Analysis of biased sampling designs using longitudinal data

Leila R. Zelnick

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

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Patrick J. Heagerty, Chair

Peter B. Gilbert

Susanne J. May

Nicole M. Hamblett

Emma S. Spiro

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Analysis of biased sampling designs using longitudinal data

Leila R. Zelnick

Chair of the Supervisory Committee:
Professor Patrick J. Heagerty
Biostatistics

With increasing availability of prospective cohort studies, registry data, and electronic health records, numerous secondary investigations are being conducted using data that were originally collected for a different primary research goal. When explorations of novel biomarkers and their associations with outcomes are of interest, it is natural to leverage existing cohorts for which stored biological specimens may be available or for which new specimens can be selectively collected and processed, yielding new exposure data. However, limited availability of specimens and limited financial resources may require investigators to target only a subset of patients for any new analyses. In such cases, the use of outcome dependent sampling (ODS) designs can provide an efficient and cost-effective way to conduct substudies leveraging existing outcomes. In ODS designs a subsample is chosen based on characteristics of the outcome variable, and for these select subjects the detailed covariate data is then collected.

Design and analysis methods that use longitudinal outcomes to guide choice of a subsample have been shown to improve efficiency over random sampling (Schildcrout et al. (2013)), but to date statistical methods have focused exclusively on using only the data from the subsample for final analysis. However, ODS research in the univariate setting (Lawless et al. (1999), Weaver and Zhou (2005), Chatterjee et al. (2003)) has shown that analyzing the incomplete data from unsubsampled individuals, in addition to those on whom the biomarker

has been ascertained, may contribute to improved estimation of target regression parameters. Once an ODS sample is obtained a variety of analysis approaches can be used to provide valid inference, but the complexity and utility of alternative analysis approaches has not been thoroughly investigated for designs with longitudinal outcome data.

This dissertation focuses on the use of ODS sampling designs and analysis in the longitudinal setting with continuous outcomes. We examine the potential efficiency gains of this family of designs/analyses from a likelihood perspective and offer robust alternatives that may be preferred under possible model misspecification. Finally, we adapt a standardization technique from the epidemiological/causal inference literature that may provide benefit in analyzing ODS of longitudinal processes and can be used to accommodate unanticipated missingness caused by participants dropping out of a study prior to its completion. In each case, the methods explored here are illustrated for a hypothetical biomarker substudy, using data from the Cystic Fibrosis Foundation Patient Registry.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to acknowledge the invaluable contributions of the committee in helping bring this dissertation to fruition. In particular, Patrick Heagerty has demonstrated in ways both implicit and explicit how to be an outstanding mentor, collaborator, and scientist. Finally, thanks to the Cystic Fibrosis Foundation for use of the patient registry data; may the work here contribute in some small way to developing therapies for this insidious disease.

DEDICATION

To my parents, John and Nancy, for their many long years of support of my education.

And to Greg and Phil, who keep the home fires burning.

Chapter 1

INTRODUCTION AND MOTIVATION

With the increasing availability of prospective cohort studies, registry data, and electronic health records, numerous secondary investigations are being conducted using data that were originally collected for a different primary research goal. For example, existing patient cohorts may provide longitudinal outcome data measured as part of a primary study, while secondary studies may be interested in new candidate predictors that are hypothesized to be associated with different outcome trajectories. When explorations of novel biomarkers and their associations with outcomes are of interest, it is natural to leverage existing cohorts for which stored biological specimens may be available or for which new specimens can be selectively collected and processed, yielding new exposure data. However, limited availability of specimens and limited financial resources may require investigators to target only a subset of patients for any new analyses. In such cases, the use of outcome dependent sampling (ODS) designs can provide an efficient and cost-effective way to conduct substudies while leveraging existing outcomes. In ODS designs a subsample is chosen based on characteristics of the outcome variable, and for these select subjects the detailed covariate data is then collected.

Design and analysis methods that use longitudinal outcomes to guide choice of a subsample have been shown to improve efficiency over random sampling [39], but to date statistical methods have focused exclusively on using only the data from the subsample for final analysis. However, ODS research in the univariate setting [16] [46] [4] has shown that analyzing the

incomplete data from unsubsampled individuals, in addition to those on whom the biomarker has been ascertained, may contribute to improved estimation of targeted regression parameters. Once an ODS sample is obtained a variety of analysis approaches can be used to provide valid inference, but the complexity and utility of alternative analysis approaches has not been thoroughly investigated for designs with longitudinal outcome data.

This dissertation will focus on the use of ODS sampling designs and analysis in the longitudinal setting with continuous outcomes. We will examine the potential efficiency gains of this family of designs/analyses from a likelihood perspective and offer alternatives for use under possible model misspecification, as well as considering practical considerations for ODS analysis under unanticipated missingness due to dropout. Specifically, the aims of this research program are:

1. To develop and characterize alternative conditional likelihood estimators that use various sources of information, such as the covariate distribution among subsampled individuals and the marginal longitudinal outcome distribution from unsubsampled individuals. We will use a likelihood framework to compare the relative efficiency of these estimators relative to a random sample for longitudinal ODS designs with continuous outcomes.

2. To develop and evaluate candidate non-likelihood-based estimators that are potentially robust to model misspecification for use with longitudinal ODS designs.

3. To develop and evaluate estimators appropriate for use with ODS designed data, subject to unanticipated missingness caused by a simple MAR nonignorable dropout structure.

Chapter 2

LIKELIHOOD-BASED ANALYSIS OF OUTCOME DEPENDENT SAMPLING DESIGNS FOR LONGITUDINAL DATA

2.1 Introduction

In a resource-limited world, efficient study design and data analysis continue to be a priority for scientific researchers. Registries and other cohorts of patients can provide cheap and accessible sources of longitudinal data to researchers; when novel biomarkers are discovered, it is natural to leverage these data sources in order to evaluate the new marker's association with longitudinal outcomes. However, limited availability of biological specimens together with financial constraints may require investigators to target only a subset of patients for further study. In such cases outcome dependent sampling (ODS) designs, which collect detailed covariate data only on a subset of individuals based on characteristics of the outcome variable, can provide an efficient and cost-effective strategy to conduct substudies that leverage existing information.

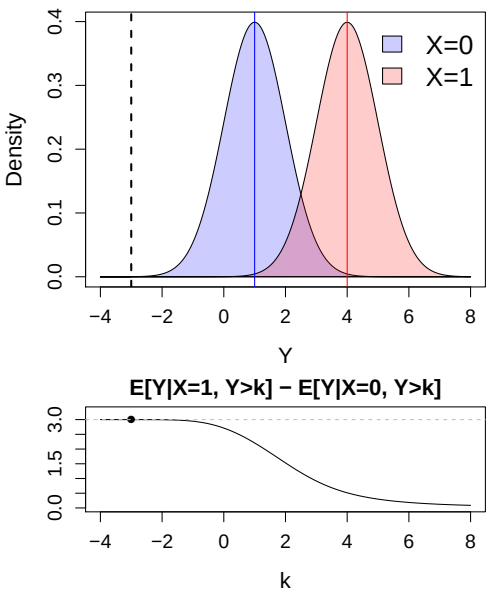
Methods that selectively subsample informative individuals have a long history of offering efficiency gains over random sampling. The most common outcome dependent sampling design, the case-control study, has been frequently used in epidemiological studies as a cost-effective way to study the association between rare binary outcomes and exposures. A classic work by Prentice and Pyke [28] showed that a retrospectively sampled case-control study may be analyzed as if the data had come from a prospectively sampled cohort without having to correct for the biased sampling design. As a result, the case-control study, analyzed using logistic regression, is one of the most commonly implemented designs in use today.

In the more general regression setting, however, analyzing outcome-dependent samples as if chosen randomly is problematic and can result in biased estimation. As an example, consider the following scenario. Suppose we have an independent and identically distributed normal outcome Y and a binary marker X such that

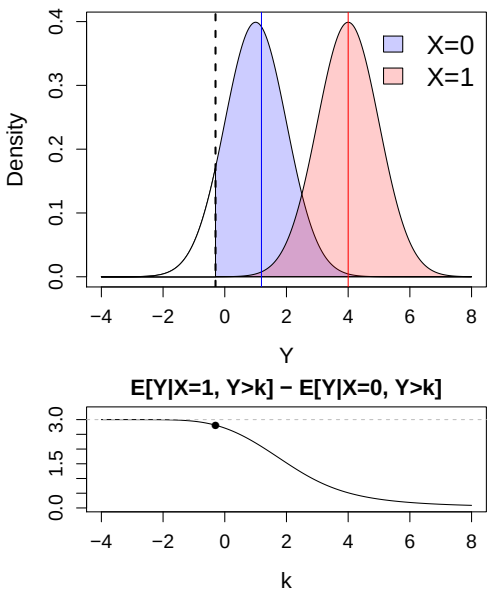
$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where $\beta_0 = 1$, $\beta_1 = 3$, and $\epsilon \sim N(0, 1)$, and suppose that the biased sampling scheme restricts sampling above a threshold k. The true parameter of interest is the average difference in outcome between those with and without the marker, or $\beta_1 = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0]$ (here, $\beta_1 = 3$). Calculating a difference in means for data collected under this biased sampling mechanism, however, instead estimates the difference in truncated means, $\mathbb{E}[Y|X = 1, Y > k] - \mathbb{E}[Y|X = 0, Y > k]$, which may be far from the true parameter of interest. Figure 2.1 illustrates how the extremity of the sampling design impacts the bias in estimation. When the sampling is nearly unrestricted (Figure 2.1(a)), the difference in mean outcomes is close to the true parameter β_1. As k increases (Figures 2.1(b) through 2.1(d)), however, the restricted or truncated marker means differ more from the unrestricted means, resulting in greater bias in $\hat{\beta}_1$. In general, the more the sampling scheme deviates from a random sample, the greater the potential negative impact of ignoring the sampling design in the analysis stage.

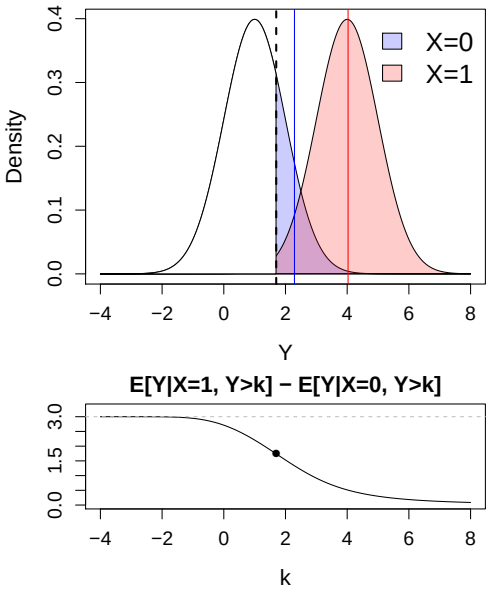
To avoid the potential pitfalls in estimation illustrated above, continuous outcomes have sometimes been dichotomized and then analyzed using logistic regression, proceeding as if the data were from a case-control study. Where the outcome is truly continuous, however, such dichotomization can be inefficient and make comparisons across studies difficult if there is not an accepted and meaningful cutpoint [9] [29] [42]. Such an analysis may have reduced power to detect and association; moreover, a logistic regression analysis of the dichotomized outcome leads to estimation of the odds ratio, which may not be the parameter of greatest



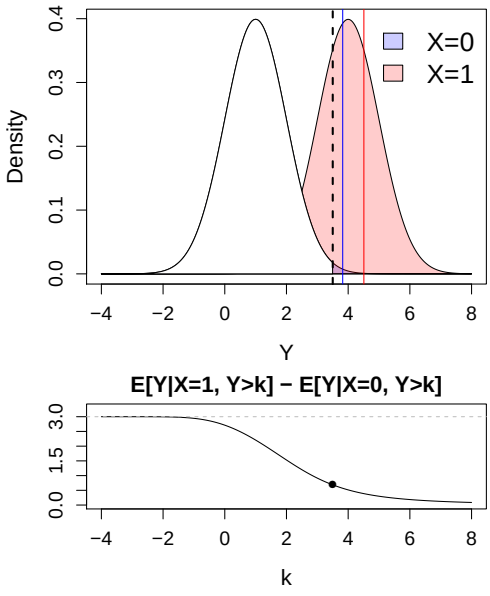
(a) When sampling is essentially unrestricted, there is no/little bias in analyzing the data ignoring the biased sampling design.



(b) As sampling becomes more restricted, the observed mean among those with the marker (blue line) does not reflect the true center of $X = 0$'s distribution.



(c) Further restriction of k results in highly biased estimates of β_X.



(d) The more extreme the biased sampling design, the more important accounting for it in the analysis becomes.

Figure 2.1: Schematic of impact of ignoring biased sampling design in analysis.

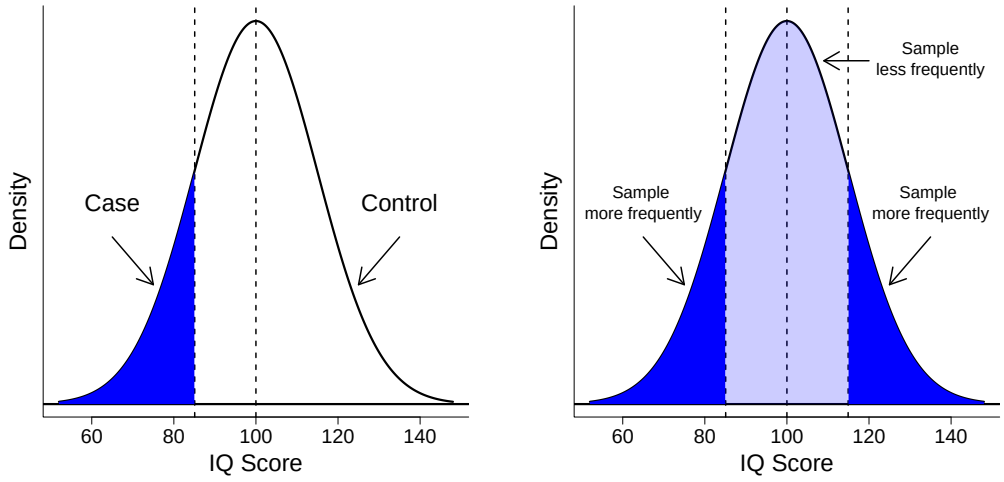
scientific interest.

For situations in which the association of an exposure with a continuous longitudinal outcome is of primary interest, ODS design and analysis strategies offer the prospect of valid inference and increased efficiency at a reduced cost compared to traditional methods. With this work we hope to illuminate the design and analysis options available to researchers, and to demonstrate scenarios that would benefit from using ODS design and analytic approaches.

2.2 Background

Outcome dependent sampling designs for univariate continuous outcomes have previously been shown to have potentially increased efficiency relative to a random sample. Zhou et al. [48] illustrated these results in the univariate setting using a study that examined the relationship between infant intelligence quotient (IQ) at 8 months of age and *in utero* exposure to the neurodevelopmental toxin polychlorinated biphenyl (PCB). Infant IQ, a continuous outcome, was available for all mother-baby pairs, while maternal serum was stored for all mothers but costly to process. Thus, the PCB levels could only be measured on a subset of mothers. Rather than dichotomize (Figure 2.2(a)) the continuous IQ outcomes below/above 85, Zhou et al. proposed a biased sampling scheme based on the outcome that would subsample differentially from each region, oversampling infants with extreme IQs (Figure 2.2(b)) and supplementing this biased subsample with a random sample of infants.

Work by Zhou et al. [49] [48] [50] and Weaver et al. [46] demonstrated the potential benefits of biased sampling in the regression setting for cross-sectional data with continuous outcomes in the following way. Assume we have taken a biased subsample of size N_S from a cohort of N individuals, and wish to draw inferential conclusions about the parameter vector $\boldsymbol{\beta}$ resulting from regressing a continuous outcome Y on an exposure X in the entire population from which the cohort was sampled. For simplicity we assume the exposure to be binary, although the arguments extend easily to exposures with an arbitrary number of



(a) Biased subsampling when IQ is dichotomized into cases (low IQ) and controls (normal IQ).

(b) Biased subsampling when IQ is retained as a continuous outcome.

Figure 2.2: Schematic of biased subsampling choices for Zhou et al. (2007) infant IQ example.

levels. Likewise, the following framework could easily accommodate the presence of additional inexpensive covariates measured on all subjects; for simplicity these are omitted here. Define the following:

- S: the indicator of being subsampled
- Y: the continuous outcome of interest, observed in all subjects
- X: the (expensive) exposure of interest, ascertained only in a subsample

Under this scenario, following the example of Zhou et al. and Weaver et al., the contribution of the biased sample to the observed data likelihood for subsampled individuals, conditional

on being sampled[1], can be written as:

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}, \boldsymbol{S}) &= \prod_{i=1}^{N_S} f(Y_i, X_i | S_i = 1; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{N_S} \frac{f(Y_i, X_i, S_i = 1 | \boldsymbol{\theta})}{P(S_i = 1)} \\ &= \prod_{i=1}^{N_S} \frac{P(S_i = 1 | Y_i, X_i) \cdot f(Y_i | X_i; \boldsymbol{\theta}) \cdot g(X_i)}{P(S_i = 1)} \\ &= \prod_{i=1}^{N_S} \frac{\omega(Y_i) \cdot f(Y_i | X_i; \boldsymbol{\theta}) \cdot g(X_i)}{P(S_i = 1)} \end{aligned} \tag{2.1}$$

With this application of Bayes rule, the likelihood of the observed biased sample can be written as a scaled version of the likelihood under random sampling. By design, the probability of being subsampled is strictly a function of the observed outcome, so $\omega(Y) \equiv P(S = 1|Y, X)$ does not depend on $\boldsymbol{\theta}$ and can be ignored in maximizing the likelihood. The scaling factor $P(S = 1)$ involves both the distributions[2] of $[X]$ and $[S|X]$, and since the latter is related to $\boldsymbol{\theta}$ under biased sampling, this term must be taken into account for valid inference. While the marginal distribution $g(X)$ is unrelated to $\boldsymbol{\theta}$, $g(X)$ is implicated in the scaling factor $P(S = 1)$ and cannot be ignored in analysis as under random sampling designs. To address this issue, Zhou et al. [49] used an empirical likelihood estimate of $g(x)$, which was combined with usual likelihood maximization techniques to obtain a semiparametric estimator of $\boldsymbol{\beta}$.

The likelihood considered by Zhou et al. in Equation 2.1 is not the only likelihood that may be analyzed in this situation. A variety of analysis strategies could be and have been

---

[1]In fact, Zhou et al. and Weaver et al. condition not on having been sampled marginally ($S_i = 1$), but on having been sampled in region k, $k = 1, 2, 3$. Here, we use the marginal conditioning, for greater consistency with the work that follows in the remainder of this dissertation and since the arguments remain unchanged.

[2]Notation: henceforth, we write $[X]$ to mean "the distribution of X".

employed to create valid estimators of regression parameters, that can all be derived from the complete data likelihood:

$$
\begin{aligned}
L(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}, \boldsymbol{S}) &= \sum_{i=1}^{N} f(Y_i, X_i, S_i; \boldsymbol{\theta}) \\
&= \prod_{S_i=1} P(S_i = 1; \boldsymbol{\theta}) \cdot f(Y_i, X_i | S_i = 1; \boldsymbol{\theta}) \cdot \prod_{S_i=0} f(Y_i | S_i = 0; \boldsymbol{\theta}) \cdot P(S_i = 0; \boldsymbol{\theta}) \\
&= \overbrace{\prod_{S_i=1} P(S_i = 1; \boldsymbol{\theta}) \cdot f(X_i | S_i = 1; \boldsymbol{\theta}) \cdot f(Y_i | X_i, S_i = 1; \boldsymbol{\theta})}^{\text{subsampled + unsampled, with covariates}} \cdot \overbrace{\prod_{S_i=0} f(Y_i | S_i = 0; \boldsymbol{\theta}) \cdot P(S_i = 0; \boldsymbol{\theta})}^{\text{subsampled + unsampled, no covariates}} \\
&= \underbrace{\prod_{S_i=1} f(X_i | S_i = 1; \boldsymbol{\theta}) \cdot \underbrace{f(Y_i | X_i, S_i = 1; \boldsymbol{\theta})}_{\text{subsampled only, no covariates}}}_{\text{subsampled only, with covariates}} \cdot \prod_{S_i=0} f(Y_i | S_i = 0; \boldsymbol{\theta}) \cdot \prod_{i=1}^{N} P(S_i; \boldsymbol{\theta})
\end{aligned}
\tag{2.2}
$$

As seen in Equation 2.2, the unconditional observed data likelihood can be factored into a number of terms that could be utilized to yield estimators of $\boldsymbol{\theta}$. Zhou et al. [49] chose to analyze the joint likelihood of outcome Y and exposure X only among individuals in whom the exposure was ascertained (termed "subsampled only, with covariates" in Equation 2.2). The same approach was taken by Neuhaus et al. [25] [26] for clustered binary outomes in case-control family studies. In that work, families of ovarian cancer patients were subsampled for further covariate information based on strata defined by features of the family outcome vector, such as the number of cancer cases in the family. Maximization of the joint likelihood of outcome and covariate was accomplished using a profile likelihood approach that Neuhaus et al. [25] generalized for use with any type of covariate.

In the approach taken by Zhou et al.[49] for cross-sectional data and by Neuhaus et al.[25] for clustered data, the authors chose to analyze the joint likelihood for those whose

exposure had been ascertained. However, in analyzing subsampled individuals only, one could further condition on the exposure X, and maximize the resulting likelihood (i.e., based on the distribution of $[Y|X, S = 1]$ instead of $[Y, X|S = 1]$) to obtain a valid estimator. This strategy sidesteps the issue of modeling or estimating $g(X)$, which would otherwise be necessary because of the biased sampling design, but also potentially loses some information through further conditioning. In the longitudinal setting, this conditional approach was taken by Schildcrout et al. [39].

For longitudinal or clustered data, the construction of the biased sampling design presents an added level of complexity compared to cross-sectional data. For a single continuous outcome, there exists a natural ordering upon which a biased subsampling scheme can be based. For clustered/longitudinal data, however, there are many outcome orderings upon which to base a preferential subsampling scheme. One simple way to define the outcome dependent sampling scheme for longitudinal data is to transform the outcome vector into a low-dimensional summary that again has a natural ordering. As mentioned above, for clustered binary outcomes Neuhaus et al. [25][26] subsampled on the number of cases in a cluster, while Schildcrout et al. [40] preferentially subsampled clusters whose members were not all 0 or 1. For continuous longitudinal outcomes, Schildcrout et al. [39] defined a class of subsampling variables Q which are generally a low-dimensional linear combination of the outcome vector $\mathbf{Y}$. This approach admits a wide range of sampling variables whose distribution is related to the distribution of $\mathbf{Y}$, and which may be analytically convenient. In particular, after performing a subject-specific regression of each subject's fully observed outcome vector on time, the authors chose to subsample based on the value of either the subject-specific intercept, subject-specific regression, or on a bivariate combination of the two.

The approaches discussed above all analyzed, either conditionally upon or jointly with covariate information, individuals whose exposure was ascertained ($S = 1$). However, as

Equation 2.2 shows, including unsubsampled individuals in the analysis would likewise yield a likelihood function that could be maximized to obtain valid estimators of $\boldsymbol{\theta}$. While unsubsampled subjects do not have the exposure measured, and cannot directly provide information on the relationship of the expensive covariate with the outcome, the observed outcomes of unsubsampled individuals provide information on the population-level mixture of covariate-specific mean outcomes. Thus, including these individuals in inference (at added computational but no additional logistical cost) has the potential to improve inference on some parameters or combinations of parameters.

The idea of including unsubsampled indviduals in the analysis of an outcome dependent sample was explored by Weaver et al. [46] for univariate data. They proposed an estimated likelihood approach, which they found to be more efficient for some parameters than the semiparametric estimator explored in Zhou et al. [49], which analyzed only subsampled individuals. Pseudoscore [4] and semiparametric MLE [16] methods that included unsubsampled individuals were found to perform similarly well in simulations. More recently, Rose and van der Laan [33] used a targeted maximum likelihood approach combined with inverse probability of censoring weighting to address two-stage sampling designs generally. Table 2.1 summarizes the relative contributions and sources of information examined by some of this body of literature.

The ODS ideas described here have partially been extended into the longitudinal setting. Neuhaus et al. [25] [26] considered ODS in the context of binary clustered data in case-control family studies. Families of ovarian cancer patients were subsampled for further covariate data collection based on strata defined by features of the family outcome vector, such as the number of cancer cases in the family. Maximization of the joint likelihood of outcome and covariate was accomplished using a profile likelihood approach that Neuhaus et al. generalized for use with any type of covariate. More recently, Schildcrout et al. [39] proposed an ODS design for continuous longitudinal outcomes based on a low-dimensional summary of

Table 2.1: Summary of selected ODS literature contributions

		Outcome	
		Univariate	Longitudinal/ clustered
Sampled only	Covariates	Zhou et al. (2002, 2007, 2014) [49] [48]	Neuhaus et al. (2002, 2006) [25] [26]
	No covariates	–	Schildcrout et al. (2013) [39]
Sampled and unsampled	Covariates	Weaver et al. (2005) [46], Chatterjee et al. (2003) [4], Lawless et al. (1999) [16]	Zelnick et al. (2016)
	No covariates	–	Zelnick et al. (2016)

the outcome vector. Subjects were subsampled based on the value of their subject-specific intercept, subject-specific slope, or a combination of the two, and an "ascertainment-corrected" likelihood that conditioned both on being sampled and on the observed covariate value was the basis for inference. Notably, in this work thoughtful sampling design decisions were shown to have potentially large impacts on regression parameter estimation efficiency.

Collectively, this body of work has demonstrated that ODS designs can provide substantial efficiency gains on regression parameters of interest compared to a random sample. Yet approaches to creating valid estimators have varied with respect to both design and analysis of the biased subsample. Decisions about which likelihood to maximize and whether to include information from the exposure and/or unsubsampled individuals have been explored in some circumstances, but not codified. For a continuous longitudinal outcome, this chapter explores the contributions of various sources of information to the estimation of key regression parameters in a likelihood framework. For simplicity we assume the ascertained exposure to be binary, although the arguments extend easily to exposures with an arbitrary number of levels. Likewise, the following framework could easily accommodate the presence of additional inexpensive covariates that are measured for all subjects; for simplicity these

are omitted here. Section 2.3 introduces six likelihood-based regression parameter estimators to be compared, while Section 2.4 presents operating characteristics and results of these. In Section 2.5, we illustrate the results with an application to the Cystic Fibrosis Foundation Registry dataset, and finally offer a discussion of our results in Section 2.6.

2.3 Methods

2.3.1 Notation and design

In this section we explore a variety of possible valid likelihood-based estimators of the regression parameters of interest arising from the usual linear mixed model for longitudinal data. Specifically, suppose we have a cohort of N subjects, each measured n_i times, so that for the $n_i \times 1$ continuous outcome vector $\mathbf{Y}_i,\ i = 1, \ldots, N$, the linear mixed model of interest as proposed by Laird and Ware [15] is

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

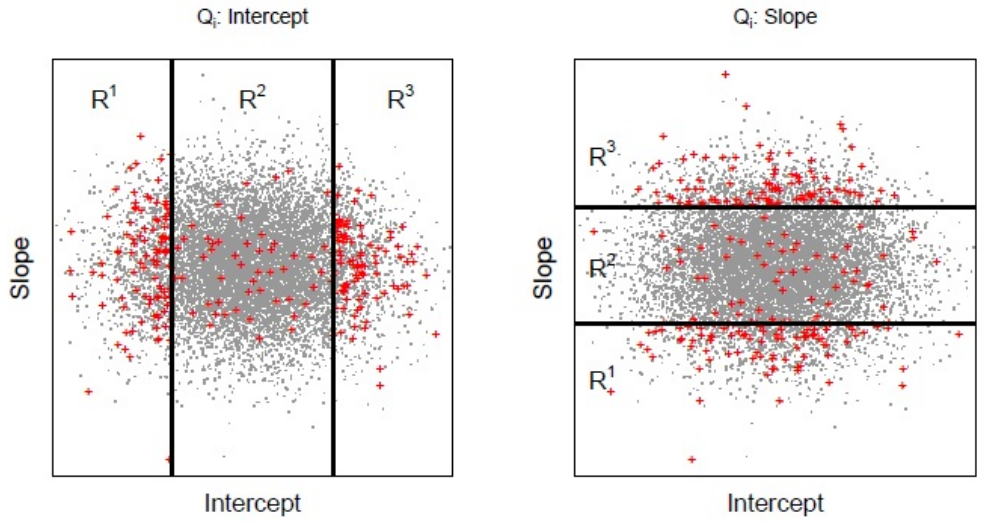
where $\mathbf{X}_i = [\mathbf{1},\ \boldsymbol{T}_i,\ \boldsymbol{M}_i,\ \boldsymbol{M}_i \times \boldsymbol{T}_i]$ is the $n_i \times 4$ design matrix. We define the vector of times $\boldsymbol{T}_i = [T_{ij}],\ j = 1, \ldots, n_i$, and M_i the retrospectively ascertained time-invariant binary covariate. The 4×1 vector $\boldsymbol{\beta}$ contains the regression parameters of interest, while $\mathbf{Z}_i$ is the design matrix for the intercept and slope random effects. The vector $\boldsymbol{b}_i = (b_{i0}, b_{i1})^T$ is assumed to be multivariate normally distributed with 2×1 mean vector $\mathbf{0}$ and covariance matrix $\mathbf{D}$ consisting of diagonal elements $(\sigma^2_{b_0}, \sigma^2_{b_1})^T$ and off diagonal covariance $Cov(b_{i0}, b_{i1}) = \rho\sigma_{b_0}\sigma_{b_1}$. The $n_i \times 1$ vector of errors $\boldsymbol{\epsilon}_i$ are assumed to be conditionally independent and normally distributed with common variance σ^2_e. Transforming the variance components to $\boldsymbol{\gamma} = \left(\log(\sigma^2_{b_0}), \log\left(\frac{1+\rho}{1-\rho}\right), \log(\sigma^2_{b_1}), \log(\sigma^2_e)\right)^T$ for ease of estimation, the parameter vector on which we focus is the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})^T$.

We assume that all N members of the cohort have completely observed outcome vector

$\mathbf{Y}$, and that M is an expensive covariate that is ascertained only for a subsample of N_S individuals. For illustration, we consider a simple binary and time-invariant covariate, such as a novel marker retrospectively measured on a stored biological sample. For subjects who are subsampled, the complete information vector on $(\mathbf{Y}, \boldsymbol{T}, M)$ is available; for the remaining $N - N_S = N_{NS}$ subjects who are not subsampled, only the vector $(\mathbf{Y}, \boldsymbol{T})$ is known. The indicator S_i denotes that subject i has been subsampled.

As described previously by Schildcrout et al. [39], a simple subsampling approach for longitudinal ODS designs is to subsample subjects based on a low-dimensional summary of the outcome vector $\mathbf{Y}_i$, often a linear combination of the longitudinal outcome, $\boldsymbol{Q}_i = \mathbf{W}_i\mathbf{Y}_i$ for some $m \times n_i$ matrix $\mathbf{W}_i$. Briefly, we consider subsampling based on a feature (either intercept or slope) of the vector $\boldsymbol{q}_i = \left(\mathbf{X}_{ti}^T\mathbf{X}_{ti}\right)^{-1}\mathbf{X}_{ti}^T\mathbf{Y}_i$, where $\mathbf{X}_{ti} = [\mathbf{1},\ \boldsymbol{T}_i]$, the result of regressing outcome vector $\mathbf{Y}_i$ on time for each cohort member. The resulting values of the sampling variable Q will fall into one of three regions: Region 1 $(-\infty, a_1)$, Region 2 $[a_1, a_2)$, or Region 3 $[a_2, \infty)$, where a_1 and a_2 are predetermined constants. Within each region, subjects are subsampled with constant probability $\omega_k(q) = P(S = 1|q \in R_k), k = 1, 2, 3$, which may differ by stratum but is assumed to be a constant chosen by design. Figure 2.3, taken from Schildcrout et al. [39], provides a graphical depiction of the subsampling scheme described above. As in previous work, we generally wish to oversample subjects with extreme values of Q. Although we choose to examine only designs based on these two simple features, other approaches to biased sampling could be equally valid provided the design is adequately considered in the analysis stage.

Figure 2.3: Graphical depiction of outcome dependent sampling scheme for longitudinal data. Regressing an individual's outcome vector on observation times produces a subject-specific intercept and slope, one of which can be used for preferential subsampling. Individuals with extreme values of the subsampling variable Q_i falling in Regions 1 or 3 are selected for subsampling (red crosses) with a higher probability than those that fall in Region 2. Figure from Schildcrout et al. (2013).



2.3.2 Likelihood

Under the longitudinal data scenario described above, the complete observed data likelihood can be written as:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}, \boldsymbol{S}) &= \prod_{i=1}^{N} f(\mathbf{Y}_i, \mathbf{X}_i, S_i; \boldsymbol{\theta}) \\
&= \prod_{S_i=1} P(S_i = 1; \boldsymbol{\theta}) \cdot f(\mathbf{Y}_i, \mathbf{X}_i | S_i = 1; \boldsymbol{\theta}) \cdot \prod_{S_i=0} f(\mathbf{Y}_i | S_i = 0; \boldsymbol{\theta}) \cdot P(S_i = 0; \boldsymbol{\theta}) \\
&= \prod_{S_i=1} \overbrace{P(S_i = 1; \boldsymbol{\theta}) \cdot f(\mathbf{X}_i | S_i = 1; \boldsymbol{\theta}) \cdot f(\mathbf{Y}_i | \mathbf{X}_i, S_i = 1; \boldsymbol{\theta})}^{\text{subsampled + unsubsampled, with covariates}} \cdot \overbrace{\prod_{S_i=0} f(\mathbf{Y}_i | S_i = 0; \boldsymbol{\theta}) \cdot P(S_i = 0; \boldsymbol{\theta})}^{\text{subsampled + unsubsampled, no covariates}} \\
&= \prod_{S_i=1} \underbrace{f(\mathbf{X}_i | S_i = 1; \boldsymbol{\theta}) \cdot \underbrace{f(\mathbf{Y}_i | \mathbf{X}_i, S_i = 1; \boldsymbol{\theta})}_{\text{subsampled only, no covariates}}}_{\text{subsampled only, with covariates}} \cdot \prod_{S_i=0} f(\mathbf{Y}_i | S_i = 0; \boldsymbol{\theta}) \cdot \prod_{i=1}^{N} P(S_i; \boldsymbol{\theta})
\end{aligned}
\tag{2.3}
$$

As seen in Equation 2.3, the unconditional observed data likelihood can be factored into several terms corresponding to several conditional likelihoods that could be utilized to yield estimators of $\boldsymbol{\theta}$. For the present we consider only complete balanced designs; $\boldsymbol{T}$ may be assumed to be independent of other variables in this case. Under random sampling, the distributions of $[\mathbf{X}|S]$ and $[S]$ are $\boldsymbol{\theta}$-free and add no information to inference; conventional regression approaches condition upon $\mathbf{X}$ and S for this reason. Under biased sampling, however, both may contain information about $\boldsymbol{\theta}$ and could potentially be incorporated in the maximization to good effect. Similarly, including unsubsampled individuals in the analysis, either conditionally upon or jointly with covariate information, would yield a likelihood function that could be maximized to obtain valid estimators of $\boldsymbol{\theta}$.

2.3.3 Analysis: Subsampled only

Equation 2.3 shows that a variety of valid likelihoods derived from the complete data likelihood could be used as a basis for inference. One simple analysis option would be to consider the conditional likelihood (which we refer to as "subsampled only, no covariates", or SO,NC)

$$\mathcal{L}_{SO,NC}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}, \boldsymbol{S}) = \prod_{S_i=1} f(\mathbf{Y}_i|\mathbf{X}_i, S_i = 1; \boldsymbol{\theta}) \tag{2.4}$$

considered by Schildcrout et al. [39], which utilizes information from subsampled individuals only, conditional upon the marker value. The resulting conditional log-likelihood (see Appendix A) can be written as a term that treats subsampled data as if it had come from a random sample, less an "ascertainment-correction" term $AC_0(\boldsymbol{x}) = AC_0(m, \boldsymbol{t}) \equiv P(S = 1|M = m, \boldsymbol{T} = \boldsymbol{t}; \boldsymbol{\theta})$, which accounts for the biased sampling design.

A second option for analyzing only subsampled individuals is to add information from the marker value by analyzing the joint conditional likelihood ("subsampled only, with covariates", or SO,WC)

$$\mathcal{L}_{SO,WC}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}, \boldsymbol{S}) = \prod_{S_i=1} f(\mathbf{X}_i|S_i = 1; \boldsymbol{\theta}) \cdot f(\mathbf{Y}_i|\mathbf{X}_i, S_i = 1; \boldsymbol{\theta}) \tag{2.5}$$

The bias induced by the sampling design can again be corrected through an ascertainment correction; however, in this case the population prevalence of the marker, p, must also be estimated. Moreover, the distribution upon which the SO,WC likelihood is based can be related to the SO,NC likelihood in the following way:

$$\begin{aligned} f(Y, \mathbf{X}|S = 1; \boldsymbol{\theta}) &= f(Y|\mathbf{X}, S = 1; \boldsymbol{\theta}) \cdot P(\mathbf{X}|S = 1; \boldsymbol{\theta}) \\ &= f(Y|\mathbf{X}, S = 1; \boldsymbol{\theta}) \cdot \frac{P(S = 1|\mathbf{X}; \boldsymbol{\theta}) \cdot P(\mathbf{X})}{P(S = 1; \boldsymbol{\theta})} \\ &= f(Y|\mathbf{X}, S = 1; \boldsymbol{\theta}) \cdot \frac{AC_0(m, \boldsymbol{t}; \boldsymbol{\theta}) \cdot P(\mathbf{X})}{P(S = 1; \boldsymbol{\theta})} \end{aligned}$$

$$= \quad f(Y|\mathbf{X}, S=1;\boldsymbol{\theta}) \cdot \frac{AC_0(m,\boldsymbol{t};\boldsymbol{\theta}) \cdot P(\mathbf{X})}{AC_1(\boldsymbol{\theta})} \tag{2.6}$$

The marginal probability of being sampled, AC_1, can be viewed as the expectation of the covariate-specific ascertainment correction $AC_0(m,\boldsymbol{t})$ taken over the distribution of M and $\boldsymbol{T}$ since

$$\begin{aligned}
AC_1(m,\boldsymbol{t}) &= P(S=1) \\
&= P(S=1,M=1) + P(S=1,M=0) \\
&= \int P(S=1,M=1,\boldsymbol{T}) + P(S=1,M=0,\boldsymbol{T})\; d\boldsymbol{t} \\
&= \int P(S=1|M=1,\boldsymbol{T}) \cdot P(M=1|\boldsymbol{T}) \cdot f(\boldsymbol{T}) \\
&\qquad +P(S=1|M=0,\boldsymbol{T}) \cdot P(M=1|\boldsymbol{T}) \cdot f(\boldsymbol{T})\; d\boldsymbol{t} \\
&= \int f(\boldsymbol{t}) \left[P(S=1|M=1,\boldsymbol{T}) \cdot P(M=1) + P(S=1|M=0,\boldsymbol{T}) \cdot P(M=0)\right]\; d\boldsymbol{t} \\
&= \int \mathbb{E}_M[S=1|M,\boldsymbol{T}] \cdot f(\boldsymbol{t})\; d\boldsymbol{t} \\
&= \mathbb{E}_{M,\boldsymbol{T}}[AC_0(M,\boldsymbol{T})]
\end{aligned}$$

It can be shown (see Appendix B) that in complete and balanced design situations (i.e., when all individuals are both observed n_i times and those observation times are the same across all individuals), the second term in Equation 2.6 is essentially a reparameterization of the marginal distribution of M. As such, this term provides no information about $\boldsymbol{\theta}$ for complete and balanced designs, and the resulting estimators SO,NC and SO,WC will be the same with respect to the target parameter $\boldsymbol{\theta}$, although SO,WC additionally estimates the marker population prevalence, p. When the design is not balanced, and cohort members may be observed at times that differ from one another, $AC_0(M=0,\boldsymbol{t};\boldsymbol{\theta})$ and $AC_0(M=1,\boldsymbol{t};\boldsymbol{\theta})$ may vary by marker/time combination, and the inclusion of covariates in inference may offer

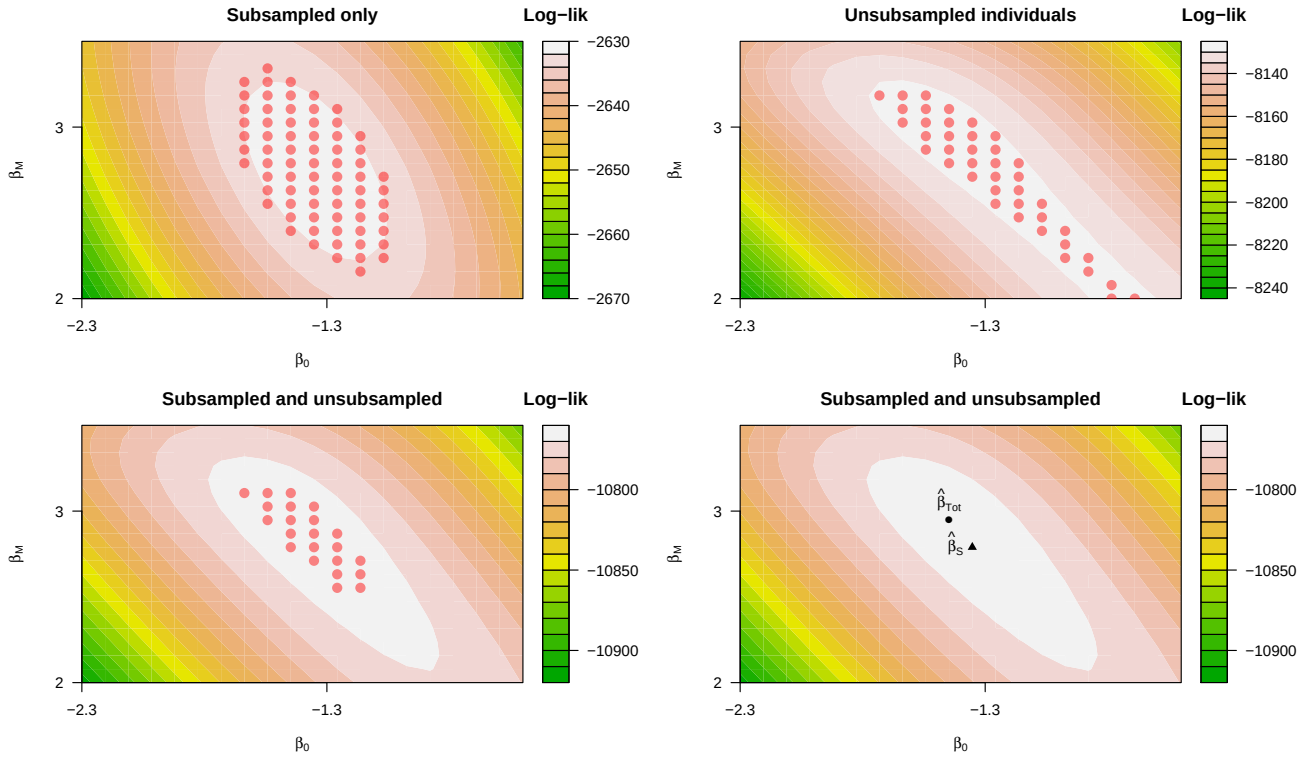
some additional information in this case.

2.3.4 Analysis: Inclusion of unsubsampled subjects

While estimators SO,NC and SO,WC exclusively utilize information from subsampled individuals, the inclusion of unsubsampled subjects may provide additional information. While unsubsampled subjects do not have the marker measured, and cannot directly provide information on the relationship of the expensive covariate with the outcome, the observed outcomes of unsubsampled individuals provide information on the population-level mixture of marker-specific mean outcomes. For example, at baseline, the mean outcomes for subjects with $M = 0$ and $M = 1$ under the usual linear mixed model are β_0 and $\beta_0 + \beta_M$, respectively. If p is the prevalence of the marker, observing the mean of Y among all cohort members at baseline would then give an estimate of $\mathbb{E}(Y|T = 0) = \beta_0 + \beta_M p$; the variance of Y is likewise related to combinations of regression parameters. Figure 2.4 shows representative contours from subsampled and unsubsampled subjects' contributions to the profile log-likelihood for these parameters, and illustrate the possible impact of including these individuals in the analysis. Notably, the log-likelihood contribution of unsubsampled subjects (top right panel of Figure 2.4) describes a ridge of linear combinations of the parameters related to the baseline mean of Y among all subjects, subject to constraints imposed by the observed variance. Adding information from the unsubsampled to the usual log-concave likelihood contributions from subsampled subjects (top left panel of Figure 2.4) has the potential to affect both estimation (i.e, orientation) and precision (the area of a 95% confidence region obtained by inversion), as seen in the lower left panel of Figure 2.4.

Incorporating the entire cohort of subsampled and unsubsampled subjects into the analysis and without including covariate information, we obtain estimator SU,NC ("subsam-

Figure 2.4: Profile log-likelihood contours showing the contribution of unsubsampled subjects. The characteristic "ridge" in the upper right panel reflects the fact that only the estimated population-level mean outcome is observed in these subjects. While many different combinations of regression parameters could give rise to the observed data, adding this information to an analysis based on subsampled subjects alone (upper left panel) may potentially improve inference for some parameters. The inclusion of unsubsampled individuals both changes the precision and orientation of a 95% confidence region obtained by inversion (lower left panel).



pled/unsubsampled, no covariates") by maximizing over the following likelihood:

$$\mathcal{L}_{SU,NC}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}, \boldsymbol{S}) = \prod_{S_i=1} f(\mathbf{Y}_i|\mathbf{X}_i, S_i = 1; \boldsymbol{\theta}) \cdot \prod_{S_i=0} f(\mathbf{Y}_i|S_i = 0; \boldsymbol{\theta})$$

Maximizing this conditional likelihood involves estimating the marginal marker prevalence, p; however, information on p is available only through the mixture distribution contributed by unsubsampled individuals. While this parameter is formally identifiable, it may not be easily estimable; to address this concern, we also evaluated another version of this estimator (denoted SU,NC+PI) that maximizes the same likelihood but uses a plug-in estimator of p, $\sum_{i=1}^{N} \frac{M_i \cdot \mathbb{I}(S_i = 1)}{N \cdot \omega(q_i)}$. Via the Central Limit Theorem, we expect the plug-in estimator to be consistent for p, since it has the proper expectation and finite variance.

$$\begin{aligned} \mathbb{E}_{M,Q,S,T}\left[M \cdot \frac{\mathbb{I}(S=1)}{\omega(q)}\right] &= \mathbb{E}_{M,Q,T}\left[\frac{M}{\omega(q)}\mathbb{E}_S[\mathbb{I}(S=1|M, \boldsymbol{T}, Q \in R_K)]\right] \\ &= \mathbb{E}_M[M] \\ &= p \end{aligned}$$

Just as estimator SO,WC added the covariate information to the conditional likelihood of estimator SO,NC, we could likewise add covariate information to estimator SU,NC. In contrast to the analyses that considered only subsampled individuals, including covariate information may prove beneficial when analyzing the entire cohort, since the covariate information from subsampled individuals can help to learn about the mixture distribution of unsubsampled subjects, which in turn informs inference about $\boldsymbol{\theta}$. Therefore, we also consider maximizing the likelihood conditioning only on sampling status ("subsampled/unsubsampled, with covariates", or SU,WC):

$$\mathcal{L}_{SU,WC}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}, \boldsymbol{S}) = \prod_{S_i=1} f(\mathbf{X}_i|S_i = 1; \boldsymbol{\theta}) \cdot f(\mathbf{Y}_i|\mathbf{X}_i, S_i = 1; \boldsymbol{\theta}) \cdot \prod_{S_i=0} f(\mathbf{Y}_i|S_i = 0; \boldsymbol{\theta})$$

Finally, we could analyze the unconditional likelihood, incorporating information from all subjects, marker values, time, and sampling status. We refer to the resulting estimator as "UC" (unconditional). Maximizing the various conditional and unconditional likelihoods using the Newton-Raphson algorithm, we have a variety of likelihood-based estimators of $\boldsymbol{\theta}$ that exploit different parts of the unconditional likelihood and thus differ in the information utilized. We compare the resulting estimators with respect to consistency and efficiency in Section 2.4.

2.4 Assessment of operating characteristics

2.4.1 Setup and data-generating mechanism

Previous work by Schildcrout et al. [39] showed large efficiency gains from an ODS design compared with a random sample of the same size, while Weaver et al. [46] demonstrated the added utility of analyzing unsampled individuals for cross-sectional data. Here, we evaluate the added incremental benefit of including information about covariates and/or information about unsubsampled subjects for longitudinal data with a continuous outcome. We compare the behavior of the likelihood-based estimators described in Section 2.3 to a random sample of the same size with respect to bias and efficiency, both analytically and through simulation.

For each replication we generated independent and identically distributed data for $N = 1000$ subjects from the linear model

$$y_{ij} = \beta_0 + \beta_T t_{ij} + \beta_M m_i + \beta_{M\times T} t_{ij} m_i + b_{0i} + b_{1i} t_{ij} + e_{ij},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5)$, $i = 1, \ldots, 1000$, $j = 1, \ldots, n_i$, where n_i was either 6 or 11 and observation times were equally spaced. The expensive time-invariant marker, m_i, had a prevalence of 10%. Random effects $\boldsymbol{b}_i = (b_{0i}, b_{1i})$ were multivariate normally distributed with mean $\mathbf{0}$ and 2×2 covariance matrix variance $\mathbf{D}$, with

variances $\sigma^2_{b_0}$ and $\sigma^2_{b_1}$ on the diagonal and covariance off-diagonal element of 0 ($\rho = 0$). Errors e_{ij} were generated to be conditionally independent and normally distributed with mean 0 and variance σ^2_e. We examined estimator performance under two variance component scenarios: one with low subject-to-subject heterogeneity ($\sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$), and one with high subject-to-subject heterogeneity ($\sigma^2_{b_0} = \sigma^2_{b_1} = \sigma^2_e = 4$). Simulation results reported here are based on 1000 replications.

Subjects were selected for marker ascertainment based on the 2×1 vector of subject-specific regression coefficients $\boldsymbol{q}_i = \left(\mathbf{X}_{ti}^T \mathbf{X}_{ti}\right)^{-1} \mathbf{X}_{ti} \mathbf{Y}_i$, where $\mathbf{X}_{ti} = [\mathbf{1}, \boldsymbol{T}_i]$. We considered two sampling schemes, either selecting subjects for subsampling based on the value of their subject-specific intercept or their subject-specific slope, both of which we derived from regressing each cohort member's outcome vector on observation times. In each case, we selected a subsample of 250 on average, with an average of 100 individuals from the lowest 20th percentile, 50 individuals from the middle 60%, and 100 individuals from the highest 20th percentile of the subsampling variable Q. Both intercept- and sloped-based outcome dependent samples were analyzed using the approaches described in Section 2.3. To ensure that estimates obeyed parameter constraints such as positive variance, we transformed the variance components as described in Section 2.3.1, and used the Newton-Raphson algorithm to maximize over the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})^T$, plus the transformed population marker prevalence, logit p, for estimators that required it. We compare the estimates resulting from each ODS design/analysis combination with the estimate obtained from a random sample of 250 individuals, and with the estimate obtained from a usual linear mixed model using all 1000 individuals from the original cohort.

2.4.2 Validity and relative efficiency: simulation

Tables 2.2 and 2.3 give the average percent relative bias for each estimator, and the efficiency of the estimator relative to a random sample of the same size (on average). Results are shown

Table 2.2: Percent bias and relative efficiency for likelihood-based ODS estimators, under low subject-to-subject heterogeneity. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator. NA for percent bias indicates that percent bias is undefined.

Estimator									
Design	**Likelihood analysis**	β_0	β_T	β_M	$\beta_{M\times T}$	γ_1	γ_2	γ_3	γ_4
Full cohort	Standard	0[3.91]	0[3.88]	-1[3.98]	0[4.02]	0[4.14]	NA[4.39]	0[4.14]	NA[3.71]
Random sample	Standard	0[1.00]	1[1.00]	0[1.00]	-1[1.00]	-2[1.00]	NA[1.00]	-1[1.00]	NA[1.00]
Intercept	SO, NC	0[1.65]	-1[1.00]	1[1.48]	-1[1.07]	-1[1.52]	NA[1.87]	1[0.95]	NA[0.92]
	SO, WC	0[1.65]	-1[1.00]	1[1.48]	-1[1.07]	-1[1.52]	NA[1.87]	1[0.95]	NA[0.92]
	SU, NC	0[2.96]	0[1.44]	2[1.54]	0[1.14]	0[3.72]	NA[3.69]	0[2.43]	NA[3.71]
	$SU, NC + PI$	0[3.42]	0[2.56]	2[1.56]	-1[1.20]	0[3.72]	NA[3.89]	0[2.87]	NA[3.71]
	SU, WC	0[3.44]	0[2.86]	2[1.52]	0[1.18]	0[3.73]	NA[3.89]	0[2.94]	NA[3.71]
	UC	0[3.43]	0[2.86]	2[1.51]	0[1.18]	0[3.83]	NA[3.90]	0[2.94]	NA[3.70]
Slope	SO, NC	0[1.08]	0[1.80]	2[1.17]	2[1.37]	-1[0.96]	NA[1.70]	1[1.36]	NA[1.03]
	SO, WC	0[1.08]	0[1.80]	2[1.17]	2[1.37]	-1[0.96]	NA[1.70]	1[1.36]	NA[1.03]
	SU, NC	0[2.69]	0[2.24]	3[1.10]	2[1.39]	0[3.65]	NA[3.50]	0[2.57]	NA[3.71]
	$SU, NC + PI$	0[3.14]	0[3.16]	2[1.18]	1[1.46]	0[3.83]	NA[3.65]	0[3.06]	NA[3.71]
	SU, WC	0[3.16]	0[3.42]	3[1.07]	2[1.42]	0[3.82]	NA[3.71]	0[3.14]	NA[3.71]
	UC	0[3.19]	0[3.43]	3[1.15]	2[1.48]	0[3.85]	NA[3.75]	0[3.33]	NA[3.71]

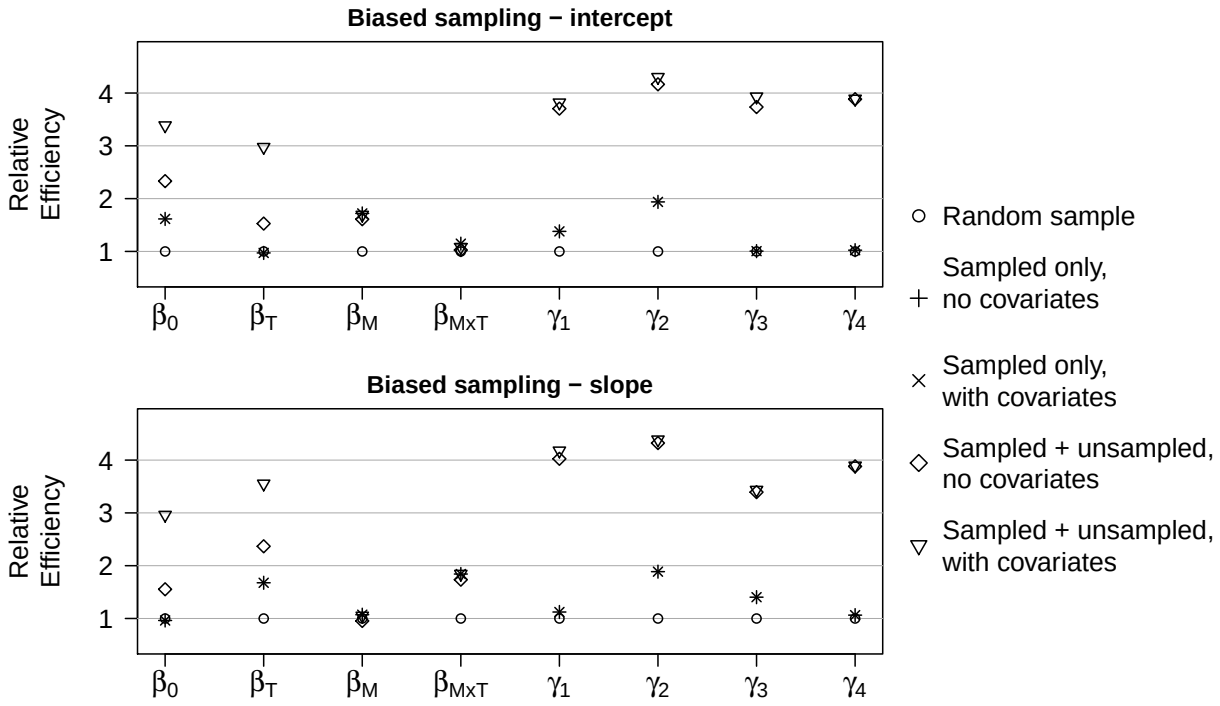
for a constant cluster size of $n_i = 6$; results when $n_i = 11$ were qualitatively similar and are not shown. For all design and analysis methods, estimates for regression and variance parameters showed little bias, generally $< 5\%$. Analysis methods that additionally estimated the population prevalence p of the expensive covariate were likewise unbiased, with the exception of Estimator SU,NC. This method experienced convergence issues related to the parameter p a substantial fraction of the time, which led to widely variable estimates of p, although the other parameters of interest continued to be correctly estimated.

Consistent with results seen by Schildcrout et al. [39], Estimator SO,NC offered major

Table 2.3: Percent bias and relative efficiency for likelihood-based ODS estimators, under high subject-to-subject heterogeneity. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator. NA for percent bias indicates that percent bias is undefined.

Estimator									
Design	**Likelihood analysis**	β_0	β_T	β_M	$\beta_{M\times T}$	γ_1	γ_2	γ_3	γ_4
Full cohort	Standard	0[3.98]	0[3.97]	0[4.09]	0[4.46]	-1[4.39]	NA[4.61]	0[4.09]	0[3.90]
Random sample	Standard	0[1.00]	1[1.00]	0[1.00]	-1[1.00]	-2[1.00]	NA[1.00]	-1[1.00]	0[1.00]
Intercept	$SO + NC$	0[1.62]	1[0.97]	0[1.72]	2[1.15]	-1[1.38]	NA[1.94]	-1[1.01]	0[1.02]
	SO, WC	0[1.62]	1[0.97]	0[1.72]	2[1.15]	-1[1.38]	NA[1.94]	-1[1.01]	0[1.02]
	SU, NC	0[2.33]	6[1.53]	0[1.61]	2[1.03]	-1[3.70]	NA[4.16]	-1[3.73]	0[3.88]
	$SU, NC + PI$	0[3.32]	-1[2.28]	-1[1.72]	2[1.11]	-1[3.80]	NA[4.28]	0[3.92]	0[3.89]
	SU, WC	0[3.39]	0[2.98]	-1[1.69]	2[1.08]	-1[3.82]	NA[4.30]	0[3.93]	0[3.89]
	UC	0[3.38]	0[2.97]	-1[1.71]	2[1.08]	-1[4.01]	NA[4.33]	0[3.92]	0[3.89]
Slope	SO, NC	0[0.96]	0[1.68]	0[1.07]	2[1.84]	-3[1.12]	NA[1.89]	-1[1.40]	0[1.06]
	SO, WC	0[0.96]	0[1.68]	0[1.07]	2[1.84]	-3[1.12]	NA[1.89]	-1[1.40]	0[1.06]
	SU, NC	0[1.55]	4[2.37]	0[0.95]	1[1.74]	-1[4.02]	NA[4.32]	0[3.39]	0[3.88]
	$SU, NC + PI$	0[2.96]	0[3.53]	0[1.07]	2[1.84]	-1[4.18]	NA[4.36]	0[3.43]	0[3.89]
	SU, WC	0[2.96]	0[3.55]	1[1.05]	2[1.84]	-1[4.18]	NA[4.39]	0[3.43]	0[3.89]
	UC	0[2.97]	0[3.56]	0[1.07]	2[1.86]	-1[4.17]	NA[4.40]	0[3.69]	0[3.89]

Figure 2.5: Relative efficiencies of ODS estimators via simulation, compared with a random sample of $N_S = 250$. For comparison, note that analyzing full cohort ($N = 1000$) would give a true relative efficiency of 4.



efficiency gains over random sampling for selected regression parameters, which depended on the subsampling design used. When subject-specific intercept was the subsampling variable, the greatest efficiency gains occurred for time-invariant parameters β_0 and β_M, while time-varying parameters β_T and $\beta_{M\times T}$ had the greatest gains when subject-specific slope was used (Figure 2.5). As previously discussed, when only subsampled individuals' information was analyzed, incorporating covariate information (Estimator SO,WC) into inference did nothing to change the relative efficiency of the estimator compared with the conditional version (Estimator SO,NC). In fact, these two estimators were numerically equivalent, up to convergence of the respective algorithms.

Adding unsubsampled individuals to the analysis produced substantial gains in efficiency for some regression parameters and all for variance components, regardless of the ODS design. For variance components, estimators that included unsubsampled subjects (Estimators SU,NC, SU,WC, and UC) recovered nearly all the information from the full cohort; for regression parameters, only β_0 and β_T had improved efficiency. Augmenting information from unsubsampled individuals without also considering covariate information (Estimator SU,NC) produced an estimator that often had convergence issues. Unlike the situation when only subsampled individuals were considered, analyzing the joint likelihood improved efficiency over the conditional likelihood when unsubsampled individuals were included (Estimators SU,WC vs SU,NC). However, using a plug-in estimator of p (Estimator SU,NC+PI) was nearly as efficient as incorporating covariate information formally into the likelihood (Estimator SU,WC). Almost no benefit was seen in analyzing the unconditional likelihood (Estimator UC) over a likelihood approach that conditioned on sampling status and time (Estimator SU,WC).

We also evaluated the benefit of ODS design and analysis for the time-specific difference in expected outcome, $\Delta_t = \mu_1(t) - \mu_0(t) \equiv \mathbb{E}(Y|M = 1, T = t) - \mathbb{E}(Y|M = 0, T = t) = \beta_M + \beta_{M\times T} \cdot t$. Percent bias and relative efficiency for Δ_t are summarized in Table 2.4. For baseline ($t = 1$) comparisons, the highest relative efficiency came from intercept-based designs; for $t = 6$, when Δ_t is more highly weighted toward the time-related parameter $\beta_{M\times T}$, the greater efficiency came from slope-based designs. In neither case did the analysis approach appear to have an impact on the relative efficiency.

2.4.3 Validity and relative efficiency: analytical

Each of the estimators examined here is a maximum likelihood estimator, and as such will be asymptotically unbiased for $\boldsymbol{\theta}$ under correct model specification. The asymptotic relative efficiency of the estimators can likewise be found through an analytical comparison of the

Table 2.4: Percent bias and relative efficiency for time-specific predicted means and predicted difference in means. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

Estimator	$\mu_0(1)$	$\mu_1(1)$	Δ_1	$\mu_0(6)$	$\mu_1(6)$	Δ_6
SO, NC_{int}	0 [1.36]	0 [1.41]	-4 [1.37]	0 [0.99]	0 [1.16]	2 [1.13]
SO, WC_{int}	0 [1.36]	0 [1.41]	-4 [1.37]	0 [0.99]	0 [1.16]	2 [1.13]
SU, NC_{int}	0 [2.20]	0 [1.38]	-4 [1.23]	0 [1.59]	0 [1.12]	2 [1.00]
$SU, NC + PI_{int}$	0 [3.46]	0 [1.44]	-5 [1.34]	-1 [2.99]	0 [1.17]	3 [1.09]
SU, WC_{int}	0 [3.49]	0 [1.40]	-5 [1.32]	0 [3.07]	0 [1.14]	2 [1.06]
UC_{int}	0 [3.50]	0 [1.40]	-6 [1.32]	0 [3.07]	1 [1.14]	3 [1.06]
SO, NC_{slope}	0 [1.14]	0 [1.13]	-4 [1.11]	0 [1.65]	0 [1.74]	2 [1.71]
SO, WC_{slope}	0 [1.14]	0 [1.13]	-4 [1.11]	0 [1.65]	0 [1.74]	2 [1.71]
SU, NC_{slope}	0 [1.90]	0 [1.13]	-2 [1.00]	-1 [2.43]	0 [1.74]	2 [1.60]
$SU, NC + PI_{slope}$	0 [3.24]	0 [1.19]	-3 [1.09]	0 [3.57]	0 [1.77]	2 [1.70]
SU, WC_{slope}	0 [3.22]	0 [1.18]	-2 [1.08]	0 [3.58]	0 [1.78]	3 [1.70]
UC_{slope}	0 [3.23]	0 [1.21]	-3 [1.11]	0 [3.58]	0 [1.80]	2 [1.72]

information in each estimator. For estimators that included only subsampled individuals, we calculated this directly; for estimators involving all cohort members we used a Monte Carlo approach to ascertain relative efficiency. In addition to the intercept- and slope-based ODS designs described above, we also investigated the relative efficiency obtained from an ODS design that used the intercept-based criterion to choose half of the subsample and the slope-based criterion to choose the other half.

The relative efficiency of the ODS designs considered here depends not only on $\boldsymbol{\theta}$ itself, but also on two key ODS design parameters, where the oversampling region is defined to be and the sampling fractions ($\omega(q)$) for each region. For a cohort of 10,000 individuals and for Estimators SO,NC and SU,WC, Figures 2.6 and 2.7 show the relative efficiency for regression parameters obtained when varying these key ODS design parameters. In Figure 2.6 the oversampling region is kept constant and we illustrate the effect of varying the number oversampled in that region; designs in Figure 2.7 sample 1000 from Regions 1 and 3 but vary how extreme those regions are. Figure 2.8 varies both these parameters, obtaining a surface of relative efficiencies. Unsurprisingly, the greatest efficiency gains occur for parameters the design has specifically targeted (i.e., β_0 and β_M for intercept-based designs, β_T and $\beta_{M\times T}$ for slope-based designs), and for designs that have a large number of individuals subsampled from the most extreme regions of Q.

Figure 2.6: Relative efficiencies of estimators SO,NC and SU,WC for various ODS designs, varying the number subsampled from the top/bottom 20th percentiles. The "bivariate" sampling design subsampled half of subjects based on subject-specific intercept and half based on subject-specific slope.

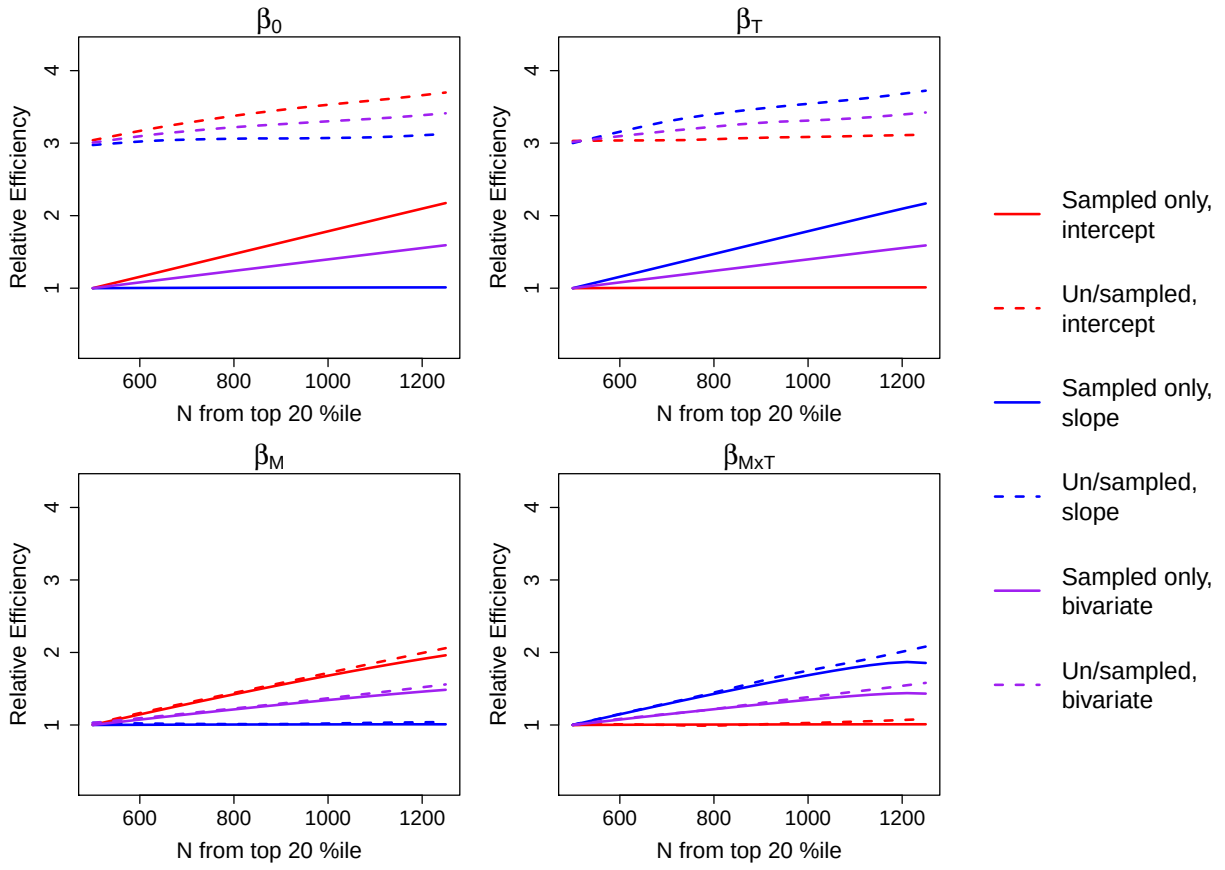


Figure 2.7: Relative efficiencies of estimators SO,NC and SU,WC for various ODS designs, varying the oversampling percentile from which 1000 subjects are subsampled. The "bivariate" sampling design subsampled half of subjects based on subject-specific intercept and half based on subject-specific slope.

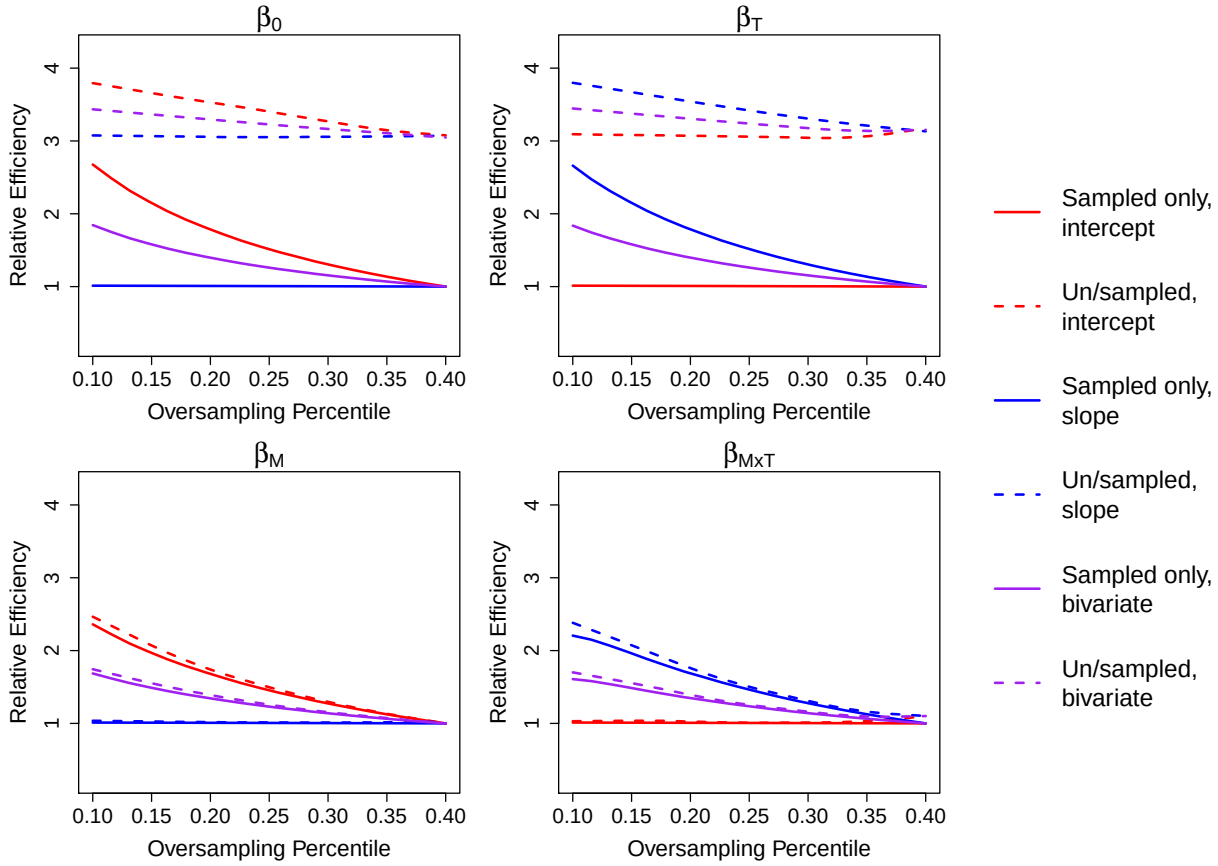
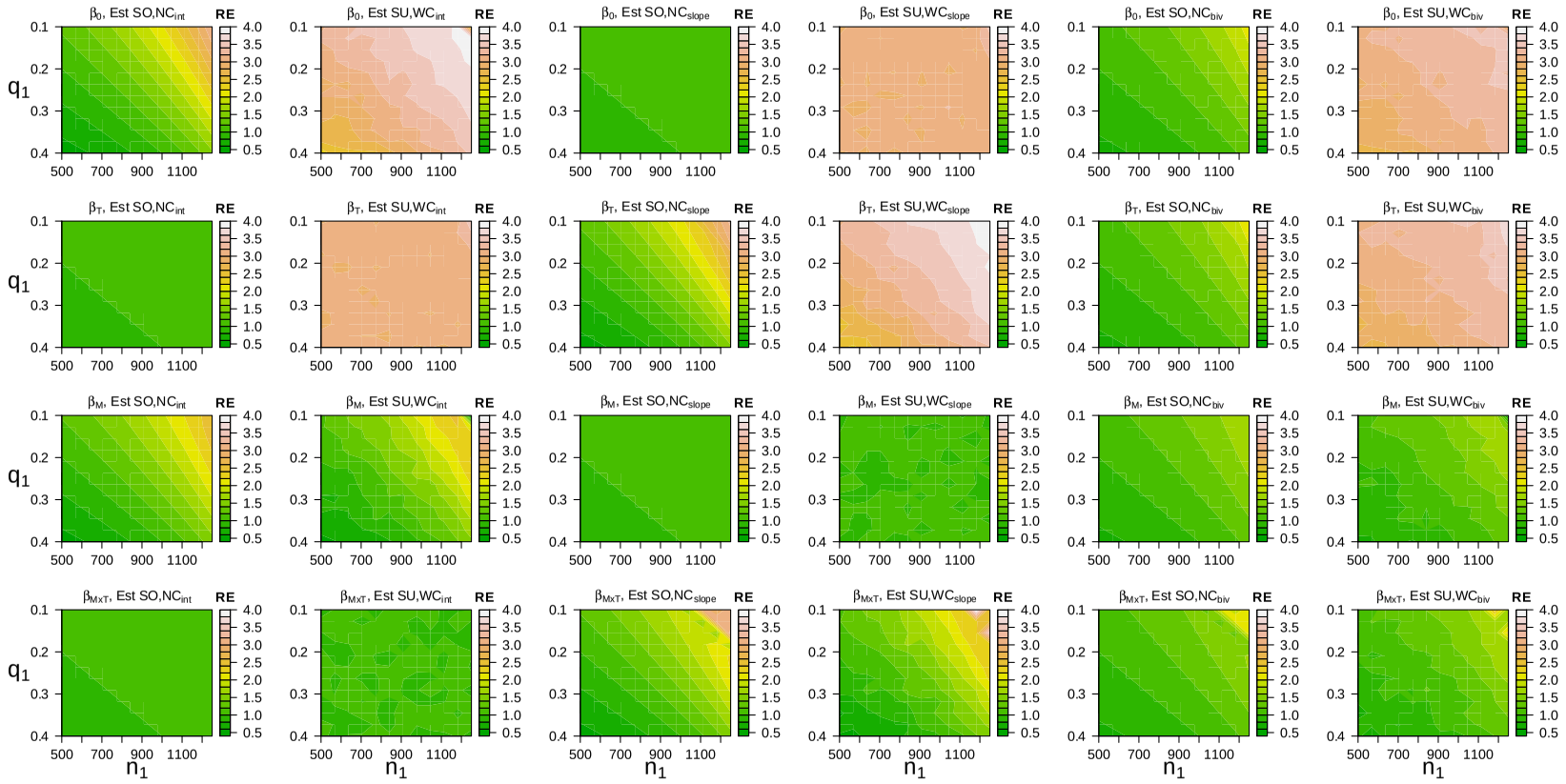


Figure 2.8: Relative efficiencies of estimators SO,NC and SU,WC for various ODS designs, varying both oversampling percentile and the number oversampled from that region. The "bivariate" sampling design subsampled half of subjects based on subject-specific intercept and half based on subject-specific slope.



2.5 Application to Cystic Fibrosis Foundation Registry data

We illustrate the relative merits of the likelihood-based estimators discussed here with an application to data from the Cystic Fibrosis (CF) Foundation Patient Registry, which collects detailed information on the health outcomes, clinical care, and demographic characteristics of patients with CF receiving care at accredited centers. For this illustration, we identified a cohort of 3,141 CF registry patients between the ages of 8 and 16 who were genotyped at diagnosis, had at least 6 equally spaced longitudinal spirometry measurements available, and whose inital spirometry measurement occurred between 1990 and 2006. Selected summary statistics for this cohort can be found in Table 2.5; the average age of participants was 8.8 years old, with the cohort split equally between boys and girls. Seventy percent of the cohort tested positive for the bacterium *Staphylococcus aureus* at baseline. We evaluated the impact of ODS design and analysis of a hypothetical substudy to investigate the longitudinal association between the presence of the bacterium *S. aureus* at baseline and FEV_1 (L), a measure of lung function.

Table 2.5: Baseline characteristics of Cystic Fibrosis Foundation Registry cohort (N = 3,141).

Characteristic	Value
N	3,141
Age (years)	8.8 (1.1)
Male sex, N (%)	1584 (50.4)
Height (cm)	125.4 (8.4)
Weight (kg)	25.9 (6.0)
Presence of *S. aureus*, N (%)	2206 (70.2)
FEV_1 (L)	1.4 (0.4)

Values are mean (SD), except as noted.

In this context, the parameters of interest include the effect of the presence of *S. aureus*

at baseline, and the difference in the slopes of lung function trajectory between those with and without baseline *S. aureus*. Since genotyping in reality was conducted on all patients for this cohort of 3,141, we can evaluate the performance of hypothetical substudies that utilize ODS design and analysis relative to the gold standard of analyzing the entire cohort. For substudies in a large cohort such as the CF registry cohort, covariate information such as an expensive or technologically complex biomarker assay will generally be ascertained only in a small subset; hence, ODS techniques can be of use in choosing and analyzing the cohort subset. We evaluated ODS design and analysis approaches for conducting a substudy of 600 patients on average, selected either by random or biased sampling based on subject-specific intercept or slope. Average parameter estimates and standard errors over 1000 resamplings of the data are presented in Table 2.6.

As shown in Table 2.6, patients infected with *S. aureus* at baseline had FEV_1 scores that were about 13 mL lower at baseline (indicating worse lung function) than patients who were not infected, although not significantly so. On average, patients' lung function tended to improve over time, probably as a result of physical growth. However, lung function for patients infected by *S. aureus* at baseline improved 11 mL less than those who were bacteria-free at baseline, a small but statistically significant effect present in the full cohort ($p = 0.03$) that was not detected by a random sample design. In fact, only subject-specific slope-based substudy designs that targeted the interaction regression parameter would have detected this difference.

In general, the inclusion of unsubsampled individuals in the analysis produced smaller standard errors for all parameters; however, the smallest standard errors overall corresponded to the design (SO,WC, slope) that both targeted the parameter of interest and incorporated unsubsampled individuals. Overall, the patterns observed in the empirical standard errors from the CF analysis tended to agree with simulation results in Section 2.4.2. Diagnostic plots of subject-specific intercepts and slopes (Figure 2.9) suggest that the distribution of

random effects were not inconsistent with bivariate normality, the violation of which may impact the expected performance of likelihood-based ODS estimators.

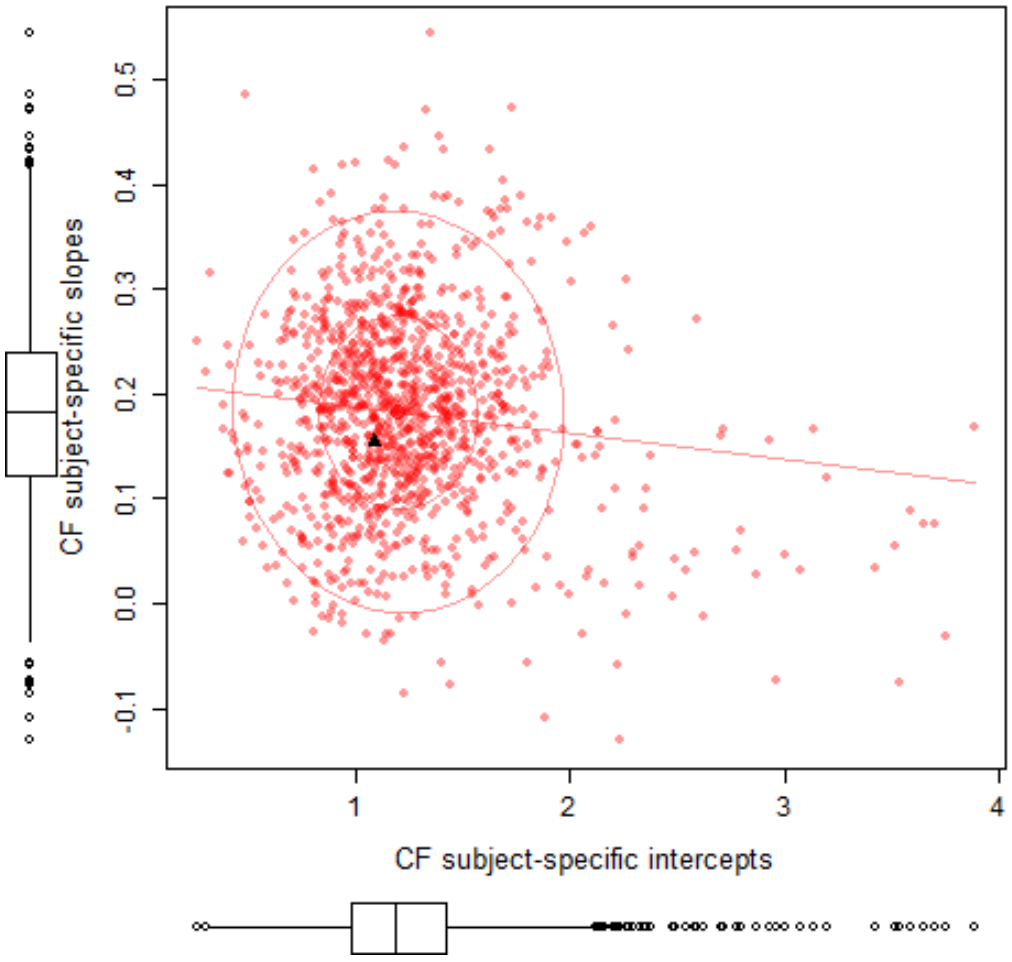
2.6 Discussion

In this chapter, we have explored the marginal utility of multiple sources of information in the analysis of ODS designs in the longitudinal data setting. We have chosen to examine a simple binary covariate, although we expect the lessons learned here to be broadly similar for a categorical or continuous covariate. In this simple case, we have shown benefit, sometimes substantial, of incorporating additional sources of information into inference. All of the likelihood-based estimators investigated here accounted for the biased sampling design though an ascertainment correction approach. Other valid strategies of estimation exist: for example, the inverse probability weighting (IPW) approach [12], which produces estimators that are valid under mild conditions but often inefficient in modest samples [32]. An exploration of estimators that balance robustness to misspecification and efficiency in the ODS setting, and comparison with likelihood-based estimators, can be found in Chapter 3.

In evaluating these estimators, we have followed the example of Schildcrout et al. [39], and conditioned on the marginal sampling status S. In contrast, some previous work (e.g., Zhou et al. [49] and Neuhaus et al. [26]) conditioned on being sampled from stratum k. Although not considered explicitly here, we expect that finer conditioning would produce a loss of information relative to the conditioning explored here.

We have observed that ODS analysis choices have the potential to improve efficiency for targeted regression parameters, sometimes dramatically, at minimal cost. However, the utility of incorporating covariate information into inference depends on the choice of subjects to analyze. When analyzing subsampled individuals only, we have shown that in the case of complete and balanced designs, there is no benefit. When there is variability in measurement times across categories of the marker, there may be a small amount of information to

Figure 2.9: Subject-specific intercepts and slopes resulting from regressing a CF patient's FEV_1 longitudinal outcome on time. The distribution of intercepts and slopes do not appear to be clearly inconsistent with bivariate normality.



be gained by adding in covariate information. When all cohort members are analyzed, however, we have observed sometimes substantial increases in efficiency, as observed covariate information among subsampled individuals allows for a more precise characterization of the mixture distribution among unsubsampled subjects.

The benefit of including unsubsampled individuals in inference, previously explored for univariate outcomes, carries over to longitudinal data, albeit for selected regression parameters only. Analysis of all subjects allowed nearly full information to be recovered for the variance components, which may be of interest in some applications. This type of analysis also improved inference on some regression parameters, although minimally for those related to the unobserved covariate; greater efficiency for these parameters will need to be addressed primarily through careful choice of ODS designs that promote efficiency for them, not through analysis. We have additionally illustrated the effects of some ODS design choices; a more thorough examination of these practical design parameters in the future will be helpful to the researcher implementing these methods. For researchers planning substudies based on existing longitudinal data, there appears to be utility in both careful design and analysis of biased sampling approaches. Overall, our results suggest that thoughtfulness at both design and analysis stages will be rewarded, sometimes substantially.

Table 2.6: Parameter estimates and empirical standard errors of likelihood-based estimators for the Cystic Fibrosis Foundation Registry dataset (N = 3,141). For "full" estimator, standard errors are derived from analysis of full Cystic Fibrosis Foundation Registry cohort. All other estimators are based on an average subsample of 600 patients, and results are averaged over 1000 resamplings. Empirical standard errors are the estimator's standard deviation over all resamplings.

Estimator	β_0			β_T			$\beta_{S.aureus}$			$\beta_{S.aureus \times T}$		
	Est.	(SE)	p	Est.	(SE)	p	Est.	(SE)	p	Est.	(SE)	p
Full cohort	1.200	(0.012)	< 0.001	0.184	(0.0029)	< 0.001	-0.013	(0.015)	0.79	-0.011	(0.0035)	0.03
Random sample	1.202	(0.025)	< 0.001	0.185	(0.0061)	< 0.001	-0.013	(0.030)	0.67	-0.012	(0.0073)	0.11
SO,NC, intercept	1.213	(0.019)	< 0.001	0.181	(0.0057)	< 0.001	-0.014	(0.023)	0.54	-0.012	(0.0066)	0.06
SU,WC, intercept	1.201	(0.014)	< 0.001	0.184	(0.0041)	< 0.001	-0.013	(0.020)	0.52	-0.011	(0.0059)	0.06
SO,NC, slope	1.191	(0.024)	< 0.001	0.185	(0.0041)	< 0.001	0.013	(0.029)	0.67	-0.013	(0.0049)	0.007
SO,WC, slope	1.185	(0.017)	< 0.001	0.185	(0.0032)	< 0.001	0.009	(0.025)	0.70	-0.013	(0.0046)	0.007

Chapter 3

ROBUST LONGITUDINAL OUTCOME DEPENDENT SAMPLING ESTIMATORS

3.1 Introduction

The likelihood-based ODS estimators described in Chapter 2 have the potential to provide increased efficiency compared to a random sample, through a combination of designs that preferentially subsample more informative individuals, and analysis methods, through which information from unsubsampled subjects and covariates may also be incorporated to good effect into inference. In these methods, the biased sampling design is accounted for by writing the conditional log-likelihood as a term which ignores the design, less an ascertainment correction which characterizes the bias induced by the sampling scheme. Since the subsampling variable Q is a linear combination of the outcome vector $\mathbf{Y}$ in these designs, an accurate characterization of the bias stems directly from a correct characterization of the distribution of $\mathbf{Y}|\mathbf{X}$. While the likelihood-based estimators considered in Chapter 2 may differ in terms of which subjects to subsample, the information to include, and the subsampling design, the methods reviewed above all have the same basic form.

Being likelihood-based, these estimators are guaranteed to be asymptotically unbiased and efficient under mild regularity conditions. Under multivariate normality of the linear mixed model's random effects, the subsampling variable Q likewise has a normal distribution; in all methods presented in Chapter 2, this was presumed to be the case. However, under violation of normality of the random effects or other model misspecification, a poor ascertainment correction may result in biased estimation. The estimation of regression parameters in the usual longitudinal linear mixed model has been shown to be moderately

robust to misspecification of normality of random effects, which might suggest the impact of misspecification for these ODS estimators to be modest. However, likelihood-based ODS methods in particular rely on characterization of the tails of the normal distribution, so these estimators might be expected to be less robust than the usual linear mixed models. In general, there continues to be great interest in development of methods that can balance the efficiency of likelihood-based methods with robustness to model misspecification.

3.2 Methods

3.2.1 Inverse probability weighted estimators

One classic solution to the challenge of creating robust estimators involves reweighting of selectively subsampled individuals. The Horvitz-Thompson estimator [12], for example, solves

$$\sum_{i=1}^{n} \frac{S_i}{\pi(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta})} U_i(\boldsymbol{\theta}) \quad = \quad 0, \tag{3.1}$$

for estimating function $U_i(\boldsymbol{\theta})$ and sampling probability $\pi = P(S_i = 1|\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta})$ bounded away from 0, where S_i is the indicator of subsampling and π is bounded away from 0. The inverse probability weighted (IPW) estimator has become a common tool in the biased sampling world, where an exposure of interest is retrospectively ascertained on a subset of participants. Each subsampled individual's contribution to the estimating equation is up-weighted by a factor of $1/\pi$, reconstructing a pseudo-population similar to that from which the target of inference $\boldsymbol{\theta}$ arises, and sandwich-based formulations of the asymptotic variance are easily obtained. So long as the estimating function U_i is unbiased, i.e., $\mathbb{E}[U(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X})] = 0$, the resulting IPW estimator $\hat{\boldsymbol{\theta}}_{IPW}$ can correct for potential selection biases, if the subsampling probability π is correctly specified (generally expected to be true in most ODS applications). However, this estimator may be extremely inefficient in modest samples [32], thus only partially achieving the desired goal of balancing efficiency with robustness to model

misspecification in the ODS setting.

3.2.2 Augmented inverse probability weighted estimators

Our results from Chapter 2, together with recently published work by Schildcrout et al. [38], suggest the potential benefit of incorporating unsubsampled individuals into likelihood-based estimators. A class of robust estimators called augmented inverse probability weighted (AIPW) estimators, proposed by Robins et al. [32] in 1994, likewise attempted to make use of all observed data. In these, the IPW estimating equations in Equation 3.1 which arise from completely observed data are supplemented by auxiliary information from partially observed individuals, who are included in the estimating equation via their expected contribution to the estimating function, given the partially observed data.

Robins et al. [32] define the class of AIPW estimators within the semiparametric framework of a conditional mean model, MAR (in the Rubin sense) missingness, and a probability of missingness π_i that is bounded away from 0, and is either known or can be modeled parametrically. Specifically, the class of AIPW estimators solves the equation

$$\sum_{i=1}^{n} \frac{S_i}{\pi} U_i(\boldsymbol{\theta}) + \left(1 - \frac{S_i}{\pi(\mathbf{Y}_i, \boldsymbol{T}_i; \boldsymbol{\theta})}\right) \phi = 0$$

for some estimating function U_i and arbitrary function of the completely observed data ϕ. Moreover, Robins et al. showed that every regular asymptotically linear estimator of $\boldsymbol{\theta}$ can be shown to be equivalent to some estimator in this class, and showed that a variety of previously proposed estimators are (typically inefficient) members of the AIPW class.

As an example of an AIPW that could be constructed for ODS data, suppose we have a matrix of covariates $\mathbf{X} = (\boldsymbol{T}, \boldsymbol{M})^T$ that includes a covariate such as time, $\boldsymbol{T}$, that is completely observed, and another covariate such as the biomarker $\boldsymbol{M}$, to be ascertained on a subset. As a special case of AIPW estimators that can always be constructed, Lipsitz et

al. [17] showed that, for data where the missing covariate is MCAR or MAR in the Rubin [35] sense, solving the weighted estimating equation (WEE)

$$\sum_{i=1}^{n} \frac{S_i}{\pi(\mathbf{Y}_i, \boldsymbol{T}_i; \boldsymbol{\theta})} U_i(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i) + \sum_{i=1}^{n} \left\{ 1 - \frac{S_i}{\pi(\mathbf{Y}_i, \boldsymbol{T}_i; \boldsymbol{\theta})} \right\} \mathbb{E}\{U_i(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}) | \mathbf{Y}_i, \boldsymbol{T}_i\} = 0 \quad (3.2)$$

produces a "doubly robust" estimator $\hat{\boldsymbol{\theta}}_{AIPW}$. Consistency of $\hat{\boldsymbol{\theta}}_{AIPW}$ is guaranteed under misspecification of the probability of the biomarker's missingness, $\pi(\mathbf{Y}_i, \boldsymbol{T}_i)$, or under misspecification of the conditional distribution of the covariates, $g(M|\boldsymbol{T})$, provided that $\mathbb{E}[U_i(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})] = 0$ for some complete data estimating function U_i (see Appendix D for derivation of double robustness). In our ODS application, missingness is determined primarily by design and π_i will generally be known, and thus correctly specified, so the doubly robust property may be of secondary importance. However, this implementation of the AIPW estimator provides an approach to robustness that utilizes all available data and may be less reliant on model assumptions than likelihood-based methods.

Because it utilizes all of the data, the AIPW estimator might be presumed to be more efficient than the standard IPW estimator. In fact, this is guaranteed to be true if both $\pi(\mathbf{Y}, \boldsymbol{T})$ and $g(M|\boldsymbol{T})$ are correctly specified [44]. While Robins et al. [32] detailed the construction of the semiparametric efficient estimator, which involves the projection of a Hilbert space onto a linear subspace, the version of the AIPW estimator outlined in Equation 3.2 need not achieve the semiparametric efficiency bound (in fact, even the optimal AIPW estimator need only asymptotically achieve semiparametric efficiency). As illustration of this fact, Lumley et al. [21] [22] give the example of case-control sampling of an outcome Y to measure covariate Z. Under the model $\text{logit}[Y_i = 1|Z_i = z] = z\theta$, unconditional logistic regression produces a maximum likelihood estimator that is also semiparametric efficient [28] [3], but not equivalent to any AIPW estimator.

Importantly, calculation of the AIPW estimator requires knowledge of the expectation

of U_i taken over the distribution of $M|\mathbf{Y}, \boldsymbol{T}$, which may not be known. Lipsitz et al. [17] recommended finding this expectation via an expectation-maximization (EM) [6] approach. Under normality of $\mathbf{Y}|M, \boldsymbol{T}$ with a binary marker M, the distribution of $M|\mathbf{Y}, \boldsymbol{T}$ has a recognizable and closed form (Appendix C) and at each iteration this expectation can be found exactly, given the current estimate $\hat{\boldsymbol{\theta}}^{(k)}$. As outlined by Lipsitz et al., a strategy for solving the system of equations given in Equation 3.2 might then be:

1. Obtain an estimate $\hat{\boldsymbol{\theta}}^{(t)}$. An IPW or naïve analysis of complete cases can be used to initialize $\hat{\boldsymbol{\theta}}^{(t)} = \hat{\boldsymbol{\theta}}^{(1)}$.

2. Using $\hat{\boldsymbol{\theta}}^{(t)}$, estimate $\mathbb{E}[U_i(\mathbf{Y}, \boldsymbol{T}, M; \boldsymbol{\theta})|\mathbf{Y}, \boldsymbol{T}]$ by calculating

$$\sum_m P(M = m|\mathbf{Y}, \boldsymbol{T}; \hat{\boldsymbol{\theta}}^{(t)}) \cdot U_i(\mathbf{Y}, \boldsymbol{T}, M = m; \hat{\boldsymbol{\theta}}^{(t)})$$

3. Treating the estimate $\mathbb{E}[U_i(\mathbf{Y}, \boldsymbol{T}, M; \hat{\boldsymbol{\theta}}^{(t)})|\mathbf{Y}, \boldsymbol{T}]$ as fixed, solve Equation 3.2 for $\hat{\boldsymbol{\theta}}^{(t+1)}$.

4. Iterate steps 1-3 until convergence, when $\hat{\boldsymbol{\theta}}^{(t)} = \hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}_{AIPW}$ is the solution to Equation 3.2.

In step 2, Lipsitz et al.'s suggested approach does not attempt to model higher order terms that may appear in $\mathbb{E}[U_i(\mathbf{Y}, \boldsymbol{T}, M; \boldsymbol{\theta})|\mathbf{Y}, \boldsymbol{T}]$; instead, the best estimate of $M|\mathbf{Y}, \boldsymbol{T}$ is used in U_i as if known. For example, for the common estimating function $U = \mathbf{X}^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})$, the true expectation $\mathbb{E}[\mathbf{X}^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})|\mathbf{Y}, \boldsymbol{T}]$ will be estimated by

$$\sum_m P(M = m|\mathbf{Y}, \boldsymbol{T}; \hat{\boldsymbol{\theta}}^{(t)}) \cdot \mathbb{E}(\mathbf{X})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbb{E}(\mathbf{X})\boldsymbol{\theta})$$

While we know these not to be the same by Jensen's inequality (i.e., $\mathbb{E}[\mathbf{X}^T\mathbf{X}] \neq \mathbb{E}[\mathbf{X}]^T\mathbb{E}[\mathbf{X}]$), the estimate used in step 2 is the linear first approximation of the true desired quantity,

which may suffice.

When the missing variable is continuous, a slight modification to the EM algorithm is necessary to avoid intractable integrals. Lipsitz et al. recommended a Monte Carlo strategy which involves, at each iteration, repeatedly sampling the missing variable M from the conditional distribution $[M|\mathbf{Y}, \boldsymbol{T}]$ of the missing covariate given the observed data L times, and then estimating $\mathbb{E}[U|\mathbf{Y}, \boldsymbol{T}]$ by $\frac{1}{L}\sum_{i=1}^{L} U_i(\mathbf{Y}, \boldsymbol{T}, M^{(i)})$. Later work by Chen [5] offered further computational strategies suitable even for continuous missing covariates with arbitrary patterns of missingness. In short, when computationally feasible, the AIPW approach is an attractive option, as it it offers protection against potential model misspecification, utilizes both complete and incomplete data, and may offer efficiency gains over the standard IPW estimator.

3.2.3 Calibration estimators

While the AIPW estimator may offer a balance of efficiency and robustness to misspecification, it can be difficult to calculate in practice, owing to the computational requirements of estimating the second term in Equation 3.2. However, the form of the AIPW estimator bears similarity to a well-known estimator in the survey-sampling literature, the calibration estimator [7] [8]. This connection has been made explicit by Breslow et al. [2] [1] and Lumley et al. [22] in recent years. In short, calibration estimators use auxiliary information from completely observed variables, or even partially unobserved variables whose total is known, to choose weights to be used in place of $1/\pi_i$ when solving estimating equations such as the IPW estimating equation in Equation 3.1. The calibration weights improve estimator precision by choosing weights that align with observed values of auxiliary information in the population, thus utilizing the partially observed information in unsubsampled individuals.

While the idea of calibration estimators has existed in various forms for decades, classic works by Deville et al. [7] [8] are typically cited as providing the formal framework for this

approach. From a population of size n, suppose that we wish to estimate the population mean of Y, T_Y, from the observation of y_i for those in a subsample. Furthermore, suppose that auxiliary information on x_i, known as the calibration variable, is also available for those in the subsample, as well as the population total of X, T_X. (The calibration variable itself in some cases may be observed in the entire population; only the population total T_X is strictly necessary for calibration.) Given the observed relationship in the subsample that

$$\sum_{i=1}^{n} x_i \quad = \quad \sum_{S_i=1} \frac{g_i}{\pi_i} x_i, \tag{3.3}$$

for some weights g_i that will make Equation 3.3 true, the calibration approach seeks to choose the g_i so that the distance between the usual IPW weights $1/\pi_i$ and the calibrated weights g_i/π_i will be minimized, for some distance function d. Lumley et al. [22] noted that using distance function $d(a,b) = a\,(\log a - \log b) + (b - a)$ gives rise to the classical raking adjustment [36], which is related to the deviance function in Poisson regression [11]. One common choice of distance function, the so-called "chi-square" distance metric

$$d\,(a,b) = \frac{(a-b)^2}{b},$$

gives rise to the generalized regression or GREG estimator [36], which can conveniently be written as

$$\hat{T}_{GREG} = \sum_{S_i=1} \frac{g_i}{\pi_i} y_i,$$

where

$$g_i = 1 + (T_X - \hat{T}_X)(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} x_i$$

Here, $\mathbf{W}$ is a diagonal matrix of inverse probabilities $1/\pi_i$, $\hat{T}_X$ is the usual Horvitz-Thompson estimator of T_X, and $\mathbf{X}$ is the design matrix for the observed subsample.

While similarities between calibration and AIPW approaches have previously been noted [31], Lumley et al. [22] in particular showed the asymptotic equivalence of the calibration estimator and the AIPW estimator, summarized below. Using calibration estimators as a convenient proxy for an AIPW estimator has been used to good effect in the case-cohort [1] and survival [11] spheres, but this connection has not been exploited in the longitudinal setting thus far. Evaluating a calibration-based estimator in the context of ODS represents a further translation of this work to provide a robust, reasonably efficient, and computationally simple estimator to an epidemiological audience.

As summarized by Lumley et al., the connection between AIPW and calibration estimators for two-phase sampling design can be seen as follows. As noted above, calibration estimators have the goal of estimating a population total $Y = \sum_{i=1}^{n} Y_i$, when Y_i is partially observed, but an auxiliary variable X_i is completely observed. The GREG estimator, which is equivalent to the calibration estimator under a chi-squared distance function [7], can be written in the form

$$T_{GREG} = \sum_{i=1}^{N} \frac{R_i}{\pi_i} y_i + \left(1 - \frac{R_i}{\pi_i}\right) \boldsymbol{x}_i \hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}}$ is the estimated regression parameter coming from regressing Y on X among those in whom both variables are observed. Replacing y_i with the estimating function U_i and the prediction $\boldsymbol{x}_i\hat{\boldsymbol{\beta}}$ with an arbitrary function $\phi(x_i)$ of the data available for all subjects, we instead obtain the most commonly implemented form of the original Robins et al. [32] AIPW estimator:

$$T = \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_i(\boldsymbol{\theta}) + \left(1 - \frac{R_i}{\pi_i}\right) \phi$$

In particular, among estimators of this form, the optimal choice of ϕ is the conditional expectation of U_i given the complete data available for everyone. However, as Lumley et al. noted, obtaining this expectation would involve iteratively both estimating $\boldsymbol{\theta}$ and constructing new calibration variables based on the estimate of $\boldsymbol{\theta}$, similar to the EM approach

to finding AIPW estimators suggested above. In contrast, a common implementation of the calibration approach is simply to construct calibration variables based on the observed data a single time. While not optimal, a reasonable approximate choice of calibration variable may come close to the true expectation in a way that, while not perfect, is easy to implement and computationally expedient.

Lumley et al. further noted that, while asymptotically equivalent, in practice the AIPW and calibration estimators differ in that the calibration estimator requires the additional estimation of $\hat{\boldsymbol{\beta}}$, while the AIPW estimator does not. However, if $\hat{\boldsymbol{\beta}}$ consistently estimates $\boldsymbol{\beta}$, the regression parameter describing the relationship between the calibration target and the partially observed auxiliary variable, the two approaches will be asymptotically equivalent. In terms of efficiency, in finite samples the calibration estimator (like the AIPW estimator) is not guaranteed to achieve the semiparametric efficiency bound, even for the optimal choice of ϕ; however, Robins et al. [32] do show the estimators to asymptotically attain semiparametric efficiency.

In the ODS setting of interest here, the total T used generically in the calibration literature that we wish to find is the sum of U_i over all cohort individuals, $i = 1, \ldots, N$, which is zero by definition when evaluated at the estimate $\hat{\boldsymbol{\theta}}$ that would be obtained with complete (full cohort) data. Translated into the ODS scenario with which we are concerned in this work, in which the outcome $\mathbf{Y}$ and time vector $\boldsymbol{T}$ which are completely observed, and M is incompletely observed, the calibration estimator with optimal choice of calibration variable can then be seen to be equivalent to solving the AIPW estimating equation

$$0 = \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_i(\boldsymbol{\theta}; M_i, \boldsymbol{T}_i) + \left(1 - \frac{R_i}{\pi_i}\right) \mathbb{E}[U_i(\boldsymbol{\theta}; M_i, \boldsymbol{T}_i) | \mathbf{Y}_i, \boldsymbol{T}_i]$$

3.2.4 Proposed calibration algorithm

The proposed implementation of a calibration estimator is as follows, for data obtained under ODS in the longitudinal setting, in which the target model is the regression parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T})$ from the usual linear mixed model. The calibration approach will improve inference insofar as the calibration variable is correlated with the incompletely observed variable whose total is being estimated. As noted by Lumley et al. [22], the completely observed auxiliary variable itself need not be highly correlated with the incompletely observed variable, and a better choice of calibration variable is the estimated influence functions from the regression of incompletely on completely observed variables (here, the regression of M on $\mathbf{Y}$); we choose this as our calibration variable here. In addition to following the guiding principles described by Lumley et al., these steps draw heavily from those proposed by Breslow et al. [1] and Fong et al. [11] for use in the case-cohort and survival settings, respectively.

1. Using the subsample only, fit marker M as a linear and additive function of the outcome vector $\mathbf{Y} = (y_1, \ldots, y_{n_i})$.

2. Using the estimated coefficients from step 1, predict marker value for all cohort members.

3. Fit the target longitudinal linear mixed model using imputed marker values for all cohort members, together with fully observed times and outcomes.

4. Extract the influence functions $\mathbf{IF_i}(\hat{\boldsymbol{\beta}}) = \mathbf{H}^{-1}(\hat{\boldsymbol{\beta}})U_i(\hat{\boldsymbol{\beta}})$ from step 3 to use as calibration variables.

5. Use the calibration variables from step 4 to obtain calibration weights g_i, where

$$g_i \;=\; 1 + \boldsymbol{\lambda}^T \cdot \mathbf{IF_i}$$

$$\text{and} \quad \boldsymbol{\lambda} \;=\; \left(\sum_{i=1}^{n} \mathbf{IF_i} - \sum_{S_i=1} \frac{\mathbf{IF_i}}{\pi_i}\right)^T \left(\sum_{S_i=1} \frac{\mathbf{IF_i} \cdot \mathbf{IF_i^T}}{\pi_i}\right)^{-1}$$

6. Solve the estimating equations $\sum_{i=1}^{N} \frac{R_i \cdot g_i}{\pi_i} U_i(\boldsymbol{\theta}) = 0$ to obtain the calibration estimator $\hat{\boldsymbol{\theta}}_{CAL}$.

7. Obtain standard error estimates of $\hat{\boldsymbol{\theta}}_{CAL}$ through the sandwich formulation of the covariance matrix [47].

This algorithm will produce a calibration estimator with asymptotic properties similar to those of the AIPW estimator, but it obtains the estimator in a way that promises to avoid the computational complexity of the AIPW estimator. It should also be noted that while we use an imputation model in Step 1 that would be exact under multivariate normality, the validity of the calibration estimator does not depend on its correct specification; however, a poorly characterized functional form, for example, may result in a calibration variable that is not highly correlated with the target population total, in turn reducing the potential efficiency gains.

For the ODS designs described in Chapter 2, we propose to examine the performance of five estimators: the two most attractive likelihood-based estimators ($\hat{\boldsymbol{\beta}}_{SO,NC}$ and $\hat{\boldsymbol{\beta}}_{SU,WC}$) and three robust estimators that attempt to address the bias-variance tradeoff: in order of expected performance, $\hat{\boldsymbol{\beta}}_{IPW}$, $\hat{\boldsymbol{\beta}}_{CAL}$, and $\hat{\boldsymbol{\beta}}_{AIPW}$. We evaluate and compare these estimators in terms of relative bias, variance, and MSE, under correct model specification and misspecification.

3.3 Assessment of operating characteristics

We investigate the operating characteristics of likelihood-based and robust semiparametric ODS estimators via simulation, under correct and incorrect model specification. For

likelihood-based estimators, we consider the estimator SO, NC examined in Chapter 2 that analyzed only subsampled individuals, conditional on covariate information, as well as the estimator SU, WC that incorporated information from both subsampled and unsubsampled individuals' outcomes jointly with covariate values. Among the estimators studied in Chapter 2, these two appeared to be the most efficient among estimators that considered subsampled individuals only and all cohort members, respectively. Among semiparametric ODS estimators, we examine the performance of the classical IPW estimator, the WEE implementation of the AIPW estimator, and the calibration estimator as described in Section 3.2. We compare these to the estimators resulting from analysis of a simple random sample, and from analysis of the entire cohort. In terms of design, we consider choosing the subsample randomly, or through an ODS design based on subject-specific intercept or subject-specific slope, as described previously in Section 2.3.1.

3.3.1 Simulation setup, under correct model specification

For simulations of estimator performance under correct model specification, we use the same framework as for simulations presented in Section 2.4.1. As before, using the true parameter vector

$$\begin{aligned}\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \left(\beta_0, \beta_T, \beta_M, \beta_{M\times T}, \log(\sigma^2_{b_0}), \log\left(\frac{1+\rho}{1-\rho}\right), \log(\sigma^2_{b_1}), \log(\sigma^2_e)\right) \\ &= \left(10, -0.25, -0.75, 0.5, \log(\sigma^2_{b_0}), \log\left(\frac{1+0}{1-0}\right), \log(\sigma^2_{b_1}), \log(\sigma^2_e)\right)\end{aligned}$$

we generated data from the usual linear mixed model for $N = 1000$ subject at $n_i = 6$ equally spaced time points. Again the binary marker M had a population prevalence of 10%, and $N_S = 250$ of the 1000 subjects on average were selected for marker ascertainment using the criteria and designs described in Section 2.3.1. Like the simulations in Chapter 2, we considered two sets of values for the variance components, one with high subject-to-subject

heterogeneity ($\sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4$) and one with low subject-to-subject heterogeneity ($\sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$). For each design, the IPW, AIPW, and calibration estimators described in Section 3.2 are compared with the best-performing likelihood-based estimators and with a random subsample of 250 individuals on average, as well as to the estimator that would have resulted from analysis of the full cohort. Simulations results are based on 2000 attempted replications; we also report the proportion of replications that failed to converge.

3.3.2 Simulation setup, under model misspecification

We consider the performance of these likelihood-based and robust estimators under two types of model misspecification: misspecification of the random effects distribution and misspecification of the error distribution. In all cases, the regression parameter vector $\boldsymbol{\beta}$ was the same as described in Section 3.3.1, and analysis of the usual linear mixed model proceeded assuming the normality of the random effects and of the error distribution. However, in these simulations at least one of these was the result of a different data-generating mechanism.

For cases in which we examined misspecification of the random effects distribution, we generated skewed, heavy-tailed, and heteroskedastic random effect distributions. In each case, we investigated the effects of both moderate and severe misspecification with a gamma distribution ($\Gamma(15, \sqrt{3})$ and $\Gamma(5, \sqrt{3})$, respectively), a t-distribution (t_{15} and t_5), and a mixture of normal distributions where the ratio of variance component standard deviations for $M_i = 1$ to $M_i = 0$ was either 1.5 or 3.0. After generating the random effects distributions, they were centered (if necessary) and scaled in order to obtain the desired variance component standard deviation values σ_{b_0} and σ_{b_1}. For each data-generating mechanism, we considered the impact of both low ($\sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$) and high ($\sigma^2_{b_0} = \sigma^2_{b_1} = \sigma^2_e = 4$) subject-to-subject heterogeneity scenarios. An examination of the impact of misspecified error distribution considered severe misspecification with a gamma, t, and mixture of heteroskedastic normal distributions, using the same severe misspecification distributions mentioned above.

3.3.3 Simulation results

Tables 3.1 and 3.5 present the relative bias and relative efficiency of ODS estimators under correct model specification. As expected, both likelihood-based and robust estimators had little bias ($\leq$ 6%). Likelihood-based ODS estimators had the same patterns of relative efficiency shown in Chapter 2: improved efficiency for targeted regression parameters related to study design, and higher efficiency for β_0 and β_T when covariate information is used in inference.

Among robust estimators, under correct specification the IPW estimator had varying performance when compared with a random sample. For an intercept-based design with low subject-to-subject heterogeneity, parameters related to level (β_0 and β_M) saw improvement with relative efficiencies of 1.48 and 1.11 respectively, while for the sloped-based design, only β_T was better than or comparable in efficiency to a random sample, respectively (1.45). Calibration and AIPW estimators showed patterns similar to the IPW estimator for β_M and $\beta_{M\times T}$, parameters that are not informed by the information relating to population-level time trends which is available for all individuals.

For β_0 and β_T, however, both calibration and AIPW estimators were able to recover a substantial portion of the information from the full cohort by using auxiliary information available for all cohort members. Each of these estimators far exceeded the relative efficiency of the robust IPW estimator for these two parameters. Finally, as expected, the calibration estimator had lower relative efficiency for these parameters than the AIPW estimator, which in turn had lower relative efficiency than the likelihood-based estimator that incorporated information from unsubsampled cohort members. A small (0.4%) percent of simulations failed due to the inability to find $\boldsymbol{\lambda}$ for construction of the calibration estimator. Results were qualitatively similar when subject-to-subject heterogeneity was high.

Under misspecification of the random effects distribution (Tables 3.2 and 3.3 for low heterogeneity, and Tables 3.6 and 3.7 for high heterogeneity), likelihood-based estimators

suffered sometimes extreme bias, especially of β_T and $\beta_{M\times T}$ under a skewed random effects distribution. Under severely heavy-tailed and heteroskedastic random effects, likelihood-based methods had sometimes double-digit percent biases for $\beta_{M\times T}$; however, when random effects were only moderately heavy-tailed or heteroskedastic, the relative bias of likelihood-based estimators was small and not noticably different from more robust methods. Among the two likelihood-based estimators examined here, the estimator that included information from unsubsampled individuals (SU,WC) tended to be substantially more affected by random effects misspecification than the one which analyzed only subsampled cohort members (SO,NC).

We also examined the impact of severe misspecification of the error distribution on these estimators. As shown in Tables 3.4 and 3.8, even likelihood-based estimators had less than 10% relative bias and appeared relatively robust to this type of misspecification. We also found that neither misspecification of random effects nor of the error distribution affected the performance of robust estimators (IPW, AIPW, or calibration), and these estimators showed the same patterns as under correct model specification, described above.

3.3.4 ODS design and analysis with robust estimators

In the simulations examined in Section 3.3.3, the IPW estimator performed less efficiently than a random sample for parameters that were not targeted by the ODS design (i.e., β_T and $\beta_{M\times T}$ for intercept-based designs, β_0 and β_M for slope-based designs), as can be seen in e.g., Table 3.1. The variance in the IPW estimator is determined by the weights used in Equation 3.1, which in turn is affected by the parameters used in the ODS design: namely, the oversampling percentile and the number oversampled in those percentiles (see Section 2.4.3). Thus, for IPW and IPW-derived estimators (such as AIPW and calibration estimators), the estimator performance – and hence, possibly the decision about which ODS estimator is most appropriate for a given situation – is inextricably linked to the design parameters used

Table 3.1: Percent bias and relative efficiency for likelihood-based and robust estimators under correct model specification and low subject-to-subject heterogeneity. Results shown summarize 2000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

	Estimator				
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Intercept	Calibration	0 (2.84)	0 (1.97)	6 (1.10)	-1 (0.53)
	AIPW	0 (3.22)	0 (2.03)	5 (1.15)	1 (0.56)
	IPW	0 (1.48)	0 (0.48)	5 (1.11)	0 (0.52)
	Likelihood, SO,NC	0 (1.73)	0 (0.99)	3 (1.72)	0 (1.04)
	Likelihood, SU,WC	0 (3.57)	-1 (2.88)	3 (1.78)	0 (1.23)
Slope	Calibration	0 (1.96)	-1 (2.81)	1 (0.65)	5 (0.72)
	AIPW	0 (2.46)	0 (3.12)	2 (0.65)	4 (0.73)
	IPW	0 (0.47)	0 (1.45)	3 (0.63)	4 (0.71)
	Likelihood, SO,NC	0 (0.95)	0 (1.75)	1 (1.15)	2 (1.20)
	Likelihood, SU,WC	0 (3.17)	0 (3.38)	1 (1.18)	2 (1.24)
Random sample	Likelihood	0 (1.00)	1 (1.00)	0 (1.00)	0 (1.00)
	Calibration	0 (2.68)	0 (2.71)	0 (0.99)	-1 (1.00)
	AIPW	0 (3.05)	0 (2.76)	1 (1.05)	0 (1.07)
Full cohort	Likelihood	0 (4.02)	0 (4.01)	1 (4.21)	0 (4.33)

Note: 0.4% of simulations failed under correct model specification.

Table 3.2: Percent bias and relative MSE for likelihood-based and robust ODS estimators under model misspecification of random effects distribution and low subject-to-subject heterogeneity. Results shown summarize 2000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative MSE for an estimator is defined as the ratio of MSEs between a random sample of $N_S = 250$ and the estimator.

		Moderate misspecification				Severe misspecification			
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$	β_0	β_T	β_M	$\beta_{M\times T}$
Skewed random effects									
Intercept	Calibration	0 (2.72)	0 (1.93)	5 (0.99)	-1 (0.53)	0 (2.79)	1 (1.73)	3 (1.11)	1 (0.52)
	AIPW	0 (3.08)	0 (1.90)	4 (1.05)	1 (0.50)	0 (3.18)	1 (1.65)	2 (1.18)	4 (0.46)
	IPW	0 (1.47)	0 (0.46)	4 (1.00)	-1 (0.51)	0 (1.46)	0 (0.45)	3 (1.11)	1 (0.50)
	Likelihood, SO,NC	1 (1.37)	-5 (0.88)	2 (1.54)	-1 (1.05)	1 (1.00)	-8 (0.71)	1 (1.73)	-1 (1.02)
	Likelihood, SU,WC	0 (3.37)	7 (1.54)	-6 (1.89)	28 (0.38)	0 (3.48)	13 (0.86)	-15 (2.01)	52 (0.16)
Slope	Calibration	0 (1.98)	0 (2.70)	0 (0.55)	4 (0.70)	0 (2.06)	0 (2.85)	1 (0.57)	5 (0.73)
	AIPW	0 (2.32)	0 (2.89)	0 (0.56)	3 (0.71)	0 (2.33)	0 (2.95)	1 (0.57)	5 (0.73)
	IPW	-0 (0.47)	0 (1.34)	1 (0.53)	3 (0.70)	0 (0.49)	0 (1.49)	2 (0.55)	4 (0.74)
	Likelihood, SO,NC	0 (0.84)	-5 (1.45)	9 (1.06)	6 (1.03)	1 (0.74)	-8 (1.12)	14 (0.99)	10 (0.91)
	Likelihood, SU,WC	0 (3.22)	1 (3.12)	-5 (1.27)	24 (0.43)	0 (3.22)	3 (2.80)	-12 (1.10)	48 (0.18)
Random sample	Likelihood	0 (1.00)	0 (1.00)	2 (1.00)	0 (1.00)	0 (1.00)	0 (1.00)	1 (1.00)	1 (1.00)
	Calibration	0 (2.54)	0 (2.61)	1 (0.99)	0 (1.00)	0 (2.78)	0 (2.47)	0 (0.99)	0 (1.00)
	AIPW	0 (2.99)	0 (2.68)	2 (1.05)	1 (1.01)	0 (3.14)	0 (2.47)	0 (1.03)	2 (0.92)
Full cohort	Likelihood	0 (3.74)	0 (3.96)	1 (3.95)	0 (3.90)	0 (4.01)	0 (3.86)	0 (4.06)	0 (4.02)
Heavy-tailed random effects									
Intercept	Calibration	0 (2.73)	1 (1.90)	4 (1.03)	1 (0.49)	-0 (2.72)	1 (1.72)	4 (1.07)	-1 (0.53)
	AIPW	0 (3.24)	0 (2.00)	4 (1.06)	2 (0.50)	0 (3.32)	0 (1.85)	4 (1.07)	2 (0.46)
	IPW	0 (1.46)	0 (0.46)	4 (1.02)	1 (0.48)	-0 (1.53)	-0 (0.48)	4 (1.08)	-0 (0.49)
	Likelihood, SO,NC	0 (1.63)	0 (0.96)	3 (1.48)	1 (0.90)	0 (1.48)	0 (0.89)	5 (1.30)	1 (0.94)
	Likelihood, SU,WC	0 (3.53)	0 (3.01)	1 (1.61)	3 (0.99)	-0 (3.43)	1 (3.01)	3 (1.30)	9 (0.65)
Slope	Calibration	0 (1.88)	0 (3.02)	-1 (0.57)	4 (0.74)	0 (1.87)	-0 (2.59)	1 (0.59)	4 (0.80)
	AIPW	0 (2.41)	0 (3.28)	-1 (0.57)	4 (0.77)	0 (2.45)	-0 (2.94)	4 (0.57)	3 (0.80)
	IPW	0 (0.48)	0 (1.48)	0 (0.55)	4 (0.74)	-0 (0.47)	-0 (1.41)	2 (0.57)	3 (0.81)
	Likelihood, SO,NC	0 (0.98)	1 (1.69)	2 (1.07)	1 (1.22)	0 (0.92)	1 (1.36)	9 (1.13)	-0 (1.27)
	Likelihood, SU,WC	0 (3.15)	0 (3.50)	1 (1.10)	1 (1.14)	0 (3.06)	-0 (3.24)	11 (0.95)	-1 (0.72)
Random sample	Likelihood	0 (1.00)	0 (1.00)	-1 (1.00)	0 (1.00)	-0 (1.00)	-0 (1.00)	-1 (1.00)	0 (1.00)
	Calibration	0 (2.50)	0 (2.68)	-1 (1.01)	0 (1.00)	0 (2.47)	0 (2.32)	-2 (0.99)	0 (0.99)
	AIPW	0 (2.97)	0 (2.77)	0 (1.03)	1 (1.06)	0 (3.17)	0 (2.52)	0 (1.01)	1 (0.91)
Full cohort	Likelihood	0 (4.02)	0 (4.11)	-1 (4.02)	0 (3.75)	0 (4.07)	-0 (3.86)	1 (4.11)	-0 (4.19)

Note: a small fraction of simulations failed under model misspecification of random effects distribution: 0.3 and 0.2% for moderate and severe skewness, 0.2% for moderately heavy tails, and 0.3% for severely heavy tails.

Table 3.3: Percent bias and relative MSE for likelihood-based and robust ODS estimators under model misspecification of random effects distribution and low subject-to-subject heterogeneity (continued). Results shown summarize 2000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the $100 \times$ the difference between estimator mean and parameter value, divided by parameter value, and relative MSE for an estimator is defined as the ratio of MSEs between a random sample of $N_S = 250$ and the estimator.

		Moderate misspecification				Severe misspecification			
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$	β_0	β_T	β_M	$\beta_{M\times T}$
Heteroskedastic random effects									
Intercept	Calibration	0 (2.64)	0 (2.07)	5 (1.26)	-2 (0.66)	0 (1.95)	0 (1.29)	7 (1.43)	1 (0.76)
	AIPW	0 (2.99)	0 (2.17)	5 (1.40)	0 (0.73)	0 (2.60)	-1 (1.52)	6 (1.97)	1 (0.93)
	IPW	0 (1.43)	0 (0.53)	5 (1.30)	-1 (0.65)	0 (1.09)	0 (0.46)	6 (1.73)	1 (0.76)
	Likelihood, SO,NC	0 (1.71)	0 (1.05)	-8 (1.97)	-2 (1.20)	0 (1.23)	1 (0.83)	-11 (2.59)	-2 (1.32)
	Likelihood, SU,WC	0 (2.99)	5 (2.07)	-4 (1.95)	7 (1.14)	0 (1.98)	7 (1.07)	1 (2.08)	13 (0.91)
Slope	Calibration	0 (1.68)	0 (3.06)	-2 (0.69)	3 (1.03)	0 (1.14)	0 (2.09)	-4 (0.90)	4 (1.22)
	AIPW	-0 (2.11)	0 (3.39)	0 (0.73)	2 (1.15)	0 (1.41)	0 (2.69)	-3 (0.99)	2 (1.54)
	IPW	-0 (0.49)	0 (1.53)	0 (0.67)	2 (1.06)	0 (0.45)	0 (1.01)	-1 (0.81)	2 (1.40)
	Likelihood, SO,NC	0 (0.94)	1 (1.81)	-7 (1.30)	-3 (1.29)	0 (0.75)	3 (1.12)	-10 (1.43)	-9 (1.46)
	Likelihood, SU,WC	0 (2.57)	2 (3.22)	-2 (1.25)	4 (1.18)	0 (1.50)	5 (1.70)	-3 (1.26)	10 (0.79)
Random sample	Likelihood	0 (1.00)	0 (1.00)	-2 (1.00)	-1 (1.00)	0 (1.00)	0 (1.00)	-1 (1.00)	0 (1.00)
	Calibration	0 (2.29)	0 (2.59)	-1 (1.00)	-1 (1.00)	0 (1.45)	0 (1.83)	1 (1.13)	1 (1.07)
	AIPW	0 (2.68)	0 (2.76)	-1 (1.13)	-1 (1.17)	0 (1.68)	-1 (1.92)	0 (1.19)	-1 (1.19)
Full cohort	Likelihood	0 (3.92)	0 (4.51)	1 (3.90)	0 (4.14)	0 (3.97)	0 (3.99)	1 (3.99)	0 (4.00)

Note: a small fraction of simulations failed under model misspecification of random effects distribution: 0.4% for moderate heteroskedasticity, and 0.7% for severe heteroskedasticity.

Table 3.4: Percent bias and relative MSE for likelihood-based and robust ODS estimators under model misspecification of error distribution and low subject-to-subject heterogeneity. Results shown summarize 2000 replications under same parameters as Table 3.1. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative MSE for an estimator is defined as the ratio of MSEs between a random sample of $N_S = 250$ and the estimator.

	Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Skewed errors	Intercept	Calibration	0 (2.99)	0 (1.84)	4 (1.08)	1 (0.53)
		AIPW	0 (3.45)	0 (1.92)	3 (1.15)	2 (0.55)
		IPW	0 (1.62)	0 (0.48)	3 (1.09)	1 (0.52)
		Likelihood, SO,NC	0 (1.89)	1 (0.98)	1 (1.60)	1 (1.06)
		Likelihood, SU,WC	0 (3.74)	0 (2.71)	-1 (1.79)	1 (1.28)
	Slope	Calibration	0 (1.97)	0 (2.56)	-3 (0.63)	4 (0.81)
		AIPW	0 (2.52)	0 (2.80)	-2 (0.64)	3 (0.84)
		IPW	0 (0.53)	0 (1.36)	-2 (0.63)	3 (0.81)
		Likelihood, SO,NC	0 (1.01)	0 (1.68)	0 (1.14)	1 (1.34)
		Likelihood, SU,WC	0 (3.31)	0 (3.10)	-3 (1.20)	1 (1.40)
	Random sample	Likelihood	0 (1.00)	0 (1.00)	-1 (1.00)	0 (1.00)
		Calibration	0 (2.61)	0 (2.52)	-1 (1.01)	0 (1.01)
		AIPW	0 (3.13)	0 (2.62)	-1 (1.04)	1 (1.07)
	Full cohort	Likelihood	0 (4.18)	0 (3.65)	0 (4.21)	0 (4.22)
Heavy-tailed errors	Intercept	Calibration	0 (2.42)	0 (1.73)	4 (1.01)	0 (0.52)
		AIPW	0 (2.91)	0 (1.89)	3 (1.07)	0 (0.55)
		IPW	0 (1.42)	0 (0.47)	3 (1.02)	0 (0.51)
		Likelihood, SO,NC	0 (1.63)	0 (0.93)	2 (1.54)	0 (0.99)
		Likelihood, SU,WC	0 (3.27)	0 (2.72)	2 (1.65)	0 (1.19)
	Slope	Calibration	0 (1.78)	0 (2.43)	0 (0.61)	5 (0.75)
		AIPW	0 (2.31)	0 (2.95)	1 (0.61)	4 (0.77)
		IPW	0 (0.48)	0 (1.30)	1 (0.60)	4 (0.75)
		Likelihood, SO,NC	0 (0.91)	0 (1.69)	-1 (1.16)	2 (1.24)
		Likelihood, SU,WC	0 (3.02)	0 (3.21)	0 (1.21)	1 (1.18)
	Random sample	Likelihood	0 (1.00)	0 (1.00)	1 (1.00)	0 (1.00)
		Calibration	0 (2.36)	0 (2.51)	1 (1.00)	0 (1.00)
		AIPW	0 (2.86)	0 (2.69)	1 (1.05)	0 (1.11)
	Full cohort	Likelihood	0 (3.69)	0 (3.70)	0 (4.09)	0 (3.85)
Heteroskedastic errors	Intercept	Calibration	0 (2.11)	1 (1.63)	4 (1.31)	2 (0.70)
		AIPW	0 (2.99)	1 (1.97)	4 (1.44)	3 (0.74)
		IPW	0 (1.37)	0 (0.49)	4 (1.36)	2 (0.69)
		Likelihood, SO,NC	0 (1.64)	0 (0.98)	-6 (2.07)	6 (1.05)
		Likelihood, SU,WC	0 (3.00)	5 (1.93)	11 (1.68)	13 (0.92)
	Slope	Calibration	0 (1.58)	1 (2.37)	3 (0.72)	4 (0.96)
		AIPW	0 (2.28)	0 (3.24)	3 (0.73)	4 (1.02)
		IPW	0 (0.51)	0 (1.44)	2 (0.70)	3 (0.98)
		Likelihood, SO,NC	0 (0.96)	0 (1.61)	28 (1.04)	-2 (1.20)
		Likelihood, SU,WC	0 (2.37)	2 (3.15)	49 (0.80)	7 (0.99)
	Random sample	Likelihood	0 (1.00)	0 (1.00)	-2 (1.00)	1 (1.00)
		Calibration	0 (1.90)	1 (2.26)	-3 (0.98)	0 (0.99)
		AIPW	0 (2.82)	0 (2.71)	-3 (1.13)	0 (1.21)
	Full cohort	Likelihood	0 (3.88)	0 (4.06)	-1 (3.99)	0 (3.91)

Table 3.5: Percent bias and relative efficiency for likelihood-based and robust ODS estimators under correct model specification and high subject-to-subject heterogeneity. Results shown summarize 2000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4\sigma^2_{b_1} = 4, \sigma^2_e = 4$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Intercept	Calibration	0 (2.46)	0 (2.06)	6 (1.04)	2 (0.51)
	AIPW	0 (3.24)	-1 (2.27)	6 (1.06)	3 (0.49)
	IPW	0 (1.44)	0 (0.51)	5 (1.05)	2 (0.49)
	Likelihood, SO,NC	0 (1.84)	-1 (0.94)	2 (1.56)	2 (0.98)
	Likelihood, SU,WC	0 (3.53)	-1 (2.96)	2 (1.54)	3 (0.90)
Slope	Calibration	0 (1.63)	-1 (2.80)	-2 (0.56)	4 (0.95)
	AIPW	0 (2.33)	-1 (3.04)	0 (0.53)	3 (0.99)
	IPW	0 (0.49)	0 (1.46)	-1 (0.55)	3 (0.98)
	Likelihood, SO,NC	0 (0.99)	0 (1.64)	-3 (1.11)	2 (1.52)
	Likelihood, SU,WC	0 (3.02)	-1 (3.36)	-2 (1.06)	2 (1.51)
Random sample	Likelihood	0 (1.00)	-1 (1.00)	-2 (1.00)	3 (1.00)
	Calibration	0 (2.31)	0 (2.73)	-2 (0.99)	3 (1.00)
	AIPW	0 (2.90)	0 (3.02)	-2 (1.01)	3 (1.00)
Full cohort	Likelihood	0 (4.01)	-1 (3.91)	-1 (4.22)	1 (4.03)

Note: 3.5% of simulations failed under correct model specification.

Table 3.6: Percent bias and relative MSE for likelihood-based and robust ODS estimators under model misspecification of random effects distribution and high subject-to-subject heterogeneity. Results shown summarize 2000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative MSE for an estimator is defined as the ratio of MSEs between a random sample of $N_S = 250$ and the estimator.

		Moderate misspecification				Severe misspecification			
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$	β_0	β_T	β_M	$\beta_{M\times T}$
Skewed random effects									
Intercept	Calibration	0 (2.41)	1 (2.14)	3 (0.97)	1 (0.52)	-0 (2.45)	0 (2.04)	2 (1.06)	3 (0.54)
	AIPW	0 (2.96)	1 (2.33)	2 (1.00)	3 (0.49)	-0 (3.22)	0 (2.23)	1 (1.10)	7 (0.50)
	IPW	0 (1.43)	3 (0.50)	3 (0.98)	1 (0.51)	-0 (1.56)	-2 (0.51)	1 (1.10)	4 (0.53)
	Likelihood, SO,NC	0 (1.55)	-7 (0.99)	1 (1.45)	0 (1.01)	0 (1.55)	-13 (0.95)	-0 (1.64)	2 (1.12)
	Likelihood, SU,WC	-0 (3.28)	9 (1.98)	-3 (1.61)	41 (0.38)	-0 (3.58)	28 (0.84)	-10 (1.85)	134 (0.12)
Slope	Calibration	0 (1.56)	0 (2.86)	1 (0.48)	9 (0.97)	-0 (1.65)	-1 (2.73)	-7 (0.48)	10 (0.96)
	AIPW	0 (2.13)	1 (3.14)	2 (0.47)	8 (0.99)	-0 (2.24)	-0 (3.08)	-7 (0.47)	11 (0.94)
	IPW	0 (0.48)	-1 (1.48)	2 (0.48)	8 (0.99)	-0 (0.50)	-1 (1.54)	-7 (0.47)	9 (0.97)
	Likelihood, SO,NC	1 (0.83)	-21 (1.35)	1 (0.90)	8 (1.46)	1 (0.72)	-37 (0.93)	-3 (0.89)	16 (1.36)
	Likelihood, SU,WC	-0 (2.98)	2 (3.29)	-7 (0.98)	19 (0.98)	-0 (3.11)	5 (2.69)	-20 (0.89)	55 (0.39)
Random sample	Likelihood	-0 (1.00)	2 (1.00)	1 (1.00)	3 (1.00)	-0 (1.00)	0 (1.00)	0 (1.00)	3 (1.00)
	Calibration	0 (2.37)	1 (2.73)	1 (0.99)	2 (0.98)	0 (2.56)	0 (2.69)	-0 (1.00)	3 (1.00)
	AIPW	0 (2.97)	1 (2.97)	1 (1.01)	3 (0.97)	0 (3.08)	0 (2.97)	-0 (1.01)	5 (0.96)
Full cohort	Likelihood	0 (3.70)	0 (3.96)	1 (3.78)	1 (3.94)	0 (4.07)	-0 (3.90)	-0 (4.03)	1 (4.09)
Heavy-tailed random effects									
Intercept	Calibration	-0 (2.49)	1 (2.07)	5 (0.97)	-2 (0.53)	0 (2.44)	2 (1.95)	3 (0.99)	-1 (0.52)
	AIPW	-0 (3.32)	0 (2.29)	4 (1.01)	0 (0.49)	0 (3.35)	1 (2.24)	4 (1.01)	5 (0.45)
	IPW	-0 (1.61)	-2 (0.46)	4 (1.00)	-1 (0.52)	0 (1.50)	-1 (0.51)	3 (1.02)	-0 (0.50)
	Likelihood, SO,NC	-0 (1.81)	-1 (0.99)	2 (1.52)	2 (0.98)	0 (1.68)	-0 (1.00)	3 (1.39)	3 (0.95)
	Likelihood, SU,WC	-0 (3.67)	1 (3.03)	2 (1.50)	4 (0.86)	0 (3.59)	3 (2.80)	1 (1.35)	19 (0.53)
Slope	Calibration	0 (1.64)	-0 (2.91)	-0 (0.49)	7 (1.02)	0 (1.49)	-1 (2.73)	2 (0.46)	5 (0.98)
	AIPW	-0 (2.28)	1 (3.17)	2 (0.47)	6 (1.03)	0 (2.18)	-0 (3.08)	6 (0.43)	6 (0.97)
	IPW	-0 (0.49)	1 (1.55)	0 (0.48)	6 (1.03)	-0 (0.48)	-1 (1.47)	3 (0.46)	5 (1.01)
	Likelihood, SO,NC	-0 (0.99)	2 (1.66)	1 (0.91)	6 (1.44)	0 (0.87)	1 (1.23)	8 (0.84)	9 (1.19)
	Likelihood, SU,WC	-0 (3.13)	0 (3.47)	1 (0.88)	4 (1.48)	0 (2.83)	-0 (3.17)	14 (0.64)	4 (1.12)
Random sample	Likelihood	-0 (1.00)	-1 (1.00)	-2 (1.00)	2 (1.00)	-0 (1.00)	0 (1.00)	-1 (1.00)	3 (1.00)
	Calibration	-0 (2.46)	0 (2.74)	-2 (1.00)	2 (0.99)	0 (2.43)	1 (2.57)	-2 (0.99)	3 (0.99)
	AIPW	-0 (3.09)	1 (3.08)	-2 (0.98)	2 (1.00)	0 (3.19)	1 (2.98)	-0 (0.97)	6 (0.94)
Full cohort	Likelihood	-0 (4.19)	1 (4.10)	-1 (3.90)	2 (3.76)	0 (4.14)	-1 (3.91)	1 (4.11)	-1 (4.06)

Note: a small fraction of simulations failed under model misspecification of random effects distribution: 3.4% for moderate and severe skewness, 3.8% for moderately heavy tails, and 4.0% for severely heavy tails.

Table 3.7: Percent bias and relative MSE for likelihood-based and robust ODS estimators under model misspecification of random effects distribution and high subject-to-subject heterogeneity (continued). Results shown summarize 2000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative MSE for an estimator is defined as the ratio of MSEs between a random sample of $N_S = 250$ and the estimator.

		Moderate misspecification				Severe misspecification			
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$	β_0	β_T	β_M	$\beta_{M\times T}$
Heteroskedastic random effects									
Intercept	Calibration	-0 (2.39)	-0 (1.75)	4 (1.10)	6 (0.59)	-0 (1.84)	1 (0.93)	1 (1.39)	8 (0.80)
	AIPW	-0 (3.37)	-1 (2.01)	2 (1.20)	6 (0.60)	-0 (2.67)	-1 (1.11)	-3 (1.76)	2 (0.86)
	IPW	-0 (1.44)	-3 (0.50)	2 (1.19)	8 (0.55)	-0 (1.18)	0 (0.48)	-1 (1.61)	6 (0.69)
	Likelihood, SO,NC	0 (1.78)	-1 (1.04)	-7 (1.84)	2 (1.09)	0 (1.42)	1 (0.85)	-15 (2.45)	2 (1.30)
	Likelihood, SU,WC	0 (3.39)	3 (2.44)	-6 (1.80)	7 (1.01)	0 (2.31)	6 (1.17)	-13 (2.35)	6 (1.23)
Slope	Calibration	-0 (1.54)	0 (2.74)	-5 (0.61)	13 (1.28)	0 (1.05)	1 (1.74)	-2 (0.80)	10 (1.47)
	AIPW	-0 (2.19)	-0 (3.05)	-3 (0.63)	10 (1.40)	-0 (1.74)	0 (2.51)	-4 (0.97)	8 (1.83)
	IPW	-0 (0.49)	0 (1.34)	-2 (0.60)	11 (1.36)	0 (0.44)	0 (0.97)	-3 (0.81)	10 (1.70)
	Likelihood, SO,NC	-0 (0.97)	1 (1.70)	-6 (1.14)	-3 (2.17)	0 (0.77)	3 (1.21)	-8 (1.49)	-11 (2.70)
	Likelihood, SU,WC	0 (2.55)	2 (2.97)	-5 (1.14)	-2 (2.11)	0 (1.50)	6 (1.75)	-14 (1.71)	-13 (2.71)
Random sample	Likelihood	0 (1.00)	-0 (1.00)	1 (1.00)	2 (1.00)	-0 (1.00)	1 (1.00)	-4 (1.00)	5 (1.00)
	Calibration	0 (2.25)	-0 (2.37)	1 (0.99)	2 (1.01)	-0 (1.67)	2 (1.45)	-3 (1.01)	7 (1.22)
	AIPW	-0 (2.80)	-1 (2.63)	0 (1.03)	1 (1.05)	-0 (2.06)	-1 (1.51)	-7 (1.13)	2 (1.17)
Full cohort	Likelihood	-0 (4.09)	-0 (4.01)	0 (3.76)	5 (3.98)	-0 (3.81)	0 (4.09)	-2 (4.08)	4 (3.85)

Note: a small fraction of simulations failed under model misspecification of random effects distribution: 6.4% for moderate heteroskedasticity, and 8.2% for severe heteroskedasticity.

Table 3.8: Percent bias and relative MSE for likelihood-based and robust ODS estimators under model misspecification of error distribution and high subject-to-subject heterogeneity. Results shown summarize 2000 replications under same parameters as Table 3.5. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative MSE for an estimator is defined as the ratio of MSEs between a random sample of $N_S = 250$ and the estimator.

	Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Skewed errors	Intercept	Calibration	0 (2.45)	2 (2.07)	4 (1.07)	4 (0.54)
		AIPW	0 (3.16)	1 (2.31)	4 (1.09)	6 (0.51)
		IPW	0 (1.53)	3 (0.51)	4 (1.09)	4 (0.52)
		Likelihood, SO,NC	0 (1.74)	6 (1.07)	1 (1.73)	3 (1.05)
		Likelihood, SU,WC	0 (3.50)	0 (3.04)	-1 (1.78)	2 (1.02)
	Slope	Calibration	0 (1.61)	-0 (3.11)	-2 (0.53)	4 (1.11)
		AIPW	0 (2.26)	-0 (3.34)	-1 (0.51)	4 (1.15)
		IPW	0 (0.50)	-1 (1.61)	-1 (0.52)	4 (1.13)
		Likelihood, SO,NC	0 (0.98)	-0 (1.88)	0 (1.09)	3 (1.69)
		Likelihood, SU,WC	-0 (3.19)	-0 (3.63)	-5 (1.11)	3 (1.69)
	Random sample	Likelihood	0 (1.00)	2 (1.00)	-3 (1.00)	1 (1.00)
		Calibration	0 (2.47)	0 (2.83)	-3 (0.99)	-0 (1.00)
		AIPW	-0 (2.98)	-0 (3.05)	-3 (1.02)	0 (1.00)
	Full cohort	Likelihood	0 (3.99)	-0 (4.09)	-0 (4.17)	0 (4.27)
Heavy-tailed errors	Intercept	Calibration	0 (2.25)	1 (1.98)	3 (1.04)	2 (0.52)
		AIPW	-0 (3.10)	-0 (2.24)	1 (1.06)	2 (0.51)
		IPW	0 (1.47)	1 (0.46)	1 (1.06)	2 (0.51)
		Likelihood, SO,NC	0 (1.66)	1 (0.90)	1 (1.58)	3 (0.98)
		Likelihood, SU,WC	-0 (3.38)	0 (2.82)	1 (1.51)	3 (0.91)
	Slope	Calibration	0 (1.44)	0 (2.81)	-4 (0.56)	9 (1.01)
		AIPW	-0 (2.33)	0 (3.10)	-3 (0.54)	8 (1.04)
		IPW	-0 (0.50)	2 (1.45)	-3 (0.55)	8 (1.03)
		Likelihood, SO,NC	-0 (0.97)	-0 (1.75)	-4 (1.00)	5 (1.60)
		Likelihood, SU,WC	-0 (3.01)	0 (3.42)	-2 (0.93)	6 (1.58)
	Random sample	Likelihood	-0 (1.00)	0 (1.00)	-2 (1.00)	3 (1.00)
		Calibration	0 (2.11)	1 (2.56)	-2 (0.99)	3 (0.99)
		AIPW	-0 (2.94)	1 (2.88)	-2 (0.99)	4 (0.99)
	Full cohort	Likelihood	-0 (4.00)	-0 (3.79)	-1 (4.09)	1 (4.00)
Heteroskedastic errors	Intercept	Calibration	0 (1.92)	1 (1.37)	6 (1.03)	-3 (0.54)
		AIPW	-0 (2.89)	-0 (2.25)	5 (1.12)	-2 (0.53)
		IPW	0 (1.02)	2 (0.42)	5 (1.10)	-3 (0.54)
		Likelihood, SO,NC	-0 (1.53)	1 (0.78)	-6 (1.92)	6 (1.01)
		Likelihood, SU,WC	0 (2.95)	7 (2.46)	-5 (1.82)	7 (1.00)
	Slope	Calibration	0 (1.28)	-1 (2.60)	2 (0.52)	2 (1.06)
		AIPW	-0 (1.86)	-0 (3.26)	-0 (0.55)	0 (1.12)
		IPW	0 (0.50)	0 (1.50)	1 (0.51)	0 (1.11)
		Likelihood, SO,NC	0 (1.04)	-0 (1.79)	10 (0.97)	0 (1.70)
		Likelihood, SU,WC	0 (2.55)	1 (3.49)	16 (0.87)	2 (1.67)
	Random sample	Likelihood	-0 (1.00)	1 (1.00)	-3 (1.00)	-1 (1.00)
		Calibration	0 (1.91)	0 (2.05)	-3 (0.99)	-2 (0.98)
		AIPW	-0 (2.67)	-0 (3.03)	-4 (1.03)	-1 (1.02)
	Full cohort	Likelihood	-0 (3.98)	0 (3.90)	-1 (4.12)	-0 (4.07)

at the subsampling stage.

To investigate the impact of ODS design parameters on the performance of the IPW estimator relative to a random sample, we varied the oversampling percentile from the 10%ile to the 40%ile and the number oversampled in each extreme region from 50 to 125 (out of 250 subsampled individuals). Resulting contours of relative efficiency for each longitudinal parameter are shown for both intercept-based (Figure 3.1) and slope-based (Figure 3.2) designs. Each relative efficiency depicted in these plots was the result of 1000 simulations using the same simulation parameters in Section 3.3.1, varying only the design parameters as described above.

As seen in Figures 3.1 and 3.2, the relative efficiency seen in both targeted and non-targeted parameters can vary widely by the combination of ODS design parameters used. The design parameters used in the simulations outlined in Section 3.3.1, which oversampled 100 individuals from each of the most extreme 20%iles, were not the most efficient if an IPW estimator (or similar) was to be chosen for analysis. While a thorough treatment of the optimal design parameters for use with various analysis strategies is beyond the scope of this work, these figures hint at the importance of considering both design and analysis in conjunction for maximal ODS efficiency gains.

3.4 Application to Cystic Fibrosis Foundation Registry data

We again use the Cystic Fibrosis Foundation Patient Registry data to illustrate the possible performance gains and tradeoffs of robust ODS methods compared with likelihood-based estimators. Using the same cohort of 3,141 CF patients described in Section 2.5, we present parameter estimates and empirical standard errors for the usual longitudinal model, over 1000 resamplings of a subset of 600 CF patients on average (Table 3.9).

Broadly, the efficiency patterns seen in simulation were illustrated in the CF registry data. Both likelihood-based and robust estimators produced similar effect estimates. Within each

Figure 3.1: Relative efficiencies of IPW estimator relative to a random sample of the same size, varying both oversampling percentile and the number oversampled from that region, for intercept-based designs. Black dots indicate the combination of design parameters used in simulations in Section 3.3.1.

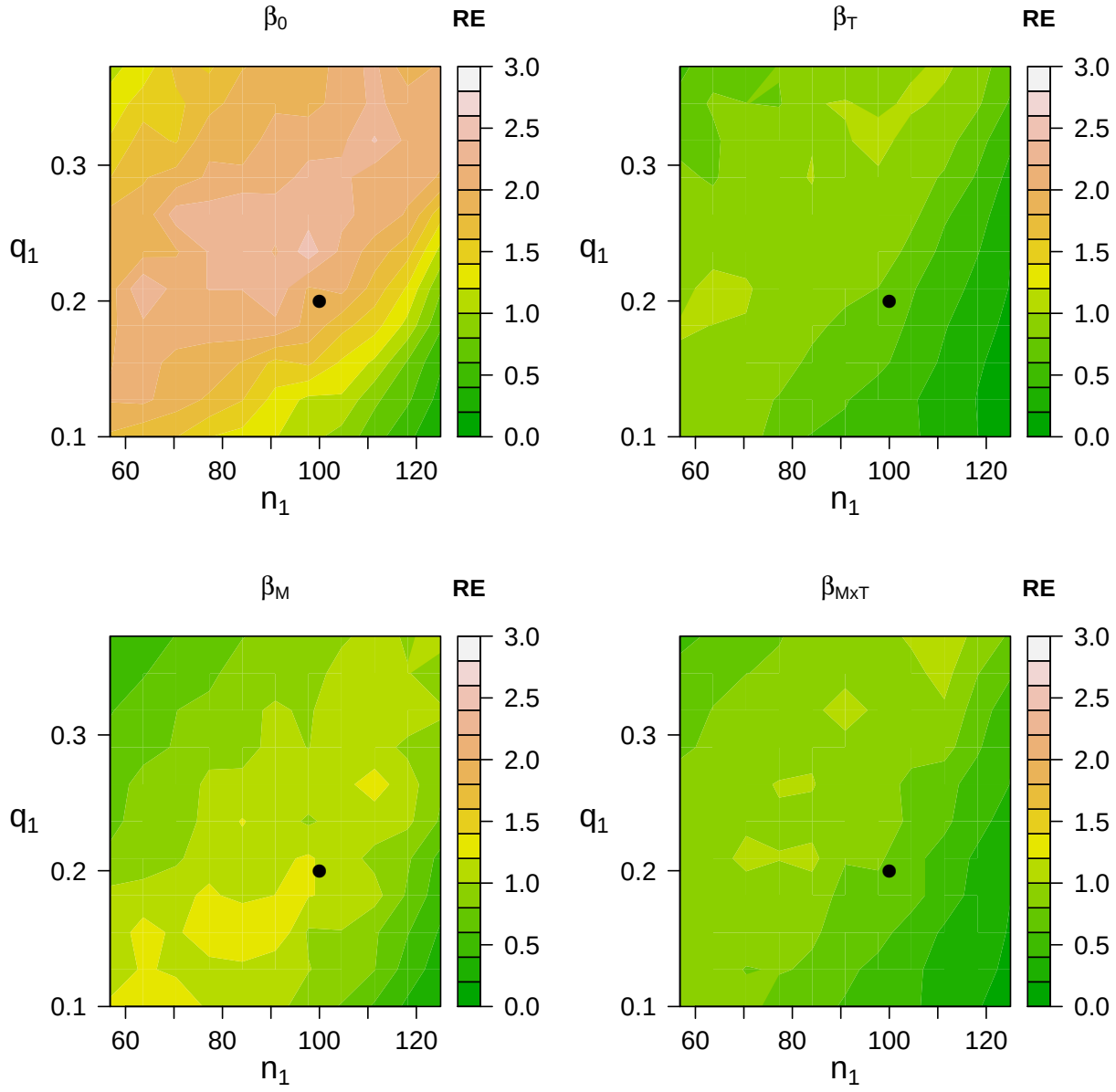
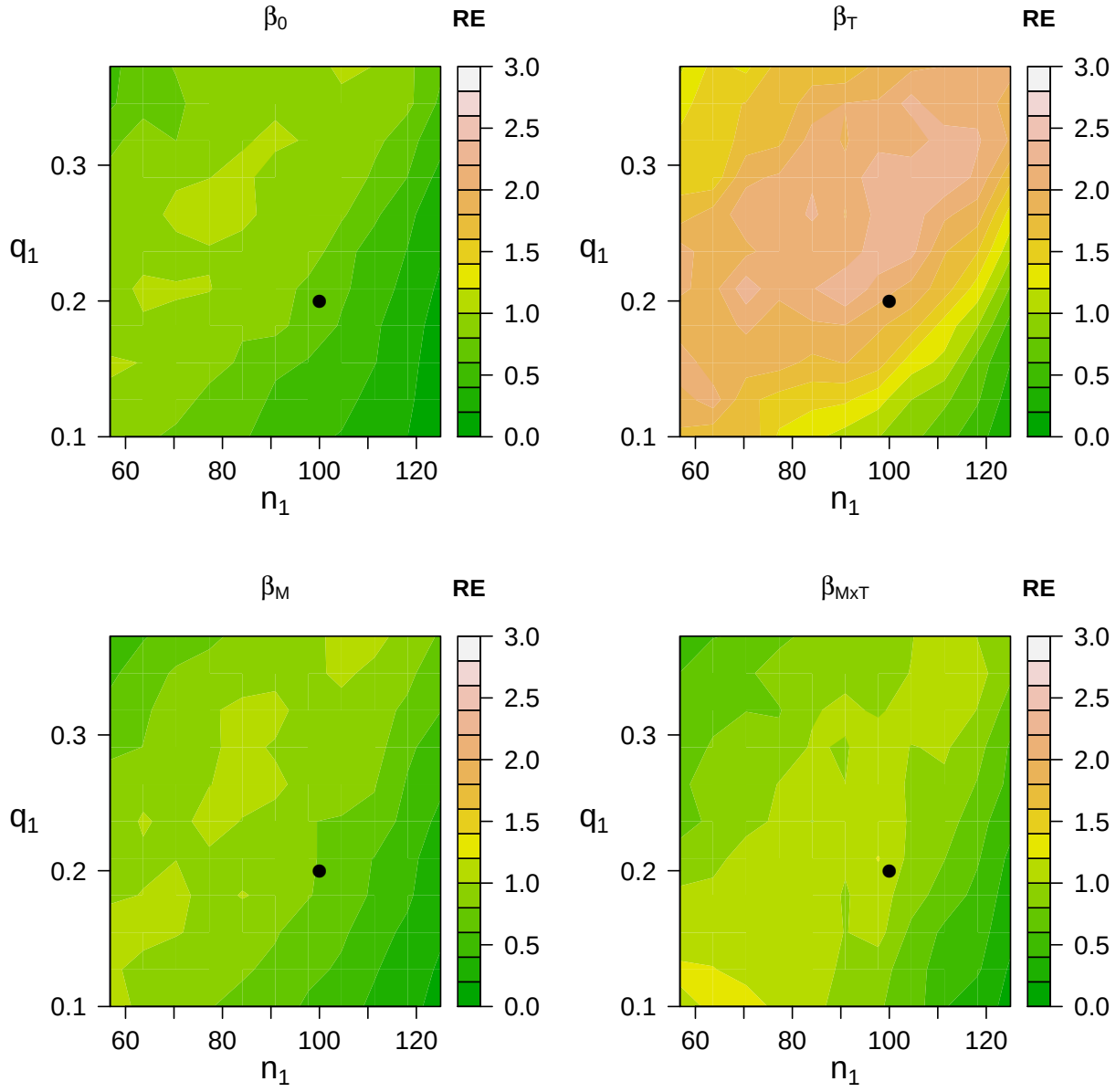


Figure 3.2: Relative efficiencies of IPW estimator relative to a random sample of the same size, varying both oversampling percentile and the number oversampled from that region, for slope-based designs. Black dots indicate the combination of design parameters used in simulations in Section 3.3.1.



design, robust methods generally had somewhat larger standard errors than the corresponding likelihood-based estimators. Among the robust estimators investigated here, the AIPW had the smallest standard errors, followed by calibration and IPW estimators, respectively (e.g., 0.0071 vs 0.0075 vs 0.0085 for intercept-based SEs of $\hat{\beta}_T$). As before, the only estimators that would have detected the small but significant marker by time interaction present in the full cohort were likelihood-based methods that targeted this parameter. As we have suggested elsewhere, a careful consideration of primary inferential targets and preferred analysis strategies would have been key to obtaining a robust yet precise answer to the main scientific question of the effect of *Staphylococcus aureus* on lung function over time.

3.5 Discussion

While likelihood-based ODS methods have the potential to realize substantial efficiency gains, researchers may be apprehensive about their use under model misspecification, or prefer to use methods that are model agnostic. In this section we have evaluated the performance of two likelihood-based ODS estimators under correct and incorrect model specification, and have suggested several alternatives that are robust to misspecification for use with ODS designs.

Likelihood-based methods were unbiased and had relative efficiency higher than that of robust estimators under correct model specification. Even when the error distribution was severely misspecified, the likelihood-based methods showed little bias, and consequently the likelihood-based estimator that incorporated information from unsubsampled cohort members had a higher relative MSE than robust methods, indicating a superior performance.

The impact of misspecification of the random effects distribution on likelihood-based ODS estimators varied by type and severity. For moderately heavy-tailed or heteroskedastic random effects distributions, likelihood-based estimators showed small amounts of bias and tended to outperform robust methods as under correct specification. For severe departures, or

when the random effects distribution was skewed, likelihood-based estimators were sometimes highly biased and did not compare favorably with robust methods. The estimator (SU,WC) that used the partial information from the full cohort seemed especially sensitive to a skewed random effects distribution, especially with respect to $\beta_{M\times T}$.

The robust estimators showed consistent performance regardless of the correct or incorrect specification of the random effects or error distributions. Interestingly, we found the classical IPW estimator to have greater efficiency than a random sample only for targeted parameters; we found this to be true for a number of simulation scenarios (not shown) which varied the amount of variability present in the data. In contrast, Zhou et al. [48] found the IPW estimator to be more efficient than a random sample but less efficient than a semiparametric ODS estimator. However, a univariate ODS design necessarily targets a single type of regression parameter (related to level), while in the longitudinal model ODS designs must balance targeting parameters related to level and to change in level. Furthermore, an exploration of the effect of design parameters in Section 3.3.4 suggests that the performance of the IPW estimator relative to an estimator derived from a random sample depends substantially on the design parameters, which may need to be taken into consideration in conjunction with the ODS analysis plan.

Both AIPW and calibration estimators utilized partially observed information on unsubsampled cohort members to improve inference. Under correct model specification, neither approach had a performance equal to that of the analagous likelihood-based estimator (SU,WC), but both represented substantial improvements over the classical IPW estimator. Of these two, the AIPW was more efficient, although it may be computationally difficult to implement for more complex patterns of missingness. The calibration estimator recovered a substantial portion of the information in the AIPW estimator, yet was far faster and easier to obtain, since the algorithm involved no EM-style iterative processes. Both calibration and AIPW inherited the low efficiency of the IPW estimator for the non-targeted parameter

related to marker value ($\beta_{M\times T}$ for intercept-targeted designs, β_M for slope-targeted designs); an ODS design that simultaneously targets both parametes may have more acceptable performance, as may a design more optimally suited to this class of estimators. Finally, we note that neither the calibration nor the AIPW estimator rely on model correctness for consistency; in both cases, choosing a poor calibration variable (in the case of the calibration estimator) or suboptimal function of the complete data (for the AIPW estimator) merely leads to a less efficient but still consistent estimator of the regression parameter $\boldsymbol{\theta}$.

Table 3.9: Parameter estimates and empirical standard errors for the Cystic Fibrosis Foundation Registry dataset (N = 3,141). For "full" estimator, standard errors are derived from analysis of full Cystic Fibrosis Foundation Registry cohort. All other estimators are based on an average (over 1000 resamplings) subsample of 600 patients. Empirical standard errors are the estimator's standard deviation over all resamplings.

Estimator		β_0		β_T		$\beta_{S.aureus}$		$\beta_{S.aureus \times T}$	
Design	Analysis method	Est (SE)	p	Est. (SE)	p	Est. (SE)	p	Est. (SE)	p
Intercept	Calibration	1.201 (0.018)	< 0.001	0.185 (0.0075)	< 0.001	-0.013 (0.024)	0.59	-0.012 (0.0100)	0.25
	AIPW	1.201 (0.017)	< 0.001	0.185 (0.0071)	< 0.001	-0.013 (0.023)	0.58	-0.012 (0.0100)	0.25
	IPW	1.200 (0.018)	< 0.001	0.184 (0.0085)	< 0.001	-0.013 (0.023)	0.58	-0.012 (0.0100)	0.25
	Likelihood, SO + NC	1.214 (0.019)	< 0.001	0.181 (0.0058)	< 0.001	-0.015 (0.022)	0.51	-0.013 (0.0069)	0.06
	Likelihood, SO + WC	1.201 (0.013)	< 0.001	0.184 (0.0043)	< 0.001	-0.013 (0.019)	0.50	-0.012 (0.0063)	0.06
Slope	Calibration	1.199 (0.030)	< 0.001	0.184 (0.0046)	< 0.001	-0.010 (0.040)	0.80	-0.011 (0.0059)	0.054
	AIPW	1.199 (0.028)	< 0.001	0.184 (0.0041)	< 0.001	-0.011 (0.040)	0.78	-0.011 (0.0058)	0.053
	IPW	1.200 (0.035)	< 0.001	0.184 (0.0046)	< 0.001	-0.011 (0.040)	0.78	-0.011 (0.0059)	0.053
	Likelihood, SO + NC	1.191 (0.024)	< 0.001	0.185 (0.0041)	< 0.001	0.013 (0.029)	0.66	-0.013 (0.0048)	0.006
	Likelihood, SO + WC	1.186 (0.016)	< 0.001	0.185 (0.0032)	< 0.001	0.009 (0.024)	0.71	-0.012 (0.0046)	0.007
Random sample	Likelihood	1.200 (0.025)	< 0.001	0.184 (0.0061)	< 0.001	-0.012 (0.030)	0.69	-0.011 (0.0071)	0.12
	Calibration	1.200 (0.023)	< 0.001	0.184 (0.0054)	< 0.001	-0.012 (0.031)	0.70	-0.011 (0.0071)	0.12
	AIPW	1.200 (0.021)	< 0.001	0.184 (0.0050)	< 0.001	-0.012 (0.030)	0.69	-0.011 (0.0070)	0.12
Full cohort	Likelihood	1.200 (0.012)	< 0.001	0.184 (0.0029)	< 0.001	-0.013 (0.015)	0.79	-0.012 (0.0035)	0.03

Chapter 4

OUTCOME DEPENDENT SAMPLING UNDER UNANTICIPATED MISSINGNESS: DESIGN AND ANALYSIS CONSIDERATIONS

4.1 Introduction

This research program has thus far focused exclusively on outcome dependent sampling designs in cohort studies with complete and balanced longitudinal outcomes. The family of ODS designs has been shown to offer the potential of valid inference and increased power relative to random sampling designs when data is complete. However, even in high-quality studies subjects may drop out of a study before its completion because of study burden, logistics, or reasons related to treatment effects. Moreover, longitudinal outcomes and time to dropout may be linked, reflecting a common underlying disease process. For example, a kidney disease patient whose kidney function (measured by eGFR) declines rapidly may feel more fatigued and be less likely to continue in a study than a patient with more modest decline.

To date longitudinal ODS research has implicitly presumed that the only missing data comes from that deliberately induced by the substudy design. When the longitudinal cohort from which we selectively subsample is subject to missingness beyond the researcher's control, the typical ODS analysis may be adversely affected and potentially result in biased estimation. In this chapter we explore two possible approaches for dealing with a commonly encountered dropout structure found in longitudinal data. Specifically, we assume that the longitudinal outcome vector exists for each cohort member (and thus the regression parameter describing $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$ continues to be meaningful as a target of inference), but may be

partially unobserved and thus not wholly available for analysis for some cohort members [14]. We propose both likelihood-based and regression standardization approaches, both practical options for handling a simple but common MAR dropout structure within the context of ODS design and analysis. Furthermore, we address accompanying design implications and issues of practical implementation for both approaches.

4.2 Design under dropout

Outcome dependent sampling designs under dropout deserve special consideration. Suppose we observe a longitudinal outcome $\mathbf{Y}$ and a dropout time D after which the longitudinal outcome is no longer observed by researchers. For this scenario we focus on a simple MAR dropout structure which excludes death; we assume that the longitudinal outcome exists, and that the usual time trends are meaningful but unobserved in some cases [14]. As in ODS situations in which the data is complete, attention must be paid both to which subjects will be subsampled and how the resulting obtained data will be analyzed.

In reference to the issue of which individuals to subsample for further covariate ascertainment, a naïve approach might be to subsample from among only those participants who completed the study (the "complete case" analysis). While simple, such an approach would be valid only under a "missing completely at random" (MCAR) data generating mechanism [35], which is unlikely to be true in most circumstances. Even when valid the complete case analysis strategy discards potentially useful though incomplete data, and its use is generally discouraged. When dropout time is related to observed outcomes or baseline patient characteristics, which we presume to be the case in many common settings, ignoring the missing data and analyzing only complete cases can lead to biased estimation. Thus, in most realistic situations, if dropout-related missing data is present, it will be important to account for it in design and analysis in order to yield correct inference.

In ODS settings, dealing with dropout will affect substudy design as well as analysis

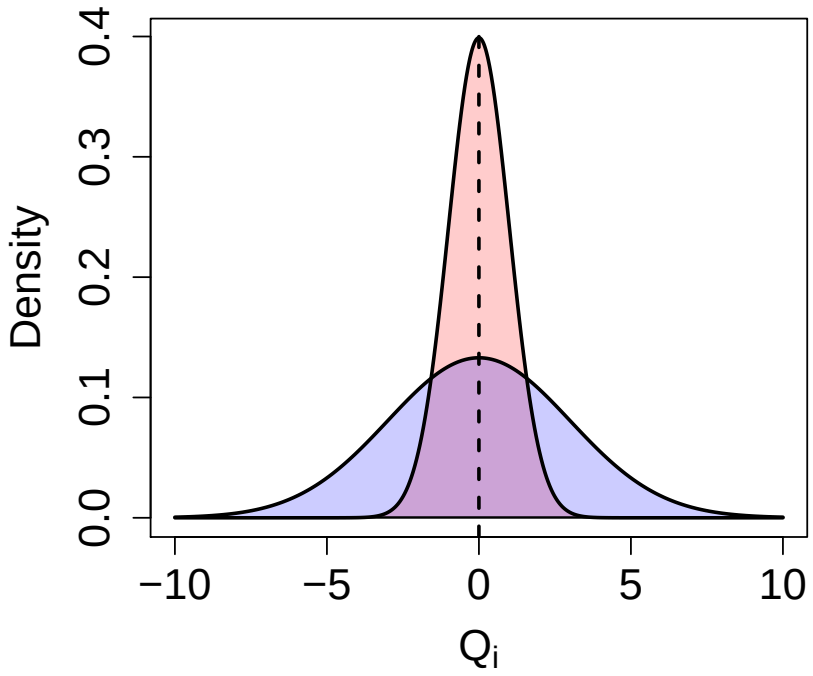
choice. The subsampling strategy used in Chapters 2 and 3 oversampled individuals with extreme values of the subsampling variable Q, previously taken to be the subject-specific intercept and/or slope. Under the usual linear mixed model with dropout, the expected value of Q will be identical whether or not a subject's complete outcome is observed. However, under dropout, subjects with fewer observations will tend to have more variable estimates of Q (Figure 4.1). Thus, when dropout is present, the strategy of subsampling based on observed marginal percentiles of Q across all dropout times will tend to oversample subjects with extreme values of Q. In turn, subjects with fewer observations will tend to be overrepresented in the subsample.

One possible subsampling strategy that accounts for dropout time is to subsample based on the value of $Q|D$, or on the percentiles of Q stratified by dropout time. Under certain assumptions such as those in the linear mixed model, Q is normally distributed conditional on dropout time, and the quantile-based choice of oversampling region seen previously can be used. The resulting subsample across all dropout strata will have q_i's from a mixture of normal distributions whose variance depends on D. Where previously the marginal distribution of Q was used in likelihood maximization, here the source of selection choice and hence selection bias depends on the distribution of Q conditional on dropout time; as such it will need to be factored into the analysis.

4.3 Likelihood-based methods

One approach to handling missing data is to characterize and model its occurrence, and where possible incorporate it into a likelihood that can then be used for valid inference. In a classic framework proposed by Rubin [35], missing data can be classified as one of three types: missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR). Suppose we have a longitudinal outcome Y_{ij} that may be observed (O) or missing (M) for person i at time j, and where the indicator R_{ij} denotes that Y_{ij} was observed,

Figure 4.1: Subjects with completely observed outcomes (red shaded region) will tend to have less variable values of Q, while those who drop out prematurely (purple shaded region) will have more variable Q's. A sampling strategy that preferentially subsamples individuals with extreme Q without regard to dropout time will tend to oversample subjects who have dropped out of the study before completion.



and covariate vector $\mathbf{X}_i$. The missingness in $\mathbf{Y}$ is defined to be MCAR if the missingness is independent of the outcome vector, i.e.,

$$P(\boldsymbol{R}_i|\mathbf{Y}_i, \mathbf{X}_i) = P(\boldsymbol{R}_i|\mathbf{X}_i)$$

In contrast, we define missingness as MAR if the probability of missing data depends on observed, but not missing, values, i.e.,:

$$P(\boldsymbol{R}_i|\mathbf{Y}_i^O, \mathbf{Y}_i^M, \mathbf{X}_i) = P(\boldsymbol{R}_i|\mathbf{Y}_i^O, \mathbf{X}_i)$$

Finally, if the missingness in $\mathbf{Y}$ depends on missing values, i.e., if $P(\boldsymbol{R}_i|\mathbf{Y}_i^O, \mathbf{Y}_i^M, \mathbf{X}_i)$ depends on $\mathbf{Y}_i^M$, we call it NMAR and non-ignorable. Likelihood-based methods can accommodate both MCAR and MAR (sometimes termed "ignorable") missingness, provided the model is correctly specified; if data is NMAR this approach is likely to be biased. However, under many circumstances it may be plausible that observed dropout times are related to observed longitudinal trajectories or underlying patient characteristics, reflecting a common disease process; in this case a likelihood-based method for accounting for dropout may be reasonable.

To see why a likelihood-based approach produces valid inference under a MAR structure, we can write the likelihood contribution of the ith individual, conditional on $\mathbf{X}_i$, for whom we observe the vector $(\mathbf{Y}_i^O, \mathbf{X}_i, \boldsymbol{R}_i)$, as:

$$\begin{aligned}
\mathcal{L}_i(\boldsymbol{\theta}, \phi|\mathbf{Y}_i^O, \mathbf{X}_i, \boldsymbol{R}_i) &= \int P(\mathbf{Y}_i^O, \mathbf{Y}_i^M|\mathbf{X}_i; \boldsymbol{\theta}) \cdot P(\boldsymbol{R}_i|\mathbf{Y}_i^O, \mathbf{Y}_i^M, \mathbf{X}_i; \phi)\, d\mathbf{Y}_i^M \\
&= \int P(\mathbf{Y}_i^O, \mathbf{Y}_i^M|\mathbf{X}_i; \boldsymbol{\theta}) \cdot P(\boldsymbol{R}_i|\mathbf{Y}_i^O, \mathbf{X}_i; \phi)\, d\mathbf{Y}_i^M \qquad (4.1) \\
&= P(\boldsymbol{R}_i|\mathbf{Y}_i^O, \mathbf{X}_i; \phi) \int P(\mathbf{Y}_i^O, \mathbf{Y}_i^M|\mathbf{X}_i; \boldsymbol{\theta})\, d\mathbf{Y}_i^M \\
&= P(\boldsymbol{R}_i|\mathbf{Y}_i^O, \mathbf{X}_i; \phi) \cdot P(\mathbf{Y}_i^O|\mathbf{X}_i; \boldsymbol{\theta}) \qquad (4.2)
\end{aligned}$$

where (4.1) is true by the MAR mechanism of missingness. Provided ϕ and $\boldsymbol{\theta}$ are "functionally distinct" [18], maximizing the likelihood across only across the second term in (4.2) will provide valid inference about $\boldsymbol{\theta}$. A similar argument shows that maximum likelihood leads to valid inference under MCAR as well.

Chapter 2 explored the contribution to inference of various components of the complete data likelihood, and produced a variety of valid estimators through maximization of several ascertainment-corrected likelihoods. A simple adaptation of the methods explored in Chapter 2 offers a straightforward approach to handling the simple but common MAR dropout structure considered here.

Assume that the complete longitudinal outcome vector $\mathbf{Y}_i$ has a multivariate normal distribution arising from the usual linear mixed model, as described previously in Section 2.3.1:

$$\mathbf{Y}_i |\, \mathbf{X}_i \sim MVN\left(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}\right),$$

where $\mathbf{X}_i = [\mathbf{1},\ \boldsymbol{T}_i,\ \boldsymbol{M}_i,\ \boldsymbol{M}_i \times \boldsymbol{T}_i]$, $\boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T})$, and $\boldsymbol{\Sigma} = \mathbf{ZDZ}^T + \sigma_e^2\mathbf{I}$. For a subject who drops out of the study before completion, the longitudinal outcome vector may be decomposed into its observed and missing components, i.e.,

$$\mathbf{Y}_i |\, \mathbf{X}_i = \left(\begin{array}{c} \mathbf{Y}_i^O \\ \mathbf{Y}_i^M \end{array} \middle|\, \mathbf{X}_i \right) \sim MVN\left[\left(\begin{array}{c} \mathbf{X}_i^O\boldsymbol{\beta} \\ \mathbf{X}_i^M\boldsymbol{\beta} \end{array} \right), \left(\begin{array}{cc} \boldsymbol{\Sigma}_{OO} & \boldsymbol{\Sigma}_{OM} \\ \boldsymbol{\Sigma}_{MO} & \boldsymbol{\Sigma}_{MM} \end{array} \right) \right]$$

and the marginal distribution of the observed portion of the outcome vector will likewise be multivariate normal, with

$$\mathbf{Y}_i^O |\, \mathbf{X}_i, D_i \sim MVN\left(\mathbf{X}_i^O\boldsymbol{\beta}, \boldsymbol{\Sigma}_{OO}\right)$$

Using the same ascertainment-correction approach as in Chapter 2, the likelihood (say, that gives rise to the estimator SO,NC that analyzes only subsampled individuals, conditional on

covariate values) under possible dropout would be

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}, \boldsymbol{S}) &= \prod_{i=1}^{N} \mathbb{I}(S_i = 1) \cdot f(\mathbf{Y}_i^O | \mathbf{X}_i, S_i, D_i; \boldsymbol{\theta}) \\
&= \prod_{i=1}^{N} \mathbb{I}(S_i = 1) \cdot \frac{\omega(q_i, d_i) \cdot f(\mathbf{Y}_i^O | D_i, \mathbf{X}_i) \cdot f(D_i | \mathbf{X}_i) \cdot f(\mathbf{X}_i)}{P(S_i = 1 | D_i, \mathbf{X}_i) \cdot f(D_i | \mathbf{X}_i) \cdot f(\mathbf{X}_i)} \\
&\propto \prod_{i=1}^{N} \mathbb{I}(S_i = 1) \cdot \frac{f(Y_i^O | \mathbf{X}_i; \boldsymbol{\theta})}{AC_0(m_i, \boldsymbol{t}, D_i; \boldsymbol{\theta})}
\end{aligned}$$

In contrast to Chapter 2, the ascertainment correction used to account for the biased sampling design will vary both by dropout time and marker type (and observations times, if not a balanced design); it can be written as:

$$\begin{aligned}
AC_0(M, \boldsymbol{T}, D) &= P(S = 1 | M, \boldsymbol{T}, D) \\
&= \sum_{k=1}^{3} P(S = 1, q \in R_k(d) | M, \boldsymbol{T}) \\
&= \sum_{k=1}^{3} P(S = 1 | q \in R_k(d), M, \boldsymbol{T}) \cdot P(q \in R_k(d) | M, \boldsymbol{T}) \\
&= \sum_{k=1}^{3} \omega_k(q, d) \int_{R_k} f(q | m, \boldsymbol{t}) dq
\end{aligned}$$

and is accommodated easily within the likelihood framework. Estimators that are unbiased and aymptotically efficient under correct model specification and mild regularity conditions can then be obtained via usual Newton-Raphson maximization techniques.

4.4 Regression standardization methods

In the ODS setting, while the researcher chooses which individuals to subsample, the data seen in subsampled individuals bears some resemblance to conventional observational data, in which the researcher does not choose the exposure that an individual receives. In both

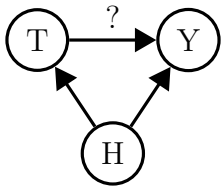
cases, the parameter of interest is the marginal effect of exposure arising from a hypothetical trial in which the researcher could randomly assign the exposure and then subsequently observe the corresponding outcome. And in both cases, the researcher instead observes only one of the possible exposures for each individual, the assignment of which is likely related to participant characteristics, or in more complicated longitudinal settings, possibly related to intermediate exposures or outcomes. Because of these similarities, the causal inference literature, which attempts to recover the aforementioned causal parameter of interest from observational data, may provide methods relevant to the ODS setting, yielding yet another avenue to valid inference.

4.4.1 G-computation and regression standardization

One tool from the causal inference literature, which allows recovery of the causal parameter of interest under certain conditions, comes in the form of G-computation [30]. Originally developed by Robins to account for time-dependent confounders, it can nonetheless be applied to cross-sectional studies or longitudinal studies with simpler covariate structures. As a hypothetical example, suppose a researcher wishes to study the effect of a drug's off-label use ($T = 1$ indicating treatment) on kidney function Y measured by eGFR; however, the researcher has access only to clinical (i.e., uncontrolled and nonrandomized) data. Moreover, the researcher notices that clinic physicians tend to prescribe the drug more often to healthier patients ($H = 1$) because of concerns about how sicker patients may tolerate the drug. Clearly, simply examining the observed marginal effect of drug on kidney function in this scenario would not produce valid inference, because health status is related both to the probability of treatment and to kidney function itself (Figure 4.2).

Moreover, suppose that, for patients of a given health status, the physician prescribes the treatment with a constant probability; alternatively, that within each stratum, we effectively have a randomized experiment. We could recover the marginal distribution by integrating

Figure 4.2: Illustration of hypothetical G-computation example. Here, the treatment T's effect on kidney function Y is confounded by health status H, which affects both the probability of being treated with the drug and kidney function itself. Analyzing the marginal effect of T on Y would not produce correct inference.



the fully conditional distribution over the distribution of $H|T$, or

$$f(Y|T) = \int_{H|T} f(Y|T,H)\ dF_{H|T}$$

If we care not about the entire distribution, but only about the mean marginal effect, we can obtain the desired expectation as:

$$\mathbb{E}[Y|T] = \mathbb{E}_{H|T}[\mathbb{E}(Y|H,T)]$$

If we further suppose that health status H is the only potential confounder, and that it is measured and accounted for in the above manner, the parameters associated with the recovered expectation $\mathbb{E}[Y|T]$ can be interpreted as the same causal parameters arising from the desired hypothetical randomized trial.

In this simple example, we have supposed that health status explained everything about the pattern of treatment; more likely, this fully explanatory set might be a collection of covariates, previous outcomes (in the longitudinal setting), or a "balancing score" [34]. Once the balancing score (or set of covariates) is obtained, then the marginal and causal parameter of interest can be recovered by first conditioning on and then integrating over the balancing

score, as shown above. In the last 15 years, this approach has been used to estimate causal effects from observational data in fields such as cardiovascular epidemiology, [43] air pollution, [23], and infectious disease. [27]

While the G-computation approach has the potential to allow for estimation of causal parameters in observational data, several conditions must be met for this interpretation to be valid. First, there must be no unmeasured confounding. Either through the set of confounders explicitly, or through the balancing score, the conditioning step should break all confounding by restriction to strata within which the exposure's effect is causal. Second, the stable unit of treatment value assumption (or SUTVA) must be met; the probability that one experimental unit (the clinic patient, in the example above) receives the treatment does not affect the probability that another will receive the treatment. Finally, the experimental treatment assumption (ETA) must be fulfilled: in all strata there must be a non-zero probability of both treatment and non-treatment.

Ideas closely related to G-computation have long been used in the epidemiological literature under the name "standardization" [41]. "Indirect" standardization applies covariate-specific disease risks derived from an unexposed population to an exposed population; "direct" standardization conversely calculates the number of cases would have occurred among the unexposed, had they been exposed. [45] [24] [13] Applying the idea of standardization to the entire population of exposed and unexposed has been shown [45] [37] to be equivalent to the G-computation methods described above. We use the term "regression standardization" for the method here as a descriptive term that relates the technique to its long lineage in the epidemiological literature.

4.4.2 Pattern mixture modeling

A similar idea to G-computation/regression standardization has been proposed in a different context, that of the development of methods to accommodate dropout. In many circum-

stances, subjects with differing dropout times may have different trends in trajectories. In order to handle dropout, Little's "pattern mixture modeling" [19] [20] focused on the distribution of $\mathbf{Y}_i$ conditional on the dropout time D_i. If $\boldsymbol{R}_i = [R_{ij}]$ is the indicator of observing Y_{ij} with $D_i = \min_k(R_{ik} = 0)$, Little observed that the complete data likelihood could be written

$$\begin{aligned} f(\mathbf{Y}_i, \boldsymbol{R}_i) &= f(\mathbf{Y}_i, D_i) \\ &= f(\mathbf{Y}_i|D_i)\ f(D_i) \end{aligned}$$

Thus, if $\mathbb{E}[Y_{ij}|X_{ij}]$ is the parameter of interest, it could be reconstructed from the partially observed data using a model of $\mathbf{Y}$ on $\mathbf{X}$, conditional on dropout time:

$$\mathbb{E}[Y_{ij}|X_{ij}, D_i = d] = \beta_0(d) + \beta_1(d)X_{ij}$$

Together with the conditional distribution of dropout times, $\pi_d = P(D_i = d|\mathbf{X}_i)$, the desired unconditional expectation can then be calculated by averaging across the distribution of dropout times, as

$$\mathbb{E}[Y_{ij}|X_{ij}] = \sum_d (\beta_0(d) + \beta_1(d)X_{ij}) \cdot \pi_d$$

If dropout times and $\mathbf{X}_i$ are not independent, π_d may need to be derived from the observed empirical distributions of $[\mathbf{X}|D]$, $[\mathbf{X}]$ and $[D]$, via Bayes theorem. Finally, we note that while convenient to write $\mathbb{E}[\mathbf{Y}|\mathbf{X}, D]$ as a linear model, as above, more flexible versions could also easily be accommodated.

4.4.3 Regression standardization for complete ODS data

Results from Chapters 2 and 3 have shown that incorporating information from unsubsampled people can potentially improve efficiency for select regression parameters. Given this

fact, our primary target of inference, the parameters that describe the complete outcome trajectories over time by marker type, could alternatively be recovered by applying ideas from the regression standardization literature [19] [20] [10] and iterating over information that is available for the full cohort. Specifically, under complete and balanced data, with regression standardization it would be possible to recover the target regression parameter $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$ by iterating over the subsampling variable Q:

$$\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbb{E}_{Q|\mathbf{X}}\left\{\mathbb{E}_{\mathbf{Y}|Q,\mathbf{X}}[\mathbf{Y}|Q,\mathbf{X}]\right\}$$

Since Q is a coarsened verison of the outcome vector $\mathbf{Y}$, it may appear somewhat unconventional to iterate over the subsampling variable Q; however, doing so allows information from the whole cohort to be incorporated into the estimator. As a moment-based estimator, some robustness to model misspecification may also be expected. Moreover, this approach may also be an attractive alternative to the likelihood-based methods, since intermediate models (such as those for $\mathbb{E}[\mathbf{Y}|Q,\mathbf{X}]$) necessary for the computation of $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$ could easily be made more flexible with nonparametric regression techniques.

For example, a simple approach to calculating $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$ in a complete and balanced case might proceed by first fitting

$$\mathbb{E}[\mathbf{Y}|\mathbf{X},Q] \;=\; \gamma_0 + \gamma_T T + \gamma_M M + \gamma_{M\times T} M\cdot T + \gamma_Q Q + \gamma_{Q\times T} Q\cdot T \tag{4.3}$$

using subsampled individuals only. Estimating $\mathbb{E}[\mathbf{Y}|\mathbf{X},Q]$ using only subsampled individuals is valid; since the ODS design assumes the probability of subsampling conditional on Q to be a constant, $\omega(q)$. Under this assumption, conditional on Q, those who are subsampled are exchangeable with those who are not subsampled. As mentioned above, the particular choice of model could be made more flexible; the above representation is that which would arise if $\mathbf{Y}$ had a multivariate normal distribution.

Next, the conditional expectation in Equation 4.3 must be averaged across the distribution of Q given covariates $\mathbf{X}$. Unlike Equation 4.3, however, which could be estimated based on subsampled individuals only, the same is not true here since

$$[Q|\mathbf{X}] \neq [Q|\mathbf{X}, S = 1] \neq [Q|\mathbf{X}, S = 0]$$

Under the ODS design, however, it *is* true that

$$[\mathbf{X}|Q] = [\mathbf{X}|Q, S = 1] = [\mathbf{X}|Q, S = 0]$$

and for a binary marker this fact can be exploited together with an application of Bayes Theorem to estimate expectations over the distribution of $Q|\mathbf{X}$. In other words, for complete designs with a binary marker we can write that

$$f(Q|\mathbf{X}) \;\; = \;\; f(Q|M) = \frac{P(M|Q;\boldsymbol{\alpha}) \cdot f(Q)}{\displaystyle\int P(M|Q;\boldsymbol{\alpha}) \cdot f(Q)\; dq}$$

We can estimate $\mathbb{E}[M|Q]$ using logistic regression among subsampled individuals (again valid because of the conditioning on Q). Combining this with the empirical distribution of Q, $\mathbb{F}_Q$, estimated from the entire cohort, we have an expedient moment-based method of averaging across the distribution of $Q|\mathbf{X}$.

If we estimate $\text{logit}(\hat{\mathbb{E}}[M|Q]) = \hat{\eta}_0 + \hat{\eta}_1 Q$ through logistic regression on subsampled individuals and estimate the marginal distribution of Q with the empirical distribution $\mathbb{F}_Q$, we can in turn estimate $\hat{\mathbb{E}}[Q|M] = \hat{\alpha}_0 + \hat{\alpha}_1 M$, where $\hat{\alpha}_0(\hat{\boldsymbol{\eta}}, \mathbb{F}_Q) = \hat{\mathbb{E}}[Q|M = 0]$, $\hat{\alpha}_1(\hat{\boldsymbol{\eta}}, \mathbb{F}_Q) = \hat{\mathbb{E}}[Q|M = 1] - \hat{\mathbb{E}}[Q|M = 0]$, and

$$\hat{\mathbb{E}}[Q|M = 1] \;\; = \;\; \sum_{i=1}^{n} \frac{\hat{P}(M = 1|q_i) \cdot q_i}{\sum_j \hat{P}(M = 1|q_j)}$$

$$= \sum_{i=1}^{n} \frac{\text{expit}(\hat{\eta}_0 + \hat{\eta}_1 q_i) \cdot q_i}{\sum_j \text{expit}(\hat{\eta}_0 + \hat{\eta}_1 q_i)} \tag{4.4}$$

$$\text{and } \hat{\mathbb{E}}[Q|M=0] = \sum_{i=1}^{n} \frac{\hat{P}(M=0|q_i) \cdot q_i}{\sum_j \hat{P}(M=0|q_j)}$$

$$= \sum_{i=1}^{n} \frac{(1 - \text{expit}(\hat{\eta}_0 + \hat{\eta}_1 q_i)) \cdot q_i}{\sum_j (1 - \text{expit}(\hat{\eta}_0 + \hat{\eta}_1 q_j))} \tag{4.5}$$

We then average Equation (4.3) across $\mathbb{E}[Q|M]$ to obtain an estimate of $\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the regression parameter vector of interest from the usual longitudinal linear mixed model:

$$\begin{aligned}
\mathbb{E}[\mathbf{Y}|\mathbf{X}] &= \mathbb{E}_{Q|\mathbf{X}}\left\{\mathbb{E}_{\mathbf{Y}|Q,\mathbf{X}_i}[\mathbf{Y}|Q,\mathbf{X}]\right\} \\
&= \mathbb{E}_{Q|\mathbf{X}}\left\{\gamma_0 + \gamma_T T + \gamma_M M + \gamma_{M\times T} M \cdot T + \gamma_Q Q + \gamma_{Q\times T} Q \cdot T\right\} \\
&= \gamma_0 + \gamma_T T + \gamma_M M + \gamma_{M\times T} M \cdot T + \gamma_Q\left(\alpha_0 + \alpha_1 M\right) + \gamma_{Q\times T}\left(\alpha_0 + \alpha_1 M\right) \cdot T \\
&= \underbrace{(\gamma_0 + \gamma_Q\alpha_0)}_{\beta_0} + \underbrace{(\gamma_T + \gamma_{Q\times T}\alpha_0)}_{\beta_T} T + \underbrace{(\gamma_M + \gamma_Q\alpha_1)}_{\beta_M} M + \underbrace{(\gamma_{M\times T} + \gamma_{Q\times T}\alpha_1)}_{\beta_{M\times T}} M \cdot T
\end{aligned} \tag{4.6}$$

Asymptotically valid standard errors for $\hat{\boldsymbol{\beta}}$ can be found using the multivariate δ-method (Appendix E).

4.5 Operating characteristics of regression standardization methods under complete data

To assess the potential benefit of regression standardization methods in the ODS setting under complete data, we performed a simulation study comparing these methods against the likelihood-based ODS methods explored in Chapter 2 and traditional random sampling likelihood-based methods. Using the same (correctly specified) data-generating mechanism and simulation setup as described in Section 2.4, we compared the percent bias and relative efficiency of ODS regression standardization methods with the best-performing likelihood based ODS methods, where performance was measured relative to a random subsample one

Table 4.1: Percent bias and relative efficiency for likelihood-based and regression standardization methods under complete data and low subject-to-subject heterogeneity. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 × the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

	Estimator				
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Full cohort	Likelihood	0 (4.18)	-0 (3.76)	0 (4.13)	0 (4.00)
Intercept	Likelihood (SO,NC)	0 (1.74)	0 (0.99)	2 (1.73)	0 (1.06)
	Likelihood (SU,WC)	0 (3.68)	-0 (2.66)	3 (1.81)	1 (1.19)
	Regression standardization	0 (3.64)	0 (1.00)	3 (1.73)	0 (1.08)
Slope	Likelihood (SO,NC)	-0 (0.97)	-0 (1.54)	2 (1.21)	2 (1.28)
	Likelihood (SU,WC)	0 (3.34)	-0 (3.05)	3 (1.19)	3 (1.28)
	Regression standardization	-0 (0.97)	-0 (3.08)	2 (1.20)	2 (1.38)

quarter of the original cohort size (Tables 4.1 and 4.2 show results for low and high subject-to-subject heterogeneity, respectively).

Under correct model specification and low subject-to-subject heterogeneity (Table 4.1), the regression standardization methods uniformly had similar or lower relative efficiency than ODS likelihood-based methods that used both subsampled and non-subsampled individuals, for the same design. For example, the regression standardization estimator paired with an intercept-targeted design had efficiencies of 3.64 and 1.73 relative to a random sample for the intercept and marker main effects, respectively; the likelihood-based estimator SU,WC had relative efficiencies of 3.68 and 1.81, respectively. For non-targeted parameters, regression standardization estimators had relative efficiencies that were similar to (and sometimes better than) a random sample. Interestingly, efficiency gains for regression standardization estimators were not seen for both β_0 and β_T, but rather only in the parameter targeted by

Table 4.2: Percent bias and relative efficiency for likelihood-based and regression standardization methods under complete data and high subject-to-subject heterogeneity. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 × the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

	Estimator				
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Full cohort	Likelihood	0 (3.85)	0 (4.06)	3 (3.83)	1 (3.53)
Intercept	Likelihood (SO,NC)	0 (1.62)	0 (0.96)	7 (1.56)	-1 (0.84)
	Likelihood (SU,WC)	0 (3.29)	0 (3.03)	7 (1.48)	0 (0.78)
	Regression standardization	0 (3.31)	0 (0.96)	7 (1.55)	0 (0.86)
Slope	Likelihood (SO,NC)	0 (0.90)	0 (1.89)	4 (0.93)	2 (1.47)
	Likelihood (SU,WC)	0 (2.95)	0 (3.63)	5 (0.86)	2 (1.45)
	Regression standardization	0 (0.90)	0 (3.63)	4 (0.94)	2 (1.49)

the design. Results under high subject-to-subject heterogeneity (Table 4.2) were qualitatively similar, although in that case regression standardization estimators had slightly lower efficiency than a random sample for non-targeted parameters. In no case, however, was the "efficiency penalty" for non-targeted parameters for the regression standardization estimator as severe as for the IPW estimator considered in Chapter 3.

We also investigated the possible robustness properties of the regression standardization estimator under the random effects misspecification described in Section 3.3.2 that had the most profoundly adverse effects on the ODS likelihood-based estimators, when the random effects followed a $\Gamma(5, \sqrt{3})$ distribution but were analyzed under likelihood-based methods as if bivariate normally distributed. Percent biases and relative MSEs for likelihood-based and regression standardization estimators are shown in Tables 4.3 and 4.4. For low subject-to-subject heterogeneity (Table 4.3), regression standardization estimators had relative biases

less than 15%. Consequently, the impact on the relative MSE was relatively modest, and most of the efficiency gains seen under correct model specification in Table 4.1 were retained (again, results under high subject-to-subject heterogeneity presented in Table 4.4 were qualitatively similar). While regression standardization methods do use assumptions about model form, etc., in the initial conditioning step, the moment-based estimation at the heart of the method appears to provide some robustness under even severe misspecification. In this way, regression standardization estimators may be thought of as having intermediate efficiency and robustness compared to the likelihood-based methods of Chapter 2 and the robust methods of Chapter 3.

Table 4.3: Percent bias and relative MSE for likelihood-based and regression standardization methods under complete data and model misspecification, under low subject-to-subject heterogeneity. Results shown summarize 1000 replications with $N = 1000$, $\boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$, and $\rho = 0$, where the true data generating mechanism was a severely skewed $\Gamma(5, \sqrt{3})$ distribution. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative MSE for an estimator is defined as the ratio of MSEs between a random sample of $N_S = 250$ and the estimator.

	Estimator				
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Full cohort	Likelihood	0 (4.17)	-0 (4.15)	-1 (4.38)	-0 (4.08)
Intercept	Likelihood (SO,NC)	1 (1.57)	-10 (1.04)	0 (1.91)	-2 (1.04)
	Likelihood (SU,WC)	-0 (3.69)	12 (2.57)	-15 (2.25)	51 (0.82)
	Regression standardization	0 (3.63)	-11 (1.05)	-2 (1.97)	-2 (1.04)
Slope	Likelihood (SO,NC)	1 (1.07)	-9 (1.64)	14 (1.20)	9 (0.95)
	Likelihood (SU,WC)	-0 (3.74)	3 (3.29)	-9 (1.14)	47 (0.57)
	Regression standardization	1 (1.09)	-1 (3.41)	13 (1.20)	5 (1.29)

Table 4.4: Percent bias and relative MSE for likelihood-based and regression standardization methods under complete data and model misspecification, under high subject-to-subject heterogeneity. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4$, and $\rho = 0$, where the true data generating mechanism was a severely skewed $\Gamma(5, \sqrt{3})$ distribution. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative MSE for an estimator is defined as the ratio of MSEs between a random sample of $N_S = 250$ and the estimator.

	Estimator				
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Full cohort	Likelihood	0 (4.11)	0 (4.04)	1 (4.05)	0 (4.32)
Intercept	Likelihood (SO,NC)	1 (1.32)	-14 (0.87)	5 (1.51)	-4 (1.02)
	Likelihood (SU,WC)	0 (3.51)	26 (0.84)	-5 (1.56)	119 (0.13)
	Regression standardization	0 (3.51)	-17 (0.86)	3 (1.57)	-4 (1.03)
Slope	Likelihood (SO,NC)	1 (0.65)	-39 (0.91)	4 (0.84)	12 (1.46)
	Likelihood (SU,WC)	0 (3.10)	4 (2.95)	-12 (0.63)	45 (0.47)
	Regression standardization	2 (0.59)	0 (3.65)	4 (0.84)	9 (1.58)

4.6 Regression standardization methods for data with unanticipated missingness

When unanticipated missingness is present, the regression standardization approach described in Section 4.4.3 can be modified to accommodate this additional complexity. In the complete case, iterating over the subsampling variable Q incorporated information from unsubsampled individuals into inference; under dropout, additionally iterating over the dropout time D will allow us to recover the parameter of interest $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$, unconditional on dropout time. Supposing that the ODS subsampling scheme has followed the dropout-stratified approach outlined in Section 4.2, we can write the target of inference as

$$\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbb{E}_{Q|\mathbf{X}}\left(\mathbb{E}_{D|Q,\mathbf{X}}\left\{\mathbb{E}_{\mathbf{Y}|D,Q,\mathbf{X}}[\mathbf{Y}|D,Q,\mathbf{X}]\right\}\right)$$

We follow the example of the complete data scenario, but allow for differing effects by dropout time:

$$\begin{aligned}\mathbb{E}[\mathbf{Y}|D,Q,\mathbf{X}] &= \gamma_0+\gamma_T T+\gamma_M M+\gamma_{M\times T}M\cdot T+\gamma_Q Q+\gamma_{Q\times T}Q\cdot T\\ &\quad+\gamma_D D+\gamma_{T\times D}T\cdot D+\gamma_{M\times D}M\cdot D+\gamma_{M\times T\times D}M\cdot T\cdot D\\ &\quad+\gamma_{Q\times D}Q\cdot D+\gamma_{Q\times T\times D}Q\cdot T\cdot D \qquad (4.7)\end{aligned}$$

We pragmatically model $\mathbb{E}[D|Q,\mathbf{X}]=\phi_0+\phi_Q Q+\phi_M M$; combined with Equation (4.7), this yields

$$\begin{aligned}\mathbb{E}[\mathbf{Y}|Q,\mathbf{X}] &= \mathbb{E}_{D|Q,\mathbf{X}}\left[(\gamma_0+\gamma_T T+\gamma_M M+\gamma_{M\times T}M\cdot T+\gamma_Q Q+\gamma_{Q\times T}Q\cdot T\right.\\ &\quad\left.+\gamma_D D+\gamma_{T\times D}T\cdot D+\gamma_{M\times D}M\cdot D+\gamma_{M\times T\times D}M\cdot T\cdot D\right.\\ &\quad\left.+\gamma_{Q\times D}Q\cdot D+\gamma_{Q\times T\times D}Q\cdot T\cdot D)\,|Q,\mathbf{X}\right]\\ &= (\gamma_0+\gamma_D\phi_0)+(\gamma_T+\gamma_{T\times D}\phi_0)T+(\gamma_M+\gamma_D\phi_M+\gamma_{M\times D}\phi_0)M\\ &\quad+(\gamma_{M\times T}+\gamma_{T\times D}\phi_M+\gamma_{M\times T\times D}\phi_0)M\cdot T+(\gamma_Q+\gamma_D\phi_Q+\gamma_{Q\times D}\phi_0)Q\\ &\quad+(\gamma_{T\times Q}+\gamma_{T\times D}\phi_Q+\gamma_{T\times Q\times D}\phi_0)Q\cdot T+(\gamma_{M\times D}\phi_Q+\gamma_{Q\times D}\phi_M)M\cdot Q\\ &\quad+\gamma_{M\times D}\phi_M M^2+\gamma_{Q\times D}\phi_Q Q^2+(\gamma_{M\times T\times D}\phi_Q+\gamma_{T\times Q\times D}\phi_M)M\cdot Q\cdot T\\ &\quad+\gamma_{M\times T\times D}\phi_M M^2\cdot T+\gamma_{Q\times T\times D}\phi_Q Q^2\cdot T \qquad (4.8)\end{aligned}$$

Using Bayes Theorem together with a logistic model of $\mathbb{E}[M|Q,\boldsymbol{T}]=\mathbb{E}[M|Q]=\text{expit}(\eta_0+\eta_1 Q)$, we can obtain $\mathbb{E}[Q|M,\boldsymbol{T}]=\alpha_0+\alpha_1 M$, as in Section 4.4.3, and $\mathbb{E}[Q^2|M,\boldsymbol{T}]=\xi_0+\xi_1 M$, similarly. Averaging $\mathbb{E}[Y|Q,\mathbf{X}]$ from Equation (4.8) across $\mathbb{E}[Q|M,\boldsymbol{T}]$, we finally have:

$$\begin{aligned}\mathbb{E}[\mathbf{Y}|\mathbf{X}] &= \gamma_0+\gamma_D\phi_0+\alpha_0(\gamma_Q+\gamma_D\phi_Q+\gamma_{Q\times D}\phi_0)+\xi_0\gamma_{D\times Q}\phi_Q \qquad (4.9)\\ &\quad+\left\{\gamma_T+\gamma_{T\times D}\phi_0+\alpha_0\left(\gamma_{T\times Q}+\gamma_{T\times D}\phi_Q+\gamma_{T\times Q\times D}\phi_0\right)+\xi_0\gamma_{T\times Q\times D}\phi_Q\right\}T\\ &\quad+\left\{\gamma_M+\gamma_D\phi_M+\gamma_{M\times D}\phi_0+\alpha_0\left(\gamma_{M\times D}\phi_Q+\gamma_{Q\times D}\phi_M\right)+\alpha_1\left(\gamma_Q+\gamma_D\phi_Q+\gamma_{Q\times D}\phi_0\right)\right.\end{aligned}$$

$$
\begin{aligned}
&\quad + \xi_1\gamma_{Q\times D}\phi_Q\} M \\
&+ \{\gamma_{M\times T} + \gamma_{T\times D}\phi_M + \gamma_{M\times T\times D}\phi_0 + \alpha_0\left(\gamma_{M\times T\times D}\phi_Q + \gamma_{T\times Q\times D}\phi_M\right) \\
&\quad + \alpha_1\left(\gamma_{T\times Q} + \gamma_{T\times D}\phi_Q + \gamma_{Q\times T\times D}\phi_0\right) + \xi_1\gamma_{T\times Q\times D}\phi_Q\} M\cdot T \\
&+ \{\gamma_{M\times D}\phi_M + \alpha_1\left(\gamma_{M\times D}\phi_Q + \gamma_{Q\times D}\phi_M\right)\} M^2 \\
&+ \{\gamma_{M\times T\times D}\phi_M + \alpha_1\left(\gamma_{M\times T\times D}\phi_Q + \gamma_{T\times Q\times D}\phi_M\right)\} M^2\cdot T
\end{aligned}
$$

In contrast to the complete data scenario, iterating across subsampling variable Q and dropout time D leads to an expectation involving M^2 and $M^2 \times T$ terms, which does not conform to the usual longitudinal model of interest. However, for a binary marker, M and M^2 are the same value, so in this special case we can interpret the expectation in Equation 4.9 as being equivalent to

$$
\begin{aligned}
\mathbb{E}[\mathbf{Y}|\mathbf{X}] &= \gamma_0 + \gamma_D\phi_0 + \alpha_0(\gamma_Q + \gamma_D\phi_Q + \gamma_{Q\times D}\phi_0) + \xi_0\gamma_{D\times Q}\phi_Q \\
&\quad + \{\gamma_T + \gamma_{T\times D}\phi_0 + \alpha_0\left(\gamma_{T\times Q} + \gamma_{T\times D}\phi_Q + \gamma_{T\times Q\times D}\phi_0\right) + \xi_0\gamma_{T\times Q\times D}\phi_Q\} T \\
&\quad + \{\gamma_M + \gamma_D\phi_M + \gamma_{M\times D}\phi_0 + \alpha_0\left(\gamma_{M\times D}\phi_Q + \gamma_{Q\times D}\phi_M\right) + \alpha_1\left(\gamma_Q + \gamma_D\phi_Q + \gamma_{Q\times D}\phi_0\right) \\
&\qquad + \xi_1\gamma_{Q\times D}\phi_Q + \gamma_{M\times D}\phi_M + \alpha_1\left(\gamma_{M\times D}\phi_Q + \gamma_{Q\times D}\phi_M\right)\} M \\
&\quad + \{\gamma_{M\times T} + \gamma_{T\times D}\phi_M + \gamma_{M\times T\times D}\phi_0 + \alpha_0\left(\gamma_{M\times T\times D}\phi_Q + \gamma_{T\times Q\times D}\phi_M\right) \\
&\qquad + \alpha_1\left(\gamma_{T\times Q} + \gamma_{T\times D}\phi_Q + \gamma_{M\times T\times D}\phi_0\right) + \xi_1\gamma_{T\times Q\times D}\phi_Q \\
&\qquad + \gamma_{M\times T\times D}\phi_M + \alpha_1\left(\gamma_{M\times T\times D}\phi_Q + \gamma_{T\times Q\times D}\phi_M\right)\} M\cdot T \\
&= \beta_0 + \beta_T T + \beta_M M + \beta_{M\times T} M\times T
\end{aligned}
$$

As with the complete data scenario, asymptotic standard errors can be found using repeated applications of the δ-method.

4.7 Operating characteristics of regression standardization methods under dropout

4.7.1 Simulation setup

We evaluated the likelihood-based and regression standardization methods for ODS data described above under conditions of dropout, via simulation. We consider the validity and possible efficiency gains of these methods under two dropout mechanisms, one MAR and one MNAR. In each simulation study, the true underlying conditional outcome distribution is multivariate normal, but the dropout mechanism depends either on the marker value alone (in the MAR scenario) or on the marker and on the entire outcome vector (in the MNAR scenario). Given the ODS sampling schemes under consideration here, we generated dropout structures that required that individuals have at least two observed outcomes. The data-generating mechanisms and simulation results are described below.

For the MAR dropout simulation, each individual's true (though possibly partially unobserved) $n_i = 6 \times 1$ outcome vector was generated from a correctly specified multivariate normal distribution. Next, each person's "potential" dropout time, related to marker, was generated; for those without the marker, potential dropout times were uniformly distributed; for those with the marker, potential dropout times of $D_i = (2, 3, 4, 5, 6)$ occurred with probability $\boldsymbol{p} = (0.24, 0.22, 0.20, 0.18, 0.16)$. (i.e., those with the marker were more likely to dropout earlier, if they do drop out). The probability of dropping out at the individual's potential dropout time was then generated as a Bernoulli random variable with probability of success expit$(-2 + m_i)$. In short, those with the marker were more likely to drop out than those without, and if they did, they were more likely to drop out earlier. For these simulations, the marker population prevalence was upped to 25%; otherwise, data was generated as described previously in Section 2.4.1.

We also conducted a simulation study of the performance of likelihood-based and regression standardization methods under a MNAR dropout mechanism that depended partially

on possibly unobserved outcomes. In this setup, potential dropout times were generated as described above, with individuals with the marker more likely to dropout earlier, if dropout occurred. However, here we generated the probability of dropping out at the individual's potential dropout time as a Bernoulli random variable that depended both on the marker value and the sum of residuals from the individual's complete outcome vector; i.e., the ith individual's probability of dropout was expit $\left(-2 + m_i + \sum_{j=1}^{n_i} (y_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta})\right)$. Thus those with above average outcomes were more likely to drop out, as were those with the marker. Since the probability of dropout is related to a possibly unobserved component of the outcome vector, the dropout mechanism is considered MNAR, under which neither likelihood-based nor regression standardization models are guaranteed to be valid. As above, the marker population prevalence was set to 25%; all other simulation parameters were as described in Section 2.4.1.

4.7.2 Simulation results

Under a MAR dropout structure and low subject-to-subject heterogeneity, regression standardization methods were unbiased, as were likelihood-based methods (Table 4.5). Likelihood-based methods exhibited the same patterns of relative efficiency as seen under complete data (Table 4.1), although relative efficiencies were lower under dropout, possibly because a smaller amount of important (i.e., marker-related) information was available for analysis. Regression standardization estimators for intercept-based designs maintained efficiencies for targeted parameters (2.30 and 1.40 for β_0 and β_M, respectively), but had efficiency worse than a random sample for non-targeted parameters. When slope-based designs were used, the regression standardization estimator showed improved efficiency only in β_T; all other parameters had similar or worse efficiency than a random sample. Results of simulations with MAR (Tables 4.5 and 4.6) and MNAR (Tables 4.7 and 4.8) were similar.

Table 4.5: Percent bias and relative efficiency for likelihood-based and regression standardization methods under MAR dropout and low subject-to-subject heterogeneity. The data generating mechanism was correctly specified, but dropout time and probability are related to marker type. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

	Estimator				
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Full cohort	Likelihood	0 (3.75)	0 (3.94)	0 (4.06)	0 (4.24)
Intercept	Likelihood (SO,NC)	0 (1.66)	0 (0.99)	1 (1.67)	0 (1.03)
	Likelihood (SU,WC)	0 (2.90)	1 (2.23)	0 (1.92)	0 (1.35)
	Regression standardization	0 (2.30)	0 (0.69)	1 (1.40)	0 (0.60)
Slope	Likelihood (SO,NC)	0 (1.05)	0 (1.65)	1 (1.08)	1 (1.34)
	Likelihood (SU,WC)	0 (2.44)	1 (2.80)	0 (1.14)	-1 (1.69)
	Regression standardization	0 (0.98)	1 (1.29)	1 (1.01)	1 (0.76)

4.8 Application to Cystic Fibrosis Foundation Registry data

We illustrate the relative merits of regression standardization methods with an application to the previously explored (Sections 2.5 and 3.4) cohort of 3,141 cystic fibrosis patients. All cohort members had six complete and equally spaced lung function measurements; we first illustrate results using the complete outcomes, and then induce some missingness to illustrate the possible utility of these methods under dropout.

Table 4.10 shows the results of analyzing the complete outcomes CF registry data using likelihood-based and regression standardization methods. Both likelihood-based and regression standardization estimators gave similar point estimates for all parameters. For regression standardization estimators, the empirical standard errors were similar to those of the

Table 4.6: Percent bias and relative efficiency for likelihood-based and regression standardization methods under MAR dropout and high subject-to-subject heterogeneity. The data generating mechanism is correctly specified, but dropout time and probability are related to marker type. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

	Estimator				
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Full cohort	Likelihood	0 (4.14)	0 (3.89)	-1 (4.39)	0 (4.14)
Intercept	Likelihood (SO,NC)	0 (1.77)	0 (0.96)	1 (1.75)	2 (0.92)
	Likelihood (SU,WC)	0 (2.90)	2 (2.12)	-1 (1.85)	2 (0.93)
	Regression standardization	0 (1.38)	2 (0.72)	2 (0.70)	3 (0.54)
Slope	Likelihood (SO,NC)	0 (1.01)	1 (1.70)	-3 (1.06)	3 (1.58)
	Likelihood (SU,WC)	0 (2.22)	2 (2.88)	-2 (1.08)	2 (1.65)
	Regression standardization	0 (0.93)	1 (2.40)	-4 (0.89)	2 (1.28)

most efficienct likelihood-based estimator (SU, WC) for targeted parameters and typically somewhat larger for non-targeted parameters. Finally, both likelihood-based and regression standardization estimators that targeted $\beta_{M\times T}$ detected the small but statistically significant interaction; these were the only estimators (other than analyzing the full cohort) that did so.

In order to illustrate the use of regression standardization and likelihood-based methods under dropout, we used the aforementioned cohort of 3,141 CF patients with six complete and equally spaced lung function measurements. Similar to the data-generating mechanism used in the simulations described in Section 4.7, we artificially induced some missingness into the CF cohort that depended on whether an individual had the *S. aureus* bacteria present at baseline. Specifically, as in Section 4.7 we generated potential dropout times of

Table 4.7: Percent bias and relative efficiency for likelihood-based and regression standardization methods under MNAR dropout and low subject-to-subject heterogeneity. The data generating mechanism is correctly specified, but dropout time and probability are related to marker type. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

	Estimator				
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Full cohort	Likelihood	0 (3.72)	0 (4.26)	0 (3.76)	0 (4.32)
Intercept	Likelihood (SO,NC)	0 (1.67)	1 (0.94)	1 (1.63)	0 (1.01)
	Likelihood (SU,WC)	0 (2.88)	2 (2.17)	0 (1.84)	0 (1.25)
	Regression standardization	0 (2.42)	0 (0.69)	1 (1.41)	0 (0.60)
Slope	Likelihood (SO,NC)	0 (0.93)	0 (1.66)	0 (1.06)	1 (1.40)
	Likelihood (SU,WC)	0 (2.19)	2 (2.90)	-1 (1.12)	-1 (1.67)
	Regression standardization	0 (0.90)	1 (1.39)	1 (1.00)	1 (0.74)

$D_i = (2, 3, 4, 5, 6)$ that occurred with probability $\boldsymbol{p} = (0.24, 0.22, 0.20, 0.18, 0.16)$. Then, the probability of dropping out at the individual's potential dropout time was generated as a Bernoulli random variable with probability of success expit$(-2 + m_i)$. This resulted in 597 (19 %) individuals of the CF cohort having some missingness; Table 4.9 shows the distribution of *S. aureus* marker by dropout time.

Table 4.11 shows the results of analyzing the CF cohort data under induced dropout. All likelihood-based and regression standardization methods gave similar point estimates, except for slope-based designs with respect to the parameter β_M. For intercept-based designs, standard errors were virtually identical to the likelihood-based estimator than analyzed both subsampled and unsubsampled individuals; regression standardization standard errors for non-targeted parameters were larger than those of this likelihood-based estimator or those

Table 4.8: Percent bias and relative efficiency for likelihood-based and regression standardization methods under MNAR dropout and high subject-to-subject heterogeneity. The data generating mechanism is correctly specified, but dropout time and probability are related to marker type. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

	Estimator				
Design	Analysis method	β_0	β_T	β_M	$\beta_{M\times T}$
Full cohort	Likelihood	0 (4.16)	0 (3.55)	-1 (4.11)	-2 (3.52)
Intercept	Likelihood (SO,NC)	0 (1.86)	0 (0.93)	0 (1.70)	0 (0.87)
	Likelihood (SU,WC)	0 (3.12)	3 (2.06)	-1 (1.75)	1 (0.86)
	Regression standardization	0 (1.49)	-8 (0.71)	2 (0.66)	2 (0.50)
Slope	Likelihood (SO,NC)	0 (1.02)	2 (1.61)	0 (1.08)	-1 (1.31)
	Likelihood (SU,WC)	0 (2.51)	1 (2.59)	0 (1.09)	-2 (1.37)
	Regression standardization	0 (0.96)	-1 (2.16)	0 (0.91)	1 (1.13)

of the random sample estimator. As seen in simulations, the regression standardization methods which used a slope-based design seemed to be more affected by dropout than the likelihood-based methods in terms of efficiency; under dropout, targeted parameters had efficiency similar to a random sample for these designs.

4.9 Discussion

As researchers leveraging existing data sources, it is nearly inevitable that we should encounter data that may contain unintentional missingness. In this chapter we have offered two approaches to handle unanticipated missingness due to dropout and not induced deliberately by the ODS design. First, we adapted the best-performing likelihood-based methods from Chapter 2 for use with data under dropout. Second, we borrowed an idea from the

Table 4.9: Induced dropout time of Cystic Fibrosis Foundation Registry cohort, by marker type. Overall, 597 (19%) patients dropped out early.

Dropout Time	*S. aureus* absent	*S. aureus* present
$t = 2$	24	145
$t = 3$	20	132
$t = 4$	33	117
$t = 5$	27	99
$t = 6$	831	1713

causal inference/epidemiological literature, which we term "regression standardization", to first condition upon and then integrate over the dropout time. Finally, since the distribution of the subsampling variable is likely impacted by dropout, we offered some practical suggestions for adapting ODS designs under dropout.

Although motivated by the practical issues of accommodating unanticipated missing data within the ODS sphere, the regression standardization approach explored here had the dual advantage of incorporating information from unsubsampled individuals. Through iteration over the subsampling variable Q, we observed similar trends in efficiency gains by this technique as were seen in Chapters 2 and 3, where using the information from unsubsampled individuals was considered for likelihood-based and robust estimators. The regression standardization method can also be used with complete data, and as a moment-based estimator shows promise as a middle ground between the approaches of Chapters 2 and 3, both in terms of efficiency and robustness to model misspecification.

Table 4.10: Parameter estimates and empirical standard errors of likelihood-based and regression standardization estimators for the Cystic Fibrosis Foundation Registry dataset (N = 3,141). For "full" estimator, standard errors are derived from analysis of full Cystic Fibrosis cohort. All other estimators are based on an average subsample of 600 patients, and results are averaged over 1000 resamplings. Empirical standard errors are the estimator's standard deviation over all resamplings.

	Estimator	β_0		β_T		β_M		$\beta_{M\times T}$	
Design	Analysis method	Est. (SE)	p	Est. (SE)	p	Est. (SE)	p	Est. (SE)	p
Full cohort	Likelihood	1.200 (0.012)	< 0.001	0.184 (0.0029)	< 0.001	-0.013 (0.015)	0.79	-0.012 (0.0035)	0.03
Random sample	Likelihood	1.200 (0.0248)	< 0.001	0.184 (0.0063)	< 0.001	-0.012 (0.0305)	0.70	-0.011 (0.0074)	0.12
Intercept	Likelihood (SO,NC)	1.214 (0.0186)	< 0.001	0.181 (0.0057)	< 0.001	-0.016 (0.0235)	0.50	-0.012 (0.0069)	0.07
	Likelihood (SU,WC)	1.202 (0.0143)	< 0.001	0.184 (0.0044)	< 0.001	-0.014 (0.0206)	0.50	-0.011 (0.0063)	0.07
	Regression standardization	1.201 (0.0137)	< 0.001	0.181 (0.0057)	< 0.001	-0.013 (0.0197)	0.48	-0.012 (0.0069)	0.08
Slope	Likelihood (SO,NC)	1.192 (0.0244)	< 0.001	0.185 (0.0042)	< 0.001	0.011 (0.0296)	0.73	-0.014 (0.0049)	0.005
	Likelihood (SU,WC)	1.186 (0.0175)	< 0.001	0.185 (0.0033)	< 0.001	0.008 (0.0256)	0.72	-0.013 (0.0047)	0.005
	Regression standardization	1.192 (0.0235)	< 0.001	0.185 (0.0032)	< 0.001	0.010 (0.0296)	0.75	-0.013 (0.0046)	0.006

Table 4.11: Parameter estimates and empirical standard errors of likelihood-based and regression standardization estimators for the Cystic Fibrosis Foundation Registry dataset (N = 3,141), under induced dropout. For "full" estimator, standard errors are derived from analysis of full Cystic Fibrosis cohort. All other estimators are based on an average subsample of 600 patients, and results are averaged over 1000 resamplings. Empirical standard errors are the estimator's standard deviation over all resamplings.

	Estimator	β_0		β_T		β_M		$\beta_{M\times T}$	
Design	Analysis method	Est. (SE)	p	Est. (SE)	p	Est. (SE)	p	Est. (SE)	p
Full cohort	Likelihood	1.203 (0.0124)	< 0.001	0.0183 (0.0030)	< 0.001	-0.012 (0.0148)	0.42	-0.012 (0.0036)	0.001
Random sample	Likelihood	1.202 (0.0248)	< 0.001	0.184 (0.0062)	< 0.001	-0.011 (0.0298)	0.72	-0.012 (0.0075)	0.11
Intercept	Likelihood (SO,NC)	1.216 (0.0195)	< 0.001	0.181 (0.0059)	< 0.001	-0.014 (0.0237)	0.66	-0.013 (0.0069)	0.08
	Likelihood (SU,WC)	1.204 (0.0143)	< 0.001	0.183 (0.0044)	< 0.001	-0.013 (0.0205)	0.55	-0.011 (0.0063)	0.06
	Regression standardization	1.202 (0.0140)	< 0.001	0.182 (0.0064)	< 0.001	-0.009 (0.0205)	0.53	-0.014 (0.0082)	0.07
Slope	Likelihood (SO,NC)	1.200 (0.0244)	< 0.001	0.185 (0.0043)	< 0.001	0.005 (0.0301)	0.89	-0.014 (0.0052)	0.07
	Likelihood (SU,WC)	1.192 (0.0165)	< 0.001	0.184 (0.0034)	< 0.001	0.003 (0.0239)	0.87	-0.013 (0.0050)	0.006
	Regression standardization	1.199 (0.0239)	< 0.001	0.184 (0.0057)	< 0.001	0.004 (0.0305)	0.90	-0.014 (0.0075)	0.008

Chapter 5

CONCLUDING REMARKS AND FUTURE WORK

5.1 Conclusion

Innovation in the face of scarcity is a fact of the modern researcher's life. Both government and other funding agencies continue to encourage research that can leverage available information from existing studies and eliminate redundancies in study administration. The evaluation of biomarkers that may employ novel and expensive technologies under sometimes austere financial conditions is a particularly relevant scenario. Acknowledging this reality, the retrospective and biased ascertainment of biomarkers and/or other selectively subsampled covariates offers both potential benefit and added complexity. This dissertation has explored methods for the design and analysis of ODS biomarker substudies of longitudinal data; both aspects appear to be worthy of careful consideration. We have shown that outcome-dependent sampling designs based on low-dimensional summaries of longitudinal data may improve inference for targeted parmeters of interest, and have demonstrated the validity and potential efficiency gains for a variety of analytical approaches.

In terms of the analysis of ODS designs, we explored three broad classes of estimators: likelihood-based, robust, and standardization-based. In Chapter 2 we explored several likelihood-based estimators, whose performance relative to a random sampling design varied by the information incorporated into the likelihood. Analyses which included information from individuals in whom the expensive covariate was not ascertained produced more efficient inference for selected parameters; maximizing a joint instead of a conditional likelihood in conjunction with the analysis of all individuals likewise improved efficiency.

In Chapter 3, we investigated several methods of estimation that are robust to model

misspecification. A classic approach, inverse probability weighting (IPW), provided protection again misspecification, but at the cost of reduced efficiency for regression parameters not targeted by the ODS design. The augmented inverse probability weighting (AIPW) estimator took a similar weighting strategy but utilized information from unsubsampled individuals, producing efficiency gains for parameters related to that information. A calibration-based approach captured a significant fraction of the efficiency gains seen by AIPW without the computational complexity, and offered an easily implemented and robust alternative to likelihood-based methods.

Finally, we explored a regression standardization estimation method that used iteration over the subsampling variable, together with integration (through a simple application of Bayes theorem) over the same subsampling variable, to incorporate the information from unsubsampled individuals in yet a third way. Efficiency gains seen with this method were intermediate to those realized by likelihood-based and those from robust methods, while regression standardization methods appeared to inherit some robustness as a moment-based method. Like likelihood-based methods, regression standardization methods were able to be adapted for use with ODS data that had unanticipated missingness due to dropout, a significant advantage when considering the practical issues of implementation for these methods.

5.2 Future work

In this dissertation, we have explored the landscape of possible analysis strategies for analyzing outcome dependent sampling designs for longitudinal data. We have proposed three broad classes of estimators in this space that offer analysis options for researchers who may feel more or less comfortable with the assumptions required by each method. The bias-variance tradeoff observed in these methods likewise emphasizes the importance of diagnostics in this space, in which data that is truly multivariate normally distributed may benefit from the

most efficient likelihood analyses.

Even the most efficient of analyses, however, will have limited use and impact if they are difficult to implement; thus, this area requires further work to translate, explain, and make implementation of these ideas easy for the typical reseacher (with no special expertise in this area). Some of the methods explored here, such as the calibration estimator and the regression standardization estimator, could likely be implemented without special software; for others, and especially for likelihood-based methods, software such as an R package is of critical importance and will be realized in the near future.

While this dissertation has focused on the analysis of ODS designs for longitudinal data, the importance of ODS design cannot be overstated. Most of this thesis has taken the ODS design parameters to be static, but it may well be that the most appropriate analysis depends on the particularly implemented design parameters; a brief exploration into the effect of design parameters on the efficiency of the IPW estimator intimated as much. Likewise, there may exist auxiliary information related to the marker of interest that could be taken into consideration in the design stage in addition to the outcome itself. A more thorough exploration of the risks and benefits associated with various ODS designs, and practical suggestions for implementation thereof, will be important for the widespread implementation of these methods.

Finally, we explored here the effect of using an ODS design for a continuous longitudinal outcome to selectively subsample a single binary marker. Some studies, however, may have multiple outcomes that are of interest and may wish to utilize more than one outcome to choose the subsample; adaptation of these methods for use with e.g. longitudinal count data may also be of interest. Conversely, the researcher may be interested in the effect of a time-varying marker on a single longitudinal outcome; the decision about which individuals to subsample and at which time points to gain maximal information will likely require additional extensions to the current research. Finally, we have examined only main and interaction

effects of marker and time. Almost certainly, any real-life problem would need to consider the presence of confounders, which would complicate the implementation of some of the analysis methods offered here.

In conclusion, the ODS design and analysis methods for longitudinal data investigated here offer the potential for sometimes substantial gains in efficiency, which can translate into meaningful savings in terms of research dollars. At the expense of added complexity at both design and analysis stages, these methods provide investigators the opportunity to leverage existing data sources to selectively ascertain a subsample, and the promise of answering meaningful scientific questions at reduced cost.

BIBLIOGRAPHY

[1] Norman E Breslow, Thomas Lumley, Christie M Ballantyne, Lloyd E Chambless, and Michal Kulich. Improved horvitz–thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statistics in biosciences*, 1(1):32–49, 2009.

[2] Norman E Breslow, Thomas Lumley, Christie M Ballantyne, Lloyd E Chambless, and Michal Kulich. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*, 169(11):1398–1405, 2009.

[3] Norman E Breslow, James M Robins, and Jon A Wellner. On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, pages 447–455, 2000.

[4] Nilanjan Chatterjee, Yi-Hau Chen, and Norman E Breslow. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168, 2003.

[5] Hua Yun Chen. Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association*, 99(468):1176–1189, 2004.

[6] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[7] Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

[8] Jean-Claude Deville, Carl-Erik Särndal, and Olivier Sautory. Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423):1013–1020, 1993.

[9] Valerii Fedorov, Frank Mannino, and Rongmei Zhang. Consequences of dichotomization. *Pharmaceutical Statistics*, 8(1):50–61, 2009.

[10] Garrett M Fitzmaurice and Nan M Laird. Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics*, 1(2):141–156, 2000.

[11] Youyi Fong and Peter Gilbert. Calibration weighted estimation of semiparametric transformation models for two-phase sampling. *Statistics in medicine*, 34(10):1695–1707, 2015.

[12] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[13] Niels Keiding. The method of expected number of deaths, 1786-1886-1986, correspondent paper. *International Statistical Review/Revue Internationale de Statistique*, pages 1–20, 1987.

[14] Brenda F Kurland and Patrick J Heagerty. Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, 6(2):241–258, 2005.

[15] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[16] JF Lawless, JD Kalbfleisch, and CJ Wild. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):413–438, 1999.

[17] Stuart R Lipsitz, Joseph G Ibrahim, and Lue Ping Zhao. A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94(448):1147–1160, 1999.

[18] Roderick J Little, Ralph D'Agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, Joseph W Hogan, Geert Molenberghs, Susan A Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

[19] Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

[20] Roderick JA Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483, 1994.

[21] Thomas Lumley. Robustness of semiparametric efficiency in nearly-correct models for two-phase samples. 2009.

[22] Thomas Lumley, Pamela A Shaw, and James Y Dai. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79(2):200–220, 2011.

[23] Kelly Moore, Romain Neugebauer, Fred Lurmann, Jane Hall, Vic Brajer, Sianna Alcorn, and Ira Tager. Ambient ozone concentrations cause increased hospitalizations for asthma in children: an 18-year study in southern california. *Environ Health Perspect*, 116(8):1063–1070, 2008.

[24] FGP Neison. On a method recently proposed for conducting inquiries into the comparative sanatory condition of various districts, with illustrations, derived from numerous places in great britain at the period of the last census. *Journal of the Statistical Society of London*, 7(1):40–68, 1844.

[25] J Neuhaus, AJ Scott, and CJ Wild. The analysis of retrospective family studies. *Biometrika*, 89(1):23–37, 2002.

[26] JM Neuhaus, AJ Scott, and CJ Wild. Family-specific approaches to the analysis of case-control family data. *Biometrics*, 62(2):488–494, 2006.

[27] Maya L Petersen, Yue Wang, Mark J Van Der Laan, David Guzman, Elise Riley, and David R Bangsberg. Pillbox organizers are associated with improved adherence to hiv antiretroviral therapy and viral suppression: a marginal structural model analysis. *Clinical Infectious Diseases*, 45(7):908–915, 2007.

[28] Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

[29] David R Ragland. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology*, 3(5):434–440, 1992.

[30] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986.

[31] James M Robins and Andrea Rotnitkzy. Discussion of: Firth, d. robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):51–52, 1998.

[32] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[33] Sherri Rose and Mark J van der Laan. A targeted maximum likelihood estimator for two-stage designs. *The international journal of biostatistics*, 7(1):1–21, 2011.

[34] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

[36] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

[37] Tosiya Sato and Yutaka Matsuyama. Marginal structural models as a tool for standardization. *Epidemiology*, 14(6):680–686, 2003.

[38] Jonathan S Schildcrout, , Paul J Rathouz, Leila R Zelnick, Shawn P Garbett, and Patrick J Heagerty. Biased sampling designs to improve research efficiency: factors influencing pulmonary function over time in children with asthma. *Annals of Applied Statistics*, 2015.

[39] Jonathan S Schildcrout, Shawn P Garbett, and Patrick J Heagerty. Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics. *Biometrics*, 2013.

[40] Jonathan S Schildcrout and Patrick J Heagerty. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*, 9(4):735–749, 2008.

[41] Jonathan M Snowden, Sherri Rose, and Kathleen M Mortimer. Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, page kwq472, 2011.

[42] Samy Suissa and Lucie Blais. Binary regression with continuous outcomes. *Statistics in medicine*, 14(3):247–255, 1995.

[43] Sarah L Taubman, James M Robins, Murray A Mittleman, and Miguel A Hernán. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology*, 38(6):1599–1611, 2009.

[44] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[45] Stijn Vansteelandt and Niels Keiding. Invited commentary: G-computation–lost in translation? *American journal of epidemiology*, page kwq474, 2011.

[46] Mark A Weaver and Haibo Zhou. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*, 100(470):459–469, 2005.

[47] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980.

[48] Haibo Zhou, Jianwei Chen, Tiina H Rissanen, Susan A Korrick, Howard Hu, Jukka T Salonen, and Matthew P Longnecker. Outcome-dependent sampling: An efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*, 18(4):461–468, 2007.

[49] Haibo Zhou, MA Weaver, J Qin, MP Longnecker, and MC Wang. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58(2):413–421, 2002.

[50] Haibo Zhou, Wangli Xu, Donglin Zeng, and Jianwen Cai. Semiparametric inference for data with a continuous outcome from a two-phase probability-dependent sampling scheme. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):197–215, 2014.

Appendix A

LIKELIHOOD-BASED ESTIMATOR USING SUBSAMPLED SUBJECTS ONLY AND NO COVARIATES

The likelihood-based estimator SO,NC analyzes subsampled individuals only and does not include covariate information, and is based on the conditional distribution of $[\mathbf{Y}|\mathbf{X}, S = 1]$, with density:

$$\begin{aligned} f(\mathbf{Y}|\mathbf{X}, S = 1; \boldsymbol{\theta}) &= \frac{f(\mathbf{Y}, \mathbf{X}, S = 1; \boldsymbol{\theta})}{f(\mathbf{X}, S = 1; \boldsymbol{\theta})} \\ &= \frac{P(S = 1|\mathbf{Y}, \mathbf{X}) \cdot f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\theta}) \cdot f(\mathbf{X})}{P(S = 1|\mathbf{X}; \boldsymbol{\theta}) \cdot f(\mathbf{X})} \\ &= \frac{\omega(q_i) \cdot f(Y|\mathbf{X}; \boldsymbol{\theta})}{P(S = 1|\mathbf{X}; \boldsymbol{\theta})} \\ &\propto \frac{f(Y|\mathbf{X}; \boldsymbol{\theta})}{AC_0(m, \boldsymbol{t}; \boldsymbol{\theta})} \end{aligned}$$

The conditional sampling probability $\omega(q) = P(S = 1|\mathbf{Y}, \mathbf{X}) = P(S = 1|q \in R)$ is assumed to be a previously chosen constant that does not depend on $\boldsymbol{\theta}$. Therefore, the conditional likelihood based on the distribution of $Y|\mathbf{X}, S = 1$ is proportional to the usual likelihood for sampled individuals under random sampling based on the distribution of $Y|\mathbf{X}$, scaled by an "ascertainment correction" factor $AC_0(m, \boldsymbol{t}) \equiv P(S = 1|M, \boldsymbol{T})$. Furthermore, this ascertainment correction can be written as:

$$\begin{aligned} AC_0(m, \boldsymbol{t}; \boldsymbol{\theta}) &= P(S = 1|M, \boldsymbol{T}; \boldsymbol{\theta}) \\ &= \sum_{k=1}^{3} P(S = 1, q \in R_k|M, \boldsymbol{T}; \boldsymbol{\theta}) \end{aligned}$$

$$
\begin{aligned}
= \quad & P(S=1|q \in R_1, M, \boldsymbol{T}) \cdot P(q \in R_1|M, \boldsymbol{T}; \boldsymbol{\theta}) \\
& \quad + P(S=1|q \in R_2, M, \boldsymbol{T}) \cdot P(q \in R_2|M, \boldsymbol{T}; \boldsymbol{\theta}) \\
& \quad + P(S=1|q \in R_3, M, \boldsymbol{T}) \cdot P(q \in R_3|M, \boldsymbol{T}; \boldsymbol{\theta}) \\
= \quad & \omega_1 \int_{R_1} f(q|m, \boldsymbol{t})\, dq + \omega_2 \int_{R_2} f(q|m, \boldsymbol{t})\, dq + \omega_3 \int_{R_3} f(q|m, \boldsymbol{t})\, dq \\
= \quad & \sum_{k=1}^{3} \omega_k \int_{R_k} f(q|m, \boldsymbol{t})\, dq
\end{aligned}
$$

Appendix B

LIKELIHOOD-BASED ESTIMATOR USING SUBSAMPLED SUBJECTS ONLY AND WITH COVARIATES

The likelihood SO,WC (Equation 2.5) differs from that of likelihood SO,NC (Equation 2.4) by the second term in Equation 2.6. Define $a_0 = AC_0(M = 0, \boldsymbol{t})$ and $a_1 = AC_0(M = 1, \boldsymbol{t})$. Under a complete and balanced design, $\mathbb{E}_{M,\boldsymbol{T}}[AC_0(m, \boldsymbol{t})] = \mathbb{E}_M[AC_0(m, \boldsymbol{t})]$, and the ith person's ascertainment correction contribution to the likelihood can be written as follows:

$$\begin{aligned}
\frac{AC_0(m,\boldsymbol{t}) \cdot P(M = m|\boldsymbol{T})}{\mathbb{E}_{M,\boldsymbol{T}}[AC_0(m,\boldsymbol{T})]} &= \frac{AC_0(m,\boldsymbol{t}) \cdot P(M = m|\boldsymbol{T})}{\mathbb{E}_M[AC_0(m,\boldsymbol{t})]} \\
&= \frac{a_1^m a_0^{1-m} \cdot p^m (1-p)^{1-m}}{pa_1 + (1-p)a_0} \\
&= \frac{(a_1 p)^m \left(a_0(1-p)\right)^{1-m}}{pa_1 + (1-p)a_0} \\
&= \frac{\left(\frac{a_1}{a_0}p\right)^m (1-p)^{1-m}}{\frac{a_1}{a_0}p + 1 - p} \\
&= \frac{\frac{a_0}{a_1 p}}{\frac{a_0}{a_1 p}} \cdot \frac{\left(\frac{a_1}{a_0}p\right)^m (1-p)^{1-m}}{\frac{a_1}{a_0}p + 1 - p} \\
&= \left(\frac{1}{1 + \frac{a_0}{a_1} \cdot \frac{1-p}{p}}\right)^m \left(\frac{\frac{a_0}{a_1} \cdot \frac{1-p}{p}}{1 + \frac{a_0}{a_1} \cdot \frac{1-p}{p}}\right)^{1-m} \\
&= \xi^m (1-\xi)^{1-m},
\end{aligned}$$

where $\xi = \left(\frac{1}{1 + \frac{a_0}{a_1} \cdot \frac{1-p}{p}}\right)$. For balanced and complete designs, a_0 and a_1 do not vary across subjects, so this term's contribution to likelihood SO,WC (Equation 2.5) can be seen as a

reparameterization of the marginal distribution of M, which contains no information about $\boldsymbol{\theta}$. Thus, adding this term to the likelihood SO, NC (Equation 2.4) will add no information to the resulting inference, and in fact will yield the same estimate, up to the chosen tolerance of the Newton-Raphson algorithm.

Appendix C

JUSTIFICATION OF THE MARKER IMPUTATION MODEL FOR AUGMENTED INVERSE PROBABILITY WEIGHTED AND CALIBRATION ESTIMATORS

A logistic imputation model was used to predict marker value based on outcome $\mathbf{Y}$ and time $\boldsymbol{T}$ in both the AIPW and calibration estimators of Chapter 3. Since the marker is binary, logistic regression is a sensible model to use in any case; however, the true form of the model may be unclear. We have modeled the log odds of marker type as a linear function of the longitudinal outcome; when under the usual longitudinal linear mixed model $[\mathbf{Y}|M, \boldsymbol{T}]$ is multivariate normally distributed, this is the true relationship, assuming that $[M|\boldsymbol{T}] = [M]$, or that marker is independent of observation times. Abbreviating $P(\mathbf{Y}|M = 1, \boldsymbol{T}) = \mathcal{L}_1$ and $P(\mathbf{Y}|M = 0, \boldsymbol{T}) = \mathcal{L}_0$ halfway down, this can be seen by writing:

$$
\begin{aligned}
P(M = 1|\mathbf{Y}, \boldsymbol{T}) &= \frac{P(M = 1, \mathbf{Y}, \boldsymbol{T})}{P(\mathbf{Y}, \boldsymbol{T})} \\
&= \frac{P(\mathbf{Y}|M = 1, \boldsymbol{T}) \cdot P(M = 1|\boldsymbol{T}) \cdot P(\boldsymbol{T})}{P(\mathbf{Y}|\boldsymbol{T}) \cdot P(\boldsymbol{T})} \\
&= \frac{P(\mathbf{Y}|M = 1, \boldsymbol{T}) \cdot P(M = 1)}{P(\mathbf{Y}|\boldsymbol{T})} \\
&= \frac{P(\mathbf{Y}|M = 1, \boldsymbol{T}) \cdot P(M = 1)}{P(\mathbf{Y}|M = 1, \boldsymbol{T}) \cdot P(M = 1) + P(\mathbf{Y}|M = 0, \boldsymbol{T}) \cdot P(M = 0)} \\
&= \frac{1}{1 + \dfrac{P(M = 0)}{P(M = 1)} \cdot \dfrac{\mathcal{L}_0}{\mathcal{L}_1}}
\end{aligned}
$$

Furthermore, assuming that $\mathbf{Y}|M, \boldsymbol{T} \sim MVN\left(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)$, we can write

$$
\begin{aligned}
\frac{P(M=0)}{P(M=1)} \cdot \frac{\mathcal{L}_0}{\mathcal{L}_1} &= \frac{P(M=0)}{P(M=1)} \cdot \exp\left\{ -\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu}_0) + \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu}_1) \right\} \\
&= \frac{P(M=0)}{P(M=1)} \cdot \exp\left\{ -\frac{1}{2}\mathbf{Y}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y} + \mathbf{Y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + \frac{1}{2}\mathbf{Y}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y} \right. \\
&\qquad \left. -\mathbf{Y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 \right\} \\
&= \exp\left\{ -\left(\ln P(M=1) - \ln(1-P(M=1)) + \mathbf{Y}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 \right. \right. \\
&\qquad \left. \left. -\frac{1}{2}\boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 \right) \right\} \\
&= \exp\left\{ \ln(1-P(M=1)) - \ln P(M=1) + \frac{1}{2}\left((2\mathbf{Y}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - (2\mathbf{Y}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 \right. \right. \\
&= \exp\left\{ -\mathbf{Y}\cdot\boldsymbol{\beta} \right\},
\end{aligned}
$$

where $\boldsymbol{\beta} = \{\beta_0, \beta_1\}^T$ and $\beta_0 = \ln\left(P(M=1)/(1-P(M=1))\right) + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1$ and $\beta_1 = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$.

Thus, under normality of $\mathbf{Y}|M, \boldsymbol{T}$, the log odds of a binary marker truly are an additive linear function of the longitudinal outcome. Note that if $\mathbf{Y}$ comes from the distribution where $M = 1$, $\frac{P(M=0)}{P(M=1)} \cdot \frac{\mathcal{L}_0}{\mathcal{L}_1}$ will be small, and hence $P(M = 1|\mathbf{Y}_i, \boldsymbol{T})$ will be large. Conversely, if $\mathbf{Y}$ comes from the distribution where $M = 0$, $\frac{P(M=0)}{P(M=1)} \cdot \frac{\mathcal{L}_0}{\mathcal{L}_1}$ will be large, and hence $P(M = 1|\mathbf{Y}_i, \boldsymbol{T})$ will be small.

Appendix D

DERIVATION OF THE DOUBLE ROBUST PROPERTY OF THE AUGMENTED INVERSE PROBABILITY WEIGHTED ESTIMATOR

Following the example of Tsiatis [44], we present here a derivation of the double robust property of the AIPW estimator considered in Chapter 3. Recall that the AIPW estimator $\hat{\theta}$ is that which solves the weighted estimating equation

$$\sum_{i=1}^{n} \frac{S_i}{\pi(\mathbf{Y}_i, \boldsymbol{T}_i, \alpha)} U_i(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i) + \sum_{i=1}^{n} \left\{ 1 - \frac{S_i}{\pi(\mathbf{Y}_i, \boldsymbol{T}_i, \alpha)} \right\} \mathbb{E}\{U_i(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}) | \mathbf{Y}_i, \boldsymbol{T}_i\} = 0 \quad \text{(D.1)}$$

where S_i is the indicator of subsampling, $\pi(\mathbf{Y}_i, \boldsymbol{T}_i, \alpha) = P(S_i = 1 | \mathbf{Y}_i, \boldsymbol{T}_i, \alpha)$, and $U_i(\boldsymbol{\theta})$ an estimating function for which $\mathbb{E}[U(\boldsymbol{\theta})] = 0$. The "double robustness" of the estimator means that $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$ if either the probability of missingness π_i or the conditional covariate distribution $g(M|\boldsymbol{T})$ is misspecified. We will show that the estimator

$$\frac{1}{n} \sum_{i=1}^{n} \frac{S_i}{\pi(\mathbf{Y}_i, \boldsymbol{T}_i, \hat{\alpha})} U_i(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, \mathbf{X}_i) + \sum_{i=1}^{n} \left\{ 1 - \frac{S_i}{\pi(\mathbf{Y}_i, \boldsymbol{T}_i, \hat{\alpha})} \right\} \mathbb{E}\{U_i(\hat{\boldsymbol{\theta}}; \mathbf{Y}, \mathbf{X}) | \mathbf{Y}_i, \boldsymbol{T}_i\}$$

has expectation zero under either of these cases.

Case 1: $\pi(\mathbf{Y}, \boldsymbol{T}, \alpha)$ misspecified, $g(M|\boldsymbol{T}; \boldsymbol{\theta})$ correctly specified

If $g(M|\boldsymbol{T}; \boldsymbol{\theta})$ is correctly specified, then

$$\mathbb{E}(U(\hat{\boldsymbol{\theta}}; \mathbf{Y}, \mathbf{X}) | \mathbf{Y}, \boldsymbol{T})) \xrightarrow{p} \mathbb{E}\left(U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}) | \mathbf{Y}, \boldsymbol{T}\right)$$

Conversely, the misspecification of π implies that

$$\pi(\mathbf{Y}, \boldsymbol{T}, \hat{\alpha}) \xrightarrow{p} \pi(\mathbf{Y}, \boldsymbol{T}, \alpha^*) \neq P(S = 1|\mathbf{Y}, \boldsymbol{T})$$

Then we can write the estimator as

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left[U_i(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, \mathbf{X}_i) + \frac{S_i - \pi(\mathbf{Y}_i, \boldsymbol{T}_i, \hat{\alpha})}{\pi(\mathbf{Y}_i, \boldsymbol{T}_i, \hat{\alpha})}\left(U_i(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, \mathbf{X}_i) - \mathbb{E}(U_i(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, \mathbf{X}_i)|\mathbf{Y}_i, \boldsymbol{T}_i)\right)\right] \\
= & \frac{1}{n}\sum_{i=1}^{n}\left[U_i(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i) + \frac{S_i - \pi(\mathbf{Y}_i, \boldsymbol{T}_i, \alpha^*)}{\pi(\mathbf{Y}_i, \boldsymbol{T}_i, \alpha^*)}\left(U_i(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i) - \mathbb{E}(U_i(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i)|\mathbf{Y}_i, \boldsymbol{T}_i)\right)\right] + o_p(1) \\
\xrightarrow{p} & \mathbb{E}\left[U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}) + \frac{S - \pi(\mathbf{Y}, \boldsymbol{T}, \alpha^*)}{\pi(\mathbf{Y}, \boldsymbol{T}, \alpha^*)}\left(U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}) - \mathbb{E}(U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})|\mathbf{Y}, \boldsymbol{T})\right)\right] \\
= & \mathbb{E}[U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})] + \mathbb{E}\left(\mathbb{E}\left[\frac{S - \pi(\mathbf{Y}, \boldsymbol{T}, \alpha^*)}{\pi(\mathbf{Y}, \boldsymbol{T}, \alpha^*)}\left\{U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}) - \mathbb{E}(U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})|\mathbf{Y}, \boldsymbol{T})\right\}\middle| S, \mathbf{Y}, \boldsymbol{T}\right]\right) \\
= & \mathbb{E}[U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})] + \mathbb{E}\left(\frac{S - \pi(\mathbf{Y}, \boldsymbol{T}, \alpha^*)}{\pi(\mathbf{Y}, \boldsymbol{T}, \alpha^*)}\left\{\mathbb{E}(U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})|S, \mathbf{Y}, \boldsymbol{T}) - \mathbb{E}(U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})|\mathbf{Y}, \boldsymbol{T})\right\}\right) \\
= & \mathbb{E}[U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})] + \mathbb{E}\left(\frac{S - \pi(\mathbf{Y}, \boldsymbol{T}, \alpha^*)}{\pi(\mathbf{Y}, \boldsymbol{T}, \alpha^*)}\left\{\mathbb{E}(U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})|\mathbf{Y}, \boldsymbol{T}) - \mathbb{E}(U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})|\mathbf{Y}, \boldsymbol{T})\right\}\right) \qquad \text{(D.2)} \\
= & \mathbb{E}[U(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X})] \\
= & 0
\end{aligned}$$

where line D.2 is true since $[M|\mathbf{Y}, \boldsymbol{T}, S = 1] = [M|\mathbf{Y}, \boldsymbol{T}, S = 0] = [M|\mathbf{Y}, \boldsymbol{T}]$. The left hand side of Equation D.1 has expectation zero; this implies that $\boldsymbol{\theta}$ is consistent for $\boldsymbol{\theta}$ when π is misspecified.

Case 2: $g(M|\boldsymbol{T})$ misspecified, $\pi(\mathbf{Y}, \boldsymbol{T}, \alpha)$ correctly specified

Similarly, when $g(M|\boldsymbol{T})$ is misspecified, this will cause

$$\mathbb{E}[U(\hat{\boldsymbol{\theta}})|\mathbf{Y}, \boldsymbol{T}] \xrightarrow{p} \mathbb{E}[U^*(\boldsymbol{\theta})|\mathbf{Y}, \boldsymbol{T}] \neq \mathbb{E}[U(\boldsymbol{\theta})|\mathbf{Y}, \boldsymbol{T}]$$

where $\mathbb{E}[U^*(\boldsymbol{\theta})|\mathbf{Y},\boldsymbol{T}]$ denotes that the expectation is taken over the misspecified distribution $g(M|\boldsymbol{T})$. However, so long as $\pi(\mathbf{Y},\boldsymbol{T})$ is correctly specified and $\mathbb{E}[U(\boldsymbol{\theta})] = 0$, then

$$\pi(\mathbf{Y},\boldsymbol{T},\hat{\alpha}) \xrightarrow{p} \pi(\mathbf{Y},\boldsymbol{T},\alpha),$$

and, as in Case 1, we can write the left hand side of Equation D.1 as

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left[U_i(\hat{\boldsymbol{\theta}};\mathbf{Y}_i,\mathbf{X}_i) + \frac{S_i - \pi(\mathbf{Y}_i,\boldsymbol{T}_i,\hat{\alpha})}{\pi(\mathbf{Y}_i,\boldsymbol{T}_i,\hat{\alpha})}\left(U_i(\hat{\boldsymbol{\theta}};\mathbf{Y}_i,\mathbf{X}_i) - \mathbb{E}(U_i(\hat{\boldsymbol{\theta}};\mathbf{Y}_i,\mathbf{X}_i)|\mathbf{Y}_i,\boldsymbol{T}_i)\right)\right] \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\left[U_i(\boldsymbol{\theta};\mathbf{Y}_i,\mathbf{X}_i) + \frac{S_i - \pi(\mathbf{Y}_i,\boldsymbol{T}_i,\alpha)}{\pi(\mathbf{Y}_i,\boldsymbol{T}_i,\alpha)}\left(U_i(\boldsymbol{\theta};\mathbf{Y}_i,\mathbf{X}_i) - \mathbb{E}(U_i^*(\boldsymbol{\theta};\mathbf{Y}_i,\mathbf{X}_i)|\mathbf{Y}_i,\boldsymbol{T}_i)\right)\right] + o_p(1) \\
\xrightarrow{p} \; & \mathbb{E}\left[U(\boldsymbol{\theta};\mathbf{Y},\mathbf{X}) + \frac{S_i - \pi(\mathbf{Y}_i,\boldsymbol{T}_i,\alpha)}{\pi(\mathbf{Y}_i,\boldsymbol{T}_i,\alpha)}\left(U(\boldsymbol{\theta};\mathbf{Y},\mathbf{X}) - \mathbb{E}(U^*(\boldsymbol{\theta};\mathbf{Y},\mathbf{X})|\mathbf{Y},\boldsymbol{T})\right)\right] \\
= \; & \mathbb{E}[U(\boldsymbol{\theta};\mathbf{Y},\mathbf{X})] + \mathbb{E}\left(\mathbb{E}\left[\frac{S - \pi(\mathbf{Y},\boldsymbol{T},\alpha)}{\pi(\mathbf{Y},\boldsymbol{T},\alpha)}\left\{U(\boldsymbol{\theta};\mathbf{Y},\mathbf{X}) - \mathbb{E}(U^*(\boldsymbol{\theta};\mathbf{Y},\mathbf{X})|\mathbf{Y},\boldsymbol{T})\right\}\middle| U,\mathbf{Y},\boldsymbol{T}\right]\right) \\
= \; & \mathbb{E}[U(\boldsymbol{\theta};\mathbf{Y},\mathbf{X})] + \mathbb{E}\left(\frac{\mathbb{E}[S|U,\mathbf{Y},\boldsymbol{T}] - \pi(\mathbf{Y},\boldsymbol{T},\alpha)}{\pi(\mathbf{Y},\boldsymbol{T},\alpha)}\left\{U(\boldsymbol{\theta};\mathbf{Y},\mathbf{X}) - \mathbb{E}(U^*(\boldsymbol{\theta};\mathbf{Y},\mathbf{X})|\mathbf{Y},\boldsymbol{T})\right\}\right) \\
= \; & \mathbb{E}[U(\boldsymbol{\theta};\mathbf{Y},\mathbf{X})] + \mathbb{E}\left(\frac{\pi(\mathbf{Y},\boldsymbol{T},\alpha) - \pi(\mathbf{Y},\boldsymbol{T},\alpha)}{\pi(\mathbf{Y},\boldsymbol{T},\alpha)}\left\{U(\boldsymbol{\theta};\mathbf{Y},\mathbf{X}) - \mathbb{E}(U^*(\boldsymbol{\theta};\mathbf{Y},\mathbf{X})|\mathbf{Y},\boldsymbol{T})\right\}\right) \\
= \; & \mathbb{E}[U(\boldsymbol{\theta};\mathbf{Y},\mathbf{X})] + 0 \\
= \; & 0
\end{aligned}$$

Thus, $\hat{\theta}$ is consistent for θ in this case as well.

Appendix E

DERIVATION OF ASYMPTOTICALLY VALID COVARIANCE MATRIX FOR THE REGRESSION STANDARDIZATION ESTIMATOR

Under complete data, the regression standardization approach to estimating $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$ iterates the expectation $\mathbb{E}[\mathbf{Y}|Q,\mathbf{X}]$ over the distribution of $Q|\mathbf{X}$. The fully conditional outcome model has

$$\mathbb{E}[\mathbf{Y}|\mathbf{X},Q] = \gamma_0 + \gamma_T T + \gamma_M M + \gamma_{M\times T} M\cdot T + \gamma_Q Q + \gamma_{Q\times T} Q\cdot T$$

Meanwhile, we use

$$\mathbb{E}[M|Q,\boldsymbol{T}] = \mathbb{E}[M|Q] = \text{expit}(\eta_0 + \eta_1 Q),$$

to derive $\mathbb{E}[Q|M] = \alpha_0 + \alpha_1 M$ and the empirical distribution $\mathbb{F}_Q$ as an estimate of Q's marginal distribution. Specifically, the estimates $\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\eta}},\mathbb{F}_Q)$ can be written as:

$$\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\eta}},\mathbb{F}_Q) = \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix} = \begin{bmatrix} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} n\cdot w_{0i}(\hat{\boldsymbol{\eta}},q_i)\cdot q_i \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} n\cdot (w_{1i}(\boldsymbol{\eta},q_i) - w_{0i}(\boldsymbol{\eta},q_i))\cdot q_i \end{bmatrix}$$

where

$$w_{0i}(\boldsymbol{\eta},\mathbb{F}_Q) = \frac{1-\text{expit}(\hat{\eta}_0+\hat{\eta}_1 q_i)}{\sum_j (1-\text{expit}(\hat{\eta}_0+\hat{\eta}_1 q_j)}$$

and

$$w_{1i}(\boldsymbol{\eta},\mathbb{F}_Q) = \frac{\text{expit}(\hat{\eta}_0+\hat{\eta}_1 q_i)}{\sum_j \text{expit}(\hat{\eta}_0+\hat{\eta}_1 q_j)}$$

We find the covariance of $\hat{\boldsymbol{\alpha}}$ using an iterated approach:

$$\text{Var}(\hat{\boldsymbol{\alpha}}) = \mathbb{E}_Q[\text{Var}_{\eta|Q}(\hat{\boldsymbol{\alpha}}|Q)] + \text{Var}_Q[\mathbb{E}_{\eta|Q}(\hat{\alpha}|Q)]$$

Then by the Central Limit Theorem we know that

$$\sqrt{n}\begin{bmatrix} \hat{\boldsymbol{\gamma}} & - & \boldsymbol{\gamma} \\ \hat{\boldsymbol{\alpha}} & - & \boldsymbol{\alpha} \end{bmatrix} \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma} \equiv \begin{pmatrix} \boldsymbol{\Sigma}_\gamma & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\alpha \end{pmatrix}\right)$$

We showed in (4.6) that the coefficients for the usual longitudinal mixed model of interest could be constructed from $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ in the following way:

$$\begin{aligned} \beta_0 &= \gamma_0 + \gamma_Q\alpha_0 \\ \beta_T &= \gamma_T + \gamma_{T\times Q}\alpha_0 \\ \beta_M &= \gamma_M + \gamma_Q\alpha_1 \\ \beta_{M\times T} &= \gamma_{M\times T} + \gamma_{T\times Q}\alpha_1 \end{aligned}$$

Then an application of the multivariate delta method yields the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ as

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, \nabla_\beta \boldsymbol{\Sigma} \nabla_\beta^T\right),$$

where

$$\nabla_\beta = \begin{bmatrix} 1 & 0 & 0 & \alpha_0 & 0 & 0 & \gamma_Q & 0 \\ 0 & 1 & 0 & 0 & 0 & \alpha_0 & \gamma_{T\times Q} & 0 \\ 0 & 0 & 1 & \alpha_1 & 0 & 0 & 0 & \gamma_Q \\ 0 & 0 & 0 & 0 & 1 & \alpha_1 & 0 & \gamma_{T\times Q} \end{bmatrix}$$

The resulting matrix $\text{Cov}(\hat{\boldsymbol{\beta}})$ can be used to obtain asymptotically valid standard errors.

A similar, if more complicated, process can be used to obtain standard errors for the

coefficients of interest when longitudinal data has unanticipated missingness. We omit details for this case here, although the process is similar and involves further applications of the multivariate delta method to obtain the desired parameters.

VITA

Leila Zelnick holds a BA in Mathematics from Williams College, a MA in Education from the University of Tulsa, and a MS in Statistics from Oklahoma State University. She served as a U.S. Peace Corps Volunteer in Bolivia from 2000 to 2002, and previously taught high school mathematics and statistics at her alma mater, Broken Arrow Senior High School in Broken Arrow, OK.

Ms. Zelnick has experience working as a biostatistician in applications as diverse as kidney and cardiovascular disease, emergency medicine, and oncology. After receiving her PhD from the University of Washington in 2015, she became Research Assistant Professor in the Department of Medicine at the University of Washington, doing collaborative research in kidney disease at the Kidney Research Institute in Seattle, WA.

She welcomes your comments to `lzelnick@uw.edu`.