A Finite Population Approach for Causal Inference in Nested Case-Control Studies

Katarina Boston Majetic

Master of Science

Department of Mathematics and Statistics

McGill University

Montréal, Québec, Canada

September 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

Acknowledgements

I would like to thank my supervisor Professor Russell Steele for recognizing my interest in statistics and taking me under his wing. With his guidance and support, I was able to solidify my interest in statistics and gain confidence.

I would also like to thank my family and friends, you know who you are, for your continual support.

Preface

Je tiens à remercier mon superviseur, le professeur Russell Steele, qui a reconnu mon intérêt pour les statistiques et m'a pris sous son aile. Grâce à ses conseils et à son soutien, j'ai pu consolider mon intérêt pour les statistiques et gagner en confiance.

J'aimerais également remercier ma famille et mes amis, vous savez qui vous êtes, pour votre soutien constant.

Abstract

The nested case-control design is employed by researchers when it is too difficult or expensive to collect and/or analyze data prospectively on rare outcomes. The sampling design is retrospective in nature but the conclusions are prospective in nature, which can lead to bias when analyzed inappropriately. Most nested case-control approaches employ logistic regression, however, in this retrospective analysis, a difficulty arises when one wants to employ causal inference methods to adjust to time-varying confounding. In this thesis, we introduce methods that allow us to use prospective causal inference methods with time-varying confounding, under a retrospective nested case-control sub-sampling scheme which requires a different approach to the classic nested case-control design. We interpret the entire cohort data set as a fixed finite population, thus, when we take our nested case-control sample, it will be viewed as a draw from the finite population. In order to account for causal effects, we use inverse probability (IP) treatment weighting on top of the sampling weights. Thus, we introduce methods to solve a nested case-control problem using finite population methods in a causal setting.

Résumé

Le plan cas-témoins emboîté est utilisé par les chercheurs lorsqu'il est trop difficile ou trop coûteux de collecter et/ou d'analyser prospectivement des données sur des résultats rares. Le plan de sondage est de nature rétrospective mais les conclusions sont de nature prospective, ce qui peut entraîner des biais en cas d'analyse inappropriée. La plupart des approches cas-témoins emboîtés utilisent la régression logistique, cependant, dans cette analyse rétrospective, une difficulté survient lorsque l'on veut utiliser des méthodes d'inférence causale pour s'ajuster aux facteurs de confusion variant dans le temps. Dans cette thèse, nous introduisons des méthodes qui nous permettent d'utiliser des méthodes d'inférence causale prospectives avec une confusion variant dans le temps, sous un schéma de sous-échantillonnage de cas-témoins emboîtés rétrospectif qui nécessite une approche différente du plan classique de cas-témoins emboîtés. Nous interprétons l'ensemble des données de la cohorte comme une population finie fixe. Ainsi, lorsque nous prélevons notre échantillon de cas-témoins emboîtés, il sera considéré comme un tirage de la population finie. Afin de tenir compte des effets causaux, nous utilisons une pondération de traitement à probabilité inverse en plus des poids de sondage. Ainsi, nous présentons des méthodes pour résoudre un problème de cas-témoins emboîtés en utilisant des méthodes de population finie dans un cadre causal.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The case-control design is often employed by researchers when it is too difficult to obtain the exposure of interest prospectively or when the outcome is rare. A nested case-control approach is a method that applies the case-control design to data that has already been collected. The design is retrospective in nature but is used to answer prospective research questions. This can lead to bias due to the need to rely on the odds ratio for inference due to the selection bias in the sampling. Using the odds ratio as a substitute for relative risk is not always appropriate and the risk difference cannot be estimated, see for example [23]. Also, it is not obvious how to estimate inverse probability weights to adjust for confounding due to the selection bias that results from the sampling design. Therefore, most researchers use regression based adjustments for confounding. In this thesis, we take a different approach to the classic nested case-control analysis. We interpret the data as a fixed finite population, thus, when we take our nested case-control sample, it will be viewed as a draw from the finite population. In order to estimate causal effects, we use inverse probability (IP) treatment weighting on top of the sampling weights.

This thesis introduces methods using finite population methods in a causal setting. This thesis uses similar ideas to those done of Morenz [16] and extends them to a causal setting. Traditional nested case-control methods employ logistic regression; however, in this retro-

spective analysis, a difficulty arises when one wants to employ causal inference methods to adjust for time-varying confounding as you cannot use regression based adjustments and you have to use inverse probability weighting, which then leads to the same selection bias problem as before. In this thesis, we introduce methods that allow us to do causal inference with time-varying confounding, using a nested case-control subsampling scheme.

We begin with a literature review in Chapter 2. This chapter provides a background on finite population sampling, generalized linear models and generalized estimating equations, nested case-control design studies and causal inference. This chapter is necessary in order to build a solid foundation of the concepts and theory that will be used in order to build our methods. We then introduce our methods for the cross-sectional case in Chapter 3. In this chapter, we can obtain results for a simple cross-sectional example in closed form. This enables us to see where finite population methods are being used in both our causal estimands and our treatment model. We then provide a simulation study in order to show how our estimators perform. The goal in this chapter is to rigorously show how our methods work both mathematically and in a simulation for the cross-sectional case. We then build on the methods introduced in Chapter 3 by extending them to the time-varying case in Chapter 4. In this chapter, we show how our methods in Chapter 3 can easily extend to multiple time points. We then discuss the performance of our methods in both the cross-sectional and time-varying case in Chapter 5 as well as the limitations we faced. Additionally, we discuss how our approach relates to what has already been done as well as future directions for our methods.

Chapter 2

Literature Review

In this chapter, we will provide the necessary foundation for us to introduce our methods to solve the nested case-control problem using finite population methods in a causal setting. We will provide a foundation on finite population sampling, generalized linear models and generalized estimating equations, nested case-control design studies and causal inference. All of these areas are crucial in order to understand the basis for our proposed methods.

2.1 Finite Population Sampling

Sampling theory is a very important area of this thesis, as it is how we will choose the data that we will analyze using methods that employ other areas of the background. This section closely follows the order and concepts of the work of Lohr in *Sampling: Design and Analysis* [15].

2.1.1 Overview

Finite population design-based inference, often used in survey contexts, differs from infinite population inference in that instead of Y being a random variable drawn from an infinite population following some distribution, we instead sub-sample from a fixed finite population.

The y's that we can observe are fixed, and the randomness comes from which subset of the y's we observe, which depends on the design D.

We first review the two different types of probability samples as defined by Lohr in *Sampling: Design and Analysis* [15].

Definition 2.1.1. ***Simple Random Sample (SRS)*** *A simple random sample is a subset of a statistical population in which each possible subset has an equal probability of being chosen. A simple random sample is meant to be an unbiased representation of a group.*

Definition 2.1.2. ***Stratified Random Sample*** *A stratified random sample is when the population is divided into subgroups known as strata. An independent SRS is then taken in each strata. The strata are formed based on individual's shared characteristics or attributes.*

2.1.2 Sampling with Unequal Probabilities

In Sampling Theory, it is known that the sampling variance can be decreased by assigning unequal probabilities to sampling units in different strata. In order to decrease variances without explicitly stratifying, we can use unequal inclusion probabilities [15]. The probabilities are deliberately varied when sampling with unequal probabilities so that we select different primary sampling units (psus) for the sample for which we obtain unbiased estimators by employing weights in the estimation [15]. A primary sampling unit is a sampling unit that is selected in the first stage of a sample, they are usually selected based on shared attributes. It is important that we control the probabilities with which we will select a given unit. Note that we will consider two different probabilities in this section, because when sampling with unequal probabilities without replacement, selecting a unit on the first draw can affect the selection probabilities for other units [15].

A finite population of N units is denoted by the index set $U = (1, ..., N)$ where samples are subsets of U. Thus we have the following probabilities with which we will select a given unit,

$$\psi_i = \text{P(unit } i \text{ selected on the first draw), and} \tag{2.1}$$

$$\pi_i = \text{P(unit } i \text{ in the sample).} \tag{2.2}$$

2.1.2.1 Unequal Probability Sampling without Replacement

We now describe the basic setting for unequal probability sampling without replacement.

Select two primary sampling units without replacement and with unequal probabilities ψ_i =P(unit i selected on the first draw). The probability that unit j is selected on the second draw depends on which unit was selected on the first draw, this is because we are sampling without replacement [15]. Note that the order of selection matters, thus we obtain

$$\text{P(unit } i \text{ chosen first, unit } k \text{ chosen second)}$$

$$= \text{P(unit } i \text{ selected on the first draw)P(unit } k \text{ chosen second|unit } i \text{ chosen first)} = \psi_i \frac{\psi_k}{1-\psi_i}.$$

Similarly, we can see that

$$\text{P(unit } k \text{ chosen first, unit } i \text{ chosen second)} = \psi_k \frac{\psi_i}{1-\psi_k}.$$

Lohr shows that for a sample of size $n = 2$ we get that

$$\text{P(units } i \text{ and } k \text{ in the sample)} = \pi_{ik} = \psi_i \frac{\psi_k}{1-\psi_i} + \psi_k \frac{\psi_i}{1-\psi_k}.$$

Therefore, Lohr defines the probability that psu i is in the sample is

$$\pi_i = \sum_{\mathcal{S}: i \in \mathcal{S}} P(\mathcal{S}), \tag{2.3}$$

where $\mathcal{S}$ is the collection of indices for units contained in a sample and $P(\mathcal{S})$ is the probability of obtaining the collection of sample indices $\mathcal{S}$.

2.1.2.2 The Horvitz–Thompson Estimator

Let us assume, as in Lohr's example, that we have a without-replacement sample of n psus, and we know the inclusion probability P(unit i in the sample) $= \pi_i$ and the joint inclusion probability, P(units i and k in the sample) $= \pi_{ik}$ [15].

Lohr states that we can calculate the inclusion probability π_i as the sum of probabilities of all samples containing the ith unit and has the property that

$$\sum_{i=1}^{N} \pi_i = n \text{ [15].} \tag{2.4}$$

Lohr also states that for the π_{ik}'s, we have that

$$\sum_{\substack{k=1 \\ k \neq i}} \pi_{ik} = (n-1)\pi_i \text{ [15].} \tag{2.5}$$

We have that π_i/n can be interpreted as the average probability that a unit will be selected on one of the draws, this is because the inclusion probabilities sum to n [15]. Let t_i denote the total of the psus in a one-stage sample where $\hat{t}_\psi$ is the average of the values of t_i/ψ_i [15]. The probabilities of selection depend on what was drawn before when the samples are drawn without replacement [15]. By dividing t_i from psu i by the average probability of selecting that unit in a draw, π_i/n instead of ψ_i, we obtain the Horvitz–Thompson (HT) estimator of the population total, determined by Horvitz and Thompson in 1952 [10]:

$$\hat{t}_{HT} = \sum_{i \in \mathcal{S}} \frac{t_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{t_i}{\pi_i}, \tag{2.6}$$

where $Z_i = 1$ if psu i is in the sample, and 0 otherwise [15].

For the Horvitz-Thompson Estimator we have the following assumptions:

1. sub-sampling is independent between the psus,
2. $t_i \perp Z_1, ..., Z_N$, and
3. $E[t_i|Z_1, ..., Z_N] = t_i$, $V[t_i|Z_1, ..., Z_N] = V_i$.

Under these assumptions we have the following for the expected value and variance for the HT estimator in one-stage sampling:

$$E[\hat{t}_{HT}] = t, \text{ and} \tag{2.7}$$

$$V_{HT}[\hat{t}_{HT}] = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + \sum_{i=1}^{N} \sum_{k\neq 1}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k, \tag{2.8}$$

where t is the population total.

There is another form of the HT variance, and it is called the Sen-Yates-Grundy (SYG) [20, 27] form of the variance and it's given by

$$V_{SYG}[\hat{t}_{HT}] = \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k\neq i}}^{N} (\pi_i \pi_k - \pi_{ik}) \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right). \tag{2.9}$$

Note that the HT expression for the variance is algebraically identical to the SYG expression for the variance. Equations (2.8) and (2.9) lead to different estimators of the variance when the inclusion, π_i, and/or joint inclusion, π_{ik}, probabilities are different [15].

Therefore, we have that the HT estimator of the variance is

$$\hat{V}_{HT}[\hat{t}_{HT}] = \sum_{i\in\mathcal{S}} (1-\pi_i) \frac{t_i^2}{\pi_i^2} + \sum_{i\in\mathcal{S}} \sum_{\substack{k\in\mathcal{S} \\ k\neq i}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} \frac{t_i}{\pi_i} \frac{t_k}{\pi_k}. \tag{2.10}$$

The SYG estimator of the variance is then

$$\hat{V}_{SYG}[\hat{t}_{HT}] = \frac{1}{2} \sum_{i\in\mathcal{S}} \sum_{\substack{k\in\mathcal{S} \\ k\neq i}} \frac{\pi_i \pi_k - \pi_{ik}}{\pi_{ik}} \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2. \tag{2.11}$$

Lohr notes that both (2.10) and (2.11) both require $\pi_{ik} > 0$ for all units in the sample in order to be well-defined [15]. The SYG estimator is generally more the stable of the two variance estimators [15].

We can pretend that the units were selected with replacement and use the with-replacement variance estimator instead in order to avoid some of the potential instability and computational complexity of the estimators (2.10) and (2.11) [15]. Let $\psi_i = \pi_i/n$, thus the with-replacement variance estimator is

$$\hat{V}_{WR}[\hat{t}_{HT}] = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\frac{t_i}{\psi_i} - \hat{t}_{HT}\right)^2 = \frac{n}{n-1}\sum_{i\in\mathcal{S}}\left(\frac{t_i}{\pi_i} - \frac{\hat{t}_{HT}}{n}\right)^2. \tag{2.12}$$

The with-replacement variance estimator, (2.12), is always nonnegative and it also does not require knowledge of the joint inclusion probabilities π_{ik} [15]. In general, the with-replacement variance estimator, (2.12), is preferred. However, when the sampling fraction n/N is large we can overestimate the variance.

2.1.2.3 Weights in Unequal-Probability Samples

The Horvitz–Thompson estimator can be written using sampling weights [15]. Lohr defines the first-stage sampling weight for psu i as

$$w_i = \frac{1}{\pi_i}. \tag{2.13}$$

Thus, Lohr shows that the HT estimator for the population total is

$$\hat{t}_{HT} = \sum_{i\in\mathcal{S}} w_i\hat{t}_i. \tag{2.14}$$

Secondary sampling units (ssus) are a random sample selected within each primary sampling unit. Thus, for a without replacement probability sample of ssus within psus, we define,

$$\pi_{j|i} = P(j\text{th ssu in the } i\text{th psu included in the sample} \mid i\text{th psu in the sample}), \tag{2.15}$$

for a without replacement probability sample of secondary sampling units (ssus) within psus [15].

Thus, we have that

$$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{\pi_{j|i}}, \tag{2.16}$$

where $\pi_{j|i}\pi_i$ is the overall probability that ssu j of psu i is included in the sample [15]. Thus, Lohr defines the sampling weight for the (i, j)th ssu as

$$w_{ij} = \frac{1}{\pi_{j|i}\pi_i}, \tag{2.17}$$

and the HT estimator of the population total as

$$\hat{t}_{HT} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij} \text{ [15].} \tag{2.18}$$

Lohr shows that the population mean is estimated by

$$\hat{\bar{y}}_{HT} = \frac{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}} = \frac{\hat{t}}{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}}. \tag{2.19}$$

Note that estimator $\hat{\bar{y}}_{HT}$ is a ratio, so in order to estimate its variance one needs to form the residuals from the estimated psu totals [15].

2.2 Generalized Linear Models and Generalized Estimating Equations

2.2.1 Generalized Linear Models (GLMs)

2.2.1.1 Overview

The ordinary linear regression model is commonly used to describe a linear relationship between the mean of a response variable and a set of explanatory variables. Generalized linear models (GLMs) extend standard linear regression models to encompass non-normal error distributions and possibly nonlinear functions of the mean. They have three components: (1) random component, (2) linear predictor, and (3) link function [2].

2.2.1.2 Exponential Dispersion Family Distributions for a GLM

In this thesis, we focus on an exponential family form that encompasses standard distributions such as the normal, Poisson, and binomial and that has general expressions for moments and for likelihood equations. We will define the three components of a GLM for the exponential family form.

The random component of a GLM consists of a response variable y with independent observations $(y_1, ..., y_n)$ from a distribution having probability density or mass function for y_i of the form

$$f(y, \theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \theta)\right\} \tag{2.20}$$

This is called an exponential dispersion family [2].The parameter θ is the natural parameter, and the parameter ϕ is the dispersion parameter. We have that $a(\phi) = \frac{\phi}{w_i}$, where the weight w_i is known, and if ϕ is known then it is an exponential family, and if ϕ is not known then it is an exponential dispersion family.

The expected value and the variance are given by:

$$E[Y] = b'(\theta) = \mu, \text{ and} \tag{2.21}$$

$$Var[Y] = b''(\theta)a(\phi). \tag{2.22}$$

We also have that the mean-variance relationship is denoted by $\nu(\mu) = b''(\theta)$. We have that our systematic component is X which is our $n \times p$ design matrix of the form

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}. \tag{2.23}$$

Our linear predictor is $\eta = X\beta$ where $\beta \in \mathbb{R}^{p\times 1}$ vector of parameters and $\eta \in \mathbb{R}^{n\times 1}$. The link function is a function g applied to each component of $E(y)$ that relates it to the linear

predictor,

$$E[Y|X] = \mu[X\beta] = \mu[\eta] \iff g[E[Y|X]] = XB,$$

$$g[E[Y|X]] = g[\mu] = XB = \eta. \tag{2.24}$$

The canonical link is the case where $g[\mu] = \theta = X\beta$.

We will focus on the canonical link, as it simplifies exposition, although other link functions could be used. This is important as although the mean may be restricted, we do not want to impose constraints on β.

2.2.1.2.1 Estimation for GLMs Our main interest within the context of GLMs are maximum-likelihood estimators and their properties.

By maximum-likelihood, we have that the likelihood for a GLM that is an exponential dispersion family is

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} f(y_i, \theta_i, \phi), \tag{2.25}$$

thus the log-likelihood for y_i is given by

$$\ell(\beta) = \sum_{i=1}^{n} log(f(y_i, \theta_i, \phi)) = \sum_{i=1}^{n} \underbrace{\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \theta) \right\}}_{\ell_i}. \tag{2.26}$$

Let $\hat{\beta} = \text{argmax}_{\beta \in \mathbb{R}^p} \ell(\beta)$ be our maximum-likelihood estimator (MLE) for the GLM $g[E[Y|X]] = g[\mu] = XB = \eta$.

We have that the score equation for the ith response is

$$\frac{\partial \ell_i(y_i, \theta_i, \phi)}{\partial \beta_j} = \frac{y_i - b'(\theta_i)}{a(\phi)} \frac{1}{b''(\theta_i)} \frac{1}{g'(\mu_i)} x_{ij} = 0, j = 1, ..., p, \tag{2.27}$$

and the score equations are

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{Var[y_i]} \frac{1}{g'(\mu_i)} x_{ij} = 0, j = 1, ..., p. \tag{2.28}$$

Note that things simplify when the canonical link is used as $g[\mu_i] = \theta_i = \nu_i$. Thus, our score equation for the ith response then becomes

$$\frac{\partial \ell_i(y_i, \theta_i, \phi)}{\partial \beta_j} = \frac{y_i - b'(\theta_i)}{a(\phi)} x_{ij} = 0, j = 1, ..., p. \tag{2.29}$$

We know that $\hat{\beta}$ is the $\hat{\beta}_n$ that solves

$$\frac{\partial \ell}{\partial \beta}(\hat{\beta}_n) = 0. \tag{2.30}$$

2.2.1.3 Models for Binomial Data

Analysts typically assume a binomial distribution for the random component of a generalized linear model (GLM) for binary responses [2].

Let $Y_1, ..., Y_n$ be independent observations where $m_i Y_i \sim Bin(\pi_i, m_i)$. We know that the exponential family for a binomial distribution is the following

$$f(y, \pi, m) = \binom{m}{y} \pi^y (1-\pi)^{m-y} = exp\left\{ \frac{ylog\frac{\pi}{1-\pi} + log(1-\pi)}{\frac{1}{m}} + log\binom{m}{ym} \right\}. \tag{2.31}$$

We have that $E(Y_i) = \pi_i$, $Var\left(\frac{1}{m_i}\pi_i(1-\pi_i)\right)$. We have that $a(\phi) = \frac{1}{m_i}$ where $m_i = w_i$ is known. The canonical link (2.24) for the binomial function is

$$g[E[Y|X]] = g(\pi) = log\frac{\pi}{1-\pi}. \tag{2.32}$$

From its exponential dispersion representation, the binomial natural parameter is the log odds, denoted logit, for which the model is referred to as logistic regression [2].

There are also other links such as the probit link ($g(\pi) = \Phi^{-1}(\pi)$) and the complementary log-log link ($g(\pi) = log(-log(1-\pi))$). However, the logit link is preferred to the probit link because: the logit link is canonical (which simplifies calculations), the logit link is explicit, and the logit link gives us a nicer interpretation as we can calculate odds ratios. In this thesis we will only consider the logit link.

2.2.1.3.1 Simple Logistic Regression If we want to measure the association of the exposure (interchangeably referred to as treatment throughout this thesis) x with the probability of outcome $\pi(x)$, then we can fit the simple logistic regression model

$$log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x, \tag{2.33}$$

using maximum likelihood.

In order to examine the probability of the outcome conditional on exposure we have the following if the predictor is continuous

$$\pi(x) = \text{logit}^{-1}(x) = \frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)}. \tag{2.34}$$

If we would like to test if the exposure is associated to the outcome, we can focus on testing β, as $\beta = 0$ implies that $log\frac{\pi(x)}{1-\pi(x)}$ does not linearly depend on x. We aren't focusing on α as it is less interesting because if the x_i's are centered ($\frac{1}{n}\sum x_i = 0$) then α is the log-odds at $\bar{x}$.

The log-odds ratio is very important for the interpretation of β, it is denoted as

$$\beta = log\frac{\pi(x+1)}{1-\pi(x+1)} - log\frac{\pi(x)}{1-\pi(x)} = log\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}}, \tag{2.35}$$

where

$$\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}}$$

is the odds-ratio.

We also have that

$$\frac{\pi(x+1)}{1-\pi(x+1)} = e^{\beta}\frac{\pi(x)}{1-\pi(x)}.$$

If $\beta > 0$, then $\pi(x)$ is increasing in x, thus exposure increases the probability of $Y = 1$. If $\beta < 0$, then $\pi(x)$ is increasing in x. If $\beta = 0$, $\pi(x)$ is equal in the two exposure groups, where the probability only depends on α. We will only test $\beta \neq 0$ as it provides a better interpretation.

2.2.1.3.2 Properties and Interpretations of Logistic Regression

This section closely follows the order and concepts of the work by Agresti in Chapter 5.2 of *Foundations of Linear and Generalized Linear Models* [2].

We can extend the simple logistic regression model (2.33), to multiple covariates, i.e.

$$log\frac{\pi_i}{1-\pi_i} = \sum_{j=1}^{p} \beta_j x_{ij}, \text{ and} \tag{2.36}$$

$$\pi_i = \frac{exp(\sum_{j=1}^{p} \beta_j x_{ij})}{1 + exp(\sum_{j=1}^{p} \beta_j x_{ij})}. \tag{2.37}$$

Note that our intercept, formerly known as α is now denoted as β_0.

Note that for β_j,

$$\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \frac{exp(\sum_{j=1}^{p} \beta_j x_{ij})}{[1 + exp(\sum_{j=1}^{p} \beta_j x_{ij})]^2} = \beta_j \pi_i (1 - \pi_i). \tag{2.38}$$

Now we consider a single binary indicator x. Agresti states that the model $logit(\pi_i) = \beta_0 + \beta_1$ then describes as 2×2 contingency table, which represents two classifications of a set of counts or frequencies, where

$$logit[P(y=1, x=1)] - logit[P(y=1, x=0)] = [\beta_0 + \beta_1(1)] - [\beta_0 + \beta_1(0)] = \beta_1 \text{ [2]}. \tag{2.39}$$

Agresti states that e^{β_1} is the odds ratio

$$e^{\beta_1} = \frac{P(y=1, x=1)/[1 - P(y=1, x=1)]}{P(y=1, x=0)/[1 - P(y=1, x=0)]} \text{ [2]}. \tag{2.40}$$

The odds $\pi_i/(1-\pi_i)$ are an exponential function of x_j and they multiply by e^{β_j} per unit increase in x_j, which then adjusts for the other explanatory variables in the model [2].

2.2.2 General Estimating Equations (GEEs)

2.2.2.1 Overview

The General Estimating Equations (GEE) approach is an extension of Generalized Linear Models (GLMs). The essence of GEEs is to simply model the mean response, as opposed to

modeling the within-subject covariance structure for correlated data [7]. Thus, for GEEs we do not need to specify the covariance structure correctly in order to obtain valid estimates of regression coefficients and standard errors [7].

Essentially, unlike a GLM, with a GEE you do not specify the underlying probability model. Instead, you specify only the mean, and the covariance function. The most common use of GEE is to fit a marginal model for longitudinal and/or clustered data analysis [7].

Assume that we observe: a response variable Y which can either be continuous or categorical, and a set of k explanatory variables $X = (X_1, ..., X_k)$ which can be discrete, continuous or a combination where X_j is a $n_i \times k$ matrix of covariates. A GEE is of a similar form as a GLM, however, a full specification of the joint distribution of the responses is not required [7]. The random component of a GEE is any distribution of the response that we can use for a GLM (such as binomial, multinomial, normal, etc.) [7]. The systematic component is a linear predictor of any combination of continuous and discrete variables [7]. The link function, like for a GLM, can be the identity, log, logit link, or others, and it must be specified for a GEE.

In addition to specifying the mean, we need to specify the covariance function for the measures. For the covariance function, we have the following assumptions. We have correlated or clustered responses, $Y_1, ..., Y_n$, meaning that the clusters are not independent [7]. We do not need to satisfy the homogeneity of variance and we can have correlated errors [7]. In order to estimate the parameters, GEEs use quasi-likelihood estimation rather than maximum likelihood estimation (MLE) or ordinary least squares (OLS) [7]. We need to specify the covariance structure for the clustered responses.

Here are the four most commonly used correlation structures: (note that the correlation structures do not depend on the subject, thus they are the same for each subject):

Independence - (correlation is independent)

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.41}$$

Exchangable (or Compund Symmetry)

$$\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}, \tag{2.42}$$

AutoRegressive Order 1 (AR 1)

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}, \tag{2.43}$$

Unstructured

Where $\rho_{ij} = corr(Y_{ij}, Y_{ik})$ for the i^{th} subject at times j and k

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}. \tag{2.44}$$

2.2.2.2 Quasi-Likelihood

GEEs are quasi-likelihood equations with estimates that are quasi-likelihood estimators. The GEE estimates are obtained by using an iterative algorithm, denoted an iterative quasi-scoring procedure, thus meaning that there are no closed-form solutions [7].

Quasi-likelihood was applied to GEEs by Liang and Zeger [28] in 1986 and works as follows.

Quasi-likelihood can be used with a variety of outcomes as it is a methodology for regression that requires few assumptions about the distribution of the dependent variable [28]. One only needs to specify the relations between the outcome mean and covariates as well as the mean and variance in quasi-likelihood as opposed to specifying the distribution as in likelihood analysis, thus, quasi-likelihood is useful in developing methods that can be used for several types of outcome variables [28].

We will use notation that resembles the notation used by Liang and Zeger. Consider the observations $(y_{ij}, \mathbf{x_{ij}})$ for times t_{ij}, where $j = 1, ..., n_i$ and subjects $i = 1, ..., K$. We have that y_{ij} is our outcome variable and that $\mathbf{x_{ij}}$ is a $p \times 1$ vector of covariates. We let $\mathbf{y_i}$ be the $n_i \times 1$ vector $(y_{i1}, ..., y_{in_i})'$ and $\mathbf{x_i}$ be the $n_i \times p$ matrix $(x_{i1}, ..., x_{in_i})'$ for the i^{th} subject. Note that Liang and Zeger drop the j subscript and treat each subject's data as a scalar as quasi-likelihood has previously been applied to the regression context where $n_i = 1 \forall i$.

Define μ_i as the expectation of y_i and let

$$\mu_i = h(\mathbf{x_i}\boldsymbol{\beta}) \tag{2.45}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vectors of parameters, and where h^{-1} is the link function (note that typically in GLM, the link function is denoted g).

In quasi-likelihood, the variance, v_i, of y_i is expressed as a known function, g (which is the variance function, typically denoted as $\nu(\mu)$ in GLM), of the expectation, μ_i, it is expressed as follows

$$v_i = g(\mu_i)/\phi \tag{2.46}$$

where ϕ is a scale parameter, which is treated as a nuisance parameter as the focus of quasi-likelihood is on methods for inference about $\boldsymbol{\beta}$. Note that ϕ here plays a similar role to a GLM where $a(\phi) = \phi$.

We have that the quasi-likelihood estimator is the solution of

$$S_k(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \mu_i}{\partial \beta_k} v_i^{-1}(y_i - \mu_i) = 0, \text{ for } k = 1, ..., p. \tag{2.47}$$

Note that (2.47) only applies if we have iid (independently and identically distributed) variables at the true μ. Liang and Zeger also note that (2.47) are actually the score equations for $\boldsymbol{\beta}$ when y_i has a distribution from the exponential family, and their solution can be obtained by an iteratively weighted least squares [28].

2.2.2.3 Quasi-Likelihood applied to Longitudinal Data

Liang and Zeger [28] in 1986 applied the quasi-likelihood approach to the analysis of longitudinal data, which we will introduce. For the vector of responses, $\boldsymbol{y}_i$, for the i^{th} subject, we must examine the mean and covariance [28].

Let $\boldsymbol{R}_i(\boldsymbol{\alpha})$ be the $n_i \times n_i$ "working" correlation matrix (working is in quotations because as expressed above, the correlation matrix in a GEE may and is often misspecified, note that even if it is misspecified the GEE approach works) [28]. In our case, we have that $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is assumed to be fully specified by $\boldsymbol{\alpha}$ which is the $s \times 1$ vector of unknown parameters, which is the same for all subjects [28].

We then obtain the following working covariance matrix

$$\boldsymbol{V}_i = \frac{\boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}}{\phi}, \tag{2.48}$$

where $\boldsymbol{A}_i^{1/2}$ is an $n_i \times n_i$ diagonal matrix with $g(\mu_{ij})$ as the j^{th} diagonal element [28].

Note that even when $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is incorrectly specified, we can obtain estimators that are consistent and have consistent variance estimates.

Liang and Zeger extended (2.47) which were our score equations to the longitudinal case as follows

$$\sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{S}_i = \sum_{i=1}^{K} \left[\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}_k}\right]' \left[\frac{\boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}}{\phi}\right]^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0, \text{ with } \boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{in})'. \tag{2.49}$$

Note that $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}_k}$ is $n \times p$. When $n_i = 1$, we have that (2.49) reduces to the quasi-likelihood equations (2.47). The regression coefficients are consistent as they are guaranteed by the

equations (2.49) when the link function is correctly specified [28]. Liang and Zeger also note that $\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}$ does not depend on the $\boldsymbol{y}$'s so that if $E[\boldsymbol{S}_i] = 0$, equations (2.49) converge to 0 and have consistent roots [28]. Also note that equations (2.49) are the score equations for $\boldsymbol{\beta}$ for Gaussian outcomes [28].

Let $\boldsymbol{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{S}_i$. Note that if we would like to adjust for confounding but don't want to make regression based adjustments, we can add inverse probability weights W_i to $\boldsymbol{U}_i$ as follows (see section 2.4.1.1.2), $\sum_{i=1}^{K} \boldsymbol{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})W_i = \sum_{i=1}^{K} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{S}_i W_i = 0$.

We have that $\hat{\boldsymbol{\beta}}_R$ is an estimator of $\boldsymbol{\beta}$ for any $\boldsymbol{R}_i(\boldsymbol{\alpha})$, and is the solution of (2.49). Liang and Zeger proved that $\hat{\boldsymbol{\beta}}_R$ is a consistent estimator of $\boldsymbol{\beta}$ [28].

Liang and Zeger introduced what is known as the "sandwich estimator" which is defined as follows

$$\begin{aligned} \boldsymbol{V}_R &= \lim_{K\to\infty} K \left(\sum_{i=1}^{K} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i \right)^{-1} \left[\sum_{i=1}^{K} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\mathrm{cov}(\boldsymbol{y}_i)\boldsymbol{V}_i^{-1}\boldsymbol{D}_i \right] \left(\sum_{i=1}^{K} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i \right)^{-1} \\ &= \lim_{K\to\infty} K \left(\boldsymbol{V}_1^{-1}\boldsymbol{V}_0\boldsymbol{V}_1^{-1} \right), \end{aligned} \tag{2.50}$$

where $\hat{\alpha}(\boldsymbol{Y}, \boldsymbol{\beta}, \phi)$ and $\hat{\phi}(\boldsymbol{y}, \boldsymbol{\beta})$ are $K^{1/2}$ consistent estimators of α and ϕ respectively.

By replacing $\mathrm{cov}(\boldsymbol{y}_i)$ by $\boldsymbol{S}_i\boldsymbol{S}_i'$ and $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and ϕ by their estimates in (2.50), without evaluating $\mathrm{cov}(\boldsymbol{y}_i)$ directly, we can estimate $\boldsymbol{V}_R$ consistently [28]. They also noted that $\hat{\boldsymbol{\beta}}_R$ does not depend on the choice of $\boldsymbol{\alpha}$ and ϕ [28].

We can also define the "sandwich estimator" from the work of Albert, Liang and Zeger [29] in 1988 as follows

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \boldsymbol{M}_0^{-1}\boldsymbol{M}_1\boldsymbol{M}_0^{-1}, \tag{2.51}$$

where $\boldsymbol{V}_i(\alpha) = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\alpha)\boldsymbol{A}_i^{1/2}$,

$$\boldsymbol{M}_0 = \sum_{i=1}^{K} \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}}, \tag{2.52}$$

and

$$\boldsymbol{M}_1 = \sum_{i=1}^{K} \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)' \hat{\boldsymbol{V}}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}} \text{ [29]}. \tag{2.53}$$

Note that in this case $\hat{\boldsymbol{\beta}}$ is the solution of

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1}(\alpha)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i). \tag{2.54}$$

Note that (2.51) is the sandwich estimator for any link.

The regression coefficients, the correlation, and the scale parameters, $\boldsymbol{\alpha}$ and ϕ, are solved for iteratively in order to solve the GEE for $\hat{\boldsymbol{\beta}}_R$ [28]. We can calculate an updated estimate of $\boldsymbol{\beta}$ by iteratively reweighted least squares given an estimate of $\boldsymbol{R}_i(\boldsymbol{\alpha})$ and of ϕ [28].

Liang and Zeger point out that ϕ does not need to be estimated since if the elements of $\boldsymbol{R}$ are multiples of the parameters, $\boldsymbol{\alpha}$, then we can estimate $\boldsymbol{\beta}$ without estimating ϕ directly and it can be accounted for in $\boldsymbol{R}_i(\boldsymbol{\alpha})$ [28].

When $\boldsymbol{R}_i$ is misspecified, $\hat{\boldsymbol{\beta}}_R$ and $\boldsymbol{V}_R$ are both robust to the misspecification, and the confidence intervals for $\boldsymbol{\beta}$ will be asymptotically correct, however, we can increase efficiency by choosing the working correlation matrix to be close to the actual one [28].

Nevertheless, the specification of the mean (μ_i), needs to be correct, however, $\boldsymbol{R}$ can be incorrect. If $\boldsymbol{R}$ is misspecified, then our estimate is consistent but not efficient, and if $\boldsymbol{R}$ is correct, then our estimate is both consistent and efficient.

2.2.2.4 Summary

A reason why GEEs are so powerful is that even if the covariance is misspecified, GEE estimates of model parameters are valid [7]. The standard errors will not be precise if the correlation structure is misspecified [7]. Whereas, a chosen model in practice is not necessarily correct, thus, the GEE approach can be preferred over a GLM approach. However, if the chosen model for a GLM is correct, the estimate will be more efficient and consistent in a

GLM approach rather than the GEE approach. Choosing whether a GLM or a GEE is more suitable for the analysis is a difficult choice which is up to the statistician as both methods are considered to be "better" depending on the situation. However, there are doubly robust methods that could be employed but they are beyond the scope of this thesis.

Typically, GEEs are used to determine confidence intervals, as the "sandwich estimator" is incredibly powerful for determining the standard errors for an estimate. They are also typically used for clustered data.

2.3 Nested Case-Control Design Studies

In order to understand the methods in this thesis, we will need to define some terms. A cohort is a group of people who share a common characteristic or experience within a defined period [14]. A cohort study is a longitudinal study that samples a cohort, performing a cross-section at intervals through time (typically a randomized collection of prospective data) [14]. Retrospective means looking back in time, thus using existing data such as medical records. Prospective means requiring the collection of new data.

In order to determine if an exposure is related to an outcome, case-control study designs are used [14]. The case-control study can be described simply in the case of binary exposure and response. First, one samples from the sub-population cases (a group that has the outcome) and then samples from the sub-population of controls (a group known to not have the outcome) [14]. Then, one compares the frequency of the exposure in the case group to the frequency of exposure in a control group [14]. By definition, a case-control study is always retrospective because it stratifies by the outcome then samples exposures rather than sampling exposures and outcomes jointly or stratifying on exposures first.

Case-control studies have certain advantages compared to other study designs as they are comparatively inexpensive to implement [14]. Case-control studies start with those who are known to have the outcome, thus, they are mainly employed for studying rare outcomes and

for investigating outbreaks as it can be more achievable to enroll enough participants [14]. As case-control studies are efficient, they are used as a preliminary analysis of a suspected risk factor for a common condition, this helps determine the need for a longitudinal study, which would be costly and time-consuming [14].

2.3.1 Nested Case-Control Studies and Logistic Regression

In order to estimate the effects of a single binary risk factor on disease risk, log odds regression models can be employed [4]. Only by considering each factor after stratification to control for effects of the others, and only by considering each level of exposure separately relative to baseline may multiple categorical risk factors be accommodated [4]. Recall the logistic regression model defined in equations (2.36) and (2.37). A key feature of the logistic model for case-control studies is that the regression coefficients have a relative risk interpretation. Breslow formulated the case-control study as:

$$\frac{P(Y=1|X=x_1)P(Y=0|X=x_0)}{P(Y=0|X=x_1)P(Y=1|X=x_0)} = \exp[(x_1-x_0)\boldsymbol{\beta}], \tag{2.55}$$

where $(x_1 - x_0)\boldsymbol{\beta}$ represents the log relative risk for a subject with exposure x_1 versus one with exposure x_0 [4].

Agresti [1] discusses conditional logistic regression and methods for comparing categorical responses for two paired samples. He lets (Y_{i1}, Y_{i2}) denote the ith pair of observations $i = 1, \cdots, n$. Let Y_{it} be the outcome for observation t for subject i, we thus have the conditional model

$$\text{logit}[P(Y_{it}=1)] = \alpha_i + \beta x_t \text{ [1]}. \tag{2.56}$$

As the effect β is defined conditional on the subject, it is called a conditional model [1].

It follows that

$$P(Y_{it}=1) = \frac{\exp(\alpha_i + \beta x_t)}{1+\exp(\alpha_i + \beta x_t)}. \tag{2.57}$$

Independence of responses for different subjects and for two observations on the same subjects are normally assumed [1]. However, Agresti notes that the responses are non-negatively associated when averaged over all subjects [1]. He also notes that dependence in matched pairs is not accounted for in the model [1]. Non-negative association through the model structure is taken into account by fitting the model [1].

Note that observations (y_{i1}, y_{i2}) in a matched pair don't always refer to the same subject [1]. Consider a binary response Y in a nested case-control setting, according to criteria that could influence the response, we have that each case $(Y = 1)$ is matched with a control $(Y = 0)$ [1]. In a nested case-control setting for subject t in matched pair i, Agresti considers the model

$$\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta x_{it}. \tag{2.58}$$

Agresti notes that the retrospective study only provides insight on the distribution of X given Y but the probabilities modeled refer to the distribution of Y given X [1]. As the odds ratio, $\exp(\beta)$, refers to the XY odds ratio that relates to both conditional distributions it can be estimated [1].

If our binary response has p predictors for nested case-control matched pairs, we can generalize the model (2.58) to

$$\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_p x_{pit},$$

where x_{hit} denotes the value of predictor h for observation t in pair i, $t = 1, 2$ [1].

2.4 Causal Inference

This section closely follows the order and concepts of the work of Hernán and Robins in their book *Causal Inference: What If* [9].

2.4.1 Cross-Sectional Setting

In this section, we will discuss the areas of Causal Inference in the cross-sectional setting that are relevant to the methods in this thesis.

If X is randomized, then correlation (association) between X and Y can be inferred to be **causal**.

2.4.1.0.1 Potential Outcomes Let X be a binary exposure and let $Y^{(x)}$ denote what you observe if the subject receives level (x), thus the potential responses are $(Y^{(x=0)}, Y^{(x=1)}) = (Y^{(0)}, Y^{(1)})$. Note that only one of these can be observed as the subject can only receive one level of the exposure. Recall that exposure is interchangeably referred to as treatment throughout the this thesis. If Y is the observed response from your sample then,

$$Y = XY^{(1)} + (1 - X)Y^{(0)}. \tag{2.59}$$

If X is randomized, then $X \perp (Y^{(0)}, Y^{(1)})$, thus the conditional expected value of Y given X is

$$E[Y|X = x] = E[Y^{(x)}]. \tag{2.60}$$

Therefore we have that $E[Y|X = 1] = E[Y^{(1)}|X = 1] = E[Y^{(1)}]$, the mean of the potential outcome $Y^{(1)}$ in the population. If X is binary, and randomized $(X \perp (Y^{(0)}, Y^{(1)}))$, we have that

$$E[Y|X = 1] = P[Y = 1|X = 1] = P[Y^{(1)} = 1]. \tag{2.61}$$

Note that if $X \not\perp (Y^{(0)}, Y^{(1)})$ we have that $E[Y|X = x] = E[Y^{(x)}|X = x] \neq E[Y^{(x)}]$ for either $X = 0$ or $X = 1$ (or both).

One can also make a weaker assumption that $(Y^{(0)}, Y^{(1)}) \perp X|L$, where L is some set of covariates.

We now define the Average Treatment Effect (ATE) to be

$$ATE = E[Y^{(1)} - Y^{(0)}] = E[Y^{(1)}] - E[Y^{(0)}]. \tag{2.62}$$

Note, that we can only observe a subject that receives *one* level of exposure, which creates a problem because we cannot compare potential outcomes within the same subject.

If $(Y^{(0)}, Y^{(1)}) \perp X$, we have that

$$\begin{aligned} E[Y|X=x] &= xE[Y^{(1)}|X=x] + (1-x)E[Y^{(0)}|X=x] \\ &= xE[Y^{(1)}] + (1-x)E[Y^{(0)}]. \end{aligned} \tag{2.63}$$

We can determine the sample average of people who received $X = 1$ as follows

$$\begin{aligned} E[\bar{Y}_1] &= E\left[\frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i}\right] \\ &= E_{X_1,\ldots,X_n} E_{Y_1,\ldots,Y_n|X_1,\ldots,X_n}\left[\frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i}\right] \\ &= E_{X_1,\ldots,X_n}\left[\frac{1}{\sum_{i=1}^n X_i}\sum_{i=1}^n E_{Y_1,\ldots,Y_n|X_1,\ldots,X_n}[Y_i X_i]\right]. \end{aligned}$$

Note that here we fix what people received and are treating it as constant. Before we continue the derivation, we have the following assumptions:

- Assumption 0: Subjects are a random sample of population
- Assumption 1: Subjects are independent
- Assumption 2: X is independent of everything including their potential outcomes, $(Y_i^{(0)}, Y_i^{(1)}) \perp X_i$ if X_i is randomized (this is why we run randomized trials)

$$
\begin{aligned}
E[\bar{Y}_1] &= E_{X_1,\ldots,X_n}\left[\frac{1}{\sum_{i=1}^n X_i}\sum_{i=1}^n E_{Y_i|X_i}[Y_i X_i]\right] \\
&= E_{X_1,\ldots,X_n}\left[\frac{1}{\sum_{i=1}^n X_i}\sum_{i=1}^n E_{Y_i^{(0)},Y_i^{(1)}|X_i}[(X_i Y_i^{(1)} + (1-X_i)Y_i^{(0)})X_i]\right] \\
&= E_{X_1,\ldots,X_n}\left[\frac{1}{\sum_{i=1}^n X_i}\sum_{i=1}^n E_{Y_i^{(0)},Y_i^{(1)}|X_i}[X_i Y_i^{(1)}]\right] \\
&= E_{X_1,\ldots,X_n}\left[\frac{1}{\sum_{i=1}^n X_i}\sum_{i=1}^n X_i E_{Y_i^{(1)}}[Y_i^{(1)}|X_i]\right] \\
&= E_{X_1,\ldots,X_n}\left[\frac{1}{\sum_{i=1}^n X_i}\sum_{i=1}^n X_i E[Y_i^{(1)}]\right] \\
&= E[Y_i^{(1)}]E_{X_1,\ldots,X_n}\left[\frac{1}{\sum_{i=1}^n X_i}\sum_{i=1}^n X_i\right] \\
&= E[Y^{(1)}].
\end{aligned}
$$

Thus the sample average of people who received $X = 1$ is, under a random assignment of X,

$$E[\bar{Y}_1] = E[Y^{(1)}]. \tag{2.64}$$

Similarly, for $\bar{Y}_0$ replace X_i with $(1 - X_i)$, thus the sample average of people who received $X = 0$ is

$$E[\bar{Y}_0] = E[Y^{(0)}]. \tag{2.65}$$

We can now extend our assumption to $(Y_i^{(0)}, Y_i^{(1)}) \perp X_i|L_i$, where L_i is some set of covariates. We want to find L such that potential outcomes are independent of exposure given a set of covariates. If we find this set of covariates, L, then we can correct for the bias in estimating the mean potential outcome from observed data.

2.4.1.0.2 Directed Acyclic Graphs (DAGs) In causal inference, expert knowledge is required and assumptions about the causal relationships between exposure, outcome and other variables are necessary [9].

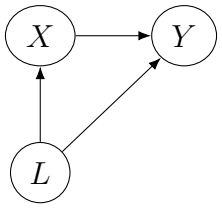
One can represent said expert knowledge and assumptions about the causal relationships through Directed Acyclic Graphs (DAGs) [9]. They help clarify conceptual problems and

enhance communication among researched by summarizing knowledge and assumptions in an intuitive way [9]. By using DAGs in causal inference, it helps one visualize your assumptions prior to your analysis.

Direct causal effects are denoted in DAGs by the presence of an arrow pointing from one variable to another, the absence of an arrow means that there is no direct causal effect from one variable to another. Note that the arrow does not specify whether the causal effect is harmful or protective [9].

In our situation we have that X is the exposure, Y is the outcome and let L be a confounder, assume that we have the following:

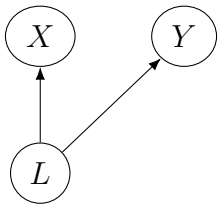
Figure 2.1 – Example of a DAG



In Figure 2.1, the arrow from X to Y means that exposure affects the probability of the outcome. Furthermore, L affects both X and Y. Note that two variables are dependent if there exists a directed path from one variable to another. Thus, in Figure 2.1 we have that X and Y, L and X, and L and Y are dependent.

Note that in Figure 2.1 we have that $Y \not\perp X$ and $Y \not\perp X|L$, and if you don't account for the presence of L in the analysis (even in X and Y are causally independent) you will get bias.

Figure 2.2 – Another Example of a DAG



In Figure 2.2, we have that L is a confounder as its a common cause of X and Y. We have that X and Y are marginally dependent, and that $Y \perp X|L$ but $Y \not\perp X$.

2.4.1.1 Inverse Probability (IP) Weighting

We want to determine the average causal effect of exposure on outcome. In this section we will describe how to estimate this affect using inverse probability (IP) weighting.

2.4.1.1.1 Basis for the Methods

We want estimate the average causal effect of exposure X on the outcome Y. We classify individuals as exposed, $X = 1$, and unexposed, $X = 0$, otherwise. The outcome Y is binary.

We have that the probability of outcome under the exposure and control can be written as $\hat{E}[Y|X = 1]$ and $\hat{E}[Y|X = 0]$, where the difference is denoted $\hat{E}[Y|X = 1] - \hat{E}[Y|X = 0]$.

We define $E[Y^{(x=1)}]$ as the probability of outcome that would have been observed if all individuals in the population had been exposed, and $E[Y^{(x=0)}]$ if all the individuals in the population had not been exposed.

We define the average causal effect on the additive scale as $E[Y^{(x=1)}] - E[Y^{(x=0)}]$, that is, the difference in outcome that would be observed if everyone had been exposed compared with unexposed [9].

The observed population difference $E[Y|X = 1] - E[Y|X = 0]$ is generally different from the causal difference $E[Y^{(x=1)}] - E[Y^{(x=0)}]$ [9]. The former will not generally have a causal interpretation if the exposed and the unexposed difference with respect to characteristics that affect outcome, due to potential confounding [9]. We let L be a confounder of the effect of X on Y and our analysis will need to adjust for L.

The exposed and unexposed also differ in their distribution of other variables, if these variables are confounders they need to be adjusted for in the analysis [9]. We let L represent a vector of the measured covariates.

2.4.1.1.2 Estimating IP Weights in the Cross-Sectional Case A pseudo-population is created when IP weighting is used where the direct effect from the confounders L to the exposure X is eliminated; therefore, all confounding is eliminated if the confounders L are sufficient to block all backdoor paths from X to Y [9]. That is, the association between X and Y in the pseudo-population consistently estimates the causal effect of X on Y [9].

Hernán and Robins informally define that the pseudo population is created by weighting each individual by the inverse of the conditional probability of receiving the exposure level ($P[X = 1|L]$, conditional probability of exposure i.e. the propensity score) that the individual received [9]. They define the individual-specific IP weights for exposure X as

$$W^X = \frac{1}{f(X|L)}, \tag{2.66}$$

where $f(X|L)$ is the probability of exposure conditional on the measured confounders, i.e. $P[X = 1|L]$ for the exposed and $P[X = 0|L]$ for the unexposed [9]. Note that only $P[X = 1|L]$ needs to be estimated for a dichotomous exposure as $P[X = 0|L] = 1 - P[X = 1|L]$ [9].

We can fit a logistic regression model for the probability of exposure with the confounders included as covariates in order to obtain parametric estimates of $P[X = 1|L]$ in each of the strata defined by L [9].

We now compute the difference, $\hat{E}[Y|X = 1] - \hat{E}[Y|X = 0]$, in the pseudo-population created by estimating the IP weights. In the pseudo-population, in the absence of confounding for the effect of X, we have that association is causation, thus a consistent estimator of the causal difference, $E[Y^{x=1}] - E[Y^{x=0}]$, is also a consistent estimator of the associated difference, $E[Y|X = 1] - E[Y|X = 0]$ [9]. To estimate $\hat{E}[Y|X = 1] - \hat{E}[Y|X = 0]$ in the pseudo-population, Hernán and Robins fit the (saturated) linear mean model $E[Y|X] = \theta_0 + \theta_1 X$ by weighted least squares, with individuals weighted by their estimated weights $\left(\frac{1}{\hat{P}[X=1|L]}\right)$ for the exposed and $\left(\frac{1}{1-\hat{P}[X=1|L]}\right)$ for the unexposed [9].

The IP weights, (2.66), adjust for confounding by L because they create a pseudo-population

in which all individuals have the same probability of having $X = 1$ and $X = 0$. Thus, X and L are independent from the pseudo-population which implies that all backdoor paths from exposure X and the outcome Y via L are eliminated.

The IP weights, (2.66), are known as nonstabilized weights. There are other IP weights known as stabilized weights that are also used that result in an unbiased estimate of the ATE. Note that in a saturated model, there is no difference in using stabilized instead of nonstabilized weights; however in a nonsaturated model, some argue it is better to use stabilized weights [9]. However, in our case we will only be focusing on nonstabilized weights as in (2.66) in order to simplify the presentation. All the methods presented in this thesis could be extended to the case of stabilized weights with minimal additional work.

2.4.1.1.3 Marginal Structural Models Hernán and Robins consider a linear model for the mean outcome under the dichotomous exposure level x, referred to as a saturated marginal structural model, that is defined as follows

$$E[Y^x] = \beta_0 + \beta_1 x \text{ [9]}. \tag{2.67}$$

Note that in (2.67) that the covariates and confounders are marginalized out. As the outcome variable of this model is counterfactual, and hence generally unobserved, it is referred to as a marginal structural mean model [9]. Models of this type cannot be fit to data of any real-world study, as we don't observe all potential outcomes for all subjects [9].

In structural mean models, the parameters for exposure correspond to average causal effects [9]. In (2.67) $\beta_1 = E[Y^{x=1}] - E[Y^{x=0}]$ as $E[Y^x] = \beta_0$ when $x = 0$ and $E[Y^x] = \beta_0 + \beta_1$ when $x = 1$. Thus, β_1 is the average causal effect of exposure on the outcome.

Recall that a pseudo-population is created when IP weighting is used. Using IP weighted least squares, the model, $E[Y|X] = \theta_0 + \theta_1 X$, is fit to the pseudo-population. We have that association is causation in the pseudo-population under their assumptions [9]. Thus, θ_1 has the same causal interpretation as the parameter β_1 from (2.67). Therefore, it follows that $\hat{\theta}_1$ is a consistent estimator of the causal effect in the population, β_1 [9].

2.4.2 Time-Varying Setting

In this section, we will discuss the areas of Causal Inference in the time-varying setting that are relevant to the methods in this thesis. This section closely follows the order and concepts of the work in Part III of *Causal Inference: What If*, by Hernán and Robins [9].

2.4.2.1 Time-Varying Exposures

2.4.2.1.1 The Causal Effect of Time-Varying Exposures Let us consider a time-fixed exposure variable X, let $X = 1$ be exposed and $X = 0$ be unexposed, at time 0 and an outcome variable Y measured K time points later. Thus, consider a time-varying dichotomous exposure X_k that may change at every k time point, where $k = 0, 1, 2, \cdots, K$ with K being the last time point. To denote exposure history we have that $\bar{X}_k = (X_0, X_1, \cdots, X_k)$ is the history of exposure from time 0 to time k. We will represent $\bar{X}_K$ as $\bar{X}$ when we refer to the entire exposure history through K. We will assume no individuals were exposed before the start of the study at time 0, namely, $X_{-1} = 0$ for all individuals [9]. An individual who is always exposed continuously has exposure history $\bar{X} = (X_0 = 1, X_1 = 1, \cdots, X_K = 1) = (1, 1, \cdots, 1) = \bar{1}$ and an individual who is never exposed has exposure history $\bar{X} = (X_0 = 0, X_1 = 0, \cdots, X_K = 0) = (0, 0, \cdots, 0) = \bar{0}$ [9].

Assume we have a value for the outcome Y the end of our study at time $K + 1$. Our goal is to estimate the average causal effect of the time-varying exposure $\bar{X}$ on the outcome Y. For the time-varying exposure, X_k, we need to take into account the effect all times k between 0 and K as the average causal effect of a time-varying exposure cannot be defined for a single time-point k [9]. Thus, the average causal effect is redefined in a time-varying setting as the difference between the counterfactual mean outcomes under two exposure strategies from time $k = 0$ to time $k = K$ [9].

2.4.2.1.2 Treatment Strategies Hernán and Robins define a treatment strategy as a rule to assign exposure at each time k [9]. Recall, the terms treatment and exposure are used interchangeably in this thesis. For example, the always expose strategy is denoted by $\bar{x} =$

$(1, \cdots, 1) = \bar{1}$, and the never expose strategy is denoted by $\bar{x} = (0, \cdots, 0) = \bar{0}$ [9]. Thus, they define an average causal effect of $\bar{X}$ on the outcome Y as the difference between the mean counterfactual outcome $Y^{\bar{x}=\bar{1}}$ under the always exposed strategy and the mean counterfactual outcome $Y^{\bar{x}=\bar{0}}$ under the never expose strategy [9]. Thus, an average causal effect of $\bar{X}$ on the outcome Y is $E[Y^{\bar{x}=\bar{1}}] - E[Y^{\bar{x}=\bar{0}}]$ [9].

Consider the time-varying covariate L_k measured at time-point k in all individuals. Let the variable L_k be dichotomous. At time zero, we could let $L_0 = 0$. We could then consider the exposure strategy of when $L_k = 0$, do not exposure and when $L_k = 1$, begin exposure and expose continuously after that time [9]. The value of the individual's evolving L_k will effect their exposure x_k at time k which is an example of a *dynamic treatment strategy* [9]. Hernán and Robins define *static treatment strategies* as non-dynamic strategies for $\bar{x}$ for which exposure does not depend on covariates [9].

The exposure strategies of interest need to be specified in order to ensure that a well-defined average causal effect of a time-varying exposure [9]. The causal effect for time-varying exposures do not have a single definition [9]. For pairs of exposure strategies, as many causal effects can be defined [9].

A *sequentially randomized experiment* is where, for an individual, exposure is randomly assigned at each time k, however, they are not as commonly used in practice [9].

2.4.2.1.3 Time-Varying Confounding Let us consider a DAG that contains time points $k = 0, 1, 2 \cdots$. Consider covariate L_k that affects subsequent exposures $X_k, X_{k+1}, \cdots$ and that affects the outcome Y through an unmeasured covariate. We need the data on $\bar{L}_k$ for all individuals in order to have no unmeasured confounding for the effect of $\bar{X}$ and to estimate the causal effects of treatment strategies, thus, $\bar{L}_k$ are defined to be time-varying confounders for the effect of time-varying exposure on the outcome [9]. However, unmeasured confounding can still be present, thus, bias can still arise when comparing treatment strategies.

If the confounder affects the exposure and the exposure affects the confounder, there is

treatment-confounder feedback, but treatment-confounder feedback is not necessary in order to have time-varying confounding [9]. Robins shows that traditional methods cannot be used to correctly adjust for time-varying confounders and treatment-confounder feedback as they may induce bias and thus not provide a valid estimate of the causal effect even with sufficient longitudinal data [17].

In general, when the joint effect of the exposure components X_k can be estimated simultaneously and without bias, then valid estimation of the effect of treatment strategies is possible [9]. However, even when data on all time-varying confounders are available, this may not be possible to achieve using stratification [9].

2.4.2.1.4 Estimating IP Weights in the Time-Varying Case Recall that the denominator of the IP weights is an subject's probability of receiving the exposure, conditional on the subject's confounder values [9]. When we have time-varying exposure and confounders we need to generalize the IP weights (2.66) [9]. We can extend the definition of the IP weights in the cross-sectional case to the time-varying case by noting that the denominator of the IP weights is now an subject's probability of receiving their exposure history, conditional on the subject's confounder history.

Hernán and Robins state that the general form of the nonstabilized IP weights in the time-varying case is

$$W^{\bar{X}} = \prod_{k=0}^{K} \frac{1}{f(X_k|\bar{X}_{k-1}, \bar{L}_K)}, \tag{2.68}$$

for $k = 0, 1, \cdots K$ [9].

A pseudo-population is created when these IP weights are used where mean of $Y^{\bar{x}}$ is the same to that in the actual population, however, we have constant randomization probabilities at each time point k [9].

We need to estimate $f(X_k|\bar{X}_{k-1}, \bar{L}_K)$ from the data in observational studies [9]. We can fit a logistic regression model to estimate the conditional probability of a dichotomous exposure $P[X_k = 1|\bar{X}_{k-1}, \bar{L}_k]$ at each time k for high-dimensional data [9]. The estimates

$\hat{f}(X_k|\bar{X}_{k-1}, \bar{L}_K)$ from these models will then replace $f(X_k|\bar{X}_{k-1}, \bar{L}_K)$ in $W^{\bar{x}}$ [9]. The resulting estimates of $E[Y^{\bar{x}}]$ and $E[Y^{\bar{x}}] - E[Y^{\bar{x}'}]$ will be biased if these estimates are based on a misspecified logistic model [9].

A model that combines information from many strategies to help estimate a given $E[Y^{\bar{x}}]$ needs to be determined as the number of unknown quantities, $E[Y^{\bar{x}}]$, can be more than the sample size [9]. One can assume, under strategy $\bar{x}$, the effect of exposure history $\bar{x}$ on the mean outcome increases linearly as a function of the cumulative exposure $cum(\bar{x}) = \sum_{k=0}^{K} x_k$ [9]. Hernán and Robins state that this is shown in the *marginal structural mean model*

$$E[Y^{\bar{x}}] = \beta_0 + \beta_1 cum(\bar{x}) \tag{2.69}$$

for all $\bar{x}$, which is a more general version of the marginal structural mean model for cross-sectional exposures discussed in Section 2.4.1.1.3.

We have that the average causal effect of the time-varying exposure, $\bar{x}$ is measured by β_1 and that $\beta_1 \times cum(\bar{x})$ is equal to $E[Y^{\bar{x}}] - E[Y^{\bar{x}=\bar{0}}]$, which is the average causal effect [9].

As discussed in the cross-sectional case in Section 2.4.1.1.3, the ordinary linear regression model is fit in order to estimate the parameters of the marginal structural model as follows

$$E[Y|\bar{X}] = \theta_0 + \theta_1 cum(\bar{X}) \tag{2.70}$$

Recall that in the pseudo-population, weighted least squares with weights being estimates $W^{\bar{X}}$ are used [9]. Similarly to the cross-sectional case, Hernán and Robins determined that $\hat{\theta}_1$ is a consistent estimator of β_1 [9].

If the marginal structural model is misspecified, the estimates of $E[Y^{\bar{x}}]$ will be incorrect [9].

Chapter 3

Causal Methods for Cross-Sectional Case Control Methods

In this chapter, we introduce our proposed methods for a simple, single time point in a nested case-control study. In this, we can illustrate the estimator in closed form and then run a simulation in order to show how our methods work in practice.

3.1 Cross-Sectional Methods

3.1.1 Causal Methods

Suppose that exposure, also denoted treatment, X, outcome Y and a set of confounders L are observed for N individuals. Let $Y^{(x)}$ denote the possibly unobserved, potential outcome that would be observed if, possibly counterfactually, exposure X were set to level $x = 0$ (unexposed) or 1 (exposed). Let Y and L be binary, and X be randomized. Thus, the probability that our outcome is $Y = 1$ given different levels of exposure is

$$E[Y^{(x)}] = P[Y^{(x)} = 1|X = x]. \tag{3.1}$$

We have two models: $X = 1|L = 0, 1$.

For the most saturated model, we will have 4 different probabilities of $Y = 1$, $P(Y^{(1)}|L = 1), P(Y^{(0)} = 1|L = 1), P(Y^{(1)} = 1|L = 0), P(Y^{(0)} = 1|L = 0)$. We want to compute the effect of treatment. Recall that from 2.4.1.0.1, we have that $Y = XY^{(1)} + (1 - X)Y^{(0)}$. As X is binary and randomized, then $X \perp (Y^{(0)}, Y^{(1)})$, thus our expected value is $E[Y|X = 1] = E[Y^{(1)}|X = 1] = E[Y^{(1)}] = P[Y = 1|X = 1] = P[Y^{(1)} = 1]$.

Assume that $(Y^{(0)}, Y^{(1)}) \perp X|L$, where L is the confounder. We have that the average treatment effect of our exposure is

$$\text{Average Treatment Effect (ATE)} = E[Y^{(1)} - Y^{(0)}] = E[Y^{(1)}] - E[Y^{(0)}]. \tag{3.2}$$

Additionally, as $(Y^{(0)}, Y^{(1)}) \perp X$, we have that $E[Y|X = x] = xE[Y^{(1)}] + (1 - x)E[Y^{(0)}]$. Note that as L is a confounder, and X is randomized, it follows that L is independent of exposure. L is also a confounder of the effect of exposure, X, on the outcome, Y.

In order to account for the causal effects, we employ IP treatment weighting, as outlined in 2.4.1.1.2, we define our unstabilized weights as follows

$$W^X = \frac{1}{f(X|L)} = \begin{cases} \frac{1}{P[X=1|L]} & \text{treated,} \\ \\ \frac{1}{P[X=0|L]} = \frac{1}{1-P[X=1|L]} & \text{untreated.} \end{cases} \tag{3.3}$$

Our goal is to estimate $E[Y^{(x)}]$ in order to determine the ATE. Our estimator is defined as follows

$$\hat{E}[Y^{(x)}] = \hat{P}[Y^{(x)} = 1|X = x]. \tag{3.4}$$

More generally, to obtain estimates of $P[X = 1|L]$, we often fit a logistic regression model for the probability of exposure with the confounders included as covariates. To estimate $\hat{E}[Y^{(x)}]$, we fit a saturated linear mean model $E[Y|X] = \theta_0 + \theta_1 X$ by weighted least squares, with the individuals weighted by their estimated weights.

3.1.2 Sampling Methods

Recall, our goal is to use finite population methods in a causal setting in order to analyze nested case-control sampled data. We now combine the causal methods and finite population methods from Chapter 2. In a nested case-control analysis, we normally include all of the cases (those with outcome $Y = 1$ in our sample) and then take a random sample of the controls. In this thesis, we focus on a SRS of all controls.

Let i denote an individual, then define $\pi_i = \mathrm{P}(\text{unit } i \text{ in the sample})$. If individual i is a case, then $\pi_i = 1$. Let j denote another individual, if j is a control, then $\pi_j = n_c/N_c$, where N_c is the total number of controls and n_c is the sample number of controls.

We use the Horvitz-Thompson estimator, defined as follows in Section 2.1.2.2

$$\hat{t}_{HT} = \sum_{i \in \mathcal{S}} \frac{t_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{t_i}{\pi_i},$$

where $\mathcal{S}$ is the sample from the cohort and $Z_i = 1$ if individual i is in the sample, and 0 otherwise.

We can rewrite the Horvitz–Thompson estimator for our problem using the sampling weights. The first-stage sampling weight for individual i in the nested case-control sample is

$$w_i = \frac{1}{\pi_i} = \begin{cases} 1 & \text{case,} \\ \frac{N_c}{n_c} & \text{control.} \end{cases} \tag{3.5}$$

Thus, the HT estimator for the cohort total is $\hat{t}_{HT} = \sum_{i \in \mathcal{S}} w_i \hat{t}_i$. Note that w_i is our sampling weight.

We want to apply the above finite population methods to our causal cohort. Recall our IP treatment weights in equation (3.3). Let U be our entire cohort, and let S be our sample from the cohort. Additionally, let $\mathbb{1}_{\{X_i = x, L_i = l_i\}}$ be the indicator function which is equal to 1 if the condition is satisfied (if $X_i = x$ and $L_i = l$) and 0 otherwise (if $X \neq x$ and/or $L_i \neq l_i$) for individual i. Recall that we're restricting in this case to binary L. We have that the

probability of exposure for each level of L for the entire cohort is

$$P[X = x|L = l] = \frac{\sum_{i\in U} \mathbb{1}_{\{X_i=x,L_i=l_i\}}}{\sum_{i\in U} \mathbb{1}_{\{L_i=l_i\}}}. \tag{3.6}$$

Note that equation (3.6) is known as the propensity score. In order to determine our weights, we want to estimate the probability of exposure for each level of L. By doing so, we determine the probability of exposure for each level of L for the sample from the cohort and multiply that by w_i to account for the sampling weights, thus giving us our HT estimator.

$$\hat{P}[X = x|L = l] = \frac{\sum_{i\in S} \mathbb{1}_{\{X_i=x,L_i=l_i\}w_i}}{\sum_{i\in S} \mathbb{1}_{\{L_i=l_i\}w_i}}. \tag{3.7}$$

By plugging in the values of w_i, we get the following

$$\hat{P}[X = x|L = l] = \frac{\sum_{i\in S_{\text{case}}} \mathbb{1}_{\{X_i=x,L_i=l_i\}} \cdot 1 + \sum_{i\in S_{\text{control}}} \mathbb{1}_{\{X_i=x,L_i=l_i\}} \cdot \frac{N_c}{n_c}}{\sum_{i\in S_{\text{case}}} \mathbb{1}_{\{L_i=l_i\}} \cdot 1 + \sum_{i\in S_{\text{control}}} \mathbb{1}_{\{L_i=l_i\}} \cdot \frac{N_c}{n_c}}. \tag{3.8}$$

In order to simplify the notation and make it more understandable, we define the following:

$$\sum_{i\in S_{\text{case}}} \mathbb{1}_{\{X_i=x,L_i=l_i\}} = n_{1xl} = \text{number of cases with } X = x \text{ and } L = l \text{ in the sample,}$$

$$\sum_{i\in S_{\text{control}}} \mathbb{1}_{\{X_i=x,L_i=l_i\}} = n_{0xl} = \text{sample number of controls with } X = x \text{ and } L = l \text{ in the sample,}$$

$$\sum_{i\in S_{\text{case}}} \mathbb{1}_{\{L_i=l_i\}} = n_{1\bullet l} = \text{number of cases with } L = l \text{ in the sample,}$$

$$\sum_{i\in S_{\text{control}}} \mathbb{1}_{\{L_i=l_i\}} = n_{0\bullet l} = \text{number of controls with } L = l \text{ in the sample,}$$

$$n_c = n_{0\bullet\bullet} = \text{number of controls in the sample.}$$

Therefore, our probability can be written as

$$\hat{P}[X = x|L = l] = \frac{n_{1xl} + \frac{N_c}{n_{0\bullet\bullet}} n_{0xl}}{n_{1\bullet l} + \frac{N_c}{n_{0\bullet\bullet}} n_{0\bullet l}}. \tag{3.9}$$

This estimated probability accounts for the sampling as well as the causal effects.

Recall that our exposure is dichotomous, $X = 0, 1$, and that we have two levels of our confounder $L = 0, 1$. Therefore, we have the following estimators

$$\hat{P}[X = 1|L = 1] = \frac{\sum_{i \in S} \mathbb{1}_{\{X_i=1, L_i=1\}} w_i}{\sum_{i \in S} \mathbb{1}_{\{L_i=1\}} w_i} = \frac{n_{111} + \frac{N_c}{n_{0\bullet\bullet}} n_{011}}{n_{1\bullet 1} + \frac{N_c}{n_{0\bullet\bullet}} n_{0\bullet 1}},$$

$$\hat{P}[X = 1|L = 0] = \frac{\sum_{i \in S} \mathbb{1}_{\{X_i=1, L_i=0\}} w_i}{\sum_{i \in S} \mathbb{1}_{\{L_i=0\}} w_i} = \frac{n_{110} + \frac{N_c}{n_{0\bullet\bullet}} n_{010}}{n_{1\bullet 0} + \frac{N_c}{n_{0\bullet\bullet}} n_{0\bullet 0}},$$

$$\hat{P}[X = 0|L = 1] = 1 - \hat{P}[X = 1|L = 1] = 1 - \frac{\sum_{i \in S} \mathbb{1}_{\{X_i=1, L_i=1\}} w_i}{\sum_{i \in S} \mathbb{1}_{\{L_i=1\}} w_i} = \frac{n_{101} + \frac{N_c}{n_{0\bullet\bullet}} n_{001}}{n_{1\bullet 1} + \frac{N_c}{n_{0\bullet\bullet}} n_{0\bullet 1}},$$

$$\hat{P}[X = 0|L = 0] = 1 - \hat{P}[X = 1|L = 0] = 1 - \frac{\sum_{i \in S} \mathbb{1}_{\{X_i=1, L_i=0\}} w_i}{\sum_{i \in S} \mathbb{1}_{\{L_i=0\}} w_i} = \frac{n_{100} + \frac{N_c}{n_{0\bullet\bullet}} n_{000}}{n_{1\bullet 0} + \frac{N_c}{n_{0\bullet\bullet}} n_{0\bullet 0}}.$$

Now, we can use our estimator for the probability of exposure for each level of L, in order to estimate the probability that our outcome is $Y = 1$ given different levels of exposure $(X = 0, 1)$.

Recall that we want to determine $E[Y^{(x)}]$, let us consider

$$E[Y^{(x)}] = P[Y^{(x)} = 1|X = x] = \frac{\sum_{i \in U} \mathbb{1}_{\{Y_i=1, X_i=x\}}}{N}, \tag{3.10}$$

where N is the number of people in the cohort. However, (3.10) does **not** work, as it will be biased due to confounding. Therefore, let us adjust for confounding using IP treatment weighting in the cohort as follows:

$$E[Y^{(x)}] = P[Y^{(x)} = 1|X = x] = \frac{\sum_{i \in U} \frac{\mathbb{1}_{\{Y_i=1, X_i=x\}}}{P[X_i=x|L_i=l_i]}}{N}. \tag{3.11}$$

Note, this is on the whole cohort. Thus, let us define our estimator accounting for the sub-sampling as follows

$$\hat{E}[Y^{(x)}] = \hat{P}[Y^{(x)} = 1|X = x] = \frac{\sum_{i \in S} \frac{\mathbb{1}_{\{Y_i=1, X_i=x\}}}{\hat{P}[X_i=x|L_i=l_i]} w_i}{N}. \tag{3.12}$$

When we plug in the values of w_i, we get the following

$$\hat{E}[Y^{(x)}] = \frac{\sum_{i \in S_{\text{case}}} \left[\frac{\mathbb{1}_{\{Y_i=1, X_i=x\}}}{\hat{P}[X_i=x|L_i=l_i]}\right] \cdot 1 + \sum_{i \in S_{\text{control}}} \left[\frac{\mathbb{1}_{\{Y_i=1, X_i=x\}}}{\hat{P}[X_i=x|L_i=l_i]}\right] \cdot \frac{N_c}{n_c}}{N}. \tag{3.13}$$

Let

$$\sum_{i \in S_{\text{case}}} \mathbb{1}_{\{Y_i=1, X_i=x\}} = n_{1x\bullet} = \text{number of cases with } Y=1 \text{ and } X=x \text{ in the sample,}$$

$$\sum_{i \in S_{\text{control}}} \mathbb{1}_{\{Y_i=1, X_i=x\}} = n_{1x\bullet} = \text{number of controls with } Y=1 \text{ and } X=x \text{ in the sample} = 0.$$

Recall that all controls have $Y=0$, therefore $\sum_{i \in S_{\text{control}}} \left[\frac{\mathbb{1}_{\{Y_i=1, X_i=x\}}}{\hat{P}[X_i=x|L_i=l_i]} \right] = 0$. Therefore, we obtain

$$\hat{E}[Y^{(x)}] = \frac{\sum_{i \in S_{\text{case}}} \left[\frac{\mathbb{1}_{\{Y_i=1, X_i=x\}}}{\hat{P}[X_i=x|L_i=l_i]} \right]}{N} = \frac{1}{N} \sum_{i \in S_{\text{case}}} \cdot \left[\frac{\mathbb{1}_{\{Y_i=1, X_i=x\}}}{\hat{P}[X_i=x|L_i=l_i]} \right]. \tag{3.14}$$

By plugging in our estimator $\hat{P}[X_i=x|L_i=l_i]$ we thus obtain

$$\begin{aligned} \hat{E}[Y^{(x)}] = \hat{P}[Y^{(x)}=1|X=x] &= \frac{1}{N} \cdot \sum_{i \in S_{\text{case}}} \frac{\mathbb{1}_{\{Y_i=1, X_i=x\}}}{\left[\frac{\sum_{i \in S} \mathbb{1}_{\{X_i=x, L_i=l_i\}} w_i}{\sum_{i \in S} \mathbb{1}_{\{L_i=l_i\}} w_i} \right]} \\ &= \frac{1}{N} \cdot \sum_{i \in S_{\text{case}}} \frac{\mathbb{1}_{\{Y_i=1, X_i=x\}}}{\left[\frac{n_{1xl} + \frac{N_c}{n_{0\bullet\bullet}} n_{0xl}}{n_{1\bullet l} + \frac{N_c}{n_{0\bullet\bullet}} n_{0\bullet l}} \right]}. \end{aligned} \tag{3.15}$$

As our exposure is dichotomous, we obtain the following estimators:

$$\begin{aligned} \hat{E}[Y^{(1)}] = \hat{P}[Y^{(1)}=1|X=1] &= \frac{1}{N} \sum_{i \in S_{\text{case}}} \cdot \left[\frac{\mathbb{1}_{\{Y_i=1, X_i=1\}}}{\hat{P}[X_i=1|L_i=l_i]} \right] \\ &= \frac{1}{N} \cdot \frac{n_{11\bullet}}{\left[\frac{n_{1xl} + \frac{N_c}{n_{0\bullet\bullet}} n_{0xl}}{n_{1\bullet l} + \frac{N_c}{n_{0\bullet\bullet}} n_{0\bullet l}} \right]}. \end{aligned}$$

$$\begin{aligned} \hat{E}[Y^{(0)}] = \hat{P}[Y^{(0)}=1|X=0] &= \frac{1}{N} \sum_{i \in S_{\text{case}}} \cdot \left[\frac{\mathbb{1}_{\{Y_i=1, X_i=0\}}}{\hat{P}[X_i=0|L_i=l_i]} \right] \\ &= \frac{1}{N} \sum_{i \in S_{\text{case}}} \cdot \left[\frac{\mathbb{1}_{\{Y_i=1, X_i=0\}}}{1 - \hat{P}[X_i=0|L_i=l_i]} \right] \\ &= \frac{1}{N} \cdot \frac{n_{11\bullet}}{\left[1 - \frac{n_{1xl} + \frac{N_c}{n_{0\bullet\bullet}} n_{0xl}}{n_{1\bullet l} + \frac{N_c}{n_{0\bullet\bullet}} n_{0\bullet l}} \right]}. \end{aligned}$$

We have thus determined our estimator as a function of counts for the probability that our outcome is $Y=1$ given different levels of exposure ($X=0,1$), taking into account different levels of our confounder ($L=0,1$).

Recall that our goal is to estimate the Average Treatment Effect (ATE) (2.62). Therefore we want to estimate the ATE for the cohort and compare that to the ATE for the sample from the cohort. We will estimate two different ATE's for the sample from the cohort, the difference is the estimation of the propensity score, $P[X = x|L = l]$. One will use the estimated propensity score from the full cohort, the other will use the estimated propensity score determined from the sampled cohort.

Additionally, we can compute the true value of the ATE, and compare that to our three estimates. For the true value of the ATE, we need to calculate the true value of $E[Y^{(x)}]$, which is defined as follows

$$\begin{aligned} E_{\text{truth}}[Y^{(x)}] &= P[Y^{(x)} = 1|X = x, L = 1]P[L = 1] + P[Y^{(x)} = 1|X = x, L = 0]P[L = 0] \\ &= P[Y^{(x)} = 1|L = 1]P[L = 1] + P[Y^{(x)} = 1|L = 0]P[L = 0]. \end{aligned} \tag{3.16}$$

Recall that L is a confounder, and as we are conditioning on it we have that $P[Y^{(x)} = 1|X = x, L = l] = P[Y^{(x)} = 1|L = l]$.

Note that in our case, as we include all of the cases in the cohort, the cases have probability 1 of being sampled, and the controls are chosen via a SRS, thus the controls have probability n_c/N_c of being sampled (where n_c = sample number of controls, N_c = cohort number of controls). Therefore if i is a case and j is a case we have that $\hat{V}_{SYG} = 0$ as $\pi_i\pi_k = \pi_{ik}$, and if i is a case and j is a control, or vice-versa we also have $\hat{V}_{SYG} = 0$ as $\pi_i\pi_k = \pi_{ik}$. Therefore, we have that the covariance is 0, and we have 0 variability as all of the cases are included in the sample. Thus, this simplifies our analysis. However, if we use the with-replacement variance estimator, (2.12), it will never equal zero, thus, it will actually get more complicated. Therefore, in our case the Sen-Yates-Grungy (SYG) estimator of the variance would be recommended, (2.11), as it simplifies if we have two cases, or a case and a control.

In the following section, we will create a synthetic cohort, and then using our methods, we will sample from it and then compute the ATE for each method.

3.2 Simulation

3.2.1 Simulation Background

For the simulation studies we used the `simcausal` `R` package [24, 25]. In our case, it was a very useful vehicle for setting up our simulation, and we were able to easily extract both true and estimated causal effects and the package enabled us to conduct transparent simulation studies that can be easily reproduced. Another benefit of using this package is that it is very easy to adapt our simulation, by adding time points, nodes, or adapting dependencies.

For the cross-sectional simulation, we started by replicating an example in the "Understanding Marginal Structural Models for Time-Varying Exposures: Pitfalls and Tips" paper by Shinozaki and Suzuki [22]. In their paper, they have an example with one time point. Their example supposes that exposure X_i, outcome Y_i, and set of covariates L_i are observed for individual $i = 1, \cdots, n$. Let Y_i^x denote the possibly unobserved, potential outcome that would be observed if, possibly counterfactually, exposure X_i were set to level $x = 0$ (unexposed) or 1 (exposed). Note, from now on they omit the subscript i to avoid confusion. Then, the average causal effect of exposure X on outcome Y may be defined as $E[Y^1] - E[Y^0]$, which compares counterfactual expectations (or risks for a binary outcome) of Y_1 and Y_0 in the same cohort along the difference-scale. We can illustrate the causal effects from their example in a DAG as follows:

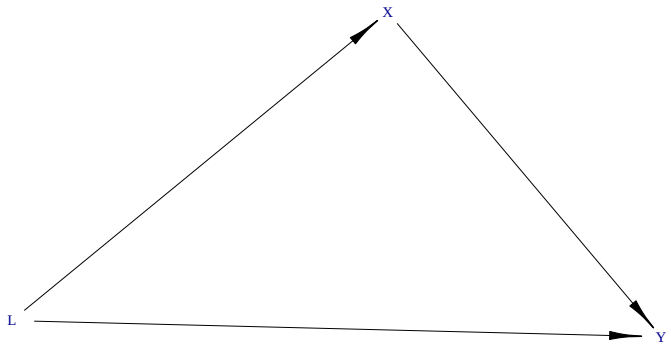


Figure 3.1 – DAG for Cross-Sectional Example as Generated from the `simcausal` Package

In their hypothetical cohort they assigned a sample size of $n = 1,240$. The following table (Table 1 from the paper), illustrates their hypothetical cohort. In our simulation we use the proportion of combinations from Table 3.1 as population level probabilities for simulation.

Table 3.1 – Cross-Sectional Hypothetical Cohort Data from the Shinozaki and Suzuki Paper used in our Simulation in 3.2.2

Stratum			
L	**X**	**P(X|L)**	**E[Y|X, L]**
1	1	0.226	0.75
1	0	0.581	0.6
0	1	0.145	0.333
0	0	0.048	0.25

In our simulation, we chose a cohort size of $N = 100,000$. The following probabilities, $P(L), P(X|L),$ and $P(Y = 1|L, X)$, were the same ones used in the paper, as shown in Table 1 in the paper. We then defined our DAG in the `simcausal` package using the structure

in Figure 3.1 and the probabilities defined in the Shinozaki and Suzuki paper and ran our simulation.

3.2.2 Simulation Results

In our simulation, we had two different versions: one keeping the cohort constant with 50 simulations, and one having 10 different cohorts with 50 simulations. We found that 50 simulations gave very stable results and allowed us to explore more scenarios.

In our results, we have the values of $E[Y^{X=1}]$ and $E[Y^{X=0}]$. We calculated the true value, the estimated value using the IP treatment weights. We also calculated the sampled value using the true propensity score as well as the sampled value estimating the propensity score from the sampled cohort. Additionally, we calculated the difference from the truth, which is the estimated value subtracted from the true value as well as the percentage difference from the truth.

Recall that the goal of our simulation is to determine how much you lose by sub-sampling. Essentially, we want to know if you achieve a close enough level of accuracy without analyzing the full cohort prospectively. For very large datasets, by not using the full cohort, one reduces analytical complexity. Note that we showed theoretically that our results using sub-sampling are unbiased due to the properties of IP treatment weighting. We now want to observe how our results vary from simulation to simulation.

To better understand our results for each treatment scheme, we looked at the number of and the proportion of participants for each treatment scheme and outcome on the simulated cohort. This is outlined in Table 3.2 below. Recall that the size of our cohort is $N = 100,000$.

We then took an average of the estimates over each replication of the sub-sampling and dataset. Recall, we did 50 replications for 10 datasets. Additionally, we looked at the percentage difference of our two sampling estimates from the estimate on the whole dataset.

Table 3.2 – Summary of Individuals for each Treatment Scheme and Outcome in the Cross-Sectional Case

X	Y	**Count**	**Proportion**
0	0	74,677	0.74677
0	1	4,161	0.04161
1	0	19,621	0.19621
1	1	1,541	0.01541

Table 3.3 – Estimated Mean Counterfactual per Treatment for each Estimator in the Cross-Sectional Case

Treatment	**True**	**Cohort Estimate**	**Sampled (Full Prop Score)**	**Sampled (Est Prop Score)**	**Sampled (Full Prop Score) % Difference from Est**	**Sampled (Est Prop Score) % Difference from Est**
$X = 0$	0.0543	0.0532	0.0532	0.0532	1.00%	1.00%
$X = 1$	0.0687	0.0707	0.0708	0.0708	1.00%	1.00%

From Table 3.3, we can see that the mean values of the estimate on the whole dataset, as well as the two sampled estimates are very close to the true value. We can also see that the percentage difference of each sampling estimate compared to the full data estimate only differs by 1.00%, thus, indicating that we do not deviate much from the estimate by sampling, in both the full and estimated propensity score case. Additionally, we can see that both sampling estimates perform almost as well as the estimate on the full data and that both sampling estimates give the same estimate.

We then look at the average standard error of the estimates over each replication and dataset.

Table 3.4 – Estimated Standard Error per Treatment for each Estimator in the Cross-Sectional Case

Treatment	**True**	**Cohort Estimate**	**Sampled (Full Prop Score)**	**Sampled (Est Prop Score)**	**Sampled (Full Prop Score) % Difference from Est**	**Sampled (Est Prop Score) % Difference from Est**
$X = 0$	0	0.000793	0.000782	0.000782	0.985%	0.985%
$X = 1$	0	0.00178	0.00170	0.00170	0.956%	0.956%

From Table 3.4, we can see that the standard error for the full estimate as well as both sampled estimates are very similar. This is also reflected in the percentage difference of each sampled estimate from the estimator, as the percentage differences are very small. Additionally, we can see that the standard error is higher for the $X = 1$ treatment scheme, however note that in Table 3.2, we can see that there is a smaller proportion of individuals who underwent the $X = 1$ treatment scheme, which naturally would give us a higher standard error. From Table 3.4 we can deduce that we only get $\sim 1.00\%$ more standard error from sampling, meaning that by sampling, we do not lose much accuracy. Additionally, there is no real difference between our sampling estimates using the full versus the estimated propensity score.

In order to fully visualize our estimated values, we have the following boxplot in Figure 3.2 to show the true and estimated values for each treatment scheme.

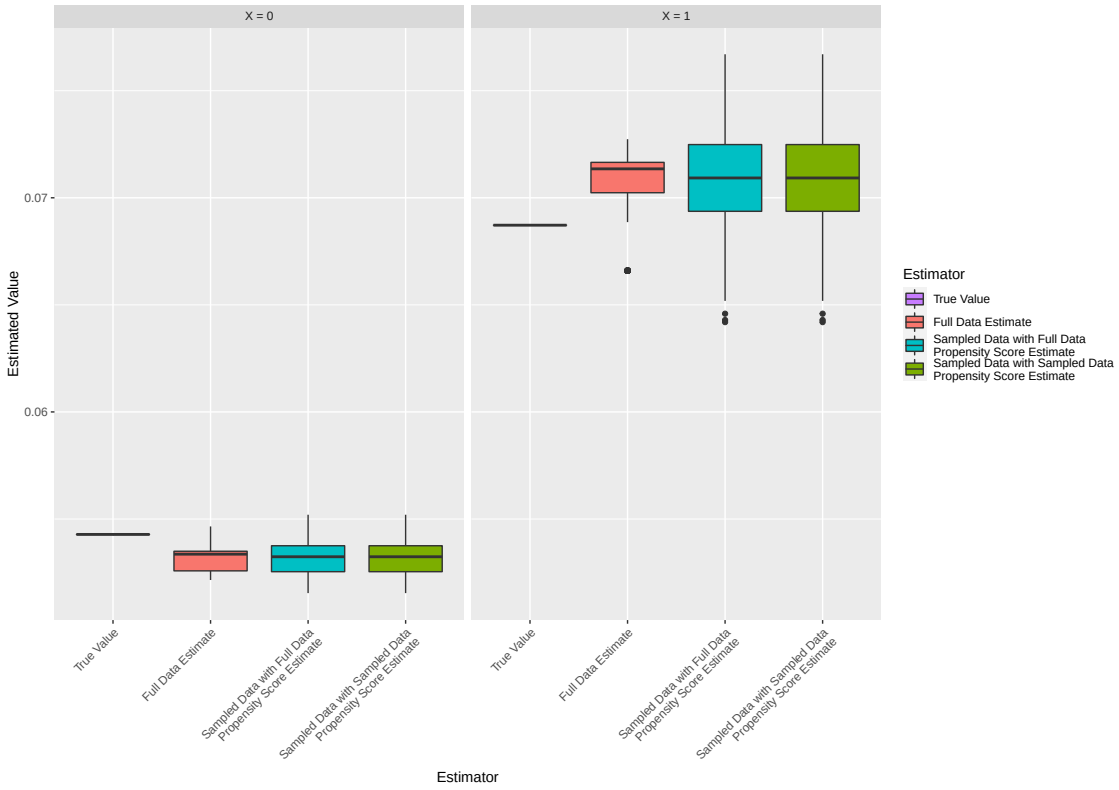


Figure 3.2 – Boxplot of the Estimates in the Cross-Sectional Case

Boxplot showing the estimated value for the truth as well as for our three estimators: the full data estimator, the sampled data with the full data propensity score estimator and the sampled data with the sampled data propensity score. The boxplots are shown for each treatment regime: $X = 0$ and $X = 1$. Recall that the estimate is of $E[Y^{(X=x)}]$.

As our aim is to determine the quality of our estimates, and to see if we can maintain that quality by sampling, we visualized the percentage difference from the truth for each estimate. This is shown in Figure 3.3.

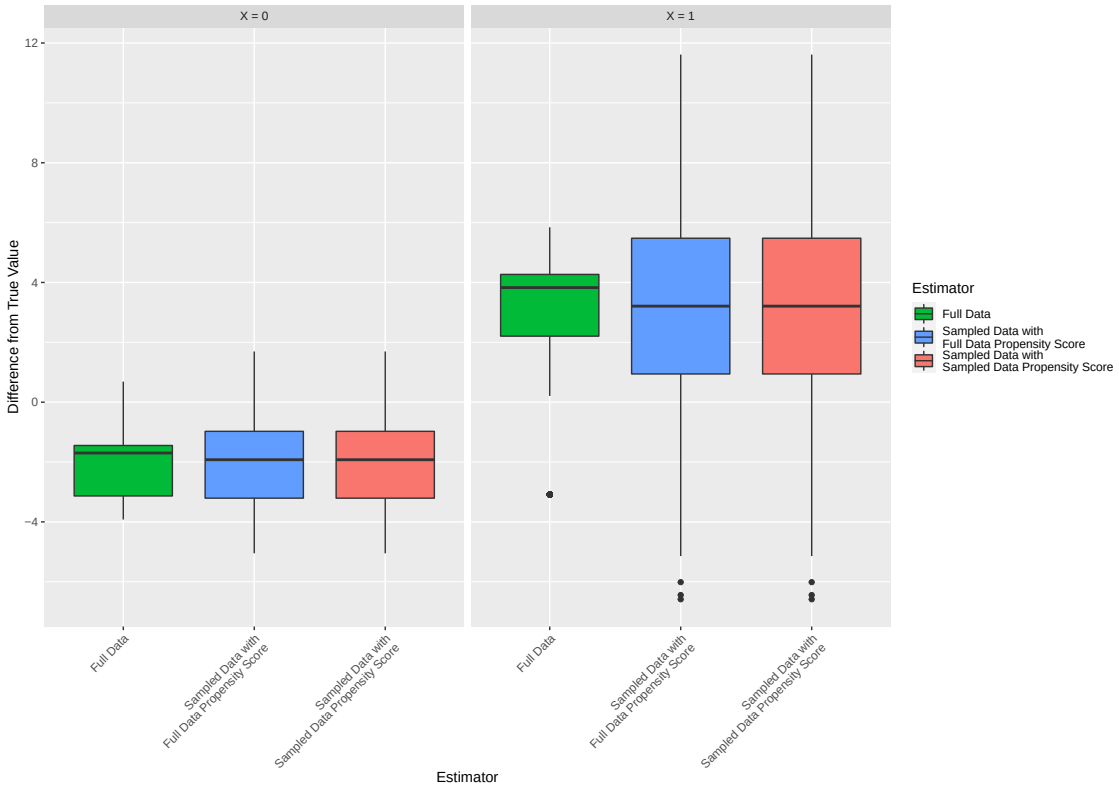


Figure 3.3 – Percentage Difference from the Truth in the Cross-Sectional Case

Boxplot showing the percentage difference from the truth of three estimators: the full data estimator, the sampled data with the full data propensity score estimator and the sampled data with the sampled data propensity score. This was calculated by the following formula: $100 \times \frac{\text{Estimated Value} - \text{True Value}}{\text{True Value}}$.

From Figure 3.3, we can see that our estimates vary more from the truth in the $X = 1$ treatment scheme, and are very close to the truth in the $X = 0$ treatment scheme. Note that both arms in the sub-sampled data have the same number of observations. However, for our data the inverse probability treatment weights for the exposure arm are more unstable. Additionally, we can see that the difference of our two sampling estimates from the truth are very close to the difference of the estimate from the full dataset from the truth. Thus, this shows us that our sampling performs almost the same as our estimate from the full dataset, meaning that we are not losing accuracy by sampling as well as we are not losing accuracy

by using the estimated propensity score versus the propensity score determined on the whole dataset.

As we did 50 replicates on 10 datasets, we want to see if there is a lot of variation between datasets. In Figure 3.4 we can see the average value over all the replications for each estimate over each dataset.

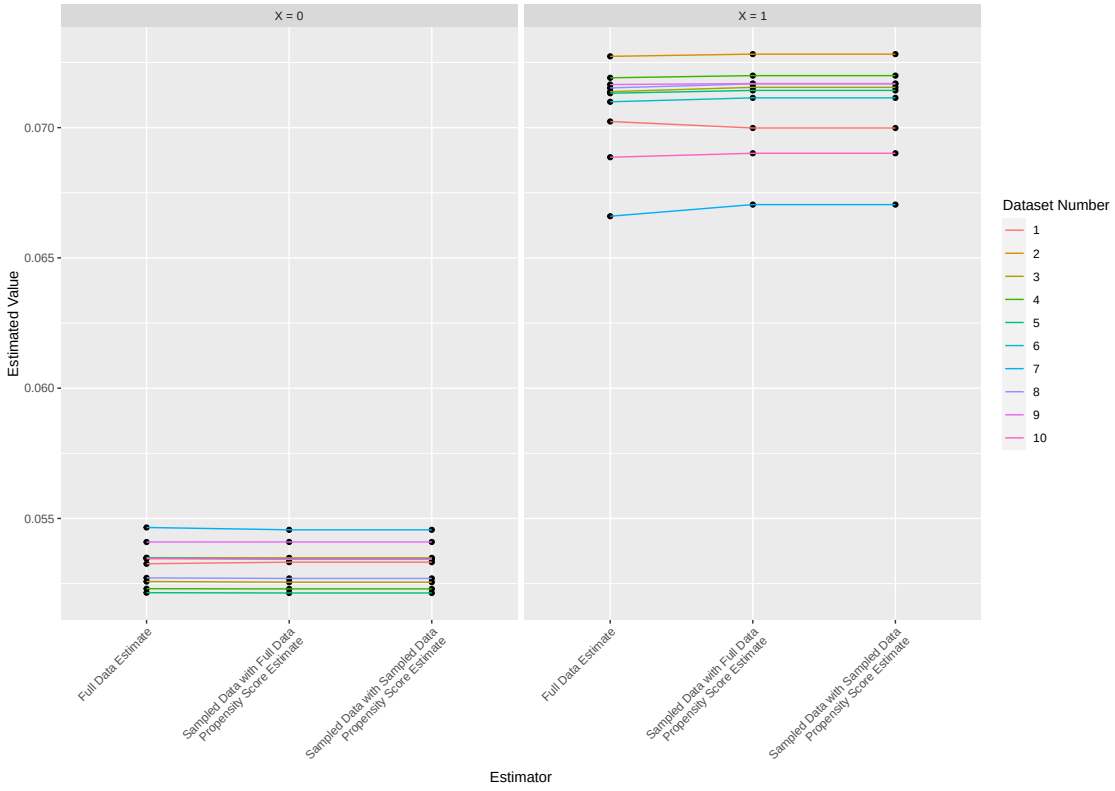


Figure 3.4 – Average Estimates per Dataset in the Cross-Sectional Case

Graph showing the estimated values for our three estimators for each dataset. For each of 10 datasets, 50 replications were run. Each line is for each if the 10 datasets in order to see how the estimated value changes for each estimator.

We can see in Figure 3.4 that there is more variation between datasets than between estimators.

Additionally, we want to determine the effect of the sample size in our sub-sampling. We initially chose to have the same number of cases as controls. In the cross-sectional case we

had $5,702$ cases out of a cohort of $N = 100,000$, thus we took a SRS of $n = 5,702$ of the $94,298$ controls. We then ran our simulation by doubling the number controls, thus for the $5,702$ cases out of a cohort of $N = 100,000$, thus we took a SRS of $n = 2 \times 5,702 = 11,404$ of the $94,298$ controls. By doubling the number of controls in the sample, the standard error per treatment per estimator is slightly smaller. However, similar to the results in Table 3.4 we still obtain only $\sim 1.00\%$ more standard error from sampling, even though we are including twice as many controls in the analysis. However, we do have a slightly lower percentage difference of the estimated values from the truth when doubling the number of controls.

These results do not come as a surprise as it is known that the number of cases drive the accuracy of the estimator, not the number of controls. Hsieh, Bloch and Larsen [12] determine, in equation (1) of their paper, that the required total sample size is determined by

$$n = (Z_{1-\alpha/2} + Z_{1-\beta})/[P1(1 - P1)\beta^{*2}], \tag{3.17}$$

where n is the required total sample size, β^* is the effect size to be tested, $P1$ is the event rate at the mean of treatment X, and Z_u is the upper uth percentile of the standard normal distribution [12]. We can see that they determine that the required total sample size depends most on the event rate, which means that it depends on the number of cases, which in our case is $Y = 1$. Due to the nested case-control sampling that we employ, it is such that all of the individuals who experience the event, the cases, are included in the sample, with probability 1 of being included, and then we take a SRS of the controls. Therefore, by doubling the number of controls, the number of cases remains unchanged, thus only slightly affecting the results.

Chapter 4

Causal Methods for Time-Varying Case Control Methods

In this chapter, we extend the cross-sectional outcome methods introduced in Chapter 3 to the time-varying case. We first introduce the methods mathematically. Then, as in the previous chapter, we run a simulation study to illustrate how our methods work in practice.

4.1 Time-Varying Methods

4.1.1 Causal Methods

Recall in Chapter 3, we assumed exposure, also denoted treatment, X, outcome Y and a set of confounders L for N individuals. Let us extend this cross-sectional example to multiple time-points. We define a time-varying setting with multiple time points $k = 0, 1, \cdots, K$. We observe a time-varying dichotomous treatment X_k that may change at every k time point of follow-up, where $k = 0, 1, 2, \cdots, K$ with K being the last time point. Denote $\bar{X}_k = (X_0, X_1, \cdots, X_k)$ as the history of exposure from time 0 to time k and $\bar{L}_t = (L_0, L_1, \cdots, L_k)$ as the confounder history from time 0 to time k. Let $\bar{X}$ denote the entire treatment history

and $\bar{L}$ denote the entire confounder history.

Recall from Chapter 2 that in a time-varying setting with multiple time points $k = 0, 1, \cdots, K$, the IP treatment weighting formula based on L for the counterfactual mean $E[Y^{\bar{x}}]$ is the average of Y among subjects with $\bar{X} = \bar{x}$ in a stabilized pseudo-population constructed by weighting each subject by their subject-specific stabilized IP treatment weights [8]. We define our unstabilized weights as follows:

$$W^{\bar{X}} = \prod_{k=0}^{K} \frac{1}{f(X_k|\bar{X}_{k-1}, \bar{L}_k)}. \tag{4.1}$$

We define the counterfactual mean, $E[Y^{\bar{x}}]$, as follows,

$$\begin{aligned} E[Y^{\bar{x}}] &= \beta_0 + \beta_1 * cum(\bar{x}) \\ &= \beta_0 + \beta_1 * \sum_{k=0}^{K} x_k. \end{aligned} \tag{4.2}$$

It follows that average causal effect, $E[Y^{\bar{x}}] - E[Y^{\bar{x}=\bar{0}}]$, is equal to $\beta_1 * cum(\bar{x})$ [9].

By extending what we had in the cross-sectional case we have that the alternative expression of $E[Y^{\bar{x}}]$

$$E\left[\left(\prod_{k=0}^{K} \frac{\mathbb{1}_{\{\bar{X}=\bar{x}\}}}{P(X_k|\bar{X}_{k-1}, \bar{L}_k)}\right) * Y\right]. \tag{4.3}$$

4.1.2 Sampling Methods

We then will apply case-control sampling to the dataset, as we did in the cross-sectional case. That is, we included every case in our sample and then took a SRS of the controls with n_c = sample number of controls and N_c = total number of controls. Thus, we had the following probabilities of inclusion, where π_i is the probability that unit i is in the cohort:

$$\pi_i = \begin{cases} 1 & \text{case,} \\ \frac{n_c}{N_c} & \text{control.} \end{cases} \tag{4.4}$$

Thus, the sampling weights are:

$$w_i = \frac{1}{\pi_i} = \begin{cases} 1 & \text{case,} \\ \frac{N_c}{n_c} & \text{control.} \end{cases} \tag{4.5}$$

As in the single time point case, our weights were the product of the sampling weights and the IP treatment weights giving us a HT estimator. It's the same in the time-varying case, thus, our weights then are:

$$W_i = \begin{cases} \prod_{k=0}^{K} \frac{1}{\hat{f}(X_k|\bar{X}_{k-1},\bar{L}_k)} & \text{case,} \\ \prod_{k=0}^{K} \frac{1}{\hat{f}(X_k|\bar{X}_{k-1},\bar{L}_k)} * \frac{N_c}{n_c} & \text{control.} \end{cases} \tag{4.6}$$

As in Chapter 3, we need to estimate the propensity score $f(X_k|\bar{X}_{k-1},\bar{L}_k) = P(X_k|\bar{X}_{k-1},\bar{L}_k)$, this will be done by we fitting a logistic regression model for the probability of exposure with the exposure and confounder history included as covariates.

We then applied the weights in (4.6) to our sample from the cohort in order to determine the treatment effects as we did in the full cohort case. Thus, we get that the estimate of the counterfactual mean is

$$\hat{E}[Y^{\bar{x}}] = \hat{E}\left[\sum_{i\in S} w_i \left(\prod_{k=0}^{K} \frac{\mathbb{1}_{\{\bar{X}=\bar{x}\}}}{\hat{P}(X_k|\bar{X}_{k-1},\bar{L}_k)}\right) * Y\right]. \tag{4.7}$$

We then want to estimate the counterfactual mean of each treatment scheme, this can give is the ATE for each treatment scheme. Therefore we want to estimate the counterfactual mean of each treatment scheme for the cohort and compare that to the counterfactual mean of each treatment scheme for the subsampled cohort. We will estimate two different counterfactual means of each treatment scheme for the sample from the cohort, the difference between the estimators is the estimation of the propensity score, $P(X_k|\bar{X}_{k-1},\bar{L}_k)$. One will use the estimated propensity score from the full cohort, the other will use the estimated propensity score determined from the sampled cohort. Additionally, we will determine the true value of the counterfactual mean of each treatment scheme and compare it to our three estimates.

In the following section, we will create a cohort, and then using our methods, we will sample it and then determine the counterfactual mean of each treatment scheme for each method.

4.2 Simulation

4.2.1 Simulation Background

As with the cross-sectional time case, we used the `simcausal` `R` package [24, 25].

For the time-varying simulation, we decided to start by replicating the simple example in the "Understanding Marginal Structural Models for Time-Varying Exposures: Pitfalls and Tips" paper by Shinozaki and Suzuki [22]. In their paper, they have an example with two time points. At baseline they do not include a confounder and just have treatment (we will denote it by X_0) which is randomized. At time point 1, there is a confounder (L_1) and another treatment (X_1). We can illustrate the causal effects from their example in a DAG as follows:

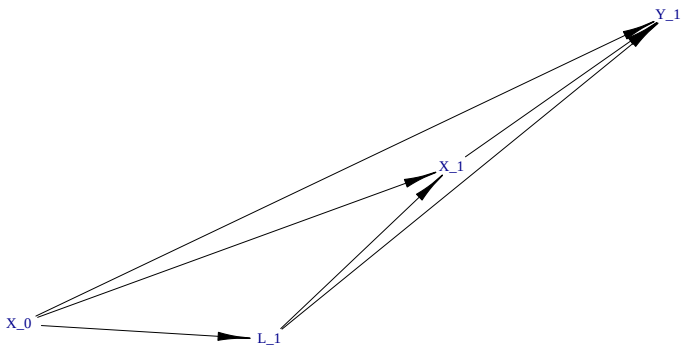


Figure 4.1 – DAG for Time-Varying Example as Generated from the `simcausal` Package

In their hypothetical cohort, they had a sample size of $n = 15,000$. Note that they de-

note treatment at baseline by X_1, and the confounder and treatment at time point 2 by X_2, and L_2, respectively. The following table (which is Table 4 from the paper), illustrates their hypothetical cohort. In our simulation we use the proportion of combinations from Table 4.1.

Table 4.1 – Time-Varying Hypothetical Cohort Data from the Shinozaki and Suzuki Paper used in our Simulation in 4.2.2

$\mathbf{X_1}$	$\mathbf{L_2}$	$\mathbf{X_2}$	**N**	$\mathbf{Y=1}$	$\mathbf{P(X_1)}$	$\mathbf{P(X_2|X_1,L_2)}$
1	1	1	720	576	0.3	0.8
1	1	0	180	108	0.3	0.2
1	0	1	1,800	720	0.3	0.5
1	0	0	1,800	900	0.3	0.5
0	1	1	5,670	4,536	0.7	0.9
0	1	0	630	567	0.7	0.1
0	0	1	840	294	0.7	0.2
0	0	0	3,360	1,008	0.7	0.8

In our simulation, we chose a sample size of $N = 100,000$. The following probabilities, $P(X_0), P(L_1|X_0), P(X_1|L_1, X_0)$ and $P(Y = 1|X_0, L_1, X_1)$, were the same ones used in the paper, as shown in Tables 2, 3 and 4 in the paper. We then defined our DAG in the `simcausal` package using the structure in Figure 4.1 and the probabilities defined in the Shinozaki and Suzuki paper and ran our simulation. In order to determine the true causal effects, we used the `simcausal` package. Note that we only estimated effects for static interventions in our specified simulation study.

In order to determine the IP weights, we followed the equation used in the paper, which when adapted to our simulation is shown in (4.8):

$$E\left[\frac{I(X_0 = x_0, X_1 = x_1)}{P(X_0)P(X_1|L_1, X_0)}Y\right]. \tag{4.8}$$

The methods we employed to determine our estimator is outlined in the previous section.

4.2.2 Simulation Results

In our simulation, we had two different versions: one keeping the dataset constant with 50 simulations, and one having 10 different datasets with 50 simulations.

In our results, we have the values of $E[Y^{\bar{x}}]$ for each static treatment regime ($\{X_0 = 0, X_1 = 0\}, \{X_0 = 0, X_1 = 1\}, \{X_0 = 1, X_1 = 0\}, \{X_0 = 1, X_1 = 1\}$). We calculated the true value, the estimated value using the IP treatment weights. We also calculated the sub-sampled estimator using the true propensity score as well as the sub-sampled value using the estimated propensity score from the sub-sampled subjects. Additionally, we calculated the difference from the truth, which is the estimated value subtracted from the true value as well as the percentage difference from the truth.

Recall that the goal of our simulation is to determine how much you lose by sub-sampling. Essentially, we want to know if you achieve a close enough level of accuracy without using the full dataset. By not using the full dataset, one saves time and money. Note that we showed that our results using sub-sampling are unbiased. We want to assess the effect of the inference when you sample. We also want to observe how our results vary from simulation to simulation.

To better understand our results for each treatment scheme, we looked at the number of and the proportion of participants for each treatment scheme and outcome on the simulated dataset. This is outlined in Table 4.2 below. Recall that the size of our simulation is $N = 100,000$.

From Table 4.2, we see that the treatment regime $\bar{X} = \{X_0 = 1, X_1 = 0\}$ has the lowest proportion.

We then took an average of the estimates over each replication and dataset. Recall, we did 50 replications for 10 datasets. Additionally, we looked at the percentage difference of our two sampling estimates from the estimate on the whole dataset.

From Table 4.3, we can see that the mean values of the estimate on the whole dataset, as

Table 4.2 – Summary of Individuals for each Treatment Scheme and Outcome in the Time-Varying Case

X_0	X_1	Y	**Count**	**Proportion**
0	0	0	25,451	0.25451
0	0	1	1,032	0.01032
0	1	0	40,238	0.40238
0	1	1	3,271	0.03271
1	0	0	12,335	0.12335
1	0	1	711	0.00711
1	1	0	16,097	0.16097
1	1	1	865	0.00865

Table 4.3 – Estimated Mean Counterfactual per Treatment for each Estimator in the Time-Varying Case

Treatment Scheme	**True**	**Cohort Estimate**	**Sampled (Full Prop Score)**	**Sampled (Est Prop Score)**	**Sampled (Full Prop Score) % Difference from Est**	**Sampled (Est Prop Score) % Difference from Est**
$X_0 = 0, X_1 = 0$	0.0672	0.0663	0.0664	0.0660	1.00%	0.996%
$X_0 = 0, X_1 = 1$	0.0634	0.0618	0.0619	0.0619	1.00%	1.00%
$X_0 = 1, X_1 = 0$	0.0530	0.0518	0.0517	0.0515	1.00%	0.995%
$X_0 = 1, X_1 = 1$	0.0489	0.0482	0.0482	0.0482	1.00%	1.00%

well as the two sampled estimates are very close to the true value. We can also see that the percentage difference of each sampling estimate compared to the full data estimate only differs by $\sim 1.00\%$, thus, indicating that we do not deviate much from the estimate by sampling, in both the full and estimated propensity score case. Additionally, we can see that both sampling estimates perform almost as well as the estimate on the full data.

We then look at the average standard error of the estimates over each replication and dataset.

From Table 4.4, we can see that the standard error for the full estimate as well as both sampled estimates are very similar. This is also reflected in the percentage difference of each sampled estimate from the estimator, as the percentage differences are very small. We have the highest percentage difference from sampling for the $\{X_0 = 0, X_1 = 1\}$ treatment scheme. However, similar to the results in Table 4.4, we can deduce that we only observe

Table 4.4 – Estimated Standard Error per Treatment for each Estimator in the Time-Varying Case

Treatment Scheme	**True**	**Cohort Estimate**	**Sampled (Full Prop Score)**	**Sampled (Est Prop Score)**	**Sampled (Full Prop Score) % Difference from Est**	**Sampled (Est Prop Score) % Difference from Est**
$X_0 = 0, X_1 = 0$	0	0.00214	0.00221	0.00202	1.03%	0.941%
$X_0 = 0, X_1 = 1$	0	0.00104	0.00121	0.00133	1.16%	1.27%
$X_0 = 1, X_1 = 0$	0	0.00113	0.00119	0.00119	1.05%	1.06%
$X_0 = 1, X_1 = 1$	0	0.00159	0.00160	0.00159	1.00%	0.997%

$\sim 1.00\%$ more standard error from sampling for the other treatment schemes, meaning that by sampling, we barely lose any accuracy for those treatment schemes. Additionally, there is no real difference between our sampling estimates using the full versus the estimated propensity score.

In order to fully visualize our estimated values, we have the following boxplot in Figure 4.2 to show the true and estimated values for each treatment scheme.

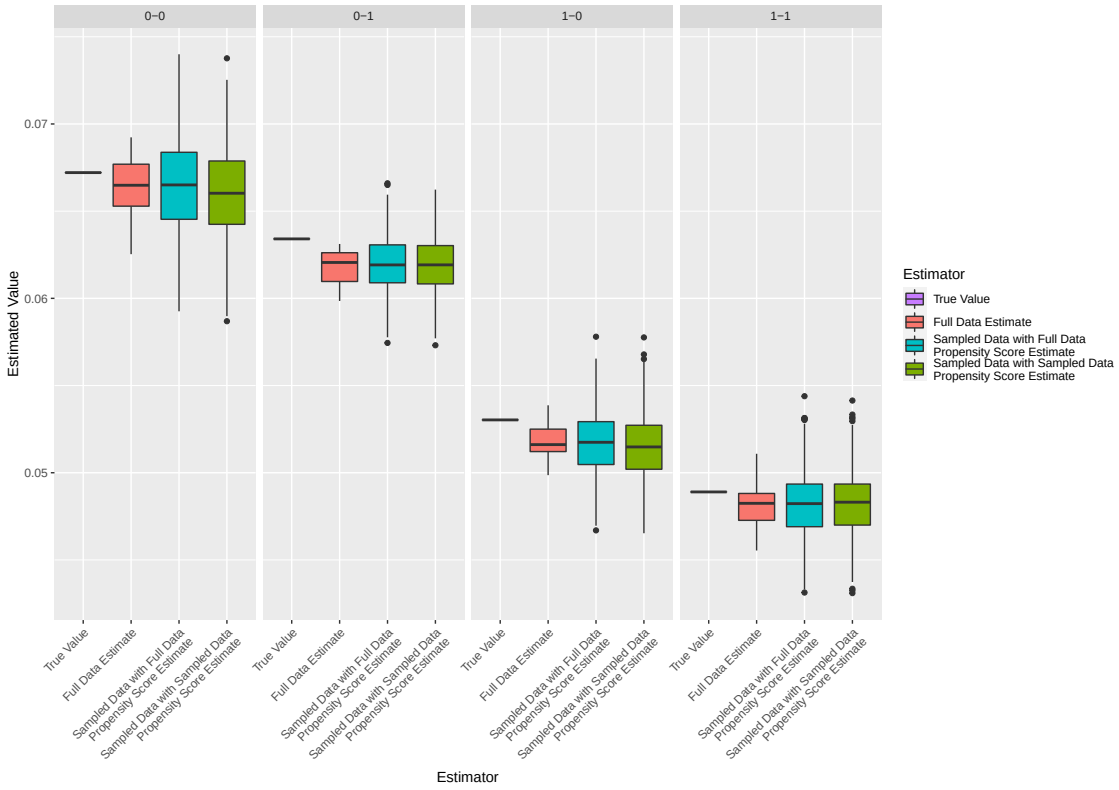


Figure 4.2 – Boxplot of the Estimates in the Time-Varying Case

Boxplot showing the estimated value for the truth as well as for our three estimators: the full data estimator, the sampled data with the full data propensity score estimator and the sampled data with the sampled data propensity score. The boxplots are shown for each treatment regime: ("$0-0$" $= \{X_0 = 0, X_1 = 0\}$, "$0-1$" $= \{X_0 = 0, X_1 = 1\}$, "$1-0$" $= \{X_0 = 1, X_1 = 0\}$, "$1-1$" $= \{X_0 = 1, X_1 = 1\}$). Recall that the estimate is of $E[Y^{\bar{x}}]$.

As our aim is to determine the quality of our estimates, and to see if we can maintain that quality by sampling, we visualized the percentage difference from the truth for each estimate. This is shown in Figure 4.3.

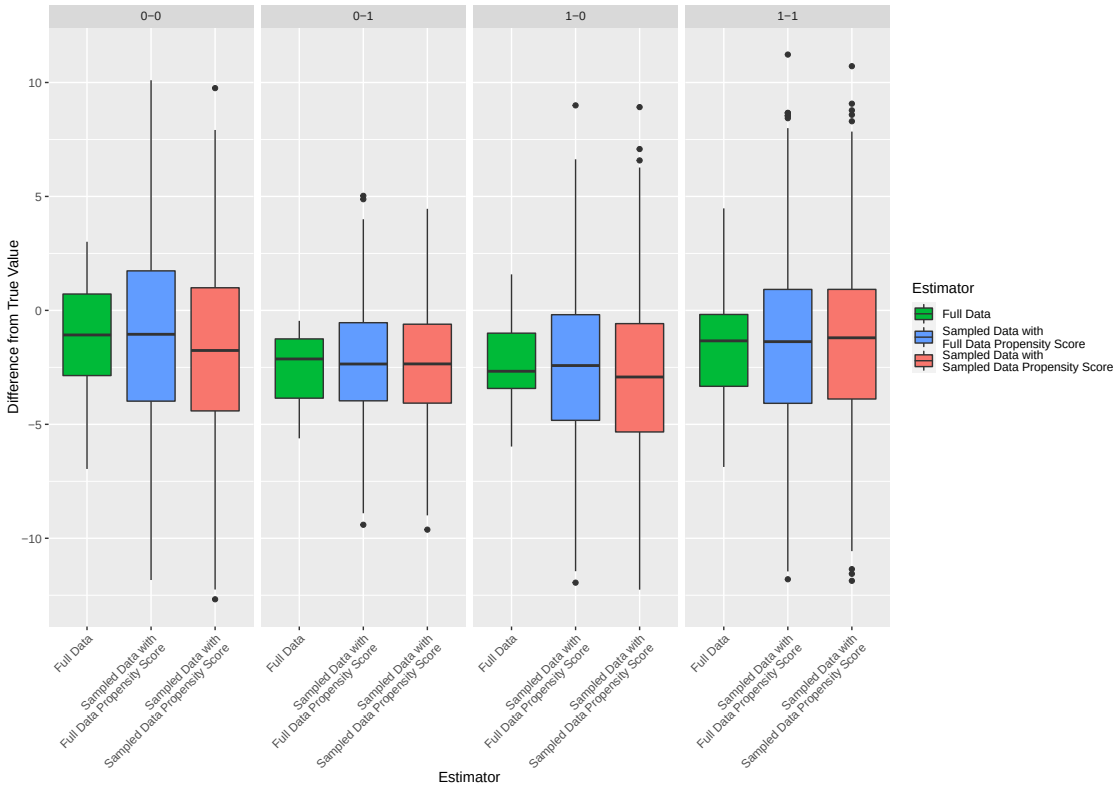


Figure 4.3 – Percentage Difference from the Truth in the Time-Varying Case

Boxplot showing the percentage difference from the truth, per treatment scheme, of the three estimators: the full data estimator, the sampled data with the full data propensity score estimator and the sampled data with the sampled data propensity score. This was calculated by the following formula: $100 \times \frac{\text{Estimated Value} - \text{True Value}}{\text{True Value}}$.

From Figure 4.3, we can see that our estimates vary more from the truth in the $\{X_0 = 1, X_1 = 0\}$ treatment scheme, and are very close to the truth in the $\{X_0 = 1, X_1 = 0\}$ treatment scheme. Additionally, we can see that the difference of our two sampling estimates from the truth are very close to the difference of the estimate from the full cohort from the truth. Thus, this shows us that our sampling performs almost the same as our estimate from the full cohort, meaning that we are not losing accuracy by sampling as well as we are not losing much accuracy by using the estimated propensity score versus the propensity score determined on the whole dataset, only precision.

As we did 50 replicates on 10 datasets, we want to see if there is a lot of variation between datasets. In Figure 4.4 we can see the average value over all the replications for each estimate over each dataset.

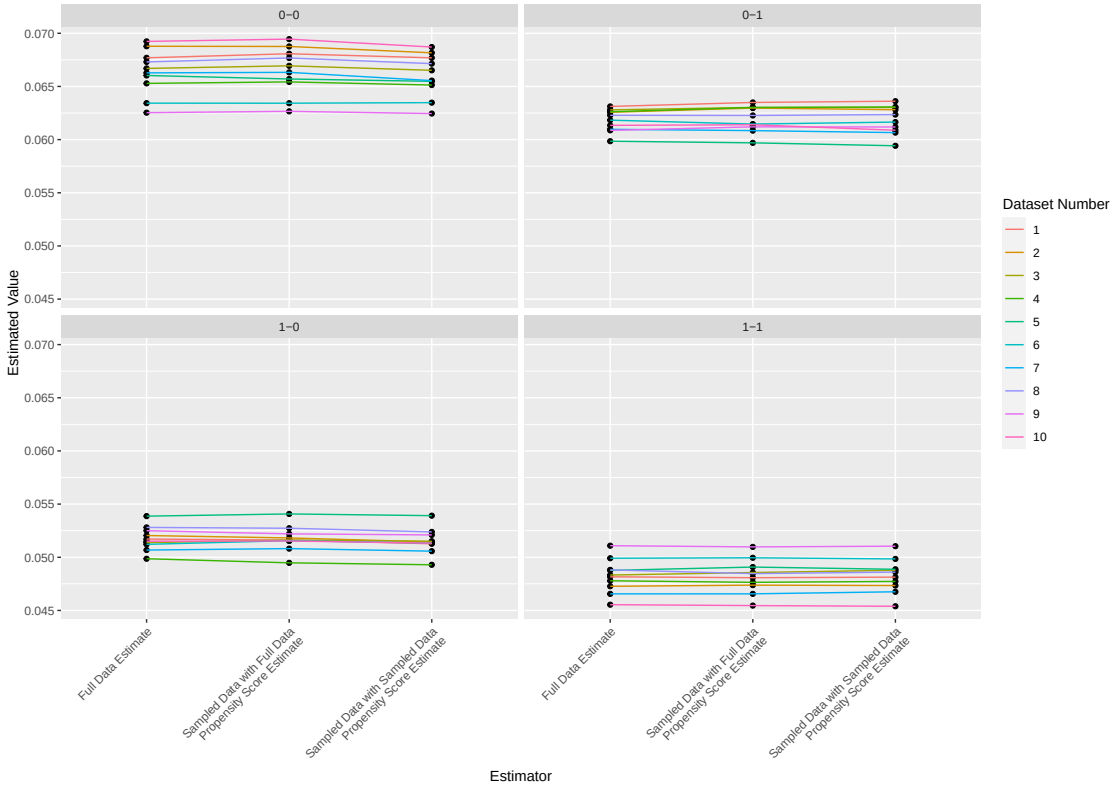


Figure 4.4 – Average Estimates per Dataset in the Time-Varying Case

Graph showing the estimated values for our three estimators for each dataset. For each of 10 datasets, 50 replications were run. Each line is for each if the 10 datasets in order to see how the estimated value changes for each estimator.

We can see in Figure 4.4 that there is a slight variation between datasets, indicating that there is variability between datasets.

Additionally, we want to determine if there is an effective sample size in our sub-sampling. We initially chose to have the same number of cases as controls. In the time-varying case we had $5,879$ cases out of a cohort of $N = 100,000$, thus we took a SRS of $n = 5,879$ of the $94,121$ controls. As with the cross-sectional case, we then ran our simulation again by

doubling the number controls, thus for the $5,879$ cases out of a cohort of $N = 100,000$, thus we took a SRS of $n = 2 \times 5,879 = 11,758$ of the $94,121$ controls. By doubling the number of controls in the sample, the standard error per treatment per estimator is slightly smaller. However, as in Table 4.4 we still only get $\sim 1.00\%$ more standard error from sampling. This indicated that even by doubling the controls, we do not gain much accuracy as when we had the same number of controls as cases. However, we do have a slightly lower percentage difference of the estimated values from the truth when doubling the number of controls.

These results does not come as a surprise as we determined in the previous chapter that the number of cases drive the accuracy of the estimator, not the number of controls.

Chapter 5

Discussion and Future Work

In Chapters 3 and 4, we developed methods to address nested case-control causal analyses using finite population methods in a causal setting in both a cross-sectional and multivariate time case. Our approach uses all of the cases of the cohort in our sample and then takes a simple random sample (SRS) from the controls. In order to adjust for the unequal probability sampling of cases and controls, we used inverse probability weighting to determine the sampling weights and applied them to our Horvitz-Thompson estimator. This resulted in the sampling weights for the cases being equal to 1, and the weights for the rest of the sample from the cohort (the sampled controls) being determined using IP treatment weighting. Once the sub-sampling was performed, we also had two different estimators, one using the propensity score that was estimated on the whole cohort and one using the propensity score that was estimated on the sub-sampled cohort. We had 3 estimators overall, that using IP treatment weighting for the causal weighting on the whole cohort, the sampled estimator using the true propensity score as well as the sampled estimator using the estimated propensity score from the sampled cohort.

From each of our simulations, in both the cross-sectional and time-varying case, we can see that there is minimal loss in accuracy when we sample our data, therefore, this is extremely important and useful result as it is a lot cheaper and less time consuming to sample. Further,

there is minimal loss from estimating the propensity score from the sampled data, rather than the full data. From our results, we can also see that increasing the sample size does not make a large difference in the accuracy or precision of the estimator. This is extremely useful, as there are situations where it would be difficult to obtain a large sample size from your cohort, therefore, we have shown that we can obtain accurate results using smaller sample sizes.

Our methods used logistic regression, however, in our analysis, we discovered that it wasn't necessary to obtain accurate estimators. Additionally, we mainly used maximum likelihood and GLMs instead of method of moments and GEEs, and we obtained the same values of the estimators as when GEEs were used. GEEs take significantly longer to run which is why we only used GLMs in this thesis.

Estimating the variance of our proposed estimators is beyond the scope of our thesis. Morenz [16] determined closed forms of the variance for his estimators, which are similar to ours, so if we were to look at the variance, we would employ similar techniques as him. Estimating the standard errors would be something to further explore in future work. However, it can be difficult to obtain reliable asymptotic variance estimators for causal problems in general; it would be even more difficult in the presence of sub-sampling. An alternate approach would be to use resampling estimators for the variance. However, one should note that using resampling for design based estimators requires different techniques than is used in standard bootstrapping estimators.

Much of the literature for analysis of nested case-control studies focuses on logistic regression models. The use of conditional logistic regression, treating the nested case-control study as a sample matched on time, is frequently discussed [19]. However, our goal was to combine the traditional analysis of nested case-control studies with causal methods, to avoid issues using the conditional logistic approach with time-varying data.

Rose and van der Laan divide the literature on the analysis of nested case-control study designs loosely into three groups [19]. The first, analyzes the nested case-control sample as a

case-control sample, ignoring the first stage of sampling the cohort, for example Barlow et al. [3]. The second, analyzes the nested case-control sample as a missing data structure, such as Robins et al. [18]. The third group straddles both of these groups, for example Breslow and Cain [5]. We fall into the third group, as does the work of Rose and van der Laan. The literature has covered similar estimators in similar contexts, however, the literature does not combine causal methods with nested case-control methods as we do, and they focus on the cross-sectional case. The paper by Kim et al. provides a link between causal analysis and using conditional logistic regression in nested case-control studies, however, they use a Cox model [13].

Zhao and Lipsitz [30], compare Breslow and Cain's estimator to their own. They conclude that when outcome and exposure are binary and the logistic link function is used, Breslow and Cain's [5, 6] estimator is preferable, since it appears more efficient than other estimators and the computation of estimates is easy using existing software [30]. Breslow and Cain have a cross-sectional estimator that requires the use of logistic regression and the use of an offset, as opposed to weighting, to account for the inclusion probability of a subject. Their estimator does not depend on the outcome, rather they condition on the inclusion probability. They also use the conditional mean, whereas Zhao and Lipsitz use the marginal mean, like we do. Zhao and Lipsitz's estimator resembles our own in the cross-sectional case but without the IP treatment weighting.

Robins et al. include a discussion of a missingness framework for the estimation of IP treatment weighted marginal causal parameters for nested case-control study designs [18]. They compare different nested case-control estimators to their semiparametric estimator.

Our estimators are very close to that of Zhao and Lipsitz, however, we use causal inverse probability treatment weighting on top of it. As mentioned, Zhao and Lipstiz show by simulation that Breslow and Cain's estimator has a lower variance, and Robins et al. discuss the efficiency of both. It would be of interest to know if the same conclusions as Robins et al. are obtained from a finite population perspective, however the loss of efficiency could be

the estimation of the random quantity which complicates the problem. Breslow and Cain's estimator is not easily extendable to the time-varying setting with causal confounding, one would need to explore this further to determine if you can add a causal weight to their estimator.

It follows that the methods developed in the literature, namely that of Robins et al. and Breslow and Cain, are much more complicated than ours, and do not appear to be as generalizable as our methods.

Bibliography

[1] A. Agresti. *Models for Matched Pairs*. John Wiley &; Sons, 2002.

[2] A. Agresti. *Foundations of linear and generalized linear models*. Wiley, 2015.

[3] W. E. Barlow, L. Ichikawa, D. Rosner, and S. Izumi. Analysis of case-cohort designs. *Journal of Clinical Epidemiology*, 52(12):1165–1172, 1999.

[4] N. E. Breslow. Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91(433):14–28, 1996.

[5] N. E. Breslow and K. C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20, 1988.

[6] K. C. Cain and N. E. Breslow. Logistic regression analysis and efficient design for two-stage studies. *American Journal of Epidemiology*, 128(6):1198–1206, 1988.

[7] P. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger. *The Analysis of Longitudinal Data*, volume 90. 01 2002.

[8] G. M. Fitzmaurice. *Longitudinal data analysis*. CRC Press/Taylor &; Francis, 2009.

[9] M. Hernán and J. M. Robins. *Causal Inference: What If*. 2021.

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

[11] F. Y. Hsieh. Sample size tables for logistic regression. *Statistics in Medicine*, 8(7):795–802, 1989.

[12] F. Y. Hsieh, D. A. Bloch, and M. D. Larsen. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, 17(14):1623–1634, 1998.

[13] Y. M. Kim, J. B. Cologne, E. Jang, T. Lange, Y. Tatsukawa, W. Ohishi, M. Utada, and H. M. Cullings. Causal mediation analysis in nested case-control studies using conditional logistic regression. *Biometrical Journal*, 62(8):1939–1959, 2020.

[14] S. Lewallen and P. Courtright. Epidemiology in practice: Case-control studies. *Community eye health / International Centre for Eye Health*, 11:57–8, 02 1998.

[15] S. L. Lohr. *Sampling: Design and Analysis*. Cengage Brooks/Cole, 2010.

[16] E. Morenz. Design based prediction of model parameters. Master's thesis, McGill University, 2017.

[17] J. M. Robins, S. Greenland, and F.-C. Hu. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94(447):687–700, 1999.

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

[19] S. Rose and M. J. van der Laan. Causal inference for nested case-control studies using targeted maximum likelihood estimation. 2009.

[20] A. R. Sen. Recent advances in sampling with varying probabilities. *Calcutta Statistical Association Bulletin*, 5(1):1–15, 1953.

[21] M. Setia. Methodology series module 2: Case-control studies. *Indian Journal of Dermatology*, 61:146, 03 2016.

[22] T. Shinozaki and E. Suzuki. Understanding marginal structural models for time-varying exposures: Pitfalls and tips. *Journal of Epidemiology*, 30(9):377–389, 2020.

[23] I. Shrier and R. Steele. Understanding the relationship between risks and odds ratios. *Clinical journal of sport medicine : official journal of the Canadian Academy of Sport Medicine*, 16:107–10, 04 2006.

[24] O. Sofrygin, M. J. van der Laan, and R. Neugebauer. simcausal package : Technical details and extended examples of simulations with complex longitudinal data. 2017.

[25] O. Sofrygin, M. J. van der Laan, and R. Neugebauer. simcausal r package: Conducting transparent and reproducible simulation studies of causal effect estimation with complex longitudinal data. *Journal of Statistical Software, Articles*, 81(2):1–47, 2017.

[26] S. Suissa. The quasi-cohort approach in pharmacoepidemiology upgrading the nested case-control. *Epidemiology*, 26(2):242–246, 2015.

[27] F. Yates and P. M. Grundy. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):253–261, 1953.

[28] S. L. Zeger and K.-Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130, 1986.

[29] S. L. Zeger, K.-Y. Liang, and P. S. Albert. Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44(4):1049–1060, 1988.

[30] L. P. Zhao and S. Lipsitz. Designs and analysis of two-stage studies. *Statistics in Medicine*, 11(6):769–782, 1992.