times q_3} \quad \mathbf{0}_{q_3 \times 2} \right], \tag{5.5}$$

where the $\mathbf{0}_{q_3 \times 2}$ matrix in $\mathbf{R}$ corresponds to the correlation parameter ρ and the error variance σ_y^2. If $\hat{\boldsymbol{\theta}}$ is the solution to the score equations $\partial \log \mathcal{L}(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$ (with likelihood as in 5.2), and $\mathbf{S}_N(\hat{\boldsymbol{\theta}})$ the robust variance estimator evaluated at $\hat{\boldsymbol{\theta}}$, then a Wald statistic, $W_N(\hat{\boldsymbol{\theta}})$, can be defined by

$$W_N(\hat{\boldsymbol{\theta}}) \;\; = \;\; (\mathbf{R}\hat{\boldsymbol{\theta}})^T (\mathbf{R}\mathbf{S}_N(\hat{\boldsymbol{\theta}})\mathbf{R}^T)^{-1} (\mathbf{R}\hat{\boldsymbol{\theta}}). \tag{5.6}$$

Under standard likelihood theory (Wald, 1943), we have that $W_N \longrightarrow_d \chi^2_{q_3}$, so that $W_N(\hat{\boldsymbol{\theta}})$ can be compared to the $(1 - \alpha)$-quantile of the $\chi^2_{q_3}$ in order to test for the presence of effect modification.

One may also utilize a similar robust Wald test in order to test for effect modification among a subset of the effect modification parameters. For example, if medication class and race category were treated as potential effect modifiers, we could test for effect modification within either one of those variables, or for both jointly. By using the robust variance estimator $\mathbf{S}_N(\hat{\boldsymbol{\theta}})$, this testing procedure will be robust to various forms of model misspecification, including error heteroscedasticity, which is likely to be encountered in practice.

5.4 Simulation Scenario 1: Bias Reduction Under Various Forms of Effect Modification

In the presence of effect modification, the TEM is not correctly specified and may not provide consistent estimates for $\boldsymbol{\beta}$. Analytically computing the finite-sample or asymptotic bias of the TEM in the presence of effect modification is not tractable, since the solutions to the score equations of interest do not possess a closed-form expression and must be obtained through computational approximation. The relative advantages of accommodating subgroup-specific effects can be more effectively elucidated by means of simulation studies over a wide range of reasonable parameters. We choose a different simulation setup in this chapter than our main simulation of Section 3.4, in order to more effectively parameterize effect modification.

In this simulation setup, let $N = 5000$ participants, and x_1, x_2, x_3 denote the set of predictors, all distributed i.i.d. $\mathcal{N}(0, 1)$. Further let $D \sim \text{Bernoulli}(p = 0.5)$. We may choose to think of D as a binary medication "dose" variable that predicts the magnitude of the treatment effect, but is not a predictor of interest (in this example, D would be a covariate specific to on-medication participants, but D could also be a variable measurable in all participants). Suppose the outcomes are generated by:

$$\begin{aligned} y_i(0) &= 50 + x_{1i} + x_{2i} + \epsilon_i \\ z_i^* &= -5 + x_{1i} + x_{3i} + 0.1y_i(0) + \gamma_i \\ &= 1.1x_{i1} + 0.1x_{2i} + x_{3i} + 0.1\epsilon_i + \gamma_i \end{aligned} \tag{5.7}$$

We set $\sigma_y = \sigma_\gamma = 50$, and $\rho = 0.5$. Thus, the covariance matrix for the total error terms in the model is given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} 50 & 30 \\ 30 & 50.5 \end{bmatrix}. \tag{5.8}$$

Relating to the notation used in Section 5.1, we have that $\boldsymbol{x}_i = (1, x_{1i}, x_{2i})^T$ and $\boldsymbol{w}_i = (1, x_{1i}, x_{2i}, x_{3i})^T$, so that $\boldsymbol{\beta} = (50, 1, 1)^T$ and $\boldsymbol{\alpha} = (0, 1.1, 0.1, 1)^T$. This setup yields an approximate 50% prevalence of medication use at each replication, and additionally, we have that $\mathbb{E}[y(0)|z_i = 1] - \mathbb{E}[y(0)|z_i = 0] \approx 6.5$ under this setup. Thus, we select the parameters in the simulation studies that follow in this chapter such that the expected marginal treatment effect (that is, $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}_i^T\boldsymbol{\eta}]$) is about 3.75, placing all simulations on a comparable scale in which medication use is modestly effective in reducing the biomarker value for participants on medication. Based on this setup, the most general form of the data generating mechanism for y can be described by:

$$y_i \quad = \quad y_i(0) - (\eta_0 + \eta_1 x_{1i} + \eta_2 x_{2i} + \eta_3 x_{3i} + \eta_4 D_i) z_i,$$

such that $\boldsymbol{v}_i = (1, x_{1i}, x_{2i}, x_{3i}, D_i)^T$. We wish to choose parameters to evaluate (a) the extent of bias that arises from using the TEM when subgroup-specific effects are present, (b) the circumstances under which the SSEM reduces bias, and (c) the efficiency cost of using of the SSEM. These results aid us in providing recommendations for when to use the updated SSEM over the original TEM. We also compare the TEM and SSEM to the simple approaches (Ignore and Adjust) in order to confirm their inappropriateness when seeking to estimate the association between $\boldsymbol{x}$ and $y(0)$, and to evaluate the extent to which TEM can still reduce bias as compared to these approaches when effect modifiers are present.

In the first scenario, we let $\boldsymbol{\eta} = (2.5, 0.5, 0.5, 0.5, 2.5)^T$. Under this setup, the effect of medication use on the biomarker varies with all predictors, as well as the dose variable, D. Conditional on $D_i = 0$, the distribution of the effects is $\mathcal{N}(2.5, 0.75)$, and conditional on $D_i = 1$, the distribution of the treatment effects is $\mathcal{N}(5, 0.75)$. The subject-specific treatment effect therefore lies between 1 and 6.5 for the vast majority of participants (and is 3.75 on average, marginally). We conduct two-thousand simulation replicates

under this setup, and compare the SSEM to the TEM. We also fit the Ignore and Adjust models to gain insights into how much of the "gain" seen from the TEM is lost when effect modification is present. Again, we do not present results on estimation of the intercept, since it is not typically of interest in association studies.

Table 5.1 presents results for estimation of $\boldsymbol{\beta}$. Specifically, we present the estimated bias (averaged across all replicates), the Monte-Carlo standard error (the standard deviation of all the estimates from each replicate), the robust standard error estimates (averaged across all replicates), and the simulated rMSE. The SSEM provides low-bias estimates of β_1 and β_2 as compared to the other approaches. In particular, estimates obtained from the Ignore and Adjust approaches are markedly biased, even under this setting in which the effect of the medication on $y(0)$ is not very large relative to the difference in mean $y(0)$ between participants on and off medication. As one might hope, the TEM appears to provide some bias reduction for estimation of β_1 relative to the Ignore and Adjust approaches under this simulation setup, suggesting that some bias correction can be achieved by using the TEM even when effect modification is present. Of note is that the bias reduction is not as strong for estimation of β_2. Likely, this is a consequence of the fact that x_1 is a strong predictor of medication use, and x_2 weakly predicts medication use (only through the underlying biomarker). Comparing the bias of the TEM to the SSEM confirms that the TEM can be sensitive to departures from the assumption of uniform treatment effects.

Estimating the additional effect modifier parameters results in a modest loss of efficiency for the SSEM. However, the potential for bias reduction in this setting is objectively large enough to justify the efficiency loss, as seen by comparing the rMSE across the methods. The improved rMSE was confirmed to be higher in follow-up simulations in which the overall expected treatment effect was higher than 3.75. The robust standard error estimates adequately estimate the simulation-based standard errors.

Table 5.1: Results from a simulation study in which effect modification arises from a variety of sources. This table specifically presents the results for estimation of the natural history association, $\boldsymbol{\beta}$.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\beta_1 = 1$				
Ignore	-0.46	9.47	9.29	47.3
Adjust	-0.61	9.31	9.16	62.0
Heckman's TEM	-0.18	11.3	11.3	21.6
SSEM	-0.008	15.2	15.8	15.2
$\beta_2 = 1$				
Ignore	-0.27	9.19	9.27	28.6
Adjust	-0.28	9.00	9.09	29.8
Heckman's TEM	-0.25	10.5	10.5	26.8
SSEM	0.00	13.6	13.5	13.6

Table 5.2: Results from a simulation study in which effect modification arises from a variety of sources. This table specifically presents the results for estimation of the treatment effect δ or the effect modification parameters given in $\boldsymbol{\eta}$ for the SSEM.

	Bias	SE $\times 10^2$	$\widehat{\text{SE}}$ $\times 10^2$	rMSE $\times 10^2$
$\delta = 3.75$				
Adjust	-1.17	18.3	18.3	117.9
Heckman's TEM	1.10	72.3	72.3	132.2
$\eta_0 = 2.5$				
SSEM	-0.006	125.2	132.0	126.2
$\eta_1 = 0.5$				
SSEM	0.00	18.1	18.0	18.1
$\eta_2 = 0.5$				
SSEM	0.007	17.9	17.8	17.9
$\eta_3 = 0.5$				
SSEM	0.007	14.8	15.1	14.8
$\eta_4 = 2.5$				
SSEM	0.004	25.2	25.2	25.2

Table 5.2 presents results for estimation of the treatment effect parameters. The effect modifier parameters are neither estimable in the Adjustment approach nor the

TEM. To estimate the bias for these approaches, we compare the estimates to the marginal treatment effect, given by $\delta = 3.75$. The Adjustment approach provides a downwardly biased estimate of the marginal treatment effect; this is consistent with results found in prior simulations (e.g., in Section 3.4). Interestingly, the TEM provides an upwardly biased estimate of the marginal treatment effect, and has an incredibly high variability associated with estimation. This demonstrates that using the TEM to estimate marginal treatment effects may not be appropriate when there is effect modification. The SSEM provides approximately unbiased estimates of all of the effect modification parameters, although of note is that the intercept is of somewhat high variability. The following scenarios focus on subgroup-specific effects generated from each predictor individually, rather than from all predictors simultaneously as in this example. We divide this into four scenarios, as illustrated in Figure 5.3.

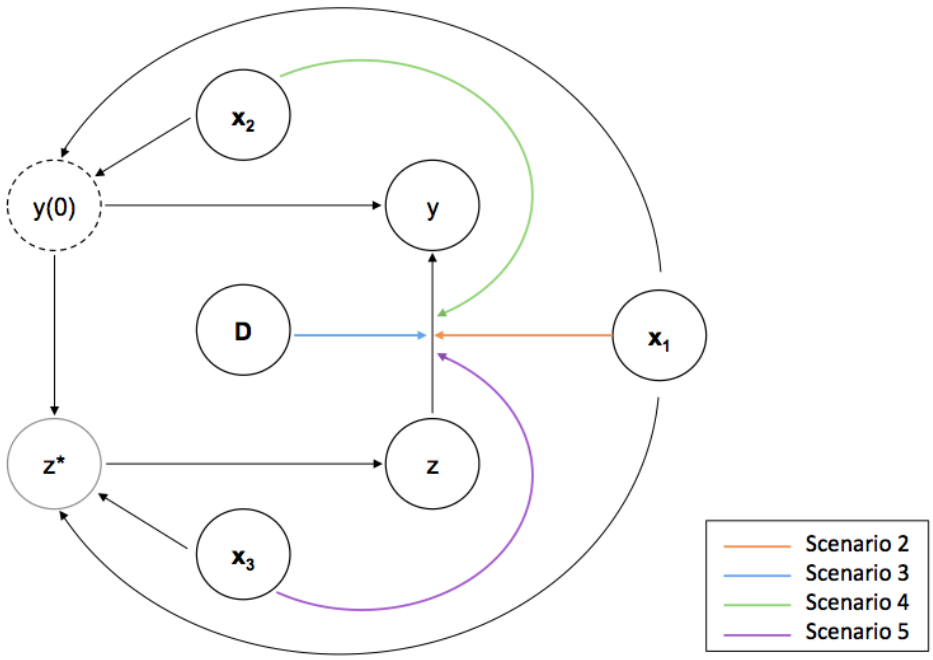

Figure 5.3: Diagram illustrating various setups for simulations seeking to evaluate the SSEM. Results from Scenario 1 were presented in Section 5.4 and incorporated all such effect modification. In the studies that follow, we consider one effect modifier at a time.

5.5 Simulation Scenarios 2 and 3: Predictors of Interest as Treatment Effect Modifiers

The purpose of this study is to evaluate bias when the single source of effect modification is a predictor of interest. Recall that x_1 is associated with $y(0)$ and z^*, and that x_2 is associated with $y(0)$ (through λ, x_2 is also weakly associated with z^*). In Scenario 2, we maintain the setup of Scenario 1, except we let $\boldsymbol{\eta} = (3.75, \eta_1, 0, 0, 0)$, and vary η_1 from 0 to 1 (with one-thousand simulation replicates at each value). For the SSEM, we do not estimate η_2, η_3, or η_4. Increasing η_1 does not alter the average effect of the medication on the biomarker (3.75).

Figure 5.4 illustrates the estimated bias and simulation rMSE for estimation of β_1 from each method, considered across the range $0 \leq \eta_1 \leq 1$. Results for estimation of β_2 are shown in Figure 5.5. For estimation of β_1, the SSEM provides approximately unbiased estimates, as expected. The TEM, on the other hand, provides approximately unbiased estimates only when $\eta_1 \approx 0$. Similar to the results from the first scenario, we find that the TEM has better performance, both in terms of bias and rMSE, than the naïve approaches. Also consistent with the previous results is that the rMSE for the TEM is slightly lower than that of the SSEM when η_1 is close to zero. However, η_1 does not have to be very large for the advantage to vanish.

We find that the TEM and SSEM provide low-bias estimates of β_2, although the Ignore and Adjust approaches provide estimates of β_2 with markedly more bias. The bias for the TEM and SSEM are nearly indistinguishable in the left panel of Figure 5.5, likely because x_2 is not associated with the effects of the medication. Because x_2 is not associated with the effects of medication, it is not very surprising that the bias does not seem to vary over the choice of β_2 for any of the approaches. Interestingly, the rMSE for estimation of β_2 in this case is higher for the SSEM than those of the other approaches. In a follow-up simulation, we generated the data such that x_1 and x_2

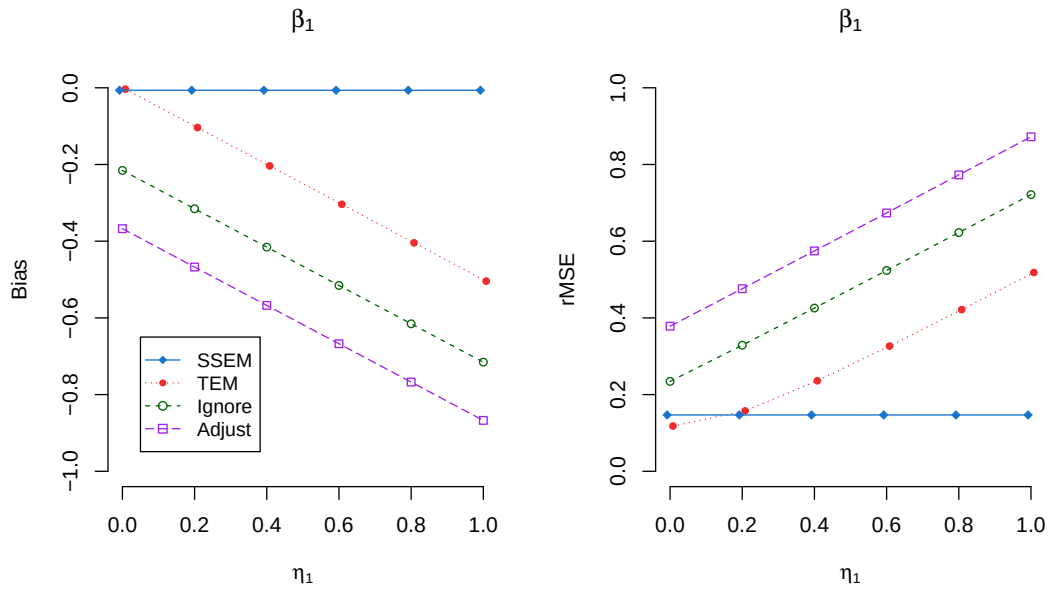

Figure 5.4: Results from simulation Scenario 2. The range of values considered for η_1 is shown on the x-axis, and the estimated bias (left) and rMSE (right) are shown. Note that the true parameter value is given by $\beta_1 = 1$.

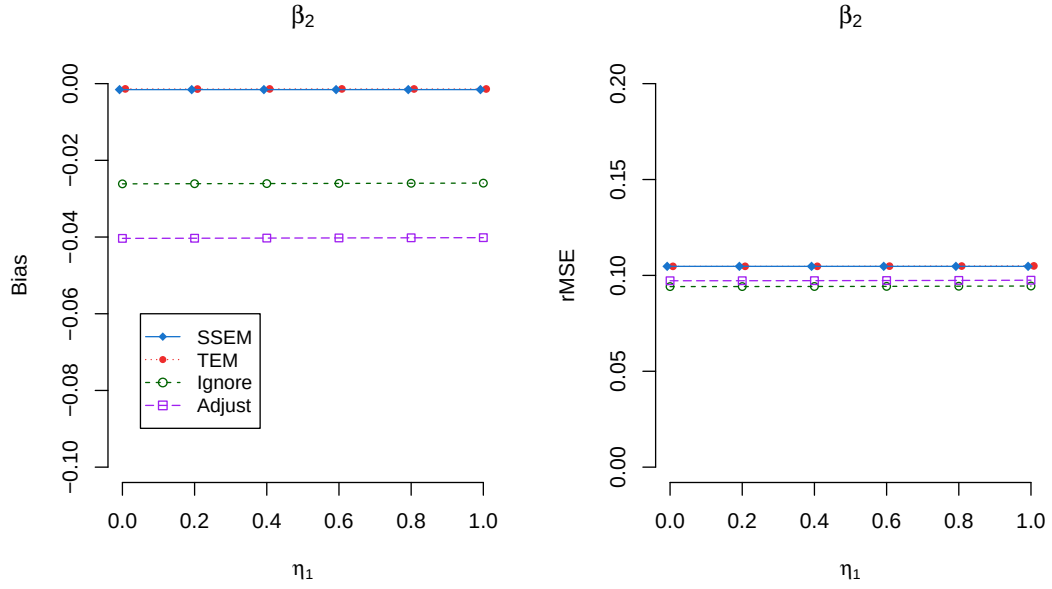

Figure 5.5: Results from simulation Scenario 2. The range of values considered for η_1 is shown on the x-axis, and the estimated bias (left) and rMSE (right) are shown. Note that the true parameter value is given by $\beta_2 = 1$.

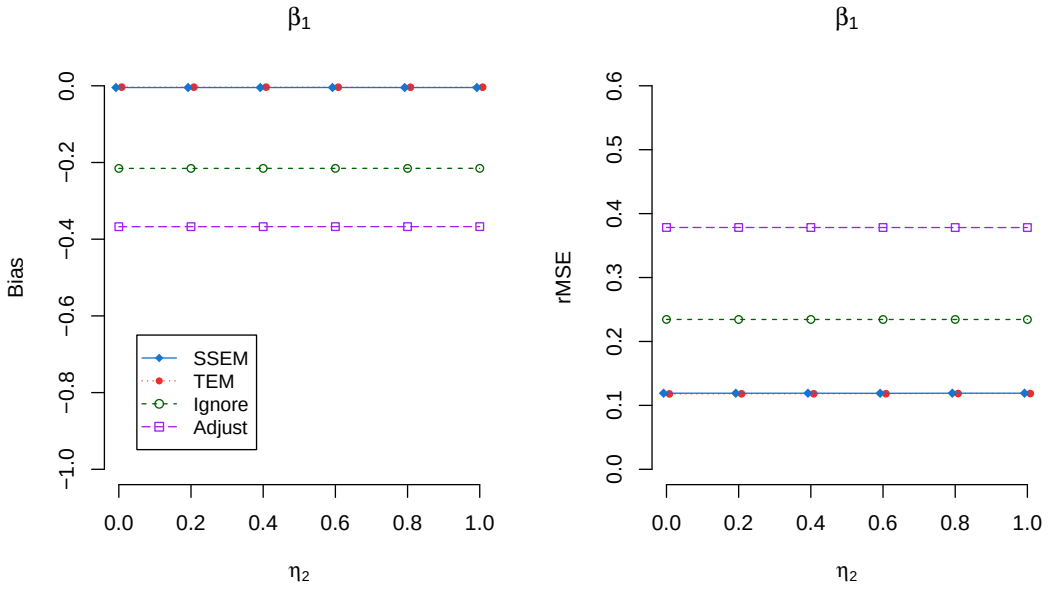

Figure 5.6: Results from simulation Scenario 3. The range of values considered for η_2 is shown on the x-axis, and the estimated bias (left) and rMSE (right) are shown. Note that the true parameter value is given by $\beta_1 = 1$.

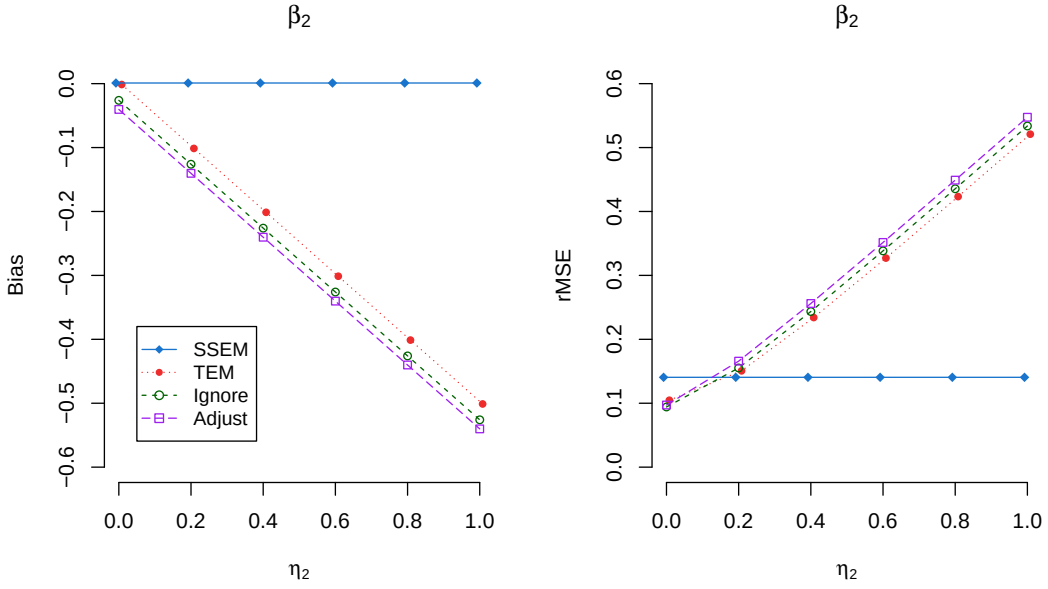

Figure 5.7: Results from simulation Scenario 3. The range of values considered for η_2 is shown on the x-axis, and the estimated bias (left) and rMSE (right) are shown. Note that the true parameter value is given by $\beta_2 = 1$.

were jointly distributed as bivariate normal, each predictor of unit variance, and with correlation 0.5. The patterns seen in Figures 5.4 and 5.5 were observed to be similar, suggesting that the different findings for β_1 and β_2 are not simply attributable to lack of correlation between the predictors.

In Scenario 3, we consider x_2 to be the sole effect modifier (i.e., we let $\boldsymbol{\eta} = (3.75, 0, \eta_2, 0, 0)^T$ and let $0 \leq \eta_2 \leq 1$). The reverse pattern was observed for β_1 and β_2; results are depicted in Figures 5.6 and 5.7. Interestingly, the TEM does not provide as much of an improvement in bias for estimating β_2 over the naïve OLS approaches as we saw in Scenario 2. This may be attributable to the fact that x_2 is only weakly associated with medication use. These two scenarios taken together suggest that if a predictor of interest is also a treatment effect modifier, it should be accounted for in the SSEM.

5.6 Simulation Scenario 4: Predictors of Medication Use as Treatment Effect Modifiers

The purpose of this simulation study is to evaluate bias under a setting in which the source of effect modification is a single factor associated with medication use only (in this case, x_3). In particular, x_3 is not associated with the biomarker. In this scenario, we let $\boldsymbol{\eta} = (3.75, 0, 0, \eta_3, 0)$, and we vary η_3 over the range $0 \leq \eta_3 \leq 1$ (with one-thousand simulation replicates at each value considered). For the SSEM, we do not estimate η_1, η_2, or η_4. As with η_1 and η_2, increasing η_3 serves only to increase the overall (marginal) variability of the effects of medication use.

Figures 5.8 and 5.9 illustrate the simulated bias and rMSE from each method for estimation of β_1 and β_2 across the range of values for η_3. The SSEM provides low-bias estimates of β_1 and β_2. The rMSE for estimation of β_1 appears to be nearly constant across η_1 values, with the TEM slightly outperforming the SSEM for small values of η_3. The TEM shows bias for estimation of β_1 (less so for β_2). The Ignore and Adjust

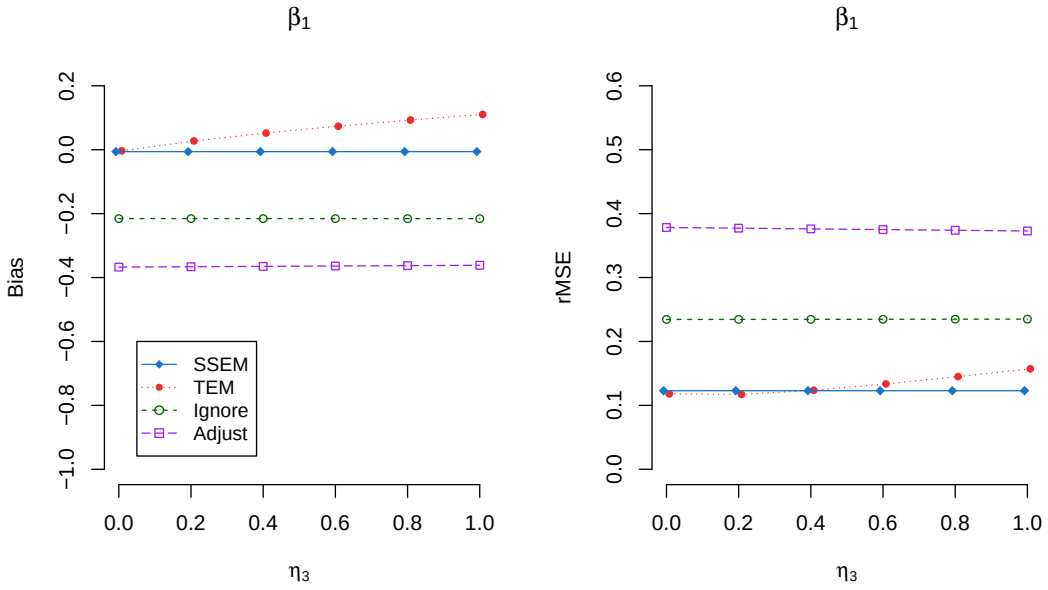

Figure 5.8: Results from simulation Scenario 4. The range of values considered for η_3 is shown on the x-axis, and the estimated bias (left) and rMSE (right) are shown. Note that the true parameter value is given by $\beta_1 = 1$.

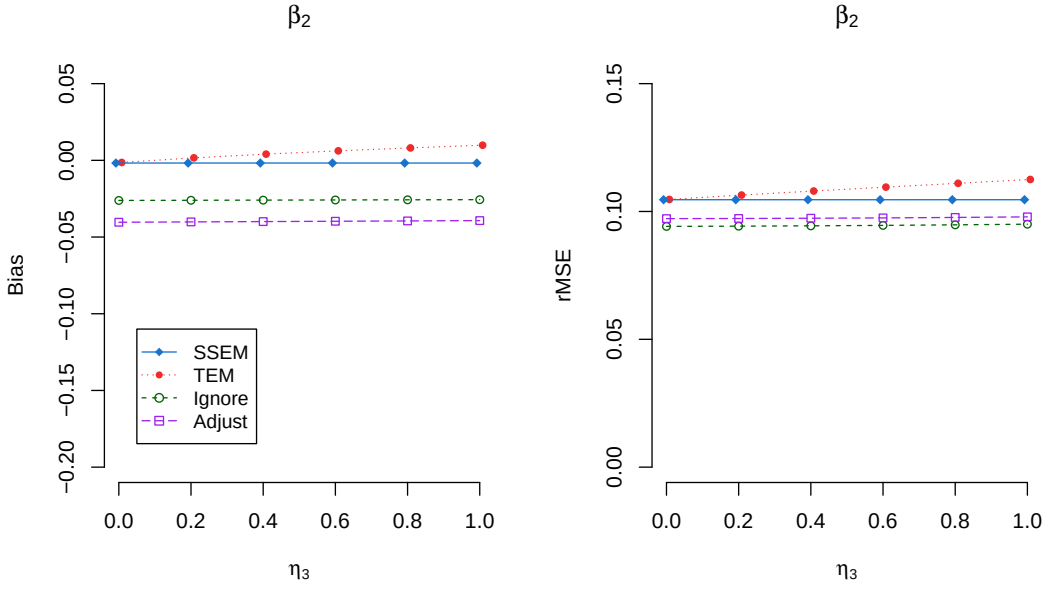

Figure 5.9: Results from simulation Scenario 4. The range of values considered for η_3 is shown on the x-axis, and the estimated bias (left) and rMSE (right) are shown. Note that the true parameter value is given by $\beta_2 = 1$.

methods perform poorly compared to the other approaches. However, with increasing η_1, the TEM eventually yields higher bias than the naïve approaches.

The two determinants of an individual subject's expected treatment effect magnitude are $\boldsymbol{v}_i^T\boldsymbol{\eta}$ and $z_i = \mathbf{1}(z_i^* > 0)$. Since x_3 is associated with both, failure to include x_3 as an effect modifier results in systematic under- or over-correction of the observed outcome (as in the TEM). This is an important result, as it suggests that we should not only account for effect modifiers when they are predictors of interest, but also if they are predictors of medication use only. In either case, bias arises from systematic under- or over-correction. Indeed, the bias from the TEM is smaller in the latter scenario because x_3. A likely explanation for this difference is that some of the bias observed in Scenarios 2 and 3 is attributable to misspecification from this Scenario: x_3 is not as strongly associated with z^* as x_1, and indeed, x_3 is not associated with $y(0)$ at all.

5.7 Simulation Scenario 5: Treatment Effect Modifiers that are Unrelated to the Underlying Biomarker and Medication Use Probability

The purpose of this study is to evaluate bias under a setting in which the source of subgroup-specific effects is the single "dose" factor D only associated with the magnitude of the medication's effect (and in particular, not with the underlying biomarker $y(0)$ or with medication use, z). In this scenario, we let $\boldsymbol{\eta} = (\eta_0, 0, 0, \eta_4)$, where η_0 ranges from 0 to 5, and $\eta_4 = 5.5 - 2\eta_0$, so that the marginal effect of medication use is constantly 3.75. We simulate one-thousand replicates at each pair of η_0 and η_4 considered. For the SSEM, we do not estimate η_1, η_2, or η_3.

Figures 5.10 and 5.11 illustrate the estimated bias and rMSE from each method for estimation of β_1 and β_2, across the range of values for η_0. Both the TEM and SSEM provide low-bias estimates of β_1 and β_2 in this setting; the Ignore and Adjust methods provide biased estimates of both parameters. These patterns were confirmed when D_i was generated from a normally distributed rather than as a binary variable. In this

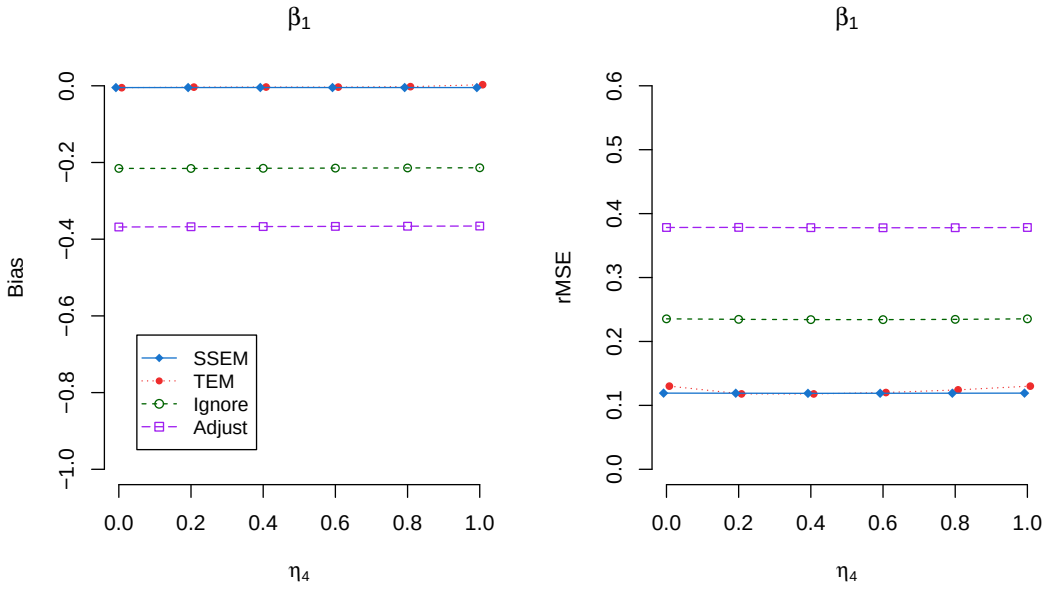

Figure 5.10: Results from simulation Scenario 5. The range of values considered for η_4 is shown on the x-axis, and the estimated bias (left) and rMSE (right) are shown. Note that the true parameter value is given by $\beta_1 = 1$.

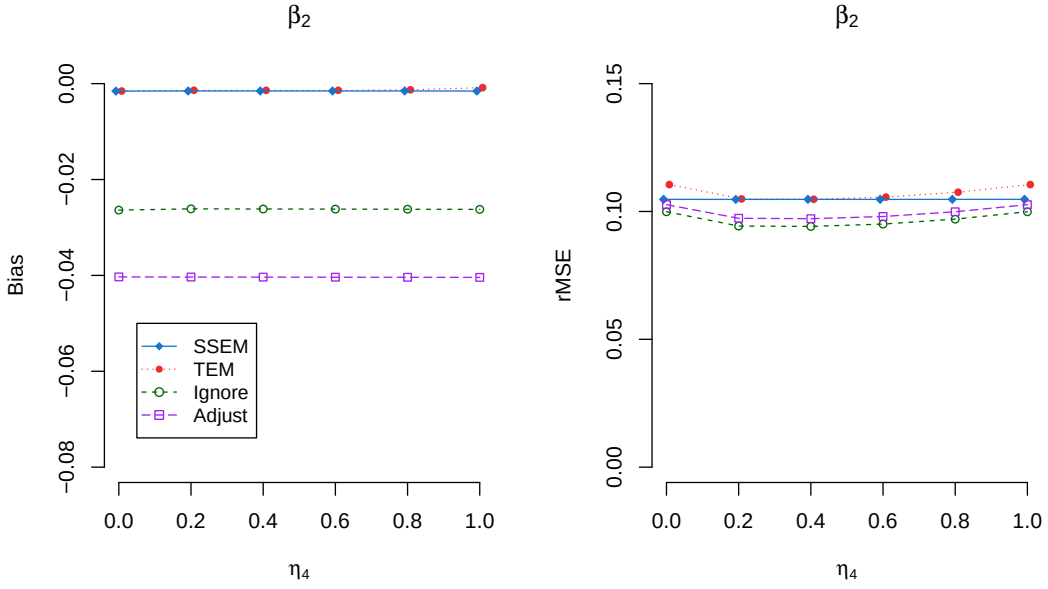

Figure 5.11: Results from simulation Scenario 5. The range of values considered for η_4 is shown on the x-axis, and the estimated bias (left) and rMSE (right) are shown. Note that the true parameter value is given by $\beta_2 = 1$.

case, this is analogous to allowing $\delta_i \sim \mathcal{N}(\delta, \sigma_\delta^2)$, as we have already done in Section 3.3. By generating D to be binary, the total errors are still approximately normally distributed, and hence it is not all that surprising that the TEM performs well. When an external treatment effect modifier is generated from, for example, a highly right-skewed Gamma distribution, the total errors for the TEM are no longer approximately normal. We followed up with a simulation study in which the external variable was given by $D \sim \text{Gamma}(\alpha = 1, \beta = 5)$, and we observed the same patterns as those in Figures 5.10 and 5.11; this is consistent with the robustness properties observed previously in Section 4.4.

This study suggests that the TEM is robust to effect modification when the effect modifier is neither a predictor of interest nor associated with medication use. If the effect modifier is highly skewed, including it in the SSEM may be advisable in order to mitigate severe departures from bivariate normality.

5.8 Asymptotically Valid Level for the Robust Wald Test

We conclude this chapter with a simulation study to examine whether the robust Wald test for presence of effect modification achieves the correct asymptotic level. We utilize the same simulation setup of this chapter, except we generate the data under the null hypothesis: $\boldsymbol{\eta} = (3.75, 0, 0, 0, 0)^T$. We compute the robust Wald statistic as in Section 5.3 at each of one-thousand simulation replicates, for sample sizes ranging from $N = 250$ to $N = 2000$. The empirical cumulative distribution function at each sample size was determined and compared to the theoretical χ_4^2 distribution of the test statistic, as illustrated in Figure 5.12 (left). Note also in the right panel of Figure 5.12 that we have plotted the empirical level with $\alpha = 0.01, 0.05$, and 0.1, as compared to the nominal levels. This study suggests that the robust Wald test proposed is asymptotically valid, with $W_N \longrightarrow_d \chi_{q_3}^2$ appearing to be true. For smaller sample sizes, the Wald test may reject the null hypothesis of no effect modification at an inappropriately high rates.

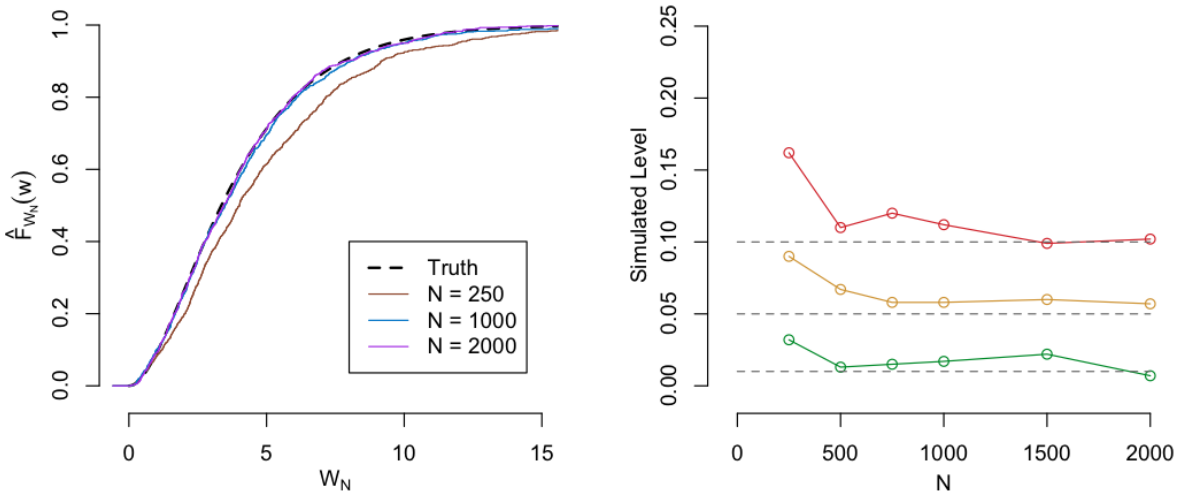

Figure 5.12: Illustration of the asymptotic validity of the robust Wald test. In the left panel, we have plotted the true χ^2_4 distribution, and the empirical cumulative distribution function of the χ^2_4 test statistic over one-thousand replications for various sample sizes. In the right panel, we have taken cross-sections of the left panel at an α level of 0.10, 0.05, and 0.01 to show that the level is asymptotically valid for common choices of significance testing levels.

5.9 Conditioning on Effect Modifiers

In the original TEM, we were required to condition on both $\boldsymbol{x}$ and $\boldsymbol{w}$ in each of the two simultaneous equations (the biomarker and medication use models). In this setting, we are required to additionally condition on the effect modifiers. That is to say that the outcome equations can be written as $\mathbb{E}[y(0)|\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{w}] = \boldsymbol{x}^T\boldsymbol{\beta}$ and $P(z = 1|\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{w}) = \Phi(\boldsymbol{w}^T\boldsymbol{\alpha})$. For example, if a variables x_1 and x_2 are associated with $y(0)$ but x_1 is also an effect modifier (e.g., $\delta_i = x_{1i}$), then x_1 should be placed in the biomarker model regardless of whether or not it is of interest. In particular, the marginal association between x_2 and $y(0)$ is a different parameter than the association between x_2 and $y(0)$ conditional on x_1.

5.10 Discussion

In this chapter, we have derived and evaluated an extension of James Heckman's original treatment effects model in order to accommodate effect measure modification for estimation of the natural history association between a predictor of interest and a biomarker outcome in the presence of endogenous medication use. In practice, the effects of medication use may systematically vary with observable factors such as medication class/dose, or with the predictors of interest. The TEM (and hence our proposed SSEM) can be presented as a system of structural equations. The maximum likelihood approach to estimating parameters of interest allows one to regard the model essentially as a missing-at-random model, in which $y(0)$ is unobservable in participants on medication. In the TEM, the appropriate correction (δ) to y_i is determined as a single value for all participants (this is the assumption of uniform effects), estimated simultaneously with the other parameters. When the effects of medication use vary with predictors of interest in the biomarker model, substantial bias can arise when using the TEM to estimate associations; this occurs because the single estimate of δ applied to on-medication participants systematically under- or over-corrects their observed biomarker values for the effects of medication use. Moreover, if the effect modifier is associated with medication use only, but not the biomarker, bias is also observed (albeit of lower magnitude in our simulations). The approach taken in the SSEM is to further condition the effects of medication use on potential effect modifiers to mitigate these problems with systematic over- or under-correction.

Viewing the model extension in this way helps explain our finding that the TEM appears to be robust to misspecification of effect modifiers associated only with the treatment effects (i.e., when covariates in $\boldsymbol{v}$ do not appear in $\boldsymbol{x}$ or $\boldsymbol{w}$). In this setting, the mean models for $y_i(0)$ and z_i^* are still correctly specified, and the variation in the effects of medication use across values of the effect modifier becomes absorbed into

the error term in the biomarker model (which can be accounted for with the robust standard error estimator). In this sense, the effect modifier would only serve to create an additional source of error.

This particular extension to Heckman's TEM is important for two reasons. Firstly, it achieves the desired goal of reducing bias in estimating biomarker associations when treatment effect magnitude varies with observable covariates. Secondly, because effect measure modification is modeled directly, we are provided with a means of understanding how medication use can work in real-world settings. For example, we may wish to determine whether the effects of medication on a biomarker are different across different races. Although this is not our primary focus, our results demonstrated that effect modifier parameters can be estimated fairly well from the SSEM.

Our simulation results further confirmed the inadequacy of the simpler approaches (ignoring medication use and adjusting for it) in the setting of endogenous medication use, and further demonstrated that the TEM can provide biased estimates of the natural association when predictors in the biomarker model modify the effects of the medication on the underlying biomarker. The proposed SSEM nearly eliminates this bias by specifically accommodating the effect modifiers when they are present. Under the setting of a modest average treatment effect, there is an efficiency loss that is inarguably small relative to the large bias reduction achieved, at least when considering measures such the MSE, although it might be considered a high cost when effect modification is not thought to be present. The robust Wald-based test might be used to provide evidence of effect modification.

Our previous work in Chapters 2-4 revealed that failure to account for endogeneity can result in bias when estimating natural associations between predictors and biomarkers. This follow-up work revealed that this improvement may be partially lost if the predictor of interest modifies the effect of medication use on the biomarker. Therefore,

this modeling framework can be a useful tool in order to estimate natural associations when effect modifiers are thought to be well understood.

On the other hand, there are settings in which effect modifiers might not be well documented. The use of the robust Wald test can be used as a complimentary tool or an exploratory data analysis procedure to first test for effect modification before applying the TEM (which presumed uniform treatment effects). Based on the results of this research, we recommend using the SSEM in conjunction with the TEM framework in order further reduce bias in estimating associations of interest when endogenous medication use is present.

When effect modifiers are present but are unrelated to the underlying biomarker or to medication use, the advantages to accounting for them is not as clear, unless understanding differential treatment effect patterns is of interest. More work is underway to evaluate the potential advantages to using this model as a tool to understand treatment effect patterns.

This chapter concludes the component of this dissertation handling the development of methods in cross-sectional data. Beginning in Chapter 6, we turn our attention to settings where repeated measures are available.

Chapter 6

ESTIMATING NATURAL HISTORY ASSOCIATIONS IN LONGITUDINAL DATA

In this chapter, we transition into a setting in which repeated measures are available on study subjects over time. Specifically, we focus on extending the methodology developed in prior chapters in order to to estimate the natural history association in longitudinal data when there is endogenous medication use.

In recent years, age-trend modeling of cardiovascular biomarkers has received a great deal of attention in the literature (Singh et al., 2012; Gurven et al., 2012; Carroll et al., 2005; Allen et al., 2014). However, results from these studies are based on analogues of the simple approaches described in Chapter 2, such as excluding on-medication observations or ignoring medication use altogether. Just as these standard approaches were not appropriate in a cross-sectional setting for estimating the natural history association, the longitudinal analogues are insufficient as they fail to address the challenges that arise from endogenous medication use.

McClelland et al. (2008) devise a multiple imputation approach in which information is obtained from participants who begin medication use during the course of the study. However, estimation of parameters of interest under the TEM framework does not appear to have been adequately addressed or explored when longitudinal data are available. There is an unmet need to formulate the longitudinal analogues of Heckman's TEM for settings in which data from cross-in participants is either sparse or unavailable. STATA allows for clustering of observations with its "cluster robust" option, although the documentation does not make it clear how this procedure is imple-

mented. Further exploration revealed that the output can be matched by implementing a procedure similar in spirit to generalized estimating equations (GEE) with working independence, whereby a marginal model is fit ignoring clustering, and a valid robust variance estimator is used *a posteriori*. This sort of working independence approach has been suggested in the literature for similar probit-response type model in order to accommodate clustering (Wooldridge, 2011).

Although such a method can provide consistent estimates under correct specification, we of course do not truly believe that cardiovascular biomarkers such as LDL cholesterol are uncorrelated over time, and there is efficiency to be gained by exploiting within-subject correlation in the biomarker (Diggle et al., 2002). In this chapter, we explore potential options for handling within-subject correlation in longitudinal settings. As will be made apparent in this chapter, extending the TEM by fully specifying correlation in the underlying biomarker and in medication use over time is analytically intractable as it demands the computation of T-fold integrals, where T is the number of observations within a cluster. We seek to put forth a reasonable set of assumptions regarding within-subject correlation and examine the extent to which efficiency is gained over the working independence model. One model extension that will receive a great deal of attention is our proposal of partial specification of the correlation structure for estimation of associations. This approach provides a computationally tractable alternative to fully parametric likelihood approaches, while still accounting for correlation in the biomarker. We will show by simulation that this approach can yield very large efficiency gains compared to the standard working independence model. Moreover, we have the important robustness property that the resulting estimators will be consistent for the natural history association irrespective of the choice of a working correlation structure.

6.1 Notation and Naïve Approaches in Longitudinal Data

As has been the case in previous chapters, let $i = 1, \ldots, N$ denote the cluster (subject); now, further let $t = 1, \ldots, T_i$ denote the observation number for the i^{th} subject. Let z_{it} denote the indicator of medication use for subject i at observation t, and let $\boldsymbol{y}_i(z) = (y_{i1}(z), \ldots, y_{iT_i}(z))^T$ denote the vector of potential biomarker values for subject i under treatment assignment z. The observed biomarker vector for subject i can be denoted by $\boldsymbol{y}_i = (y_{1i}, \ldots, y_{iT_i})^T$, which under the consistency assumption is equivalent to the potential biomarker value under treatment actually received $\boldsymbol{y}_i = (y_{i1}(z_{i1}), \ldots, y_{iT_i}(z_{iT_i}))^T$. Let $\boldsymbol{x}_{it} = (1, \boldsymbol{x}_{1,it}, \ldots, \boldsymbol{x}_{p,it})^T$ be a complete covariate vector of variables predicting the biomarker for subject i at observation t (allowing for an intercept). The entire covariate matrix for subject i will be denoted by $\mathbf{X}_i$. If $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ denotes the corresponding vector of unknown regression coefficients, the parameter of interest in this longitudinal case in described by the mean model $y_{it}(0) = \boldsymbol{x}_{it}^T\boldsymbol{\beta} + \epsilon_{it}$, where the ϵ_{it} are all i.i.d., of mean zero, and variance σ_y^2.

Naïve approaches in the cross-sectional setting carry over into the longitudinal setting for marginal modeling. We briefly describe how they could be implemented with marginal modeling approaches such as GEE. Letting $\mathbf{W}_i$ denote some weight matrix (not to be confused with covariates predicting medication use), we define the naïve estimator which ignores medication use altogether as the estimate that solves the estimating equations

$$\mathbb{G}_N(\boldsymbol{\beta}; \mathbf{W}) \;=\; \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{W}_i (\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta}) = \mathbf{0} \tag{6.1}$$

A working independence structure would result in a choice of the identity matrix as the weight matrix. In other cases, the weight matrix could be estimated from the residuals of the working independence model, and under some modest regularity conditions and

$\sqrt{N}$-consistency of the estimated weight matrix, could be used to gain a potentially more efficient estimator (the most efficient choice for the weight matrix is the inverse of the true covariance matrix).

In the "Exclude" approach, the estimating equations are instead given by

$$\mathbb{G}_N(\boldsymbol{\beta}; \mathbf{W}) = \sum_{i,t:z_{it}=0} \mathbf{X}_i^T \mathbf{W}_i (\boldsymbol{y}_i - \mathbf{X}_i \boldsymbol{\beta}) = \mathbf{0}, \tag{6.2}$$

and in the "Adjust" approach, we modify the "Ignore" approach with inclusion of a covariate z_{it} in the biomarker model. For reasons similar to why these approaches were inadequate for estimating the natural history association, these approaches are also inappropriate in the longitudinal setting.

6.2 The Longitudinal Endogeneity Model with Working Independence

We mimic the approach described by Wooldridge (2011) for other probit-response models in order to construct a valid estimator for $\boldsymbol{\beta}$ without specification of a covariance structure. We begin with the original version of the TEM as in Section 3.1, under the same restrictions satisfying the principal assumption. Now, the structural equations are given by

$$y_{it}(0) = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \epsilon_{it} \tag{6.3}$$

$$z_{it}^* = \boldsymbol{w}_{it}^T \boldsymbol{\alpha} + \lambda y_{it}(0) + \gamma_{it}, \tag{6.4}$$

and as before, z_{it}^* has total variance σ_z^2, set equal to 1 due to weak identifiability of $\boldsymbol{\alpha}$ and ρ, and $z_{it} = \mathbf{1}(z_{it}^* > 0)$ denotes the observed medication use status for subject i at time t, with $y_{it} = y_{it}(0) - \delta z_{it}$. Let $\boldsymbol{\theta}_0$ denote the true parameter value in the data generating mechanism. The likelihood for subject i at observation t is given by

$$L_{it}(\boldsymbol{\theta}) = \frac{1}{\sigma_y} \phi \left(\frac{y_{it} - \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \delta z_{it}}{\sigma_y} \right) \int_{Q_{it}} dF_{z_{it}^* | y_{it}}(z_{it}^*), \tag{6.5}$$

where $Q_{it} = (-\infty, 0]$ if $z_{it} = 0$ and $(0, \infty)$ if $z_{it} = 1$. Under correct mean-model specification, standard likelihood theory gives us that $\boldsymbol{\theta}_0$ maximizes $\mathbb{E}_{\boldsymbol{\theta}}[\log L_{it}(\boldsymbol{\theta})]$. Hence, $\boldsymbol{\theta}_0$ must also maximize $\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{N}\sum_{t=1}^{T_i}\log L_{it}(\boldsymbol{\theta})\right]$. The estimating equations $\mathcal{U}(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{t=1}^{T_i}\partial \log L_{it}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$ is merely an M-estimation problem for $\boldsymbol{\theta}$, with T_i fixed, and asymptotics in N. Hence, $\hat{\boldsymbol{\theta}} \longrightarrow_p \boldsymbol{\theta}_0$. In turn, aggregating the individual likelihoods gives rise to the following estimating equations:

$$\begin{aligned}\mathbb{G}_N(\boldsymbol{\theta}) &= \sum_{i=1}^{N}\sum_{t=1}^{T_i}\frac{\partial}{\partial\boldsymbol{\theta}}\left\{\log\frac{1}{\sigma_y}\phi\left(\frac{y_{it}-\boldsymbol{x}_{it}^T\boldsymbol{\beta}+\delta z_{it}}{\sigma_y}\right)\right. \\ &\qquad \left. + \log\Phi\left((-1)^{1-z_{it}}\frac{\boldsymbol{w}_{it}^T\boldsymbol{\alpha}+\rho(y_{it}-\boldsymbol{x}_{it}^T\boldsymbol{\beta}+\delta z_{it})/\sigma_y}{\sqrt{1-\rho^2}}\right)\right\} = \mathbf{0}.\end{aligned} \tag{6.6}$$

Estimates of $\boldsymbol{\theta}$ can be obtained from Newton-Raphson type procedures. The estimating equations above are not score equations because they are not derived from a full likelihood; the within-subject covariance structure is not modeled in these estimating equations, nor must it be known for consistent estimation. We only demand the assumptions put forth in the cross-sectional version of the model, and consistency is achieved without any specification of a correlation structure within clusters. Also, we make no assumptions about exogeneity of z in this approach.

Fisher information-based standard errors are not appropriate, and the likelihood ratio test is invalid for testing hypotheses regarding $\boldsymbol{\theta}$. A robust variance-covariance estimator can be used to account for correlation within clusters (White, 1980), from which Wald based confidence intervals can be constructed for parameters. Letting $H_N(\hat{\boldsymbol{\theta}})$ denote the Hessian $\partial\mathcal{U}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$,

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = \left(H_N(\hat{\boldsymbol{\theta}})\right)^{-1}\left[\sum_{i=1}^{N}\mathcal{U}_i(\hat{\boldsymbol{\theta}})\mathcal{U}_i(\hat{\boldsymbol{\theta}})^T\right]\left(H_N(\hat{\boldsymbol{\theta}})\right)^{-T} \tag{6.7}$$

provides a valid covariance for $\hat{\boldsymbol{\theta}}$. This approach for estimating $\boldsymbol{\beta}$ has advantages of

computational convenience in the sense that it does not require any modeling within-subject correlation. We refer to this model for estimation of $\boldsymbol{\beta}$ as the longitudinal endogeneity model (LEM) with working independence. Indeed, our extension to allow for effect modifiers easily generalizes to this model, in which the term δz_{it} would become $\boldsymbol{v}_{it}^T \boldsymbol{\eta} z_{it}$. Going forward, we will restrict our discussion to allow for effect modifiers.

6.3 Full Specification of Covariance Structure

The only correlation modeled in the working independence model is the correlation between $y_{it}(0)$ and z_{it}^*–the biomarker and medication use at corresponding times. We present an example model in which we fully specify an exchangeable correlation structure over time. Suppose that $\text{Corr}(\epsilon_{it}, \gamma_{it}) = \rho$, $\text{Corr}(\epsilon_{it}, \epsilon_{it'}) = \rho_y$, $\text{Corr}(\gamma_{it}, \gamma_{it'}) = \rho_z$, and $\text{Corr}(\epsilon_{it}, \gamma_{it'}) = \rho_{y,z}$. A covariance matrix for subject i is a $2T_i \times 2T_i$ matrix:

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{11} & \boldsymbol{\Sigma}_i^{12} \\ \boldsymbol{\Sigma}_i^{21} & \boldsymbol{\Sigma}_i^{22} \end{bmatrix} \tag{6.8}$$

with $T_i \times T_i$ blocks $\boldsymbol{\Sigma}_i^{11}$, $\boldsymbol{\Sigma}_i^{12}$, $\boldsymbol{\Sigma}_i^{21}$, and $\boldsymbol{\Sigma}_i^{22}$ given by

$$\boldsymbol{\Sigma}_i^{11} = \sigma_y^2 \begin{bmatrix} 1 & \rho_y & \cdots & \rho_y \\ \rho_y & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_y \\ \rho_y & \cdots & \rho_y & 1 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_i^{22} = \begin{bmatrix} 1 & \rho_z & \cdots & \rho_z \\ \rho_z & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_z \\ \rho_z & \cdots & \rho_z & 1 \end{bmatrix}, \text{ and}$$

$$\boldsymbol{\Sigma}_i^{12} = \boldsymbol{\Sigma}_i^{22} = \sigma_y \begin{bmatrix} \rho & \rho_{y,z} & \cdots & \rho_{y,z} \\ \rho_{y,z} & \rho & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{y,z} \\ \rho_{y,z} & \cdots & \rho_{y,z} & \rho \end{bmatrix} \tag{6.9}$$

The likelihood over for subject i over the observations $1, \ldots, T_i$ is given by:

$$\begin{aligned} L_i(\boldsymbol{\theta}) &= p(\boldsymbol{y}_i|\boldsymbol{z}_i)p(\boldsymbol{z}_i) = p(\boldsymbol{y}_i|z_{i1}, \ldots, z_{iT_i})p(z_{i1}, \ldots, z_{iT_i}) \\ &= p(\boldsymbol{y}_i|z^*_{i1} \in Q_{i1}, \ldots, z^*_{iT_i} \in Q_{iT_i})p(z^*_{i1} \in Q_{i1}, \ldots, z^*_{ik} \in Q_{iT_i}) \\ &= \int_{Q_{iT_i}} \cdots \int_{Q_{i1}} p(\boldsymbol{z}^*_i|\boldsymbol{y}_i)p(\boldsymbol{y}_i)dz^*_{i1} \ldots dz^*_{iT_i}, \end{aligned} \tag{6.10}$$

where again, $Q_{it} = (-\infty, 0]$ if $z_{it} = 0$ and $(0, \infty)$ if $z_{it} = 1$. Then, the total likelihood is given by

$$\mathcal{L}_{\mathbf{V},\mathbf{W},\mathbf{X},\mathbf{Y},\mathbf{Z}}(\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{y}_i) \int_{Q_{iT_i}} \cdots \int_{Q_{i1}} p(\boldsymbol{z}^*_i|\boldsymbol{y}_i)dz^*_{i1} \ldots dz^*_{iT_i} \tag{6.11}$$

where $p(\boldsymbol{y}_i)$ is the multivariate normal density function with mean parameter given by $\boldsymbol{\mu}^y_i = \mathbf{X}_i\boldsymbol{\beta} - (\mathbf{V}_i\boldsymbol{\eta}) \circ \boldsymbol{z}_i$ and covariance parameter given by $\boldsymbol{\Sigma}^{11}_i$; additionally, $p(\boldsymbol{z}^*_i|\boldsymbol{y}_i)$ is the conditional multivariate normal density function for $\boldsymbol{z}^*_i|\boldsymbol{y}_i$ with mean parameter given by $\boldsymbol{\mu}^z_i = \mathbf{W}_i\boldsymbol{\alpha} + \boldsymbol{\Sigma}^{12}_i(\boldsymbol{\Sigma}^{22}_i)^{-1}(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta} + \mathbf{V}_i\boldsymbol{\eta})$ and covariance parameter given by $\boldsymbol{\Sigma}^{11}_i - \boldsymbol{\Sigma}^{12}_i(\boldsymbol{\Sigma}^{22}_i)^{-1}\boldsymbol{\Sigma}^{21}_i$.

As can be seen in in the above expressions for the likelihood, this demands the computation of T_i-fold integrals, where the range of integration is one of 2^{T_i} possible hyperoctants in T_i-dimensional real space. The integral of this multivariate density has no closed form expression, making this approach generally infeasible when there are more than a couple of observations within clusters. In our example, we chose an exchangeable correlation structure, although others are possible; generally, only independence for $\boldsymbol{\Sigma}^{22}_i$ and an assumption that $\rho_{y,z} = 0$ will allow the multiple integral to collapse into a product of single integrals. In any case, fully specifying the correlation structure is analytically cumbersome in the sense that there are three levels of correlation to consider: (i) the correlation in $\boldsymbol{y}_i(0)$ over time, (ii) the correlation in $\boldsymbol{z}^*_i$ over time, and (iii) the correlation between $\boldsymbol{y}_i(0)$ and $\boldsymbol{z}^*_i$ at different times.

6.4 Partial Specification of Covariance Structure

We now propose a model to account for the correlation in $\boldsymbol{y}_i$ over time, but bypass modeling correlation in $\boldsymbol{z}_i^*$ over time and correlation between $\boldsymbol{y}_i(0)$ and $\boldsymbol{z}_i^*$ at different times. This acts as a compromise between working independence and fully parametric likelihood specification.

Let $\boldsymbol{\Theta}$ denote the true parameter space that defines the data generating mechanism, and $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0, \sigma_{y0}, \rho_0) \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ denote the true parameter value in the parametric sub-model which is presumed to hold. As an alternative to full specification of the correlation structure, we propose modifying the independence-based estimating equations to account for within-subject correlation in the biomarker error terms. Let $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma_y, \rho, \boldsymbol{\rho}_y)$ denote the set of all parameters in the estimating equations, where $\boldsymbol{\rho}_y$ describes a working correlation structure in the biomarker errors. If $\boldsymbol{\Sigma}_i^{11} = \sigma_y^2 \mathbf{R}_i(\boldsymbol{\rho}_y)$ denotes a working covariance model for the biomarker errors, an individual's contribution to the estimating function is given by:

$$\mathcal{U}_i(\tilde{\boldsymbol{\theta}}) \;=\; \frac{\partial}{\partial \tilde{\boldsymbol{\theta}}} \left\{ \log p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{y}_i) + \sum_{t=1}^{T_i} \log \int_{Q_{it}} dF_{z_{it}^*|y_{it}}(z_{it}^*) \right\},$$

where $p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{y}_i)$ is the multivariate normal density function with covariance matrix given by $\boldsymbol{\Sigma}_i^{11}$. We have the important robustness property that estimation of $\boldsymbol{\theta}_0$ is still valid, even under misspecification of the correlation structure.

Theorem: Let $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ denote the true parameter in the parameter sub-model of interest (presumed correct). Then, even under incorrect specification of the working correlation, the solution to the estimating equations $\mathcal{U}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$ is consistent for $\tilde{\boldsymbol{\theta}}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\rho}_{y0})$, where $\boldsymbol{\theta}_0$ is of interest.

Proof: We sketch the proof when $\mathbf{R}_i(\rho_y)$ depends on a single parameter ρ_y (e.g., exchangeable, AR-1, or exponential correlation structure). Let $\tilde{\boldsymbol{\theta}}_0$ denote the solution to $\mathbb{E}[\mathcal{U}(\tilde{\boldsymbol{\theta}})] = \mathbf{0}$; our goal is to show that $\tilde{\boldsymbol{\theta}}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\rho}_{y0})$. It suffices to show that $\mathbb{E}_{\boldsymbol{\theta}_0,\rho_{y0}}[\mathcal{U}(\tilde{\boldsymbol{\theta}})] = \mathbf{0}$ for some ρ_{y0}. Let $\mathbf{S}_i$ denote the true correlation matrix. We partition the estimating functions for $\tilde{\boldsymbol{\theta}}$.

$$
\begin{aligned}
\mathcal{U}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{\theta}}) &= \sum_{i=1}^{N}\sum_{t=1}^{T_i} \frac{\partial}{\partial \boldsymbol{\alpha}} \log \int_{Q_{it}} dF_{z_{it}^*|y_{it}}(z_{it}^*), \\
\mathcal{U}_{\rho}(\tilde{\boldsymbol{\theta}}) &= \sum_{i=1}^{N}\sum_{t=1}^{T_i} \frac{\partial}{\partial \rho} \log \int_{Q_{it}} dF_{z_{it}^*|y_{it}}(z_{it}^*), \\
\mathcal{U}_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\theta}}) &= \frac{1}{\sigma_y^2}\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{R}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i^y) + \sum_{i=1}^{N}\sum_{t=1}^{T_i} \frac{\partial}{\partial \boldsymbol{\beta}} \log \int_{Q_{it}} dF_{z_{it}^*|y_{it}}(z_{it}^*), \\
\mathcal{U}_{\boldsymbol{\eta}}(\tilde{\boldsymbol{\theta}}) &= -\frac{1}{\sigma_y^2}\sum_{i=1}^{N} \mathbf{V}_i^T \mathbf{R}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i^y) + \sum_{i=1}^{N}\sum_{t=1}^{T_i} \frac{\partial}{\partial \boldsymbol{\eta}} \log \int_{Q_{it}} dF_{z_{it}^*|y_{it}}(z_{it}^*), \qquad (6.12) \\
\mathcal{U}_{\sigma_y}(\tilde{\boldsymbol{\theta}}) &= \sum_{i=1}^{N} \left\{ -\frac{1}{\sigma_y} + \frac{1}{\sigma_y^3}(\boldsymbol{y}_i - \boldsymbol{\mu}_i^y)^T \mathbf{R}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i^y) \right\} \\
&\quad + \sum_{i=1}^{N}\sum_{t=1}^{T_i} \frac{\partial}{\partial \sigma_y} \log \int_{Q_{it}} dF_{z_{it}^*|y_{it}}(z_{it}^*), \\
\mathcal{U}_{\rho_y}(\tilde{\boldsymbol{\theta}}) &= \sum_{i=1}^{N} \left\{ \frac{1}{2\sigma_y^2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i^y)^T \mathbf{R}_i^{-1} \left[\frac{\partial \mathbf{R}_i}{\partial \rho_y}\right] \mathbf{R}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i^y) \right. \\
&\quad \left. - \frac{1}{2|\mathbf{R}_i|} \times \text{trace}\left(\text{adj}(\mathbf{R}_i) \frac{\partial \mathbf{R}_i}{\partial \rho_y} \right) \right\}
\end{aligned}
$$

Here, $\mathcal{U}_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{\theta}})$ and $\mathcal{U}_{\rho}(\tilde{\boldsymbol{\theta}})$ are the same as they are under the working independence model (in particular, they are independent of ρ_y). Thus under a valid parametric sub-

model, $\mathbb{E}_{\boldsymbol{\theta}_0,\rho_y}[\mathcal{U}_{\rho_y}(\tilde{\boldsymbol{\theta}})] = \mathbb{E}_{\boldsymbol{\theta}_0,\rho_y}[\mathcal{U}_{\alpha}(\tilde{\boldsymbol{\theta}})] = \mathbf{0}$ for any ρ_y, so choose $\rho_y = \rho_{y0}$. Under working independence, $\mathbf{R}_i$ is the identity matrix; under correct mean model specification, the expectation of the first (and hence the second) sums in each estimating function $\mathcal{U}_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\theta}})$ and $\mathcal{U}_{\boldsymbol{\eta}}(\tilde{\boldsymbol{\theta}})$ are zero at $\boldsymbol{\theta}_0$. Hence, $\mathbb{E}_{\boldsymbol{\theta}_0,\rho_y}[\mathcal{U}_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\theta}})] = \mathbb{E}_{\boldsymbol{\theta}_0,\rho_y}[\mathcal{U}_{\boldsymbol{\eta}}(\tilde{\boldsymbol{\theta}})] = 0$ regardless of ρ_y, so choose $\rho_y = \rho_{y0}$. At $(\boldsymbol{\theta}_0, \rho_y)$, the expectations of $\mathcal{U}_{\sigma_y}(\tilde{\boldsymbol{\theta}})$ and $\mathcal{U}_{\rho_y}(\tilde{\boldsymbol{\theta}})$ are given by

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}_0}[\mathcal{U}_{\sigma_y}(\tilde{\boldsymbol{\theta}})] &= \sum_{i=1}^{N} \text{trace}(\mathbf{S}_i - \mathbf{R}_i^{-1}\mathbf{S}_i)/\sigma_{y0} \\ \mathbb{E}_{\boldsymbol{\theta}_0}[\mathcal{U}_{\rho_y}(\tilde{\boldsymbol{\theta}})] &= \frac{1}{2}\sum_{i=1}^{N} \text{trace}\left(\mathbf{R}_i^{-1}\frac{\partial \mathbf{R}_i}{\partial \rho_y}\left(\mathbf{R}_i^{-1}\mathbf{S}_i - \mathbf{I}\right)\right) \end{aligned} \tag{6.13}$$

The equations $\mathbb{E}_{\boldsymbol{\theta}_0}[\mathcal{U}_y(\tilde{\boldsymbol{\theta}})] = 0$ and $\mathbb{E}_{\tilde{\boldsymbol{\theta}}_0}[\mathcal{U}_{\rho_y}] = 0$ share a root in ρ_y, defining the value of ρ_{y0}. Thus, $\mathbb{E}_{\boldsymbol{\theta}_0,\rho_{y0}}[\mathcal{U}(\tilde{\boldsymbol{\theta}})] = \mathbf{0}$ as desired. Hence, the solution $\hat{\boldsymbol{\theta}}$ to these estimating equations is consistent for $(\boldsymbol{\theta}_0, \rho_{y0})$ as desired. Q.E.D.

Once again, we note that the resulting estimating equations are not necessarily score equations since we have not modeled the correlation in medication use over time, nor the correlation between medication use and the underlying biomarker at different times. A robust variance-covariance estimator should therefore still be used for valid standard errors for parameters of interest. We refer to this model for estimation of $\boldsymbol{\beta}$ as the longitudinal endogeneity model (LEM) with the working correlation structure as indicated. Accounting for correlation in the biomarker terms, as presented by this model, accounts for a major source of variability in $\hat{\boldsymbol{\beta}}$; this proposed model also bypasses the need to compute T_i-fold integrals.

Diggle et al. (2002) have noted that if the covariates are non-deterministic, consistent estimation of marginal parameters relies on the following potentially unrealistic "full-covariate conditional mean" assumption: the expectation of an outcome at a given time conditional on the covariate vector, $\boldsymbol{x}_{it}$, is equal to the expectation conditional on the

entire covariate matrix, $\mathbf{X}_i$. That is, unless one utilizes a working-independence model, in which the full-covariate conditional mean assumption does not need to be satisfied for consistent estimation of marginal parameters (Pepe et al., 1994). In examples such as age-trend modeling, we bypass this issue since age is deterministic. This challenge can also be avoided when estimating associations with time-stable exposures such as gender and race. If one has random time-varying covariates in the biomarker model (i.e., in many situations other than age-trend modeling), it may be advisable to use the working-independence model instead of our proposed model. In our simulations and applied examples, we will focus on examples in which covariates are either deterministic, or in which the full-covariate conditional mean assumption is known to be satisfied.

6.5 Systematic Dependence on Subject History

In the standard cross-sectional version of Heckman's TEM, modeling the latent medication use variable as explicitly depending on the underlying biomarker (that is, including the term $\lambda y_i(0)$ in the medication use model) can be handled by simply taking any and all covariates of $\mathbf{X}$ and placing them into $\mathbf{W}$. This can be seen as follows: $z^* = \boldsymbol{w}^T\boldsymbol{\alpha} + \lambda y(0) + \gamma = \boldsymbol{w}^T\boldsymbol{\alpha} + \lambda(\boldsymbol{x}^T\boldsymbol{\beta} + \epsilon) + \gamma$. Regrouping terms gives a larger covariate vector $\tilde{\boldsymbol{w}}$ in which all covaraites of $\boldsymbol{x}$ are included, and a total error term $\tilde{\gamma} = \lambda\epsilon + \gamma$. In this way, λ need not be estimated. In longitudinal data, the options are more numerous as we may accommodate dependence of medication use on the biomarker at the current time, or on past history. The dependence itself can be time-varying if the criteria for treatment change or as medication use becomes more popular over time. In this section, we describe a few ways one could model the dependence of medication use on a subject's history and how covariates should be specified in the model to accommodate each type of dependence. We focus on examples depending on whether or not they are homogeneous or nonhomogeneous, and whether or not the dependencies correspond only at cross-sections, or if there is a telescope dependence.

6.5.1 Homogeneous Corresponding Dependencies

In this case, suppose that medication use depends on the underlying biomarker, but only at the concurrent time. Specifically, suppose that $z_{it}^* = \boldsymbol{w}_{it}^T\boldsymbol{\alpha} + \lambda y_{it}(0) + \gamma_{it}$. The dependence is also modeled as homogeneous in that the parameter λ is the same at each time point, independent of t. This example generalizes the modification proposed in the cross-sectional version of Heckman's TEM described by Spieker et al. (2015). To accommodate this dependence, one places all covariates of $\boldsymbol{x}_{it}$ in $\boldsymbol{w}_{it}$ to create a new covariate vector $\tilde{\boldsymbol{w}}_i$ and a corresponding parameter $\tilde{\boldsymbol{\alpha}}$: $z_{it}^* = \tilde{\boldsymbol{w}}_{it}\tilde{\boldsymbol{\alpha}} + \tilde{\gamma}_i$. Figure 6.1 depicts a DAG summarizing the homogeneous corresponding dependencies model.

6.5.2 Non-homogeneous Corresponding Dependencies

We now consider a case to generalize the previous approach, in which the dependence of medication use on the corresponding underlying biomarker value is time-varying (including the term $\lambda_t y_{it}(0)$ in the medication use model at each time point, t). Expanding out the term, we have that:

$$\begin{aligned} z_{it}^* &= \boldsymbol{w}_{it}^T\boldsymbol{\alpha} + \lambda_t y_{it}(0) + \gamma_{it} \\ &= \boldsymbol{w}_{it}^T\boldsymbol{\alpha} + \lambda_t(x_{it}^T\boldsymbol{\beta} + \epsilon_{it}) + \gamma_{it} \\ &= \boldsymbol{w}_{it}^T\boldsymbol{\alpha} + \boldsymbol{x}_{it}^T\boldsymbol{\alpha}_t + \tilde{\gamma}_{it}, \end{aligned}$$

where $\boldsymbol{\alpha}_t = \lambda_t\boldsymbol{\beta}$, and the λ_t's are real-valued and unknown, and may vary freely. This method is only sensible when the data are true panel data and not irregular. Without further restrictions on $\boldsymbol{\alpha}_t$, there should be a sufficiently large number of independent clusters at each time t to provide enough information about $\boldsymbol{\alpha}_t$. Figure 6.1 depicts a DAG summarizing the non-homogeneous corresponding dependencies model.

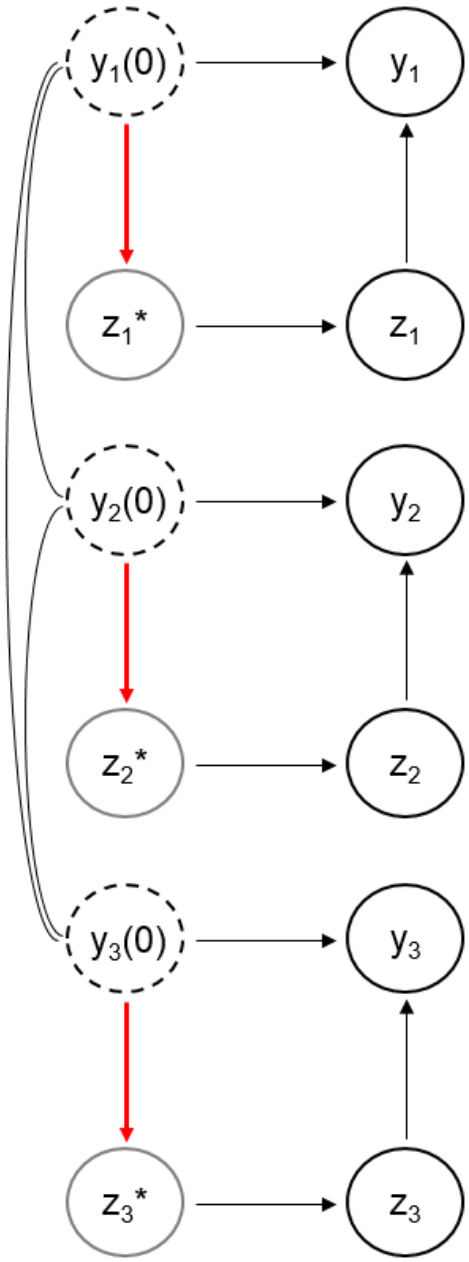

Figure 6.1: Directed Acyclic Graph depicting the setting of homogeneous corresponding dependencies. The lines connecting $y_1(0)$ to z_1^*, $y_2(0)$ to z_2^*, and $y_3(0)$ to z_3^* can be presumed to correspond to equivalent parameters. The association between $\boldsymbol{x}$ and $y(0)$ and between $\boldsymbol{w}$ and z^*, as well as the effect modifiers $\boldsymbol{v}$ are all implied in this figure, although omitted from the graphic to facilitate clarity.

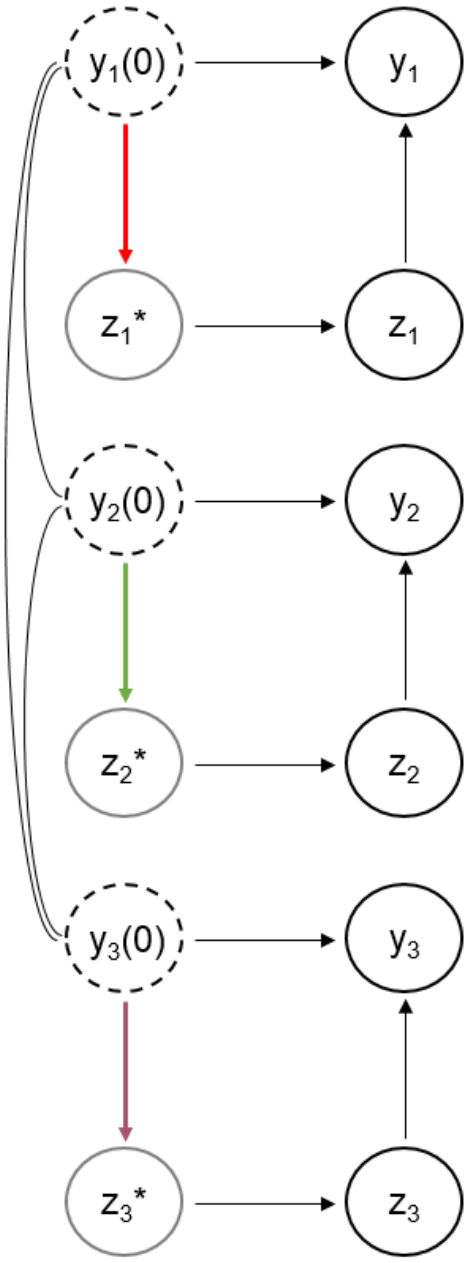

Figure 6.2: Directed Acyclic Graph depicting the setting of non-homogeneous corresponding dependencies. The lines connecting $y_1(0)$ to z_1^*, $y_2(0)$ to z_2^*, and $y_3(0)$ to z_3^* are permitted to correspond to different parameters, so $\lambda_t = \lambda(t)$ is time-varying. The association between $\boldsymbol{x}$ and $y(0)$ and between $\boldsymbol{w}$ and z^*, as well as the effect modifiers $\boldsymbol{v}$ are all implied in this figure, although omitted from the graphic to facilitate clarity.

6.5.3 Homogeneous Telescope Dependence

Now suppose that medication use depends on the history of biomarker values rather than just the biomarker at the corresponding time (that is, z^*_{it} depends not only on $y_{it}(0)$, but also $y_{i,t-1}, \ldots,$ and y_{i1}). In this example, the dependence is assumed to be homogeneous at each particular time point; expanding the term appropriately, we have that

$$\begin{aligned} z^*_{it} &= \boldsymbol{w}^T_{it}\boldsymbol{\alpha} + \sum_{s=1}^{t} \lambda_s y_{is}(0) + \gamma_{it} \\ &= \boldsymbol{w}^T_{it}\boldsymbol{\alpha} + \sum_{s=1}^{t} \lambda_s \boldsymbol{x}^T_{is}\boldsymbol{\beta} + \sum_{s=1}^{t} \lambda_s \epsilon_{is} + \gamma_{it} \\ &= \boldsymbol{w}^T_{it}\boldsymbol{\alpha} + \sum_{s=1}^{t} \boldsymbol{x}^T_{is}\boldsymbol{\alpha}_s + \tilde{\gamma}_{it}, \end{aligned}$$

where $\boldsymbol{\alpha}_s = \lambda_s\boldsymbol{\beta}$. Under this formulation, the dependence of medication on the biomarker history is presumed to be homogeneous over time, though the dependence of observation t on observation t' may be be time-varying. The medication use model for subject i at time t should include covariates $\boldsymbol{w}_{it}$, $\boldsymbol{x}_{i1}$, $\ldots$, and $\boldsymbol{x}_{it}$ only. Figure 6.3 depicts a DAG summarizing the non-homogeneous telescope dependencies model.

6.5.4 Non-homogeneous Telescope Dependence

Finally, we generalize the previous case so that medication use depends on the history of underlying biomarker values in a time-dependent fashion:

$$\begin{aligned} z^*_{it} &= \boldsymbol{w}^T_{it}\boldsymbol{\alpha} + \sum_{s=1}^{t} \lambda_{st} y_{is}(0) + \gamma_{it} \\ &= \boldsymbol{w}^T_{it}\boldsymbol{\alpha} + \sum_{s=1}^{t} \lambda_{st} \boldsymbol{x}^T_{is}\boldsymbol{\beta} + \sum_{s=1}^{t} \lambda_{st} \epsilon_{it} + \gamma_{it} \\ &= \boldsymbol{w}^T_{it}\boldsymbol{\alpha} + \sum_{s=1}^{t} \boldsymbol{x}^T_{is}\boldsymbol{\alpha}_{st} + \tilde{\gamma}_{it}, \end{aligned}$$

where $\boldsymbol{\alpha}_{st} = \lambda_{st}\boldsymbol{\beta}$, $s = 1, \ldots, t$. Under this formulation, the medication use model for subject i at time t should include covariates $\boldsymbol{w}_{it}$, $\boldsymbol{x}_{i1}$, ..., and $\boldsymbol{x}_{it}$ only. As with the non-homogeneous corresponding dependencies method, this method is only sensible in true panel data and demands a sufficiently large number of independent clusters at each time t. Figure 6.4 depicts a DAG in order to summarize the homogeneous telescope dependencies model. Importantly, we assume in each of these methods that the treatment effect at each time point is based only the medication use status at that time, rather than on the entire history of medication use.

6.6 Simulation: Efficiency Gains with Correct Partial Covariance Specification

We conduct a set of simulation studies to elucidate efficiency gains in accounting for correlation in the underlying biomarker. For these studies, we consider a study of $N = 1000$ independent clusters, each with $T_i = 4$ observations, under the homogeneous corresponding dependencies scenario for the longitudinal endogeneity model. Let $\boldsymbol{x}_{1,i}, \boldsymbol{x}_{2,i}, \boldsymbol{x}_{3,i}$ each be independent multivariate normal covariates each with mean zero, common unit variance, and with an exchangeable correlation structure with correlation 0.6 between distinct pairs (where, for example, $\boldsymbol{x}_{1,i} = (x_{1,i1}, \ldots, x_{1,i4})^T$). Suppose that the data generating mechanism for the underlying biomarker is given by

$$\begin{aligned} y_{it}(0) &= 10 + x_{1,it} + x_{2,it} + \epsilon_{it} && (6.14) \\ z^*_{it} &= -0.1 + x_{1,it} + x_{3,it} + \lambda y_{it}(0) + \gamma_{it} && (6.15) \\ &= 1.1x_{1,it} + 0.1x_{2,it} + x_{3,it} + (0.1\epsilon_{it} + \gamma_{it}). && (6.16) \end{aligned}$$

where $\lambda = 0.1$, so that $\boldsymbol{\beta} = (10, 1, 1)^T$ and $\boldsymbol{\alpha} = (0, 1.1, 0.1, 1)$. Set the error variance in the biomarker at $\sigma^2_y = 36$, and let γ_{it} be i.i.d. error terms with unit variance. Let

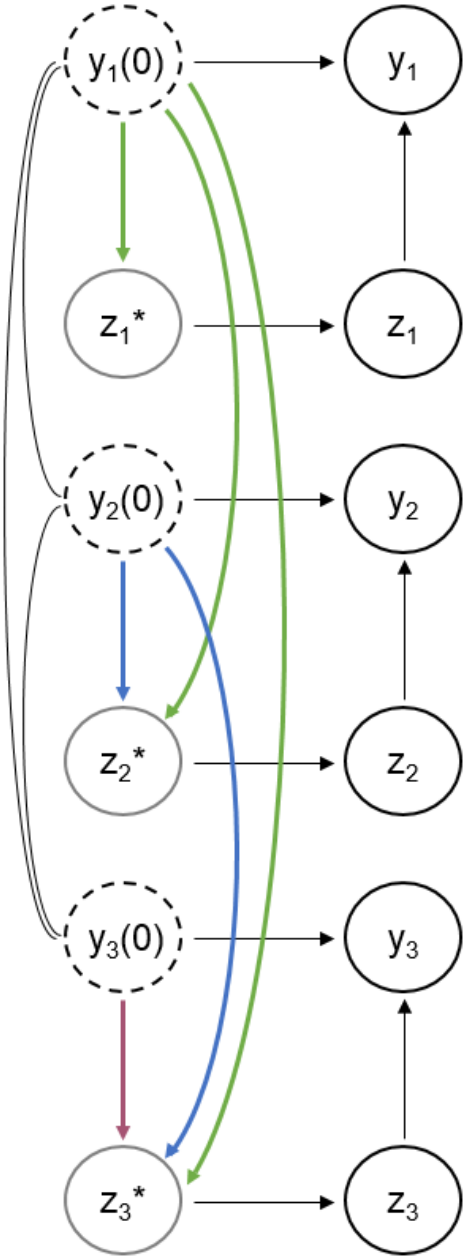

Figure 6.3: Directed Acyclic Graph depicting the setting of homogeneous telescope dependencies. Note that in this case, z_i^* in influenced by all prior values of the underlying biomarker, but $y_{it}(0)$ is presumed to influence $z_{it'}^*$ in the same way for all $t' \geq t$. The association between $\boldsymbol{x}$ and $y(0)$ and between $\boldsymbol{w}$ and z^*, as well as the effect modifiers $\boldsymbol{v}$ are all implied in this figure, although omitted from the graphic to facilitate clarity.

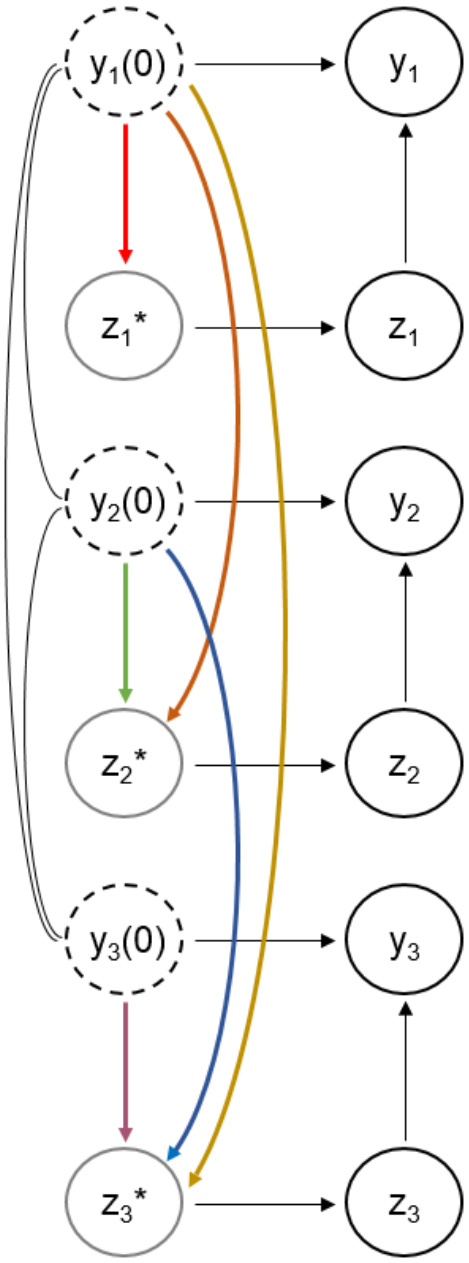

Figure 6.4: Directed Acyclic Graph depicting the setting of non-homogeneous telescope dependencies. Note that in this case, z_i^* in influenced by all prior values of the underlying biomarker, but $y_{it}(0)$ may influence $z_{it'}^*$ differently for each $t' \geq t$. The association between $\boldsymbol{x}$ and $y(0)$ and between $\boldsymbol{w}$ and z^*, as well as the effect modifiers $\boldsymbol{v}$ are all implied in this figure, although omitted from the graphic to facilitate clarity.

$C_i \sim \text{Bernoulli}(p = 0.5)$ denote a medication class variable, so that

$$y_{it} = y_{it}(0) - (0.8 + 0.4 \times C_i) z_{it}. \tag{6.17}$$

In the first setting, we vary ρ_y over a range of values from 0.1 to 0.9 under an exchangeable correlation structure, and fit the longitudinal endogeneity model with working independence, working exchangeable, and working AR-1 structures, all under the homogeneous corresponding dependencies model including x_1, x_2, and x_3 in the medication use model. We compute the bias, Monte-Carlo standard error, and the estimated sandwich standard errors for $\hat{\boldsymbol{\beta}}$ in each model over two-thousand simulation realizations. Results are summarized in Figure 6.5 for estimation of β_1 and β_2. Bias is near zero across all values of ρ_y, consistent with the theoretical findings of Section 6.4. As the correlation increases, the standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ increase for the working independence model, and decrease for the LEM accounting for correlation; the efficiency for the models accounting for correlation relative to the working independence model improves dramatically with higher ρ_y. Of the three models fit, the working exchangeable model (which is correctly specified) is uniformly the most efficient, although the AR-1 model is nearly as efficient. The patterns are the same for estimation of both β_1 and β_2. Of note also is that the robust standard error estimators estimate the Monte-Carlo based standard errors very well in all cases.

Figure 6.6 depicts results when instead the true correlation structure in the biomarkers is AR-1. Similar patterns are observed in this case, although with the AR-1 model providing the most efficiency for estimating $\boldsymbol{\beta}$.

6.7 Simulation: Efficiency Loss when Medication Use is Correlated

In the prior setup, the medication use error terms were uncorrelated over time. Our theoretical results from Section 6.4 suggest that consistent estimates of $\boldsymbol{\beta}$ can be ob-

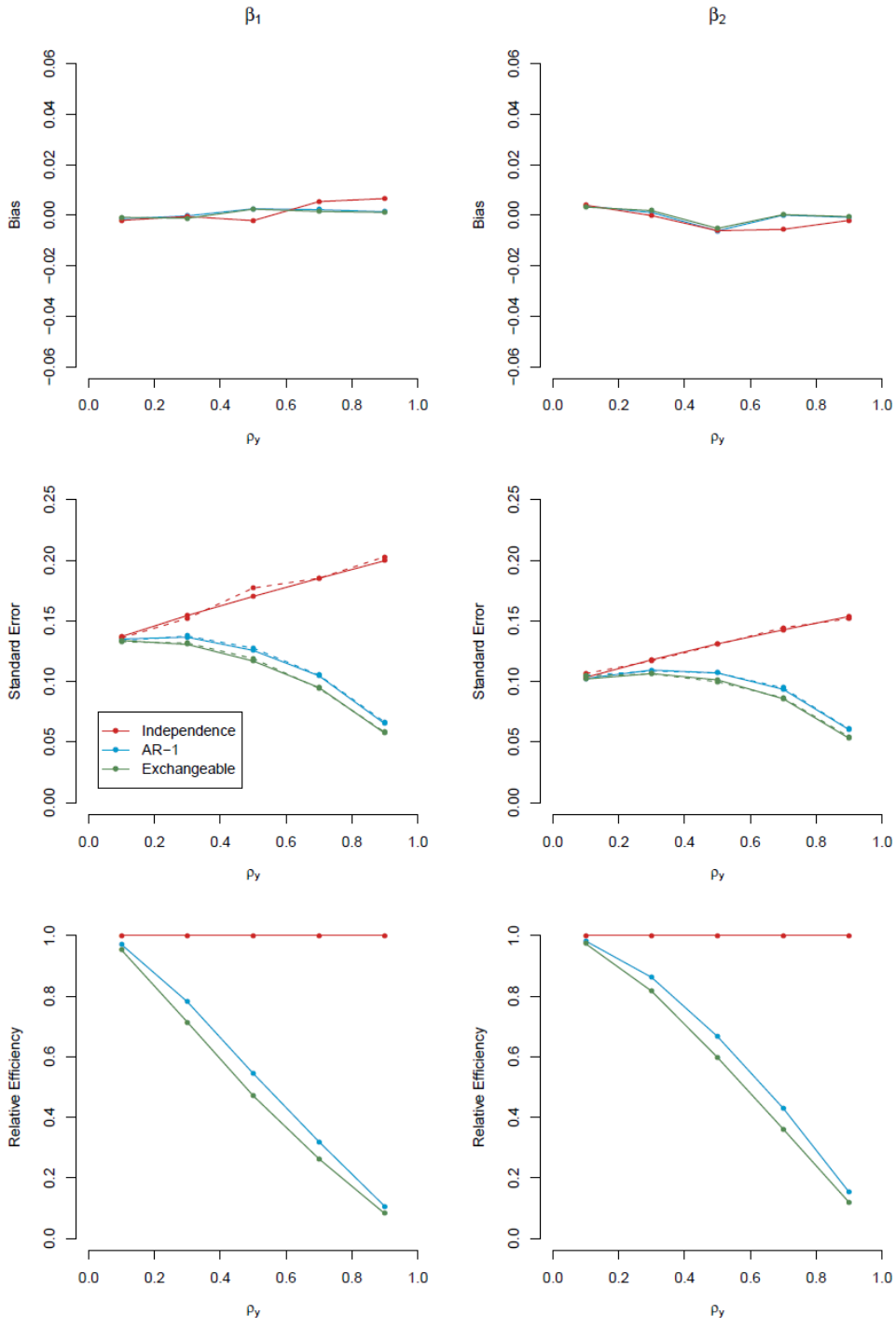

Figure 6.5: Simulation results for estimating β_1 (left) and β_2 (right) under exchangeable structure: On the x-axis is the within-subject correlation in the biomarker, and on the y-axis is bias (top), Monte-Carlo standard error (middle), and efficiency (bottom). In the middle panel, the robust standard error estimates are shown in dashed lines.

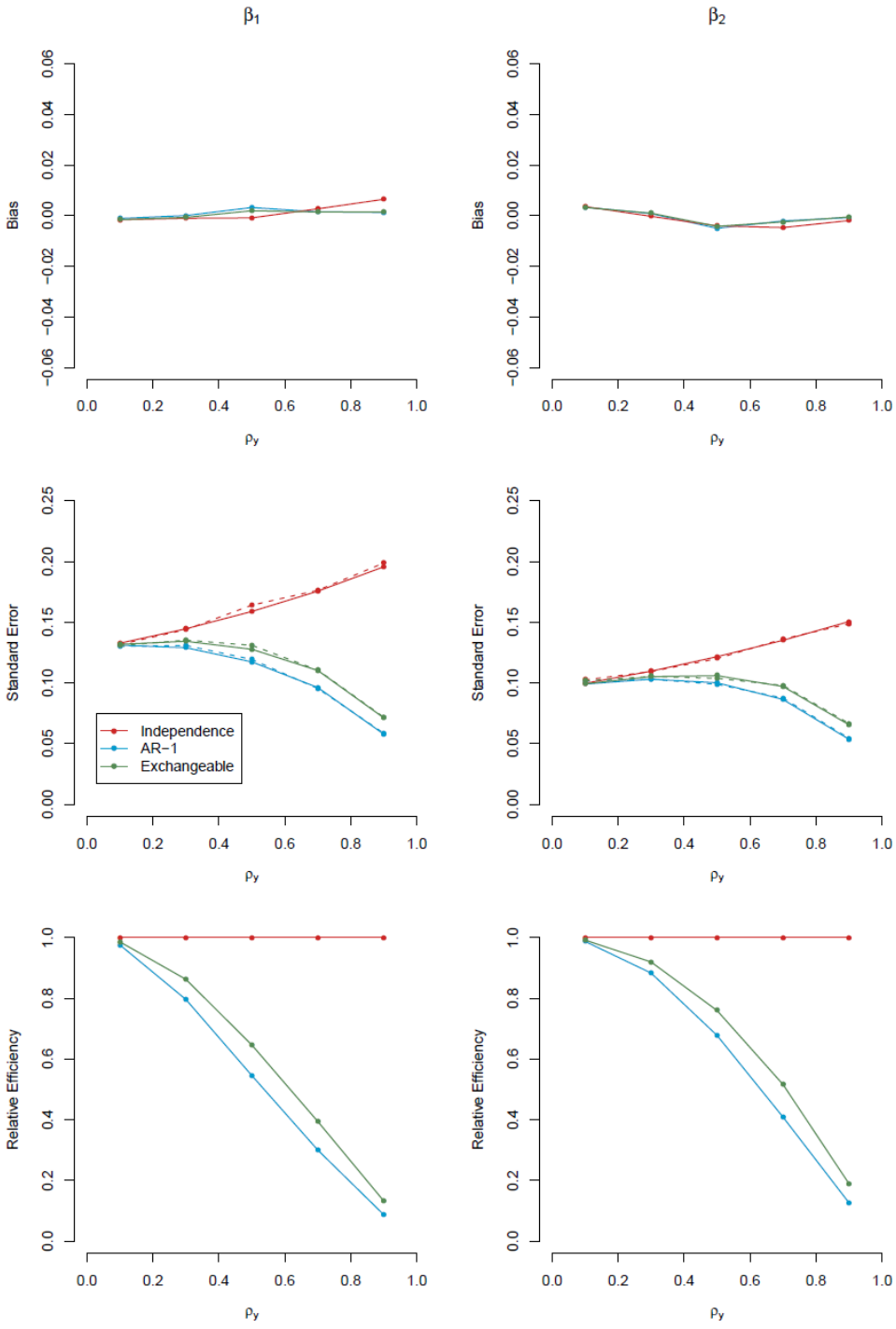

Figure 6.6: Simulation results for estimating β_1 (left) and β_2 (right) under AR-1 structure: On the x-axis is the within-subject correlation in the biomarker, and on the y-axis is bias (top), Monte-Carlo standard error (middle), and efficiency (bottom). In the middle panel, the robust standard error estimates are shown in dashed lines.

tained with the partial specification of correlation structure, even if medication use is correlated over time. In this scenario, we fix $\rho_y = 0.5$ from the previous setup, and alter the within-subject correlation, ρ_z, in the medication use model over a range of values from 0.1 to 0.7 (under an exchangeable structure). Results for estimation of β_1 and β_2 under the working independence, working exchangeable, and working AR-1 models are shown in Figure 6.7 when the biomarker errors in reality have an exchangeable correlation structure. Consistent with what we would expect, bias is near zero for all three approaches. The standard errors are lowest in the exchangeable model, which is consistent with results from the previous study, although for estimation of β_1, the standard errors tend to increase with increasing ρ_z, whereas the standard errors for $\hat{\beta}_2$ appear not to be heavily dependent on ρ_z; this is likely because x_2 in this simulation setup is not a very strong predictor of z_i^*. At low values of ρ_z, there is still a substantial efficiency gain in accounting for the correlation in the biomarker. That efficiency gain weakens as ρ_z increases; the value of ρ_z must be quite high for efficiency gains from the working correlation model to be negated. Of note is also that at correlation levels beyond $\rho_z \approx 0.8$, it is not uncommon to run into convergence issues, although this result is consistent with prior findings in the cross-sectional setting (Marchenko et al., 2012).

Figure 6.8 depicts the analogous results when the true biomarker errors are generated under an AR-1 structure. The major difference in this setting is that there is a very substantial loss of efficiency in using the exchangeable working correlation structure for estimation of β_1 as ρ_z increases. At moderately high values of ρ_z (> 0.55, approximately), we have that the working independence model performs better than the working exchangeable model. The robust variance estimator provide estimates that represent the true repeat-sample variability very well in all cases.

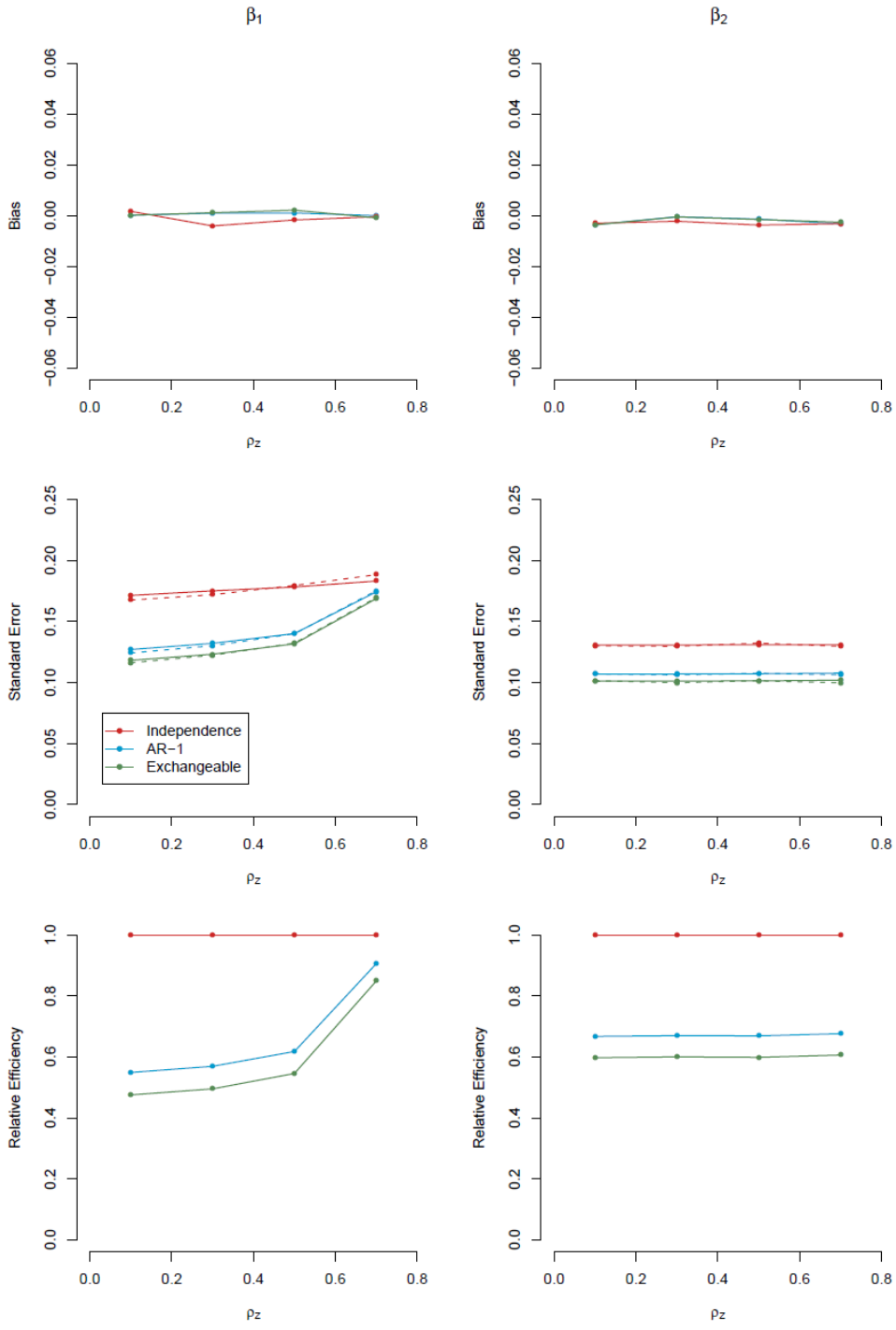

Figure 6.7: Simulation study results: On the x-axis is the within-subject correlation in the biomarker, and on the y-axis is bias (left), standard error (center), and efficiency (right). In the center panel, the robust standard error estimates are shown in dashed lines, and estimate the Monte-Carlo standard errors well.

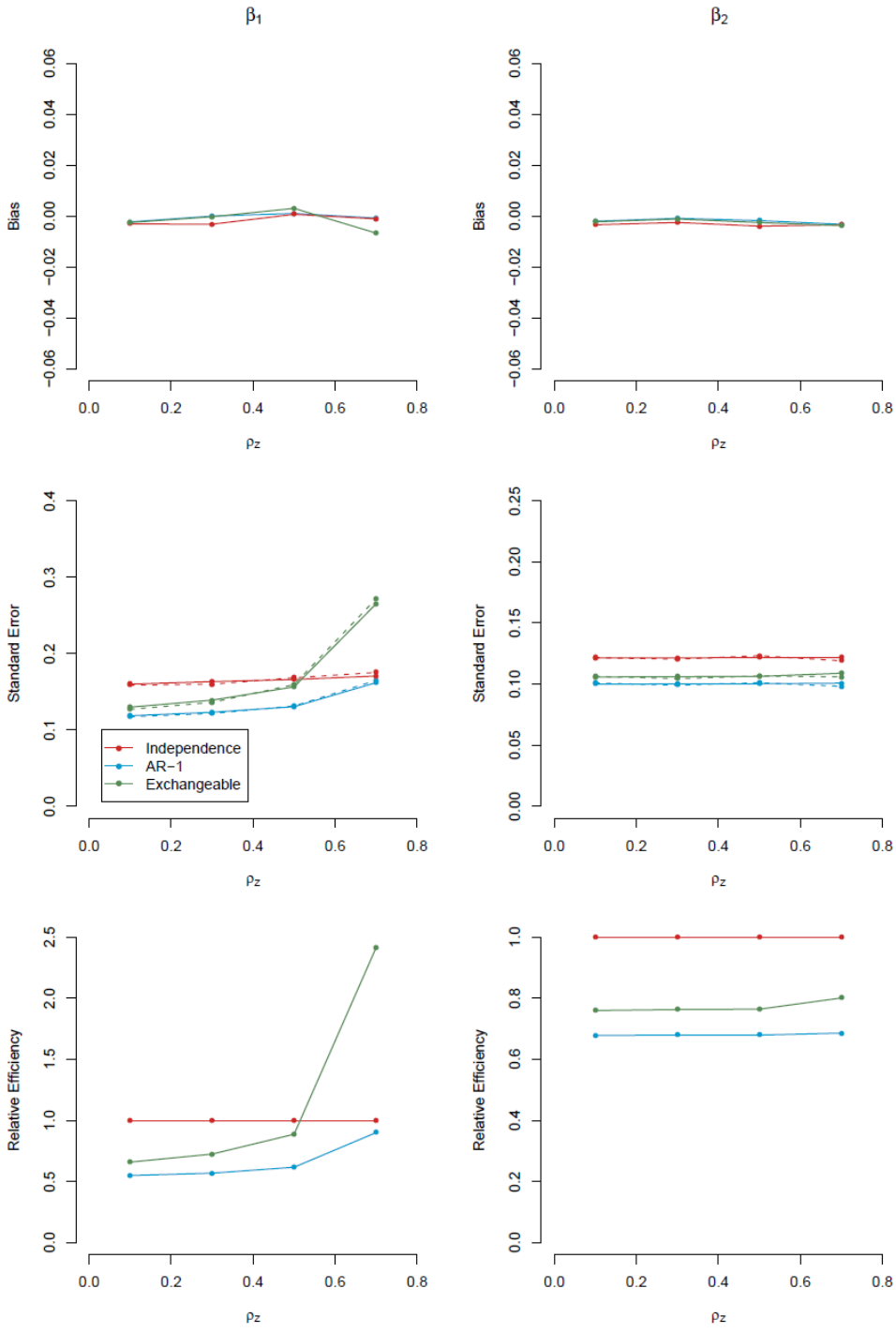

Figure 6.8: Simulation study results: On the x-axis is the within-subject correlation in the biomarker, and on the y-axis is bias (left), standard error (center), and efficiency (right). In the center panel, the robust standard error estimates are shown in dashed lines, and estimate the Monte-Carlo standard errors well.

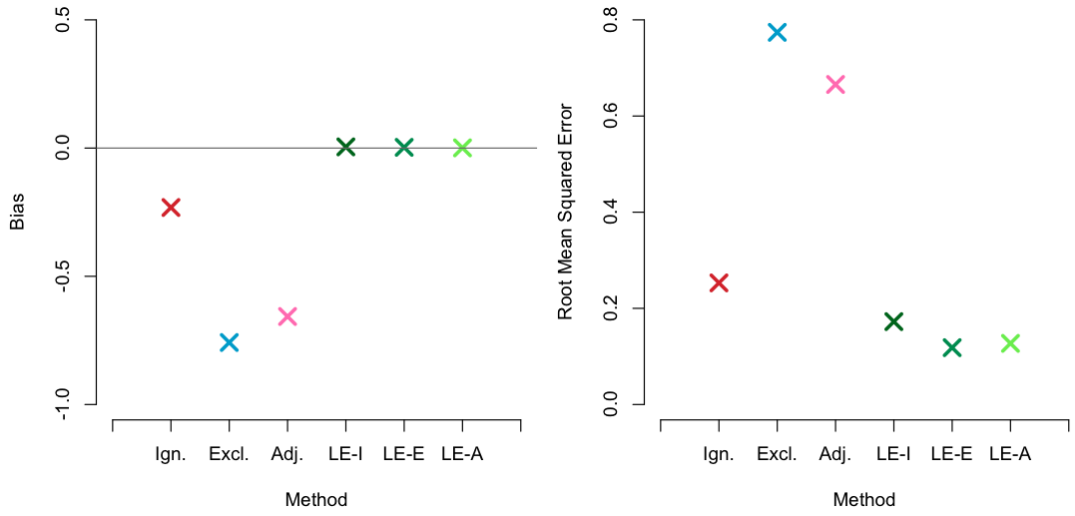

Figure 6.9: Simulation study results: Bias and root mean-squared error for estimation of β_1 for each of the six approaches (three naïve: Ignore, Exclude, and Adjust, and the longitudinal endogeneity model under working independence, exchangeable, and AR-1).

6.8 Simulation: Comparison of LEM to Naïve Approaches

In this section, we conduct a simulation study to compare three versions of the longitudinal endogeneity model (with working independence, exchangeable, and AR-1 correlation structures) to the naïve approaches outlined in Section 6.1 (Ignore, Exclude, and Adjust). We take the simulation setup of Section 6.6, fixing the underlying biomarker correlation, ρ_y, to be 0.5, under the exchangeable correlation. Figure 6.9 depicts results for the estimated bias and simulated root mean squared error over one-thousand simulation replicates.

Consistent with results in Section 6.6, the LEM shows near-zero bias regardless of the choice of working correlation structure. The naïve approaches show substantial bias. Although the each version of the LEM was shown to have higher variability than each of the naïve methods (results not shown), the root mean squared errors for all versions are still superior to those of the naïve models.

6.9 Discussion

Just as in the cross-sectional setting, endogenous medication use acts as a contaminant when estimating natural history associations between predictors and biomarker outcomes in longitudinal data. A subject's natural history of the biomarker value is distorted by the effects of medication use. Since medication users differ from nonusers in their expected underlying off-medication biomarker value, naïve approaches to account for this distortion are not appropriate. Utilizing a working independence models is a simple means of extending the cross-sectional treatment effects model to accommodate clustering while bypassing computational difficulties. Although valid, such methodology misses an opportunity to exploit information gained from within-subject correlation. Indeed, the biomarker values are understood to be correlated within clusters, as is medication use. In addition, medication use at a given time t can be correlated with a biomarker at another time t'. Models that fully specify correlation structure are computationally taxing as they involved multiple integrals that cannot be evaluated in closed form. We have also considered utilizing an inverse probability weighting approach, although studies suggested that the data generation mechanism for the TEM is incompatible with the assumptions of inverse probability weighting.

In this chapter, the model we have given the most attention to can be seen as a compromise between the working independence model and fully parametric specification of a likelihood. We retain the working independence component within the medication use model, but allow for the biomarker to be correlated within clusters according to some working correlation structure. Importantly, this model provides consistent parameter estimates regardless of misspecification of working correlation structure. Simulations indeed confirmed consistency of parameter estimates, and also demonstrated that working correlation structures that are closer to the true data generation mechanism are more efficient. Generally speaking, choosing to model correlation structure in terms of one

single parameter is sufficient for most problems, especially when correct specification is not necessary for consistent estimates.

Medication use may also depend on prior history in a systematic manner. In longitudinal data, there are multiple ways that this may be addressed. We briefly described different ways one might allow for medication use to systematically depend on the underlying biomarker. Non-homogenous dependency is likely only reasonable in the setting of regular, balanced data in which everyone is observed at the approximately the same times. One must be judicious in implementation of these more complex models as they estimate more parameters. A sufficient number of independent subjects should also be available at each observation time to justify the use of compartmentalizing parameters in that way, or identifiability can break down. Moreover, if covariates are approximately time-stable, the simpler homogeneous corresponding dependencies model should generally be adequate. Since we often believe medication use to explicitly depend on the underlying biomarker, we recommend this approach over the longitudinal endogeneity model which only accommodates error correlation and not systematic dependence. This was reflected in our choice of simulation studies and in our application.

We recommend the use of the longitudinal endogeneity model with a sensible choice of working correlation when accounting for endogenous medication use in data with repeated measures over time. Further work is also warranted to increase the flexibility of this model in terms of dependence on covariate history.

6.9.1 Lagged Treatment Effects

The LEM seeks to exploit within-subject correlation in the underlying biomarker. In order for this to be appropriate, we are presuming that treatment effects at each time point are based only on the medication use status at those time points, rather than based on the entire history of medication use statuses. We conducted a follow-up simulation

based on our original simulation setup as in Section 6.6 such that a treatment effect of 0.5 is based on medication use at the concurrent time, and an additional 0.5 is based on medication use at the prior time. We found that the LEM with working independence approach provided low-bias estimates, whereas the LEM with working exchangeable and working AR-1 correlation structures were markedly biased. This could be a justification for using working independence in settings where the effect of medication use change based on duration of use. However, this can also potentially be accounted for by including duration of use as an effect modifier.

Chapter 7

LDL CHOLESTEROL AND LIPID-LOWERING DRUGS: AN APPLICATION TO THE MULTI-ETHNIC STUDY OF ATHEROSCLEROSIS

In this chapter, we illustrate how our methodology can be applied to data from the Multi-Ethnic Study of Atherosclerosis. We compare the results from the models we have developed in this dissertation to alternative approaches. The Multi-Ethnic Study of Atherosclerosis (MESA) is a multi-site cohort study of 6,814 men and women ages 45-84 years. This study was designed to provide insights regarding the prevalence and progression of subclinical cardiovascular disease. Subjects were recruited from six U.S. communities (Baltimore, MD; Chicago, IL; Forsyth County, NC; Los Angeles County, CA; Northern Manhattan, NY; and St. Paul, MN), and were free of clinical cardiovascular disease at entry. The demographic breakdown is as follows: 47% men; 38% white, 28% African-American, 22% Hispanic, and 12% Chinese-American. All subjects provided written informed consent. Subjects were followed for five exams over 10 years. Details of the sampling, recruitment, and data collection have been reported elsewhere (Bild et al., 2002).

We specifically use our motivating example of LDL cholesterol as the outcome of interest. First, we consider the Exam 1 data set as an example of a naturally occurring cross-sectional data set; we consider a simple demographics model to compare several approaches considered in this dissertation. We then consider an age-trend example in LDL from Exams 1 through 5, modeled within each race category.

7.1 A Simple Demographics Model for LDL at Baseline

Using only the Exam 1 data, we consider a demographics model in which we include age, gender, and race category as predictors of interest. At baseline, the prevalence of lipid-lowering drug use was 16.1% (1,072/6,658); the majority of drugs used were statins (988/1,072), but others include fibrates, niacin, and bile-acid resins. Health insurance status and Framingham Risk Score (FRS) are likely strong predictors of medication use. Also note that participants with diabetes are often placed on lipid-lowering drugs (American Diabetes Association, 2004). Differential effects of lipid-lowering drugs across the races would also be consistent with prior findings (Morris and Ferdinand, 2009; Yood et al., 2006). We fit the following six models:

1. "Ignore": OLS linear regression irrespective of medication use.

2. "Exclude": OLS linear regression excluding participants on any lipid lowering drugs.

3. "Adjust": OLS linear regression adjusting for medication use status (binary).

4. "TEM 1": Treatment effects model with health insurance, FRS, and diabetes status in the medication use model.

5. "TEM 2": Treatment effects model, as stated above but also with age, gender, and race category in the medication use model.

6. "SSEM": Subgroup-specific effects model, with medication class (statins vs. other), age, gender, race, and diabetes as potential effect modifiers.

Using a complete-case analysis yields a total of $N = 6,658$ subjects. Results for estimating the association between the predictors of interest and LDL are shown in

Table 7.1: Results from LDL-demographic example in MESA. Presented are the estimates and 95% confidence intervals for all coefficients in the biomarker model from the six approaches considered. Results are expressed as "Estimate [95% CI]".

	Ignore	Exclude	Adjust
Intercept	126.0 [122.9, 129.0]	123.0 [119.8, 126.1]	123.4 [120.4, 126.5]
Age	-0.13 [-0.51, 0.24]	-0.035 [-0.43, 0.36]	-0.051 [-0.43, 0.33]
Female	REF	REF	REF
Male	-1.06 [-2.78, 0.66]	-0.92 [-2.71, 0.88]	-1.05 [-2.75, 0.66]
White	REF	REF	REF
Black	-0.65 [-2.58, 1.27]	-2.25 [-4.27, -0.22]	-0.87 [-2.79, 1.05]
Hispanic	2.35 [0.35, 4.36]	1.07 [-1.01, 3.15]	1.67 [-0.32, 3.66]
Chinese	-2.02 [-4.16, 0.12]	-2.84 [-5.06, -0.63]	-2.59 [-4.71, -0.47]
	TEM 1	TEM 2	SSEM
Intercept	130.4 [125.3, 135.6]	117.1 [111.9, 122.2]	118.6 [113.2, 124.0]
Age	-0.073 [-0.15, 0.0019]	0.16 [0.083, 0.24]	0.15 [0.065, 0.23]
Female	REF	REF	REF
Male	-2.27 [-3.85, -0.69]	-1.01 [-2.64, 0.62]	-1.15 [-2.89, 0.59]
White	REF	REF	REF
Black	-0.19 [-0.21, 1.67]	-1.44 [-3.52, 0.65]	-2.73 [-4.95, -0.51]
Hispanic	1.62 [0.40, 3.64]	-0.096 [-2.30, 2.11]	-0.41 [-2.75, 1.93]
Chinese	-2.96 [-5.23, -0.70]	-4.09 [-6.59, -1.58]	-4.48 [-7.13, -1.83]

Table 7.1. First, we note that the age-LDL association is found to show a negative trend in all naïve approaches, contrary to what we might expect (although none of them is significantly different from zero). The same is true for the TEM 1 model. The TEM 2 and SSEM model, on the other hand, show a significant positive association between age and LDL.

The TEM 1 model is the version in which we only utilize predictors of medication use in the medication use model, and the TEM 2 model is the way we would fit the model if we believed that medication use was influenced by underlying values of LDL directly. If we accept that the TEM 2 and SSEM models provide less biased estimates, this finding suggests that covariates in the biomarker model should be included in the

Table 7.2: Results from LDL-demographic example in MESA. Presented are the estimates and 95% confidence intervals for all coefficients estimating treatment effect parameters from the TEM and SSEM. Results are expressed as "Estimate [95% CI]".

	TEM 1	TEM 2	SSEM
Main Effect	47.2 [43.8, 50.5]	53.9 [50.5, 57.2]	56.1 [40.4, 71.7]
Age	×	×	-0.12 [-0.32, 0.083]
Female	×	×	REF
Male	×	×	-1.23 [-4.91, 2.45]
White	×	×	REF
Black	×	×	-7.76 [-12.3, -3.26]
Hispanic	×	×	-1.26 [-6.73, 4.21]
Chinese	×	×	-1.48 [-7.50, 4.55]
Diabetes	×	×	-0.053 [-4.57, 4.47]
Other LDL Drugs	×	×	REF
Statins	×	×	9.96 [3.83, 15.4]

medication use model if medication use is thought to be influenced by the underlying biomarker. Consistent with this discrepancy is that the association between gender and LDL was only found to be significant in the TEM 1 model, and the difference LDL between Hispanics and whites was found to be significantly different in the TEM 1, but not in the TEM 2. As a follow-up to the Adjustment model, we considered an extended model in which we adjusted for health insurance, FRS, and diabetes and found that the result were very similar to the original Adjusted model.

Results on the association between LDL and race are otherwise are fairly similar across the TEM 1, TEM 2, and SSEM, with the following exception: the SSEM shows a significant difference in mean LDL values between blacks and whites. This is not suggested by the TEM 1 or TEM 2. If we turn our attention to results on the treatment effect parameter estimates for the TEM and SSEM models (Table 7.2), we might be able to account for this discrepancy. Interestingly, the SSEM suggest that the effect of medication use is very different between whites and Blacks (larger than the difference

between any two pairs of groups), with greater lipid lowering appearing in whites than in Blacks. This is the coefficient in the biomarker model that showed the largest difference from from the otherwise similar TEM 2. This is consistent with the results of our simulation studies in Chapter 5, in which the estimate of the natural history association between the predictor and outcome could be misleading if the predictor is also an effect modifier that is not accounted for.

Note that the TEM only provides a single marginal estimate of the treatment effect (Table 7.2). Comparing the results from the TEM 1 and TEM 2, the confidence intervals for the marginal treatment effect difference between the two models barely touch, although we note that the difference in estimates for this particular example is not of great clinical relevance. The estimated effect is consistent with what has been found in previous studies, along with the result that statins produce an estimated 9.96 mg/dL higher-magnitude treatment effect as compared to other drugs such as fibrates, niacin, bile-acid resins, and cholesterol absorption inhibitors (Safeer and Lacivita, 2000).

The robust Wald test for presence of effect modification yields an overall p-value of $p = 0.002$. Testing each individual factor, medication class and race yielded strong evidence of effect modification ($p < 0.001$ and $p = 0.008$, respectively), whereas age, gender, and diabetes status did not achieve significance.

7.2 LDL Age Trends in Longitudinal Data

We now focus on an example to estimate LDL age trends using the MESA data from Exams 1 through 5. We will illustrate two major points. First, we wish to confirm that estimated efficiency gains from modeling correlation structure truly align with those seen from simulation studies in a real-world applied example. Secondly, we wish to demonstrate how this modeling framework improves estimation over the current naïve approaches currently utilized for estimating trends that do not properly address endogenous medication use.

We consider the longitudinal endogeneity model, with homogeneous corresponding dependencies. In the biomarker model, we model the expected LDL with linear and quadratic age terms, stratifying our analysis by race category. In each case, the medication use model includes the following variables: age, Framingham risk score, health insurance status (1 = yes, 0 = no), and presence of diabetes. The medication use indicator was taken to be the use of any lipid lowering drugs, with subgroups defined by statins or other lipid-lowering drugs (fibrates, niacin, bile-acid resins, etc.). We fit both the working independence and working exchangeable versions of this model; results are shown in Table 7.3. The exchangeable correlation structure shows lower standard error estimates for each coefficient in each model, across all races, with efficiency gains ranging from 37% to 49% within each coefficient (Table 7.4). The level of efficiency gain is consistent with the levels observed in simulation studies.

Now, we compare the longitudinal endogeneity model using the working exchangeable structure with three other more traditional approaches: (1) ignoring medication use altogether and fitting a GEE model with a working exchangeable correlation structure, but ignoring medication use altogether, and (2) a similar GEE model excluding any observations from analysis for which participants are on medication, and (3) a similar GEE model adjusting for medication use status. Note that the setup of the longitudinal endogeneity model is the same as in the first example. Table 7.5 presents the results, which are also depicted graphically in Figure 7.1, stratified by race.

This analysis demonstrates why accounting for endogenous medication use is important when estimating natural biomarker trends in longitudinal data. The analysis ignoring medication use altogether leads to the conclusion that underlying LDL trends downward with age. Medication use prevalence increases with age in all races. For example, the prevalence of medication use (for lipid lowering drugs in particular) is 10.4% among those less than 55 years old, 22.9% among those between 55 and 65 years old,

Table 7.3: Results from LDL age trends example in MESA Exams 1-5. Presented are the estimates and 95% confidence intervals for the parameters from the longitudinal endogeneity model under working independence and working exchangeable correlation structures. Results are expressed as "Estimate [95% CI]".

	Whites	Blacks
Independence		
Intercept	117.6 [83.9, 151.3]	102.6 [57.9, 147.3]
Age	0.38 [-0.64, 1.40]	0.65 [-0.72, 2.02]
Age^2	-0.0034 [-0.011, 0.0044]	-0.0052 [-0.016, 0.0054]
Exchangeable		
Intercept	118.2 [91.4, 145.0]	77.8 [45.7, 109.9]
Age	0.19 [-0.61, 0.99]	1.38 [0.40, 2.36]
Age^2	0.00026 [-0.0058, 0.0063]	-0.0096 [-0.017, -0.0020]
	Hispanics	Chinese
Independence		
Intercept	46.5 [2.99, 90.0]	17.0 [-40.4, 74.4]
Age	2.47 [1.14, 3.80]	3.33 [1.57, 5.09]
Age^2	-0.019 [-0.029, -0.0088]	-0.026 [-0.040, -0.012]
Exchangeable		
Intercept	63.4 [30.9, 95.9]	32.9 [-10.8, 76.6]
Age	1.94 [0.94, 2.94]	2.69 [1.32, 4.06]
Age^2	-0.014 [-0.0218, -0.0062]	-0.019 [-0.030, -0.0084]

Table 7.4: Results from LDL age trends example in MESA Exams 1-5. Presented are the estimates of the percentage efficiency gain from using the working exchangeable correlation structure as compared to the working independence structure, in the longitudinal endogeneity model.

	Whites	Blacks	Hispanics	Chinese
Intercept	37%	48%	44%	42%
Age	38%	49%	44%	40%
Age^2	40%	48%	41%	39%

Table 7.5: Results from LDL age trends example in MESA Exams 1-5. Presented are the estimates and 95% confidence intervals for the parameters from the longitudinal endogeneity model under working exchangeable and the longitudinal "Ignore" and "Exclude" approaches. Results are expressed as "Estimate [95% CI]".

Model	Whites	Blacks
Ignore		
Intercept	100.1 [71.9, 128.3]	82.7 [50.2, 115.2]
Age	1.15 [0.29, 2.01]	1.5 [0.50, 2.50]
Age^2	-0.014 [-0.020, -0.0075]	-0.015 [-0.023, -0.0074]
Exclude		
Intercept	58.8 [30.2, 87.4]	43.5 [12.9, 74.1]
Age	2.09 [1.21, 2.97]	2.44 [1.48, 3.40]
Age^2	-0.017 [-0.024, -0.010]	-0.019 [-0.026, -0.012]
Endogeneity		
Intercept	118.2 [91.3, 145.0]	77.8 [45.7, 109.9]
Age	0.19 [-0.61, 0.99]	1.38 [0.40, 2.36]
Age^2	0.00026 [-0.0058, 0.0063]	-0.0096 [-0.0172, -0.0020]
	Hispanics	Chinese-American
Ignore		
Intercept	62.6 [26.3, 98.9]	36.0 [-10.3, 82.3]
Age	2.37 [1.25, 3.49]	2.96 [1.53, 4.39]
Age^2	-0.024 [-0.032, -0.016]	-0.027 [-0.038, -0.016]
Exclude		
Intercept	24.8 [-8.52, 58.1]	-16.6 [-60.50, 27.30]
Age	3.28 [2.24, 4.32]	4.32 [2.93, 5.71]
Age^2	-0.027 [-0.035, -0.019]	-0.034 [-0.045, -0.023]
Endogeneity		
Intercept	63.4 [30.9, 95.9]	32.9 [-10.8, 76.6]
Age	1.94 [0.94, 2.94]	2.69 [1.32, 4.06]
Age^2	-0.014 [-0.022, -0.0062]	-0.019 [-0.030, -0.0084]

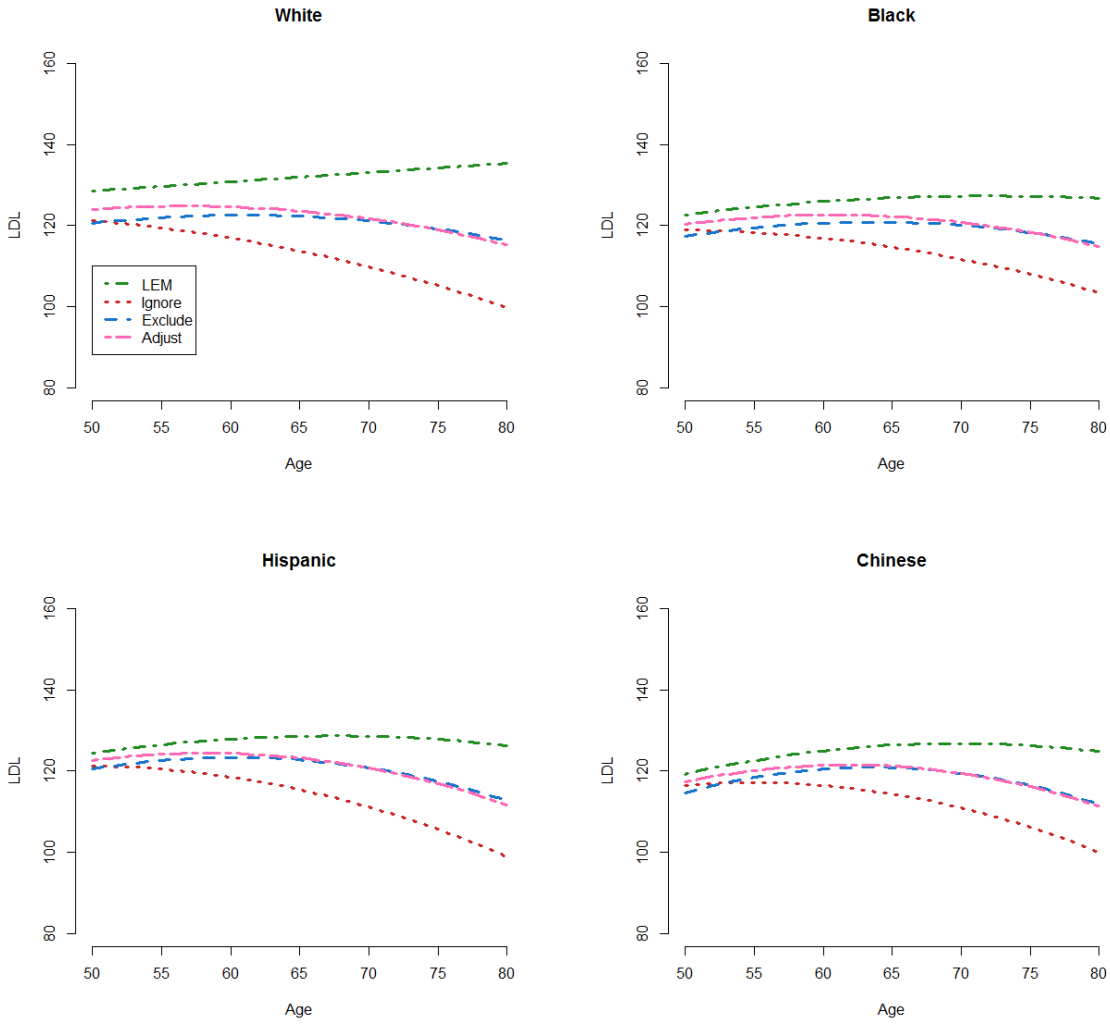

Figure 7.1: Application to MESA showing quadratic models for age-LDL association, stratified by race.

32.5% among those between 65 and 75 years old, 35.1% among those between 75 and 85 years old, and 38.6% among those older than 85 years old. Hence, the finding that ignoring medication use results in an estimated downward trend is not surprising. Excluding all observations for which participants are on medication suggest that over the ages of 50-75, LDL is either slightly increasing or approximately stable, and then takes a more steep downward decline. Aside from the loss of efficiency incurred by excluding a sixth to a third of the data from analysis, this is inappropriate since medication use is not at occurring random.

The longitudinal endogeneity model suggests that underlying LDL values are either increasing over the ages (in whites) or approximately stable (African-Americans, Hispanics, and Chinese-Americans). The minor drop-off in the latter three groups may be explainable by a deaths in these groups (where death occurs at a higher rate in individuals with high underlying LDL). Another possibility is that these three groups are more prone to experience signs of liver failure at very high ages, in which case LDL is understood to decrease.

This analysis also suggests that the longitudinal endogeneity model is generally as efficient as the alternative approaches. This finding is in contrast to the pattern seen in cross-sectional data, in which the treatment effects model which shows markedly less efficiency. In longitudinal data, coefficients' standard errors generally decrease with increasing within-subject correlation. The naïve estimate of within-subject correlation obtained from ignoring medication use is expected to be an underestimate, since many participants' biomarker values are altered by medication throughout the course of the study, creating more variability in their observed trajectories. Thus, while the longitudinal endogeneity model estimates more parameters, it provides a higher within-subject correlation estimate that more accurately reflects correlation in the underlying biomarker. In the Exclude method, only off-medication observations are considered,

and so one might expect the within-subject correlation of the observed off-medication values to be higher than considering all observed values. Hence the loss of efficiency induced by removing part of the data can be partially balanced out with a higher within-subject correlation.

As is the case in age-trend analyses, loss to follow up and death fundamentally changes the estimand; when we talk about the natural history association, we are forced to restrict our attention to the natural history conditional on having survived to the observed ages, as opposed to the natural history association is the association that would have been observed had all participants survived.

Our application to LDL data from MESA provided results that were consistent with both our simulation studies, and prior results in cross-sectional data. Accounting for correlation appears to improve efficiency across every parameter, in comparison to the working independence model. The naïve general linear model approaches show what we believe to be a severe downward bias; this finding in particular is consistent with a comparison of the cross-sectional treatment effects model with ordinary least squares approaches. The fact that the longitudinal endogeneity model provides estimates that are approximately consistent with prior understood results has important implications of its potential utility going forward. For instance, in the setting of novel inflammatory markers that are not fully understood, our model provides a way to target its natural underlying association with predictors of interest.

7.3 Discussion

Our applications to MESA provides results that are consistent with our simulation studies: (a) coefficients corresponding to predictors that strongly modify treatment effect are likely to be the most biased if effect modification is ignored, and (b) the estimates are ostensibly less biased when endogeneity is adequately accounted for (this is suggested by the finding LDL was found to have a positive trend with age in the TEM

and SSEM, whereas the simpler approaches failed to confirm this expected result). We note that the finding of differential treatment effects in Blacks is not necessary a result of an inherent biologic characteristic–the models considered are incapable of providing evidence to support or reject this hypothesis. This result could potentially be an artifact of lower adherence rates in this subgroup, for example. Although the source of this difference is unclear, the importance of accounting for race category as a potential effect modifier is not any less important. If the result is in fact simply an artifact of lower adherence, then failure to account for effect modification would still result in systematically over-correcting the biomarker for the effects of medication use in Black participants–the treatment effect is understood to be smaller if adherence rates are lower.

Chapter 8
DISCUSSION AND FUTURE DIRECTIONS

This dissertation has focused on estimation of associations between predictors and biomarker outcomes in the presence of endogenous medication use. Treating the effect of medication use on the biomarker as a contaminant, we were able to devise a set of methodology with a reasonable set of assumptions in order to recover the natural history of the biomarker that would have been observed in the absence of medication use.

First, we focused our attention on cross-sectional methodology; there is a wide body of literature in existence seeking to address this very challenge, although (a) most of the applied work in the literature appears to ignore this methodology, and (b) even if applied, these simple modifications to naïve approaches appear to be inadequate. The TEM, as proposed by James Heckman, provided us with a framework in which we could estimate the natural history association. We have shown that this model is fairly robust to departures from several of its main assumptions.

Although the TEM is indeed able to withstand departures from the assumption of bivariate normal errors for the purpose of estimating the natural history association (a result that has not been previously understood), the presence of effect modification can result in biased estimates of the natural history association. We have devoted a substantial amount of time to addressing this challenge by allowing the effects of medication use to systematically vary in expectation with observed covariates, regardless whether the effect modifiers are associated with the underlying biomarker, medication use status, or neither. Indeed, it turns out that the former two cases are the cases

with the most potential for bias reduction, although identifiability of effect modifier parameters is a welcome feature of the model.

Methodology to extend the TEM framework to accommodate repeated measures has been largely unexplored. The sparse literature on related selection models simply suggests the use of a working-independence type model, which can be very helpful in settings where the full-covariate conditional mean assumption is not thought to hold. However, this working-independence model has not been formally presented in the literature, and efficiency gains can be very substantial by accounting for correlation in the underlying biomarker when the full-covariate conditional mean assumption is thought to hold. Our research has focused primarily on addressing the latter of these two concerns in Chapter 6.

We would additionally like to comment further on the importance of distinguishing between settings in which the observed biomarker or the natural history of the biomarker is of interest. For scientific and clinical problems involving prediction of further adverse cardiac events, the observed biomarker could very well of greater relevance than the hypothetical off-medication value for treated participants. However, when we seek to estimate how groups of individuals differing in primordial or long-term predictors (such as age, race, or a genetic exposure) differ in their biomarker outcomes, consideration of the natural history is of greater scientific relevance.

Consider, for instance, the setting of cardiovascular biomarkers such as systolic blood pressure and LDL cholesterol. These are typically the target of reduction in participants with high underlying values. If the medication taken is effective in (at least partially) restoring participants' values to the values of healthy participants, failure to account for medication use when selecting a model will, in general, result in an attenuated estimate of the association of interest. If medication use is prevalent and sufficiently effective, one might even be led to the erroneous conclusion age is not associated with cardiovascular

biomarkers (as seen in our application to MESA). If the goal of the study is to identify important predictors of biomarkers, ignoring medication use will hence lead to severely under-powered study designs.

8.1 Future Directions

In this section, we discuss two future extensions to this research that could be of interest.

8.1.1 Heavy-Tailed Errors

In Chapter 4, we found that the TEM was not very sensitive to heavy tailed errors for estimating the natural history association (although the estimate of the marginal treatment effect was fairly sensitive). However, convergence issue arose when the tails were sufficiently heavy (i.e., when the degrees of freedom was sufficiently low).

Prior work has examined an extension of Heckman's TEM to allow for heavy tailed errors by modeling the error terms with a bivariate t-distribution (Marchenko et al., 2012). This could be worth exploring for the purposes of estimating the natural history association to evaluate whether the natural history association can be estimated well, with few convergence issues, if the data have heavier-tailed errors than the bivariate normal distribution.

8.1.2 Finite Mixture Modeling

Recently, there has been growing interest in understanding age trends when it is believed that there are multiple trajectory classes from which participants may originate. Specifically, it can be of interest to determine whether a finite number of latent trajectory classes can be identified and estimated. Moreover, we may wish to determine whether membership to these classes is associated with observable covariates or risk of subsequent adverse events. Nagin (1999) described a latent class mixture modeling (LCMM) framework that can be used to directly estimate the curves and to select the appro-

priate number of latent classes, typically by the Bayesian Information Criterion (BIC). Such an approach is also capable of linking class membership to time-stable covariates. Additionally, one may also predict subject-specific posterior probability estimates of belonging to these latent classes. Predicted probability classes may also be used to associate class membership with risk of subsequent adverse events. Although LCMM approach has received attention in recent years as an attempt to determine underlying biomarker trajectories (most notably SBP), the challenge of endogenous medication use is one that has largely been handled inappropriately or ignored altogether in the literature (Loucks et al., 2011; de Groot et al., 2014; Allen et al., 2014; Wills et al., 2011). Nagin motivates his work in understanding developmental trajectory classes with examples from youth criminal involvement and social behavior. For the data used in the original literature, it is unlikely that endogenous medication use severely compromises estimation of associations. Although, given the rise of psychotropic medication use in children, endogeneity may be a more substantial challenge in this setting for modern data (Leslie et al., 2010).

If a participant is on medication at certain points in the study, his or her observed trajectory is either completely unobservable (participants on medication for the entire study) or distorted (participants who commence and/or cease medication use). The challenges we discussed regarding endogeneity of medication use and effect modification extend to this setting. Simple approaches such as adjusting for medication use and excluding on-medication participants from analysis easily generalize to this setting. Just as in the cross-sectional setting, endogeneity renders these modifications insufficient for estimating underlying trajectories in the sense that the trajectories are distorted, and as a result: (i) the number of underlying curves can be inappropriately selected, (ii) class membership can fail to be appropriately linked to observable covariates or risk of subsequent events. Thus, the potential pitfalls of simple approaches for LCMM can

be more complicated and difficult to characterize than in the cross-sectional setting, in which bias and efficiency of estimators are the two primary characteristics under consideration. Of note is that both bias and efficiency are simple enough to characterize empirically, and sometimes even simple enough to characterize analytically in closed-form. Describing the distortion of LCMM curves quantitatively can be challenging. It is of interest to utilize the models developed in the cross-sectional setting in order to extend the current latent class mixture modeling approach. Such an approach allows estimation of underlying trajectory curves when there is endogenous medication use in the population.

8.1.3 Multiple Classes of Medication

In our methodology, we have focused on examples in which medication use z_i was taken to be 1 if on any medication and 0 if not, generally ignoring the fact that many times participants are on multiple medications simultaneously. In Chapter 5, we were able to accommodate this by including effect modifier parameters. However, it could be worthwhile to explore other ways in which multiple classes of medication use may be accommodated. In the case where there are two classes of medication, one might be able to use two medication use equations, each with a distinct medication use variable (z_{1i}^* and z_{2i}^*, for example). Then, the error terms in the biomarker model, together with those in the latent medication use model could be taken to be a trivariate normal distribution. One might be concerned in this case about identifiability of parameters, and perhaps it would be necessary to impose certain restrictions on the total 3×3 covariance matrix for each subject in order to achieve identifiability.

Another potential way one might be able to accommodate multiple classes of medication use would be to split up the single latent medication use equation into classes based on a more flexible (perhaps data-driven) partition of the real line rather than

the dichotomized version as in the TEM. This might be more appropriate if it is believed that a higher risk covariate profile is more likely to result in being placed on one medication over another. Otherwise, a continuous ordering may be inappropriate.

8.1.4 Estimating Natural Quantiles

This dissertation has focused primarily on regression techniques for specific types of association studies in the presence of endogenous medication use. In practice, we also could want the natural history of the biomarker for other purposes; for example, we may wish to establish a normal range of underlying biomarker values, which cannot be established in a straightforward way with the naïve data. Methods such as LMS regression (Cole and Green, 1992) can be used to normalize an outcome and perform quantile regression based on the Box-Cox transformation. It would be of interest to evaluate whether this technique could be carried over with the likelihood of the treatment effects model and the extensions we have proposed.

8.2 Concluding Remarks

In this research, we have argued and illustrated through simulation and application the failure of traditional methods in estimating the natural history association between predictors and underlying biomarkers in observational data. We have taken the existing structural equation framework presented by James Heckman in 1978, which has been historically used for estimating treatment effects, and applied it in a novel setting.

Our findings revealed that when this modeling framework is applied to estimate natural history associations, there is a high level of robustness to departures from assumptions, including distributional assumptions on the error terms. This is not the case when the marginal treatment effect is the estimand of interest. The observed sensitivity to departures from the assumption of uniform treatment effects was addressed by explicitly accounting for effect measure modification. This extension was found to

achieve substantial bias reduction when there are systematic differences in the expected treatment effect magnitude across values of the predictor of interest, or predictors of medication use. Moreover, in the setting where repeated measures are available over time, we have presented alternative model formulations in order to gain efficiency in estimating longitudinal biomarker trends.

Throughout this research, we have identified a number of future pathways and avenues for extending these methods to further improve estimation of associations in the presence of endogenous medication use.

BIBLIOGRAPHY

Allen NB, Siddique J, Wilkins JT, Shay C, Lewis CE, Goff DC, Jacobs DR Jr., Liu K, Lloyd-Jones D. Blood pressure trajectories in early adulthood and subclinical atherosclerosis in middle age. *The Journal of the American Medical Association* 2014; 311(5): 490-497.

Amemiya T. Tobit models: A survey. *Journal of Econometrics* 1984; 24(1-2): 3-61.

American Diabetes Association (Position Statement). Dyslipidemia management in adults with diabetes. *Diabetes Care* 2004; 27: S68–S71.

Angrist JD, Krueger AB. Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives* 2001; 15(4): 69-85.

Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacobs, Jr. DR, Kronmal R, Liu K, Nelson JC, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-ethnic study of atherosclerosis: objectives and design. *American Journal of Epidemiology* 2002: 156(9): 871–881.

Bowden RJ, Turkington DA. *Instrumental Variables.* Cambridge: Cambridge University Press; 1984.

Brand E, Wang JG, Herrmann SM, Staessen JA. An epidemiological study of blood pressure and metabolic phenotypes in relation to the Gbeta3 C825T polymorphism. *Hypertention* 2003; 21(4): 729-737.

Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Instrumental variable analysis of secondary pharmacoepidemiologic data. *Epidemiology* 2006; 17(4), 373-374.

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *American Journal of Epidemiology* 2006; 163(12), 1149-1156.

Carroll MD, Lacher DA, Sorlie PD, Clean JI, Gordon DJ, Holz M, Grundy SM, Johnson CL. Trends in serum lipids and lipoproteins of adults 1960-2002. *The Journal of the American Medical Association* 2005; 294(14), 1773–1781.

Chen Y, Chen Y, Li X, Post W, Herrington D, Polak J, Rotter J, Taylor K. The HMG-CoA reductase gene and lipid and lipoprotein levels: the Multi-Ethnic Study of Atherosclerosis. *Lipids* 2009 44(8): 733-743.

Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL, Jones DW, Materson BJ, Oparil S, Write JT Jr., Roccella EJ. Seventh report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension* 2003; 42(6): 1206-1252.

Choesterol Treatment Trialists' Collaborators. Efficacy of cholesterol-lowering therapy in 18 686 people with diabetes in 14 randomised trials of statins: a meta-analysis. *Lancet* 2008; 371(9607): 117-125.

Cole TJ, Green PJ. Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine* 1992; 11(10): 1305–1319.

Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; 168(6): 656-664.

de Groot S, Post MW, Hoekstra T, Valent LJ, Faber WX, van der Woude LH. Trajectories in the course of body mass index after spinal cord injury. *Archives of Physical Medicine and Rehabilitation* 2014; 95(6): 1083-1092.

Delaney JAC, Platt RW, Suissa S. The impact of unmeasured baseline effect modification on estimates from an inverse probability of treatment weighted logistic model. *European Journal of Epidemiology* 2009; 24(7): 343-349.

Diggle P, Heagerty P, Liang K, Zeger, S. *Analysis of longitudinal data.* New York: Oxford University Press.

Freedman DA, Sekhon JS. Endogeneity in probit response models. *Political Analysis* 2010; 18(2): 138-150.

Gurven M, Blackwell A, Rodríguez D, Stieglitz J, Kaplan H. Does blood pressure inevitably rise with age?: Longitudinal evidence among forager-horticulturalists. *Hypertension* 2012; 60(1), 25-33.

Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 1976; 5(4): 475-492.

Heckman JJ. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 1978; 46(4): 931-959.

Heckman J. Sample selection bias as a specification error. *Econometrica* 1979; 40(1): 153-161.

Heckman JJ, Urzua S, Vytlacil E. Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics* 2006; 88(3), 389-432.

Hernán MA, Hernández-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; 15(5): 615-625.

Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006; 17(4): 360-372.

Huynh Q, Blizzard C, Sharman J, Magnussen C, Dwyer T, Venn A. The cross-sectional association of sitting time with carotid artery stiffness in young adults. *BMJ Open* 2014; 4(3): e004384. doi:10.1136/bmjopen-2013-004384.

Iwai N, Baba S, Mannami T, Ogihara T, Ogata J. Association of sodium channel gamma-subunit promoter variant with blood pressure. *Hypertension* 2001; 38(1): 86-89.

Jorgensen NW, Sibley CT, McClelland RL. Using imputed pre-treatment cholesterol in a propensity score model to reduce confounding by indication: results from the multi-ethnic study of atherosclerosis. *BMC Medical Research Methodology* 2013; 13: 81.

Kramer H, Han C, Post W, Goff D, Diez-roux A, Cooper R, Jingouda S, Shea S. Racial/ethnic differences in hypertension and hypertension treatment and control in the multi-ethnic study of atherosclerosis (MESA). *American Journal of Hypertension* 2004; 17(1): 963-970.

Lefebvre G, Delaney JAC, Platt RW. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine* 2008; 27(18): 3629-3642.

Leslie L, Raghavan R, Zhang J, Aarons G. Rates of psychotropic medication use over time among youth in child welfare/child protective services. *Journal of Child and Adolescent Psychopharmacology* 2010; 20(2): 135-143.

Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73(1), 13-22.

Loucks EB, Abrahamowicz M, Xiao Y, Lynch JW. Associations of education with 30 year life course blood pressure trajectories: Framingham Offspring Study. *BMC Public Health* 2011; 11: 139.

Maddala GS. *Limited-dependent and qualitative variables in econometrics.* Cambridgeshire: Cambridge University Press; 1983.

Marchenko YV, Genton MG. A Heckman selection-t model. *Journal of the American Statistical Association* 2012; 107(497): 304–317.

Matsubara M, Kikuya M, Ohkubo T, Metoki H, Omori F, Fujiwara T, Suzuki M, Michimata M, Hozawa A, Katsuya T, Higaki J, Tsuji I, Araki T, Ogihara T, Satoh H, Hisamichi S, Nagai K, Kitaoka H, Imai Y. Aldosterone synthase gene (CYP11B2) C-334T polymorphism, ambulatory blood pressure and nocturnal decline in blood pressure in the general Japanese population: the Ohasama Study. *Journal of Hypertension* 2001; 19(12): 2179-2184.

McClelland RL, Kronmal RA, Haessler J, Blumenthal RS, Goff DC Jr. Estimation of risk factor associations when the response is influenced by medication use: an imputation approach. *Statistics in Medicine* 2008; 27(24): 5039-5053.

Morris A, Ferdinand KC. Hyperlipidemia in racial/ethnic minorities: Differences in lipid profiles and the impact of statin therapy. *Clinical Lipidology* 2009; 4(6), 741-754.

Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, De Ferranti S, Després J, Fullerton HJ, Howard VJ, Huffman MD, Judd SE, Kissela BM, Lackland DT, Lichtman JH, Lisabeth LD, Liu S, Mackey RH, Matchar DB, McGuire DK, Mohler ER, Moy CS, Muntner P, Mussolino ME, Nasir K, Neumar RW, Nichol G,

Palaniappan L, Pandey DK, Reeves MJ, Rodriguez CJ, Sorlie PD, Stein J, Towfighi A, Turan TN, Virani SS, Willey JZ, Woo D, Yeh RW, Turner MB. Executive summary: Heart disease and stroke statistics-2015 update: A report from the American Heart Association. *Circulation* 2015; 131(4): 434-441.

Nagin DS. Analyzing developmental trajectories: a semiparametric, group-based approach. *Pyschological Methods* 1999; 4(2): 139-157.

National diabetes statistics report, 2014. *Medical Benefits* 2015; 31(14): 3-4.

Nwankwo TS, Yoon S, Burt V, Gu Q. Hypertension among adults in the United States: National Health and Nutrition Examination Survey, 2011-2012. *NCHS Data Brief* 2013; 133: 1-8.

O'Donnell CJ, Lindpaintner K, Larson MG, Rao VS, Ordovas JM, Schaefer EJ, Myers RH, Levy D. Evidence for association and genetic linkage of the angiotensin-converting enzyme locus with hypertension and blood pressure in men but not women in the Framingham heart study. *Circulation* 1998; 97(18): 1766-1772.

Pearl J. Causal inference in statistics: a review. *Statistical Surveys* 2009; 3: 96-146.

Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation* 1994; 24(4): 939–951.

R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing: 2013. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rice T, Rankinen T, Province MA, Chagnon YC, Pérusse L, Borecki IB, Bouchard C, Rao DC. Genome-wide linkage analysis of systolic and diastolic blood pressure: the Quebec family study. *Circulation* 2000; 102(16): 1956-1963.

Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; 11(5): 550-560.

Safeer RS, Lacivita CL. Choosing drug therapy for patients with hyperlipidemia. *American Family Physician* 2000; 61(11), 3371-3382.

Schunkert H, Hense HW, Döring A, Riegger GA, Siffert W. Association between a polymorphism in the G protein beta3 subunit gene and lower renin and elevated diastolic blood pressure levels. *Hypertension* 1998; 32(3): 510-513.

Sethi AA, Nordestgaard BG, Tybjaerg-Hansen A. Angiotensinogen gene polymorphism, plasma angiotensinogen, and risk of hypertension and ischemic heart disease: a meta-analysis. *Arteriosclerosis Thrombosis and Vascular Biology* 2003; 23(7): 1269-1275.

Singh GM, Danaei G, Pelizzari PM, Lin JK, Cowan MJ, Stevens GA, Farzadfar F, Khang Y, Lu Y, Riley LM, Lim SS, Ezzati M. The age associations of blood pressure, cholesterol, and glucose: Analysis of health examination surveys from international populations. *Circulation* 2012; 125(8), 2204-2211.

Spieker AJ, Delaney, JAC, McClelland, RL. Evaluating the treatment effects model for estimation of cross-sectional associations between risk factors and cardiovascular biomarkers influenced by medication use. *Pharmacoepidemiology and Drug Safety* 2015; 24(12), 1286-1296.

Spieker AJ, Delaney, JAC, McClelland, RL. A method to account for treatment effect measure modification when estimating cross-sectional associations between biomarkers and risk factors. *In submission.*

StataCorp. Stata Statistical Software: Release 2013. College Station, TX: Stata-Corp LP.

Tate RF. Correlation between a discrete and a continuous variable. Point-biserial correlation. *Annals of Mathematical Statistics* 1954; 25(3): 603-607.

Telser LG. Iterative estimation of a set of linear regression equations. *Journal of the American Statistical Association* 1964; 59(307): 845-862.

Tobin J. Estimation of relationships for limited dependent variables. *Econometrica* 1958; 26(1): 24-36.

Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in Medicine* 2005; 24(19): 2911-2935.

Wang C, Dominici F, Parmigiani G, Zigler CM. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics* 2015; 71, 654-665.

Wang Y, Fang Y. Adjusting for treatment effect when estimating or testing genetic effect is of main interest. *Journal of Data Science* 2011; 9: 127-138.

Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* 1943; 54, 426-482.

White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; 48(4): 817-838.

Wills AK, Lawlor DA, Matthews FE, Aihie Sayer A, Bakra E, Ben-Shlomo Y, Benzeval M, Brunner E, Cooper R, Kivimaki M, Kuh D, Muniz-Terrera G, Hardy R. Life course trajectories of systolic blood pressure using longitudinal data from eight UK cohorts. *PLoS Med* 2011; 8(6): e1000440. doi:10.1371/journal.pmed.1000440.

Wooldridge JM. *Econometric analysis of cross section and panel data.* Cambridge: MIT Press; 2011.

Wooldridge JM. *Introductory econometrics: a modern approach.* Mason: South-Western; 2013.

Yood MU, McCarthy BD, Kempf J, Kucera GP, Wells K, Oliveria S, Stang P. Racial differences in reaching target low-density lipoprotein goal among individuals treated with prescription statin therapy. *American Heart Journal* 2006; 152(4), 777-784.

Appendix A

R CODE: HECKMAN'S TREATMENT EFFECTS MODEL

```
## Parameter: theta = (alpha, beta, var.y, rho, delta)
##
## Data
#### W: Design matrix of covariates predicting medication use
#### X: Design matrix of covariates prediction underlying biomarker
#### Y: Observed biomarker outcome
#### Z: Observed medication use status
##
## Comment: Place all variables of X into W to allow z to depend on y(0)


negloglik.heck <- function(theta, W, X, Y, Z) {
    ## Size of sub-parameters
    nw <- dim(W)[2]
    nx <- dim(X)[2]
    alpha <- theta[1:nw]
    beta <- theta[(nw + 1):(nw + nx)]
    var.y <- theta[nw + nx + 1]
    rho <- theta[nw + nx + 2]
    delta <- theta[nw + nx + 3]

    ## Fitted values
    Yhat <- X %*% beta - delta * Z
    Zstarhat <- W %*% alpha
```

```
    #Likelihood for all subjects
    all <- dnorm(Y - Ystarhat, mean = 0, sd = sqrt(var.y))

    #Conditional mean and variance
    mean.z.y <- W %*% alpha + (rho/sqrt(var.y))*(Yhat)
    var.z.y <- 1 - rho^2

    #Untreated and Treated Likelihoods
    tx <- 1 - pnorm(0, mean = mean.z.y, sd = sqrt(var.z.y))
    utx <- pnorm(0, mean = mean.z.y, sd = sqrt(var.z.y))
    like <- log(all) + log(utx)
    like[Z==1] <- log(all)[Z==1] + log(tx)[Z==1]
    negll <- -sum(like)
    negll
}
```

Appendix B

R CODE: SUBGROUP-SPECIFIC EFFECTS MODEL

```
## Parameter: theta = (alpha, beta, eta, var.y, rho)
##
## Data
#### V: Design matrix of covariates predicting treatment effect size
#### W: Design matrix of covariates predicting medication use
#### X: Design matrix of covariates prediction underlying biomarker
#### Y: Observed biomarker outcome
#### Z: Observed medication use status
##
## Comment: Place all variables of X into W to allow z to depend on y(0)


negloglik.ssem <- function(theta, V, W, X, Y, Z) {
    ## Size of sub-parameters
    nw <- dim(W)[2]
    nx <- dim(X)[2]
    nv <- dim(V)[2]
    alpha <- theta[1:nw]
    beta <- theta[(nw + 1):(nw + nx)]
    eta <- theta[(nw + nx + 1):(nw + nx + nv)]
    var.y <- theta[nw + nx + nv + 1]
    rho <- theta[nw + nx + nv + 2]

    ## Fitted values
```

```
    Yhat <- X %*% beta - V %*% eta * Z
    Zstarhat <- W %*% alpha

    #Likelihood for all subjects
    all <- dnorm(Y - Ystarhat, mean = 0, sd = sqrt(var.y))

    #Conditional mean and variance
    mean.z.y <- W %*% alpha + (rho/sqrt(var.y))*(Yhat)
    var.z.y <- 1 - rho^2

    #Untreated and Treated Likelihoods
    tx <- 1 - pnorm(0, mean = mean.z.y, sd = sqrt(var.z.y))
    utx <- pnorm(0, mean = mean.z.y, sd = sqrt(var.z.y))
    like <- log(all) + log(utx)
    like[Z==1] <- log(all)[Z==1] + log(tx)[Z==1]
    negll <- -sum(like)
    negll
}
```

Appendix C

R CODE: LONGITUDINAL ENDOGENEITY MODEL

```
## Parameter: theta = (alpha, beta, eta, rho.y, rho, sd.y)
##
## Data
#### V: Design matrix of covariates predicting treatment effect size
#### W: Design matrix of covariates predicting medication use
#### X: Design matrix of covariates prediction underlying biomarker
#### Y: Observed biomarker outcome
#### Z: Observed medication use status
#### id: Vector of IDs corresponding to observations
#### corstr: String specifying working correlation: "AR1" or "Exchangeable"
##
## Comment: Place X variables in W for homogeneous corresp. dependancies
## Comment: Working independence code equivalent to SSEM

negloglik.lem <- function(theta, V, W, X, Y, Z, id, corstr) {
    ## Size of sub-parameters
    nw <- dim(W)[2]
    nx <- dim(X)[2]
    nv <- dim(V)[2]
    alpha <- theta[1:nw]
    beta <- theta[(nw + 1):(nw + nx)]
    eta <- theta[(nw + nx + 1):(nw + nx + nz)]
    rho.y <- theta[nw + nx + nz + 1]
```

```
rho  <- theta[nw + nx + nz + 2]
sd.y <- theta[nw + nx + nz + 3]
atrho <- (atan(rho) + pi/2)/pi
atrho.ep <- (atan(rho.ep) + pi/2)/pi
var.z.y <- 1 - atrho^2

#Records for IDs
numvis <- as.numeric(table(id))
uids <- unique(id)
loglik <- 0
bidx <- 1

for (i in 1:length(uids)) {
   eidx <- bidx + numvis[i] - 1
   idxs <- bidx:eidx
   bidx <- bidx + numvis[i]
   Ti <- numvis[i]
   if (corstr == "AR1") {exp.mat <- abs(outer(1:Ti, 1:Ti, "-"))}
   if (corstr == "Exchangeable") {exp.mat <- matrix(1, nrow = Ti,
           ncol = Ti)}
   SigmaY <- sd.y^2 * (matrix(rep(c(1, rep(atrho.ep, Ti)),
           Ti), nrow = Ti)[1:Ti,1:Ti])^exp.mat
   cX <- X[idxs,]
   cZ <- Z[idxs]
   cY <- Y[idxs]
   cV <- V[idxs]
   cW <- W[idxs,]
   temp <- cY - cX %*% beta + (cV %*% eta) * cZ
```

```
        tempmat <- matrix(c(temp), nrow = 1, byrow = TRUE)
        SigmaY.inv <- solve(SigmaY)
        ldsy <- log(det(matrix(SigmaY, nrow = Ti)))
        lpy <- (-Ti/2) * log(2*pi) - ldsy/2 + (-1/2) * diag(tempmat
                %*% SigmaY.inv %*% t(tempmat))
        mean.z.y <- cW %*% alpha + atrho * temp / sd.y
        propz0 <- pnorm(0, mean = mean.z.y, sd = sqrt(var.z.y))
        propz1 <- 1 - pnorm(0, mean = mean.z.y, sd = sqrt(var.z.y))
        prop <- propz0
        prop[cZ == 1] <- propz1[cZ == 1]
        lprop <- log(prop)
        loglik <- loglik + sum(lprop) + lpy
    }
    negll <- -1 * loglik
    negll
}
```

VITA

Andrew Spieker was raised in West Hartford, CT, and earned a Bachelor of Science in Mathematics from Northeastern University in Boston, MA. He worked for the Beth Israel Deaconess Medical Center in the Department of Neurology before moving to Seattle in 2012. He earned a Masters degree in Biostatistics in 2015, and a Doctor of Philosophy in Biostatistics in 2016, both from the University of Washington in Seattle, WA.