Methods for Analyzing Anonymized Clustered Data in The Presence of Competing Events

Pavel Slavinov Slavchev

Department of Mathematics and Statistics,
McGill University, Montreal
June, 2018

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

DEDICATION

This document is dedicated to the graduate and undergraduate students of McGill University and other academic institutions.

ABSTRACT

Causal inference methods allow one to draw a conclusion about a causal connection between the occurrence of an event and an outcome. When analyzing longitudinal data, one must also take into consideration correlation within the same subject's history and the possibility of competing events. The former may be harder to achieve if the data is anonymized because no subject identifier is available. Further, using a causal inference framework, and adjusting for the two features mentioned can be a challenging task. For example, in an anonymized birth registry data one may be interested in the causal effect of a treatment on the gestational age of a live birth. Yet, the presence of competing events (deliveries other than live birth) and relationship between deliveries from the same mother are two factors that must be accounted for. In this thesis, we propose an algorithm to create a subject identifier in an anonymized longitudinal data, and a new method to simulate, and analyze longitudinal data under a causal framework in the presence of competing events.

Résumé

Les méthodes d'inférence causale permettent de tirer une conclusion sur un lien causal entre l'occurrence d'un événement et un résultat. Lors de l'analyse des données longitudinales, il faut également prendre en compte la corrélation dans l'histoire du même sujet et la possibilité d'événements concurrents. Le premier peut être plus difficile à réaliser si les données sont anonymisées, car aucun identificateur de sujet n'est disponible. En outre, l'utilisation d'un cadre d'inférence causale, et l'ajustement pour les deux caractéristiques mentionnées peuvent être une tâche difficile. Par exemple, dans une base anonymisée de données sur les naissances, on peut s'intéresser à l'effet causal d'un traitement sur l'âge gestationnel d'une naissance vivante. Pourtant, la présence d'événements concurrents (accouchements autres que la naissance vivante) et la relation entre les accouchements d'une même mère sont deux facteurs qui doivent être pris en compte. Dans cette thèse, nous proposons un algorithme pour créer un identifiant de sujet dans une donnée longitudinale anonymisée, une nouvelle méthode pour simuler et analyser ces données sous un cadre de travail causal en présence d'événements concurrents.

Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

I would like to thank Dr. Russell Steele and Dr. Ian Shier for their time, patience, and guidance through the research of this thesis. A special thanks to Dr. Nathalie Auger, and Dr. Ashley Isaac Naimi for their advises and comments. Additionally, I would like to thank my friends Kapilthave Rajaratnam, Nasser Akliouat, and Steve Recine for reminding me that aside from working hard, it is also important to relax. Last of all, I would like to thank my family and close ones for their support and encouragement.

Chapter 1

Introduction

Following the development of a fetus is essential in order to ensure its well-being and reduce the odds, or prepare for the occurrence, of any complications. On average, gestation in singleton pregnancies lasts 40 weeks from the first day of the last menstrual period to the estimated date of delivery [1]. Deliveries before certain weeks of gestation are considered fatal for the fetus. Sometimes, the mother may not go into labour, yet the child could be deceased e.g. stillbirths [17], or other complications may arise such as a miscarriage or ectopic pregnancy. Studying the causal effect of factors such as mother's diet, exposure to harmful substances (e.g. tobacco smoke and/or alcohol), genetic history, socioeconomic status, and many more can allow the development of more effective prevention programs and better understanding of their impact on the gestation period. A common factor across mothers is the access to prenatal care; "prenatal care is widely accepted as an important public health intervention" [7]. Yet, its efficiency and role are still vague and unclear. Nonetheless, existing studies suggest that access to prenatal care is beneficial on many levels, including the outcome of the delivery [7] [31] [28] [14].

The National Survey of Family Growth (NSFG) has a longitudinal birth registry database collected between 2006-2010 and 2011-2013 [13]. Information about the birth history of 17,352 women was gathered, and we will consider three measurements in this thesis: the delivery outcome, the pregnancy duration, and the number of weeks pregnant at first prenatal care.

This thesis is concerned with estimating the causal effect of starting prenatal care on the gestational age of a live birth in the presence of competing pregnancy events. Chapter 2 and 3 introduce basic terminology and useful quantities in survival analysis, competing events, and causal inference. Chapter 4 discuses an approach to analyze longitudinal data under a causal inference framework in the presence of competing events. Chapter 5 introduces a new technique for simulating longitudinal data using a Structural Nested Accelerated Failure Time Model (SNAFTM) in the presence of competing events, shows simulation results, discusses a clustering algorithm that may be applied to an anonymized data, and establishes the finals results of the main analysis. Chapter 6 concludes with a summary of the main results and some suggestions for further research.

Chapter 2

Survival Analysis

In this chapter, we review the essential theory, notation, and definitions of survival analysis. This chapter will be the basis for other sections.

2.1 What is Survival Analysis?

Survival analysis is a branch of statistics that focuses mainly on modeling time to event or events [27]. An example of a time to event problem would be time to death from a particular disease or the failure time of a mechanical system such as an engine. Another commonly asked question in survival analysis is : how do particular circumstances or characteristics increase or decrease the probability of survival?

2.2 Survival Time Distribution and Censoring

We are now ready to define a random variable, T, that will be the core of almost any terminology and definitions to follow.

Definition 2.2.1 *T denotes a positive – $T \geq 0$ – random variable representing time to event of interest. T is often referred to as a failure time random variable [23].*

In Definition 2.2.1, T can be either a continuous variable ($T \in \mathbb{R}^+$) or discrete in which case

$$T = (t_1, t_2, ..., t_n)$$

where $t_1 < t_2 < \cdots < t_n$. Note, for T to be useful it requires three ingredients: a well defined time scale, a well defined event, and an origin or zero time in the study [10]. Whether it represents the start of the study or the age of individuals when they enter the study, the time origin needs to be carefully set [12]. Using standard probabilistic notation, T will stand for the random variable for a person's survival time while t will stand for a specific value of T.

Censoring is when the survival time is not known exactly [27]. The reasons why the survival time is *incomplete* can be several, but generally the three main reasons are : (1) a person does not experience the event before the end of the study, (2) a person is lost to follow-up, and (3) individuals withdraw before the end of the study due to some reason other than the event of interest. These three forms of censoring are what is known as right-censoring.

We denote by C the time to the censoring event. Note, C is a non-negative random variable.

An individual's survival time is right-censored if the time to event is greater than or equal to some censoring time C i.e. all we know is that the event has not happened before the censoring time [27]. Let R be a right-censored failure time random variable

such that, for individual i, R is defined by:

$$R_i \equiv \min(T_i, C_i).$$

In other words, R_i is the observed response. In addition to R_i, we also define a failure indicator which we will denote by δ_i. So that for individual i, let

$$\delta_i = \begin{cases} 1 & \text{if the event was observed, } T_i \leq C_i; \\ 0 & \text{if the response was censored, } \ T_i > C_i. \end{cases}$$

A right-censored individual can be of Type I, II, or III [26]. Type I is when the study is pre-designed to end after K years of follow; hence, whoever did not experience the event after K years of follow-up is censored. Type II, is similar to Type I in the sense that the study ends after K years of follow up, but not all subjects have the same censoring time. Type III is when the study ends after certain number of events has been reached.

Left censoring occurs more rarely than right censoring, nonetheless it is something to be careful of. It is when the event of interest has already occurred before study enrolment or before an observation period during the study. In this case, the definition of δ_i changes. Similarly to the random variable R_i, we define a left-censored failure time random variable, L_i, such that

$$L_i \equiv \max(T_i, C_i)$$

and

$$\delta_i = \begin{cases} 0 & \text{if the response was censored, } \ C_i \leq T_i; \\ 1 & \text{if the event was observed, } C_i > T_i. \end{cases}$$

Lastly, there is interval censoring. This occurs when the survival time is known to fall into a particular interval, but its precise value is unknown. If L_i^* and R_i^* are two time

points such that $L^* < R^*$, then

$$T_i \in (L_i^*, R_i^*).$$

Note, L_i^* and R_i^* have nothing in common with L_i and R_i, respectively, even if we used the same letter; the asterisk is used to distinguish them. Aside from having these 3 categories, censoring comes in two distinct types : (1) non-informative/independent and (2) informative/dependent censoring. Non-informative censoring is when each subject has a censoring time, C_i, that is statistically independent of their failure time, T_i, conditional on values of covariates. On the other hand, informative censoring is when the probability of censoring depends on the outcome the subject would have had in the absence of censoring [35] . In survival analysis, censoring must generally be non-informative.

Using sections 2.1 and 2.2, we can illustrate a representative graph of a typical survival study. Figure 2.1 demonstrates the 3 categories of censoring. A patient would be right-censored if he is alive at the end the study, he decides to quit, or his record was lost (subject 3,4, and 5). A subject would be left-censored if he enrolls and experiences the event before the end of the study when measurements are taken (subject 2). Finally, interval censoring would occur if the event of interest happened after the beginning of a study, but before the end of the study (subject 6). It also tells the analyst if an individual was censored or if he experienced the event (or not), and gives certain idea of the survival time for an individual.

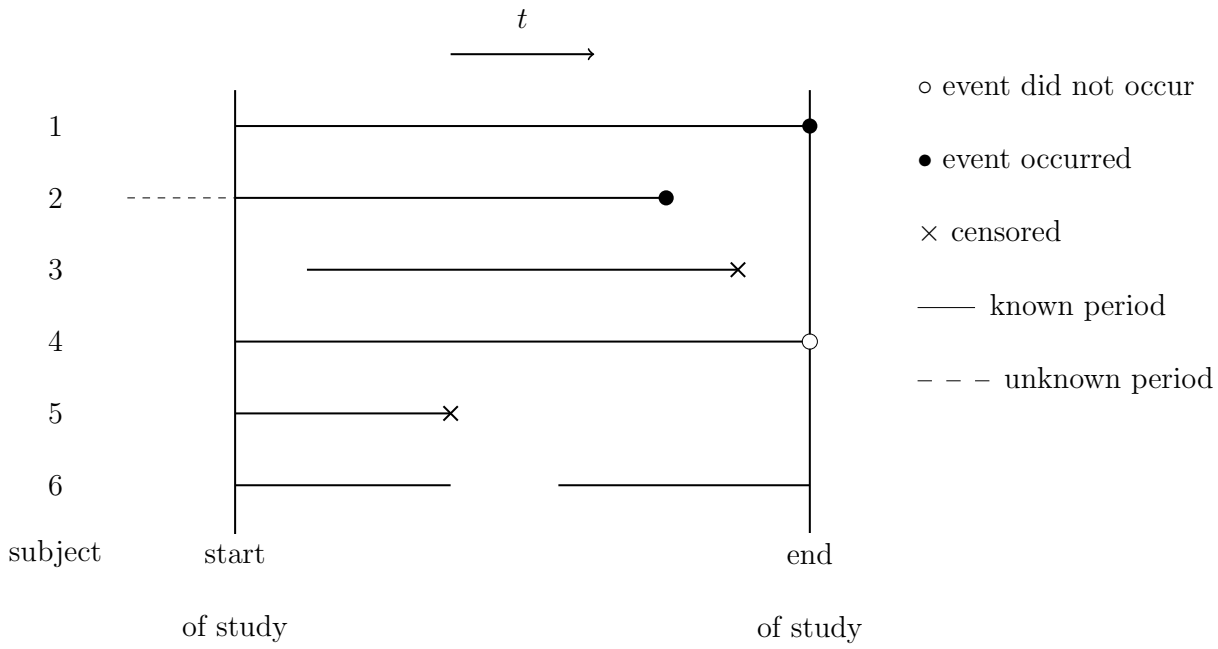


Figure 2.1: Graphical Representation of a Survival Study

2.3 Common Functions in Survival Analysis

In sections 2.1 and 2.2, we established the basic principles of survival analysis, but no analytic principles, as no forms of analyses have yet been introduced. We defined the relevant quantities: events, survival times, and censoring times.

In this section, we define and derive important functions in survival analysis, and establish relationships between them. The key point is that if you specify any of the functions to be derived, you specify all of them. We focus mainly on continuous random variables, but state some results for discrete quantities as well.

For a random variable T, the cumulative distribution function (CDF), $F_T(t)$, or just

$F(T)$, is the probability that T will take a value less than or equal to t, namely

$$F(t) = P(T \leq t).$$

The primary object of interest in survival analysis is the survival function denoted by S or $S(t)$.

Definition 2.3.1 *The survival function, $S(t)$, gives the probability that a subject will survive past time t.*

We can easily define $S(t)$ in terms of $F(t)$ as

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t). \tag{2.1}$$

Depending on the settings, some might define the survival function as $P(T \geq t)$ i.e. bigger than or equal. We will use strictly bigger in this thesis. Note also that Definition 2.3.1 assumes that there is a single event at time t otherwise 2.1 will not hold. From 2.1 and using basic properties of any CDF, we can conclude that for a single event the survival function satisfies the following properties:

(1) It is non-increasing;

(2) $S(t = 0) = 1$, i.e. the probability of surviving at time 0 is 1 since the subject is not at risk yet;

(3) $\lim_{t \to \infty} S(t) = 0$ i.e. if we wait long enough, a subject will surely experience the event in question.

If T is continuous, $S(t)$ is a smooth function and resembles to what is shown in Figure 2.2. Howerever, if T is discrete we often obtain curves similar to Figure 2.3 i.e., a step function [27].

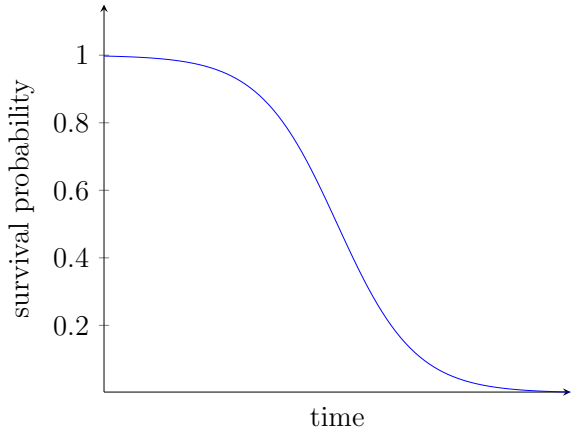


Figure 2.2: Continuous Survival Time Curve

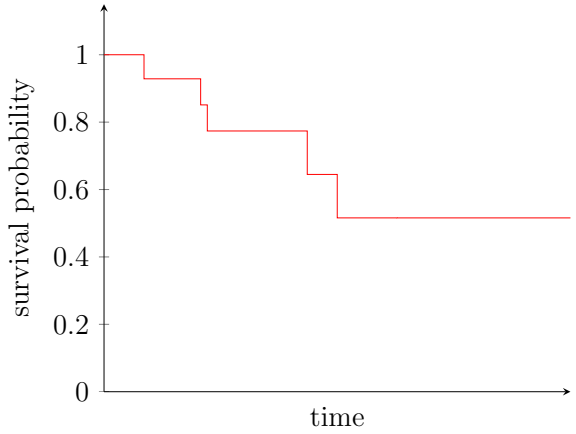


Figure 2.3: Discrete Survival Time Curve

From basic statistics, we know that for a continuous random variable, the derivative of its CDF is equal to the probability density function (PDF), or in other words

$$f(t) = \frac{d}{dt}F(t). \tag{2.2}$$

Furthermore, using 2.2 and 2.1, we have that

$$\frac{d}{dt}S(t) = \frac{d}{dt}\left(1 - F(t)\right)$$
$$\frac{d}{dt}S(t) = -\frac{d}{dt}F(t)$$

$$-\frac{d}{dt}S(t) = f(t). \tag{2.3}$$

Note that another common definition of $f(t)$ is

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}. \tag{2.4}$$

If $f(t)$ is discrete such that $T = (t_1, t_2, ..., t_n)$, then

$$f(t) = P(T = t)$$

$$f(t) = \begin{cases} f_j & \text{if } t = t_j,\ j = 1, 2, ..., n \\ 0 & \text{if } t \neq t_j,\ j = 1, 2, ..., n. \end{cases}$$

We will mainly consider T to be a continuous random variable as stated earlier. Similarly to 2.2, the CDF can be obtained from the PDF

$$F(t) = \int_{-\infty}^{t} f(u)\, du. \tag{2.5}$$

Since T is a positive random variable, we have that

$$F(t) = \int_{0}^{t} f(u)\, du. \tag{2.6}$$

Using 2.1 and 2.6, we obtain that

$$S(t) = 1 - F(t) = 1 - \int_{0}^{t} f(u)\, du = \int_{t}^{\infty} f(u)\, du. \tag{2.7}$$

In the discrete case, we would have

$$\begin{aligned} S(t) &= \sum_{u>t} f(u) \\ &= \sum_{t_j>t} f(t_j) \\ &= \sum_{t_j>t} f_j. \end{aligned} \tag{2.8}$$

The second object of great importance in survival analysis is the hazard function: $h(t)$. "The hazard function, $h(t)$, gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t." [27]

One way to understand the hazard function is to consider car's speedometer. A speedometer gives the driver's instantaneous velocity. Likewise, the hazard function gives the instantaneous potential of experiencing the event at time t given survival up to time t.

Mathematically, the hazard function can be expressed as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq \Delta t + t \,|\, T \geq t)}{\Delta t}, \tag{2.9}$$

where the numerator of 2.9 can be read as :

$$P(t \leq T \leq \Delta t + t \,|\, T \geq t) =$$

$$P(\text{individual fails in time interval } t, \Delta t + t \underbrace{|}_{\text{given}} \text{ he survived until time } t).$$

Looking at 2.7 and 2.9, a second distinction can be made between the survival and hazard function. $S(t)$ is a probability while $h(t)$ is not. Although the numerator of 2.9 is a probability, the denominator affects the interpretation of $h(t)$. Because Δt is in the denominator, $h(t)$ is not a probability, but rather a rate. Hence, while $0 \leq S(t) \leq 1$, $h(t)$ can take any value between 0 and infinity. For this, the hazard function is sometimes referred to as a conditional failure rate [27].

Going back to the expression in 2.9, let us expand it further. Notice that the top expression is a conditional probability, so

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T \leq \Delta t + t \,|\, T \geq t) \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t \leq T \leq \Delta t + t \,,\, T \geq t)}{P(T \geq t)} \end{aligned}$$

$$= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t \leq T \leq \Delta t + t)}{P(T \geq t)},$$

using 2.4 and 2.1,

$$= \frac{f(t)}{S(t)}. \tag{2.10}$$

The jump from the second to third equality i.e.,

$$P(t \leq T \leq \Delta t + t \cap T \geq t) = P(t \leq T \leq \Delta t + t),$$

can be easily seen by drawing a time-line (see Figure 2.4).

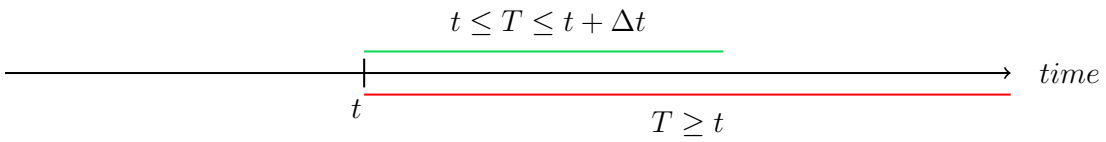


Figure 2.4: Visualization of Probability Event

Clearly, the part where the events intersect is the green segment of the time-line which is equivalent to $P(t \leq T \leq t + \Delta t)$. If T were discrete, then the hazard function is

$$\begin{aligned} h(t_j) &= P(T = t_j \,|\, T \geq t_j) \\ &= \frac{P(T = t_j)}{P(T \geq t_j)}, \end{aligned}$$

but using 2.8 and the definition of T (look below 2.4) we have

$$\begin{aligned} &= \frac{f(t_j)}{S(t_{j-1})} \\ &= \frac{f(t_j)}{\sum_{m\,:\,t_m > t_j} f(t_m)}. \end{aligned} \tag{2.11}$$

Using the expression in 2.11 (or the one above it which is equivalent), we can obtain a relation between the hazard and survival function:

$$h(t_j) = \frac{f(t_j)}{S(t_{j-1})}$$

$$= \frac{S(t_j) - S(t_{j-1})}{S(t_j)}$$
$$= 1 - \frac{S(t_{j-1})}{S(t_j)} \tag{2.12}$$

so that

$$1 - h(t_j) = \frac{S(t_{j-1})}{S(t_j)}. \tag{2.13}$$

In fact, we can derive an even more concrete relation between the hazard and the survival function. Using 2.13, the fact that $S(t_1) = 1$, and $S(t_j) = S(t)$, we have

$$\begin{aligned} S(t_j) = S(t) &= \frac{S(t_2)}{S(t_1)} \times \frac{S(t_3)}{S(t_2)} \times \cdots \times \frac{S(t_j)}{S(t_{j-1})} \\ &= [1 - h(t_1)] \times [1 - h(t_2)] \times \cdots \times [1 - h(t_j)] \\ &= \prod_{i=1}^{j} [1 - h(t_i)]. \end{aligned} \tag{2.14}$$

In simple words, this result states that in order to survive to time t_{j+1} one must first survive t_1, then one must survive t_2 given that one survived t_1, and so on, finally surviving t_j given survival up to that point.

Note that since $S(T) = 1 - F(T)$, we have that $1 - F(t_j) = \prod_{i=1}^{j} [1 - h(t_i)]$. Using 2.14 and the definition of the hazard, we have the following relation between the hazard function and the probability mass function :

$$h(t_j) = \frac{f(t_j)}{S(t_{j-1})}$$
$$S(t_{j-1}) \times h(t_j) = f(t_j)$$
$$h(t_j) \times \prod_{i=1}^{j-1} [1 - h(t_i)] = f(t_j) \tag{2.15}$$

Returning to the case where T is continuous, we can establish similar relationships between the survival and hazard function, and the probability density function. Taking

2.3 and 2.10, we obtain:

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= -\frac{S'(t)}{S(t)} \\ &= -\frac{d}{dt}(\ln S(t)) \end{aligned} \tag{2.16}$$

Using the above, we can find the survival function from the hazard via :

$$\begin{aligned} \frac{d}{dt}(\ln S(t)) &= -h(t) \\ \int_0^t \frac{d}{dt}(\ln S(t)) &= -\int_0^t h(u)\,du \\ \ln S(t) - \ln S(0) &= -\int_0^t h(u)\,du \\ \ln S(t) - \ln(1) &= -\int_0^t h(u)\,du \\ \ln S(t) &= -\int_0^t h(u)\,du \\ S(t) &= \exp\left\{-\int_0^t h(u)\,du\right\} \end{aligned} \tag{2.17}$$

Earlier, we showed that

$$h(t) = \frac{f(t)}{S(t)}$$

From this equality, we can obtain 3 other results. First,

$$h(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} \tag{2.18}$$

Second, using 2.17 :

$$\begin{aligned} f(t) &= h(t) \times S(t) \\ &= h(t) \times \exp\left\{-\int_0^t h(u)\,du\right\} \end{aligned} \tag{2.19}$$

Third,

$$h(t) = \frac{f(t)}{S(t)}$$

$$= \frac{f(t)}{\int_t^\infty f(u)\,du} \tag{2.20}$$

Using 2.17 and the fact that $F(t) = 1 - S(t)$, we see that

$$F(t) = 1 - \exp\left\{-\int_0^t h(u)\,du\right\}. \tag{2.21}$$

As stated above, the hazard function provides an instantaneous risk of experiencing the event at some time t. If we wish to know the cumulative hazard from time 0 to time t, then we need to *sum* or integrate over all the hazards.

The cumulative hazard, $H(t)$, represents the sum of risks between time 0 and some time t. The mathematical definition of the cumulative hazard is

$$H(t) = \int_0^t h(u)\,du \tag{2.22}$$

where $h(u)$ stands for the hazard function. The term in 2.22 was encountered in the derivations of 2.21. Above, we assumed that T is continuous. If T is discrete, then $H(t)$ is defined as

$$H(t) = \sum_{j\,:\,t_j<t} \ln(1 - h(t_j)). \tag{2.23}$$

Using 2.21, we can establish several relationships between $H(t)$ and $f(t)$, $F(t)$, $h(t)$, and $S(t)$. From 2.17, we immediately see that

$$S(t) = \exp\{-H(t)\}$$

so,

$$-\ln S(t) = H(t). \tag{2.24}$$

From 2.21, namely,

$$F(t) = 1 - \exp\{-H(t)\}$$

we have that

$$-\ln\{1-F(t)\} = H(t). \tag{2.25}$$

From the above it follows that

$$\begin{aligned} -\ln\left\{1-\int_0^t f(u)\,du\right\} &= H(t) \\ -\ln\left\{\int_t^\infty f(u)\,du\right\} &= H(t), \end{aligned} \tag{2.26}$$

and

$$\begin{aligned} F(t) &= 1-\exp\{-H(t)\} \\ f(t) = \frac{d}{dt}F(t) &= \frac{d}{dt}\{-\exp(-H(t))\}. \end{aligned} \tag{2.27}$$

From 2.22, we obtain

$$\frac{d}{dt}H(t) = \frac{d}{dt}\int_0^t h(u)\,du = h(t)-h(0) = h(t). \tag{2.28}$$

A quantity of importance in survival analysis or statistics in general is often the mean of the random variable. Recall, that T represents the time to event so its mean is the expected mean time for the event to happen. Mathematically, that is

$$\begin{aligned} \mu &= E[T] \\ &= \int_0^\infty t f(t)\,dt, \end{aligned}$$

using 2.3

$$= -\int_0^\infty tS'(t)\,dt.$$

Using integration by parts with $u = t$ and $dv = S'(t)\,dt$, we have that

$$= -tS(t)\Big|_0^\infty + \int_0^\infty S(t)\,dt$$

$$E[T] = -\lim_{t\to\infty} tS(t) + \int_0^\infty S(t)\,dt$$
$$= \int_0^\infty S(t)\,dt. \tag{2.29}$$

The limit term equals to 0 since

$$0 \le \lim_{t\to\infty} tS(t) = \lim_{t\to\infty} t\int_t^\infty f(u)\,du \le \lim_{t\to\infty} \int_t^\infty uf(u)\,du = 0$$

therefore,

$$\lim_{t\to\infty} tS(t) = 0.$$

So from 2.29, we can obtain the mean life time by integrating the survival function.

2.4 Proportional & Accelerated Models

Cox proportional hazards and accelerated failure time models are perhaps the first type of models an analyst will fit when given survival data. In the next two sections, we briefly introduce each type of model and some basic facts about them.

2.4.1 Proportional Hazards (PH) Model

The most commonly used Proportional Hazard (PH) model is the Cox Proportional Hazards models, proposed by Cox himself [9]; hence, we will interchangeably use the terms PH and Cox model.

The Cox Proportional Hazards model, $h(t \,|\, \boldsymbol{x})$, for subject i is given by

$$h_i(t \,|\, \boldsymbol{x}_i) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i) \tag{2.30}$$

where $'$ stands for the transpose, $h_0(t)$ is referred to as the baseline hazard function, which is the hazard function for an individual for whom all the variables included in the

model are zero, $\boldsymbol{x}_i$ is a $p \times 1$ vector of explanatory variables, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of of regression coefficients.

Note that $h(t \,|\, \boldsymbol{x})$ is dependent on the form of $h_0(t)$; hence, $h_0(t)$ determines the structure of the model i.e., parametric or semi-parametric [26]. If no assumptions are made about the form of $h_0(t)$, then the Cox model is semi-parametric; otherwise, it is parametric. Note that, for now, we are assuming the covariates are time independent. Also, for the Cox model to be adequate for inference, the proportional hazard assumption needs to hold, at least approximately. Let us formulate this mathematically. First, a more general form of 2.30 is

$$h_i(t \,|\, \boldsymbol{x}) = h_0(t)\Psi(\boldsymbol{x}_i)$$

where $\Psi(\cdot)$ is some link function [35]. The choice of log-linear link function – $\Psi(\boldsymbol{x}_i) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)$ – is useful for several reasons, but most importantly for the interpretation of β_j for $j = 1, ..., p$. Suppose we fix time, t, and we are looking at two subjects i and j with explanatory vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively, then observe that

$$\frac{h_i(t \,|\, \boldsymbol{x}_i)}{h_j(t \,|\, \boldsymbol{x}_j)} = \exp\left[\boldsymbol{\beta}'(\boldsymbol{x}_i - \boldsymbol{x}_j)\right],$$

i.e., the right-hand side of these expressions does not depend on t. If subjects i and j are identical on all but the k-th characteristic, then the above becomes

$$\frac{h_i(t \,|\, \boldsymbol{x}_i)}{h_j(t \,|\, \boldsymbol{x}_j)} = \exp\left[\boldsymbol{\beta}'(\boldsymbol{x}_i - \boldsymbol{x}_j)\right] = \exp\left[\beta_k(x_{ik} - x_{jk})\right],$$

moreover if $x_{ik} - x_{jk} = 1$ i.e., there is a unit change in x_k, then

$$\frac{h_i(t \,|\, \boldsymbol{x}_i)}{h_j(t \,|\, \boldsymbol{x}_j)} = \exp\left[\boldsymbol{\beta}'(\boldsymbol{x}_i - \boldsymbol{x}_j)\right] = \exp\left[\beta_k(x_{ik} - x_{jk})\right] = \exp(\beta_k),$$

and this is known as the hazard ratio for covariate k which we observe is time-independent. The interpretation of β_k is now straightforward: each regression coefficient summarizes

the proportional effect on the hazard of absolute changes in the corresponding covariate. Observe also that the effect of covariates is multiplicative with respect to the hazard.

Unfortunately, for estimation the typical MLE approach will not be suitable without assumptions about the baseline hazard function. Instead, a partial likelihood approach is used to obtain $L(\boldsymbol{\beta})$ which will be some function of $\exp[\boldsymbol{\beta}'\boldsymbol{x}_i]$ and hence the score equations can be calculated, without making any distributional assumptions about $h_0(t)$.

The parametric Cox PH model is somewhat simpler as it specifies a specific distribution for the baseline hazard function. Common choices of distributions are the Exponential, Weibull, Logistic, Normal, and Gamma distribution. In this setting, using full likelihood approach to make inference about the β's can typically be done in the usual fashion.

For brevity, we have not discussed any model diagnostics e.g. residuals or validation rules for the PH assumptions or what happens when the covariates are time dependent, but these are aspects the analyst must consider. We now proceed with a brief introduction to Accelerated Failure Time (AFT) Models.

2.4.2 Accelerated Failure Time (AFT) Models

Cox PH models measure the effect of covariates on the hazard. It may also be interesting to measure the effect of certain variable(s) on the survival time. In such case an Accelerated Failure Time (AFT) model should be used [26] [38].

The Accelerated Failure Time model for subject i is given by

$$\log(T_i) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i + \sigma\varepsilon_i \tag{2.31}$$

where T_i is the survival time (as defined in Definition 2.2.1) for subject i, σ is a scale

parameter, β_0 is the intercept, $\boldsymbol{x}_i$ is a $p \times 1$ vector of explanatory variables, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients.

Similarly to the Cox PH model, the model in 2.31 can be parametric or semi-parametric depending on the assumptions about the error term, ε_i. Common parametric distributions for ε_i are the same typically used as in the Cox PH.

Interpretation of the coefficients under AFT models is somewhat simpler than PH models. They capture the direct effect of explanatory variables on the survival time. In other words, "AFT models describe [a] 'stretching out' (lengthen) or contraction (shorten) of survival time as a function of [the magnitude of the] predictor variables" [27]. Mathematically speaking, the above is explained by taking a similar approach as in the PH context i.e., by looking at the ratio of survival times for two subjects. First , from 2.31, we have that

$$T_i = e^{\beta_0} e^{\boldsymbol{\beta}' \boldsymbol{x}_i} e^{\sigma \varepsilon_i}.$$

Then, if subject i and j are similar with respect to all other covariates, but the k-th, and $x_{ik} - x_{jk} = 1$, we have that

$$\frac{T_i}{T_j} = \exp\left[\beta_k (x_{ik} - x_{jk})\right] = \exp(\beta_k)$$

$\exp(\beta_k)$ is known as the *time ratio* or *acceleration factor*. It tells the analyst how "fast" a subject moves down the survival curve for a unit change in the predictor variable.

In the current context, the parameters of interest are β_0, $\boldsymbol{\beta}$, and σ, so that

$$\begin{aligned} L(\beta_0, \boldsymbol{\beta}, \sigma) &= \prod_{i=1}^{n} f(t_i)^{\delta_i} [1 - F(t_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} f(t_i)^{\delta_i} [S(t_i)]^{1-\delta_i} \end{aligned}$$

Recall, δ_i is an indicator of whether the event occurred or not. $f(t_i)$ is the density of ε_i.

Now, we need to specify $S(t_i)$ in terms of ε_i. We have

$$\begin{aligned} S(t_i) &= P\left(T_i \geq t\right) \\ &= P\left(\log(T_i) \geq \log(t)\right) \\ &= P\left(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i + \sigma\varepsilon_i \geq \log(t)\right) \\ &= P\left(\varepsilon_i \geq \frac{\log(t) - \beta_0 - \boldsymbol{\beta}'\boldsymbol{x}_i}{\sigma}\right) \\ &= S_{\varepsilon_i}\left(\frac{\log(t) - \beta_0 - \boldsymbol{\beta}'\boldsymbol{x}_i}{\sigma}\right). \end{aligned}$$

Hence, using the above, $L(\beta_0, \boldsymbol{\beta}, \sigma)$ becomes

$$L(\beta_0, \boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} f_{\varepsilon_i}(\boldsymbol{\theta}_i)^{\delta_i}[S_{\varepsilon_i}(\boldsymbol{\theta}_i)]^{1-\delta_i}$$

where

$$\boldsymbol{\theta}_i = \frac{\log(t_i) - \beta_0 - \boldsymbol{\beta}'\boldsymbol{x}_i}{\sigma}.$$

Once the distribution of ε_i is specified a more concrete form for the likelihood can be obtained. Again, we shall not comment on model diagnostics, but there exist several methods to verify the goodness of fit of an AFT model.

2.5 Competing Events

Competing risk models are used when there are at least two possible events that a subject can experience, but only one such event type can actually occur. In other words, if a study focuses on investigating some outcome/event, then other events may prohibit the main event from occurring. For example, a patient can die from lung cancer or from a stroke, but not both so the events compete with each other.

An important assumption throughout this chapter was that the censoring mechanism was independent of the outcome mechanism i.e, we have non-informative censoring. When

the event of interest is considered to be censored by the occurrence of another event, the former assumption is potentially violated and we would have informative censoring.

One common way to model competing events is to study one event type at a time [27]. Hence, other (competing) failure types are treated as censored in addition to those who are lost to follow-up and/or withdrawal.

Analysis is typically done using a Cox model i.e., studying the cause specific hazard for each failure type. We therefore define the Cause-Specific Hazard Function.

Definition 2.5.1 *Suppose that we have $\mathcal{F}$ distinct failure types, then the Cause-Specific Hazard Function is defined as*

$$h_f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_f < t + \Delta t \,|\, T_f > t)}{\Delta t}$$

where $f = 1, ..., \mathcal{F}$ and the random variable T_f denotes the time-to failure from event type f.

Following Definition 2.5.1 and the standard definition of the hazard function, $h_f(t)$ gives the instantaneous failure rate at time t for event type f, given not failing from event f by time t [27]. Consequently, we can define the Cox Proportional Hazard-Specific model.

Definition 2.5.2 *Assuming we have $\mathcal{F}$ distinct failure types, the Cox Proportional Hazard - Specific model, $h_f(t \,|\, \boldsymbol{x})$, for subject i is given by*

$$h_{i,f}(t \,|\, \boldsymbol{x}_i) = h_{0,f}(t) \exp(\boldsymbol{\beta}_f' \boldsymbol{x}_i)$$

where $f = 1, ..., \mathcal{F}$, $h_{0,f}(t)$ is the baseline cause-specific hazard, $\boldsymbol{x}_i$ is a $p \times 1$ vector of explanatory variables, and $\boldsymbol{\beta}_f$ is a $p \times 1$ vector of of regression coefficients subscripted by f to indicate the effects of the predictors for the f-th event-type [24].

In the end, $\mathcal{F}$ distinct analyses must be performed, one for each failure type, and accordingly $\mathcal{F}$ model validations. Another approach to study competing events is by generalizing the Kaplan Meier (KM) estimator to include competing risks [24]. Let

$$t_{f1} < t_{f2} < ... < t_{fk_f}$$

denote the k_f distinct failure times for failures of type f, n_{fi} denote the number of subjects at risk just before t_{fi} and let d_{fi} denote the number of deaths due to cause f at time t_{fi}. Then, the cause specific KM estimator is given by

$$\hat{S}_f(t) = \prod_{i\,:\,t_{fi}<t} \left(1 - \frac{d_{fi}}{n_{fi}}\right).$$

In a similar fashion, most of the quantities defined and derived in this chapter can be formulated to a cause specific setting. But there are some drawbacks to this method, first and most obvious, if $\mathcal{F}$ is large, say $\mathcal{F} \geq 10$, then 10 analyses need to be performed. As already mentioned, we require the assumption of independent competing events to ensure non-informative censoring, but with large $\mathcal{F}$ this assumption may not hold.

2.6 Summary

In this chapter we defined basic terminology in survival analysis and introduced commonly encountered functions, e.g., the survival, hazard, and cumulative hazard functions. We established several relationships between them when we have a continuous random variable, and then briefly discussed two models to analysis a time to event data that is to say the Cox proportional hazards and the accelerate failure time models. Lastly, we introduced what competing risks are and some very fundamental methods to analyze them. We now jump to a new topic that will be connected with elements seen in this chapter.

Chapter 3

Causal Inference

In this chapter, we introduce the basics of causal inference. We start by establishing a notation widely used in the context of causal inference, we then introduce certain conditions that will be needed later, and then review three methods of estimating what is known as the average causal effect.

3.1 What is Causal Inference?

In most studies, the main interest is some outcome, measured by a random variable Y, and how a certain set of predictors affect that outcome. In the context of survival analysis, we often subject some individuals, the population of interest, to a certain treatment or exposure which we will refer to as A. The aim of causal inference is to identify the impact of exposure/treatment on some outcome. For instance, suppose a patient can go under two different kinds of heart transplants, measured by a random variable, A. Assume we are interested in the outcome of time to death, measured by Y. The question of interest could be: does the specific heart transplant received cause a patient to die sooner or later?

Note, when we say "treatment", we do not necessarily mean a medical prescription, but rather a general "action" or intervention taken upon an individual. For instance, if we are interested in the effect of quitting smoking on body weight, than the "treatment" is quit smoking ($A = 1$) and not quit smoking ($A = 0$).

3.2 Notation

We now introduce a notation that will be used throughout the coming sections. Only a dichotomous treatment will be considered as the analysis with continuous or multi level treatment is beyond the scope of this thesis. A may have more than two levels, but causal inference aims mainly at the contrast between two levels as discussed below. Hence, we let A be a binary variable (1: treated, 0: untreated). We also let the outcome, Y, be a dichotomous outcome variable (1: death, 0: survival); Y could be continuous, but for convenience we will use a dichotomous outcome. We let $Y^{a=1}$ (read Y under treatment $a = 1$) be the outcome variable that would have been observed if the subject had received the treatment value $a = 1$. Likewise, $Y^{a=0}$ (read Y under treatment $a = 0$) is the outcome variable that would have been observed when receiving the treatment value $a = 0$. $Y^{a=1}$ and $Y^{a=0}$ are referred to as *counterfactual outcomes* [19]. Note that Y^a is not the same as $Y \mid A = a$. The latter is the outcome given treatment a was observed while the former reads, as explained, the outcome that would have been observed if treatment a was assigned.

3.2.1 Individual and Average Causal Effects

With the above in hand, we can now define the "quantity," we want to essentially measure. Suppose we are interested in whether a specific heart transplant is effective. Let $A = 1$ denote that an individual received the specific heart transplant and $A = 0$, he did not. Following the surgery, an individual may or may not die. So does the surgery cause death or, in other words, does going under the surgery have a causal effect on the outcome? Suppose that we have two individuals, call them P_1 and P_2 and suppose that P_1 had the specific heart transplant and died after some time t. While if he had not received the transplant, he would have lived; therefore, the treatment had a causal effect. On the other hand, P_2 also had the transplant, but did not die. Assume that P_2 would have lived even if he had not received the transplant i.e., no causal effect. Hence, the *causal effect for an individual* is present when the treatment A has a causal effect on an individual's outcome Y, namely $Y_i^{a=1} - Y_i^{a=0} \neq 0$ [19].

In our example, the treatment has a cause effect on P_1 while it does not on P_2. We are less interested in the individual causal effect, but more on the average causal effect in the population.

We can define *the average causal effect* of treatment A on outcome Y as $E[Y^{a=1} - Y^{a=0}]$ where $E[\,\cdot\,]$ stands for expectation or the average over a population.

We can view the average causal effect as an averaged contrast of receiving a heart transplant and not receiving one. If the treatment were not dichotomous, then we need to specify the particular contrast of interest. Note also that, depending on the nature of the outcome (continuous or discrete, respective computation of $\mathbb{E}[Y^a]$ must be taken i.e., one involving summation or integration.

One can define different measures of causal effect, depending on the problem. For instance, below we have three examples

(i) $P[Y^{a=1} = 1] - P[Y^{a=0} = 1]$,

(ii) $\dfrac{P[Y^{a=1} = 1]}{P[Y^{a=0} = 1]}$, and

(iii) $\dfrac{P[Y^{a=1} = 1]/P[Y^{a=1} = 0]}{P[Y^{a=0} = 1]/P[Y^{a=0} = 0]}$

which are referred to as the causal risk difference, risk ratio, and odds ratio, respectively.

Before, we proceed any further, we want to establish a clearer distinction between $\mathbb{E}[Y^a]$ and $\mathbb{E}[Y \mid A]$. Causal inference focuses on the counterfactual outcomes, Y^a, and at finding the average causal effect. But observe that, we never know fully what is Y^a so what one might be able to model $P[Y|A = 1]$ and $P[Y|A = 0]$ or the outcome under the true treatment which is observable within our data. Figure 3.1 (reproduced from Hernán and Robins, *Causal Inference* [19]) illustrates graphically the difference the counterfactual and actual outcomes.

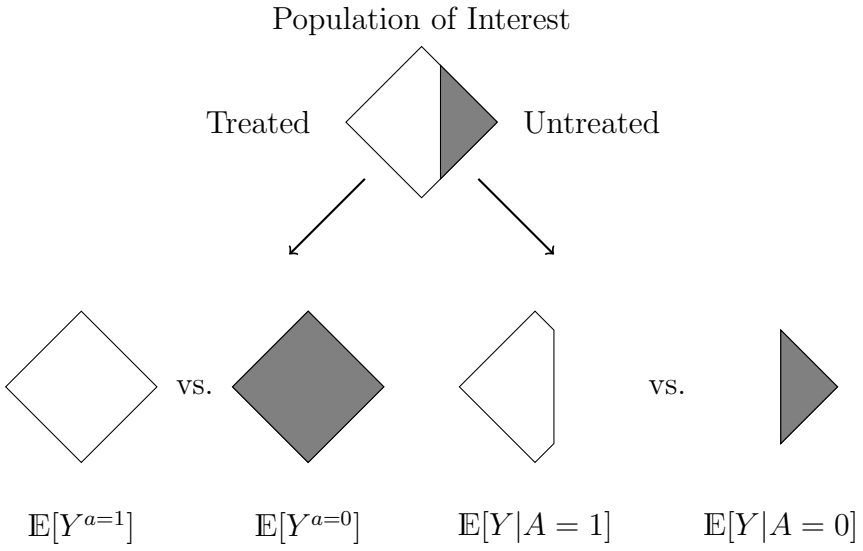


Figure 3.1: Distinction between $E[Y^a]$ and $E[Y|A = a]$ [19]

3.2.2 Conditions

In this section, we will introduce conditions that will be needed and assumed to hold in order for the models that will be discussed later to be reliable. One can view these conditions as the assumptions one makes when fitting a model.

3.2.2.1 Well-defined interventions

Before any analysis is being performed, we need to be sure that the treatment or intervention in question are well-defined, where well-defined means that any differences in treatment are ignorable with respect to expectation in the outcome. Consider the heart transplant question introduced earlier. Suppose that the question we aim to answer is: "Does heart transplant increase the risk of death?" Although this seems a straight-forward question, the challenge is that there is more than one way to perform a heart transplant (e.g. different surgical approaches, warm versus cold heart transplants). In order to consider this intervention well-defined, one must assume the differences in the transplant method used in the study were restricted to methods that do not affect the expectation of the outcome.

So the important message to bear is that well-defined intervention is crucial and must not be simply put aside.

3.2.2.2 Exchangeability

Definition 3.2.1 *Exchangeability is defined as when the counterfactual outcome risk under every exposure value a is the same in the exposed and in the unexposed* [19].

Mathematically, Definition 3.2.1 is equivalent to

$$P[Y^a = 1|A = 1] = P[Y^a = 1|A = 0]$$

A consequence of these conditional risks being equal in all subsets defined by treatment status in the population is that they must be equal to the marginal risk under treatment value a in the whole population, or in other words

$$P[Y^{a=1} = 1|A = 1] = P[Y^{a=1} = 1|A = 0] = P[Y^a = 1].$$

Because the counterfactual risk under treatment value a is the same in both groups $A = 1$ and $A = 0$, we say that the actual treatment A does not predict the counterfactual outcome Y^a. So an equivalent definition of exchangeability is :

Definition 3.2.2 *Exchangeability is when the counterfactual outcome and actual treatment are independent* [19].

Definition 3.2.2 will be mathematically expressed as

$$Y^a \perp\!\!\!\perp A \quad \forall a$$

Note, it is important to notice that $Y^a \perp\!\!\!\perp A$ and $Y \perp\!\!\!\perp A$ are two different things. The former was already defined above. The second expression, $Y \perp\!\!\!\perp A$, claims independence between observed outcome and treatment. Note also that one does not imply the other.

3.2.2.2.1 Conditional Exchangeability

Observe that so far no predictor(s) have been discussed or included in any definition or topic discussed. For instance, if we go back to the heart transplant example, considering the age of individuals, smoking status, their sport life, or other health factors will be

essential for the analysis. Ideally, we are including variables in our data. We will denote the set of variables by L and l would define a specific stratum of L. In relation to the previous section, we have the following defintion

Definition 3.2.3 *Conditional Exchangeability is defined as* [19].

$$Y^a \perp\!\!\!\perp A|L$$

Note that in the above expression, we assume independence holds within each level of L, but this may not be the case. In such situations, we would write

$$Y^a \perp\!\!\!\perp A|L = l$$

which is conditional exchangeability within a specific level of $L = l$.

A natural question to ask now is: when does exchangeability or conditional exchangeability actually hold? From Definition 3.2.3, one must verify that

$$P[Y^a = 1|A = a, L = l] = P[Y^a = 1|A \neq a, L = l],$$

but for individuals who did not receive the treatment the value of Y^a is unknown, so the right hand side is unknown. So does this mean we can never guarantee exchangeability? Realistically speaking, yes. Nonetheless, at least conditional exchangeability must be assumed to be true in order to identify parameters from observed data.

In fact, if one performs a randomized experiment, exchangeability would indeed hold. Moreover, under such an experiment with A dichotomous,

$$P[Y^{a=1} = 1] = P[Y = 1|A = 1] \text{ and}$$

$$P[Y^{a=0} = 1] = P[Y = 1|A = 0];$$

hence, we can compute the counterfactuals which were previously unknown. Note, A does not necessarily need to be dichotomous, it can have more than 2 levels, but we then need to specify the contrast that will be considered as mentioned previously in Individual and Average Causal Effects.

3.2.2.3 Positivity

Suppose that we perform a study with two treatments, and we assign all subjects to either $A = 1$ or $A = 0$. It should be clear that under such study, it will be impossible to compute *any* average causal effect or test in general if the treatment is effective since we lack subjects in one of the treatment levels. So, in order to be able to estimate any effect, we need to ensure that subjects are assigned to each level of the treatment groups. In other words, we must ensure that there is a probability greater than zero – a positive probability for each subject to be assigned to each of the treatment levels. This is referred to as *positivity*.

Definition 3.2.4 *Given a vector, L, of covariates, positivity defined as* [19]*:*

$$P[A = a|L = l] > 0 \quad \forall l \text{ with } P[L = l] \neq 0$$

As with exchangeability, we cannot generally test if positivity holds in our study. Positivity is often violated in clinical data unless some restrictions are applied to the data. Referring back to the heart transplant scenario, suppose we have a single variable, L, that takes value 1 if a person is in critical condition and 0 if he is not, then $P[A = 1|L = 0]$ is likely to be 0 since a person who is not in critical condition would almost certainly not receive a heart transplant.

3.2.2.4 Confounding

Confounding is present in many areas of statistics and it has an important role in causal inference. It can be viewed from the perspective of Directed Acyclic Graphs, but in the following section confounding will be very briefly introduced from both theoretical and graphical view.

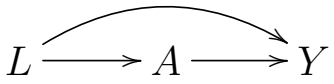


Figure 3.2: *Example of a backdoor path*

Confounding is the bias that arises when the treatment and the outcome share a cause [19]. Figure 3.2 is a graphical illustration of confounding. The diagram shows two sources of association between treatment and outcome: (1) the path $A \to Y$ that represents the causal effect of A on Y, and (2) the path $A \leftarrow L \to Y$ between A and Y that is induced by the common cause L. This is known as a *backdoor path* [19]. Figure 3.2 is called a Directed Acyclic Graph (DAG). DAG's help the analyst to visualize the relation between all the variables and identify backdoor paths and other characteristics. A full discussion of Directed Acyclic Graphs and confounding is beyond the scope of this thesis, so we will very briefly introduce the topic and its use.

Suppose, we are interested on the effect of some drug, A, on the risk of some heart related issue (say stroke), Y. The effect will be confounded if the drug is more likely to be prescribed to individuals with a certain condition L (say, heart disease).

Hence, eliminating confounding is essential. Under what conditions can confounding be eliminated? A result from graph theory known as the *backdoor criterion* guarantees that the causal effect of A on Y is identifiable if all backdoor paths between them can

be blocked by conditioning on variables that are not affected by non-descendants of treatment A [19]. More simply, we will assume to have measured enough variables, L, so that when we condition on L, we will be sure to have blocked all backdoor paths. The backdoor criterion aims at 3 questions:

1. Does confounding exist?

2. Can confounding be eliminated?

3. What variables are necessary to eliminate confounding?

The moral from the above is that in order to *remove* all confounding, one simply conditions on L. Unfortunately, sometimes conditioning on a variable(s) may cause additional problems as shown in Figure 3.3:

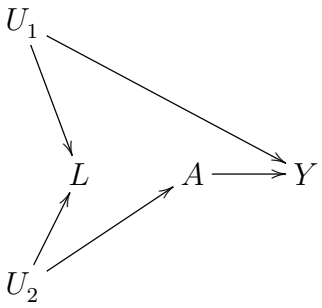


Figure 3.3: Opening of backdoor by conditioning on L

If we do not condition on L, there is no confounding at all, but if we condition on L, we will open backdoor paths via U_1 and U_2; this is known as collider stratification bias. The objective here is to illustrate that conditioning on a variable is not always necessary. Hence, drawing a DAG can help deciding whether conditioning on a variable is required or not.

For further theory and propertieson DAGs, the reader may refer to Chapter 6 and 7 of *Causal Inference* by Miguel A. Hernán and James M.Robins [19].

3.3 G-methods preview

In section 3.2.2.4, confounding was introduced and we discussed broadly how one adjusts for confounding. In the coming sections, we answer the latter problem more formally by introducing the class of G-methods which includes standardization, IP weighting, and g-estimation. These are methods that exploit conditional exchangeability in subsets defined by L to estimate the causal effect of A on Y in the entire population or in any subset of the population [19].

G-methods require conditional exchageability given the measured covariates L in order to produce consistent parameter estimates. Furthermore, Inverse Probability (IP) weighting and standardization require positivity and well-defined interventions. Recall, we defined the average causal effect as

$$\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}]$$

where $\mathbb{E}[Y^{a=1}]$ is the mean outcome that would have been observed if all individuals in the population had received the treatment and $\mathbb{E}[Y^{a=0}]$ is the mean outcome that would have been observed if all individuals in the population had not received the treatment. In the coming sections, we establish how the G-methods estimate the above difference.

3.3.1 IP Weighting and Estimating IP Weights

IP weighting adjusts for confounding by creating what is known as a pseudo-population. By creating such a population, the goal is remove the arrow from L to A in Figure 3.2. Thus, if we measure enough confounders, L, we should be able to block all the back-door paths from A to Y, and consequently remove all the confounding in the pseudo-population. That is, the association between A and Y in the pseudo-population consis-

tently estimates the causal effect of A on Y.

A pseudo-population is twice as big as our original population where we simulate the outcome of each individual if he is assumed to have been and not have been treated [19]. Below, we make a visualization of a pseudo-population. Figure 3.4 represents a hypothetical population of 20 individuals where for simplicity Y, A, and L are dichotomous variables. $Y = 1$ means the outcome was observed while $Y = 0$ it was not. $A = 1$ means a subject received the treatment and $A = 0$ he did not. L has two levels for some confounder. The numbers represent the number of individuals within each group while the numbers in parentheses are the probability of being in that specific strata. For example,

$$P[Y = 1|A = 0, L = 0] = \frac{1}{4}$$

The pseudo-population would be created by assuming that everybody had received the treatment and had not as mentioned above. So Figure 3.4 *transforms* to Figure 3.5.

If one looks carefully, the outcomes in the pseudo-population can be obtained by weighting each individual by the inverse of the conditional probability of receiving the treatment level that he indeed received.

For example, take again the branch where $L = 0, A = 0$, and $Y = 1$, in the pseudo-population we have 2 individuals for whom the outcome was observed given they did not receive the treatment and $L = 0$.

From Figure 3.4, $P[A = 0|L = 0] = 0.5$ so we see that $1/0.5 = 2$ where "1" really represents the number of individuals for whom $Y = 1$ given $A = 0$ and $L = 0$, and "2" is for the same branch ($Y = 1$ given $A = 0$ and $L = 0$) in Figure 3.5.

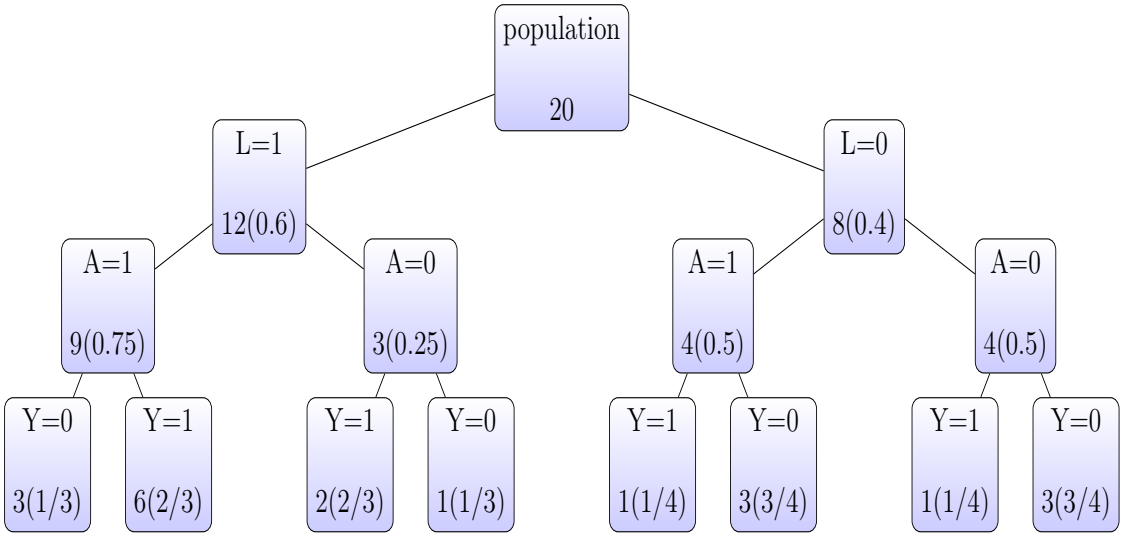


Figure 3.4: Hypothetical population

Hence, we define the individual IP weights for treatment A as

$$W^A = \frac{1}{f(A|L)}$$

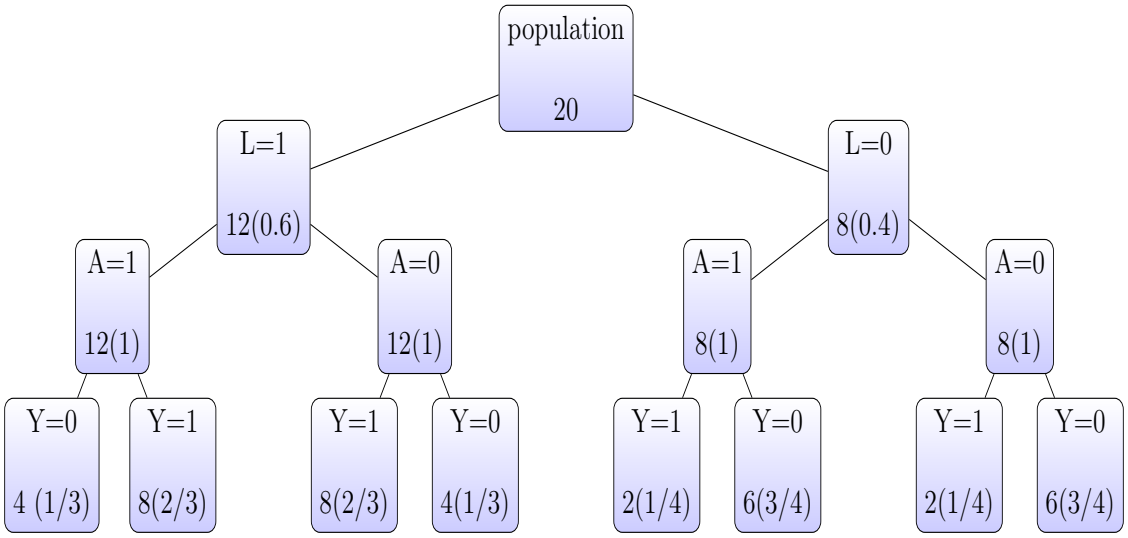


Figure 3.5: pseudo-population

where $f(A|L)$ is the probability of receiving a treatment level A conditional on the measured confounders, L i.e., $P[A = 1|L]$ for treated. Likewise for untreated subjects, we

would have $P[A = 0|L]$. Since A is a binary variable, $P[A = a|L]$ can be estimated by fitting a logistic model. $f(A|L)$ is also referred to as the propensity score.

Next, we approximate $\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$ in the pseudo-population created by using the estimated IP weights. By creating the pseudo-population, we remove the confounding bias and $\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$ is a consistent estimator of $\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}]$ [19].

To estimate $\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$, we fit the following model to the observed data

$$\mathbb{E}[Y|A] = \beta_0 + \beta_1 A \tag{3.1}$$

by weighted least squares, with individuals weighted by their estimated IP weights; $1/\hat{P}[A = 1|L]$ for subjects who received the treatment and $1/\hat{P}[A = 0|L]$ for subjects who did not received the treatment. Under conditional exchangeability, the IP weighted mean is equal to $\mathbb{E}[Y^a]$ – see proof below.

Assuming no censoring, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{I(A=a)Y}{f[A|L]}\right] &= \mathbb{E}\left[\frac{I(A=a)Y^a}{f[A|L]}\right], \text{ via consistency} \\
&= \mathbb{E}\left\{\mathbb{E}\left[\frac{I(A=a)Y^a}{f[a|L]}\middle|L\right]\right\}, \text{ via } \mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|L]] \\
&= \mathbb{E}\left\{\mathbb{E}\left[\frac{I(A=a)}{f[a|L]}\middle|L\right]\mathbb{E}[Y^a|L]\right\}, \text{ via conditional exchangeability} \\
&= \mathbb{E}\left\{\mathbb{E}[Y^a|L]\right\}, \text{ since } \mathbb{E}\left[\frac{I(A=a)}{f[a|L]}\middle|L\right] = 1, \text{ so} \\
&= \mathbb{E}[Y^a].
\end{aligned}$$

The left hand side of the first line is the mean under the IP weighted model. Hence, establishing the result.

3.3.1.1 Stabilized IP Weights

In the above section, we introduced the notion of pseudo-population, and how it is used to estimate the average causal effect. We created such population by weighting by the probability of being treated and not treated conditional on the set of covariates L i.e, weight by the IP weights given by

$$W^A = \frac{1}{f(A|L)}.$$

An issue with the above weights is that individuals with a propensity score close to 0 – those extremely unlikely to be treated – will have a very large weight; thus, making the weighted estimator unstable. We can stabilize the weights by using a more general form of W^A, that being

$$W^A = \frac{f(A)}{f(A|L)} \tag{3.2}$$

where $f(A)$ is the probability of receiving the treatment or the marginal probability of treatment i.e., $P[A = 1]$. If there is more than one treatment regime, then $f(A)$ is just $P[A = a]$. For binary treatment, $P[A = a]$ is estimated by fitting a logistic model with no predictors. The weights in 3.2 are usually referred to as *stabilized weights* which we will denote by SW^A [19]. W^A are usually referred to as *non-stabilized weights*. Generally, using SW^A will result in a more efficient causal estimator than using W^A [36, 37].

Previously, we established one method to estimate $\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}]$ that is by creating a pseudo-population, and using the fact that $\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$ is a consistent estimator of the desired quantity. We estimated $\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$ by fitting $\mathbb{E}[Y|A] = \beta_0 + \beta_1 A$ via weighted least squares. What if we want to directly fit (or

estimate) $\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}]$? Let us propose the following model

$$\mathbb{E}[Y^a] = \beta_0 + \beta_1 a. \tag{3.3}$$

Note that this model is unique in the sense that its outcome is counterfactual; hence, unobservable. Such models are referred to as *structural mean models* [19]. Observe that the model in 3.3 does not include any covariates in which case we will call such model an *unconditional* or *marginal structural mean model*. From 3.3, we see that

$$\mathbb{E}[Y^{a=0}] = \beta_0$$

so

$$E[Y^{a=1}] = \beta_0 + \beta_1 \Rightarrow E[Y^{a=1}] = E[Y^{a=0}] + \beta_1 \Rightarrow E[Y^{a=1}] - E[Y^{a=0}] = \beta_1;$$

hence, estimating β_1 is equivalent to estimating the average causal effect.

In the previous subsection, we introduced marginal structural models. If we want to estimate the average causal effect, the model in 3.3 will suffice. Yet, if we want to assess a causal effect that depends on some covariate, then we need to include that covariate in the model. By "effect," we mean, for example, the fact of being man or woman or being smoker versus not, etc. If we will symbolize our covariate by X, then one simple model would be

$$E[Y^a|X] = \beta_0 + \beta_1 a + \beta_2 X a. \tag{3.4}$$

Note that the expression in 3.4 is not general, but rather one possibility of a fit. We would estimate β_j $(j = 0, 1, 2)$ in 3.4 by fitting a linear regression model [19] :

$$E[Y|A, X] = \tilde{\beta}_0 + \tilde{\beta}_1 a + \tilde{\beta}_2 X a \tag{3.5}$$

via weighted least squares using either the weights being W^A or SW^A.

3.3.1.2 Censoring

Until now, we assumed that every subjects in our study remained in it until the outcome could be measured. This may not be case necessarily, in other words, we have censoring. In fact, all the analyses above were referring to the uncensored scenario because those are the only ones with known outcome, Y. Letting C represent the censoring indicator with $C = 0$ for being uncensored and $C = 1$ for being censored, then instead of fitting 3.3, namely

$$E[Y|A] = \beta_0 + \beta_1 A,$$

we can only fit

$$E[Y|A, C = 0] = \beta_0 + \beta_1 A, \tag{3.6}$$

to the observed data. The target average causal effect would then become

$$E[Y^{a=1,c=0}] - E[Y^{a=0,c=0}] \tag{3.7}$$

where $E[Y^{a=1,c=0}]$ would be read as the average outcome if everybody had received the treatment *and* nobody had been censored. Similar interpretation follows for $E[Y^{a=0,c=0}]$. We can estimate 3.7 constitently still using IP weights, except that now we will need to adjust for the effect of the censoring as well [19] which is done by using the weights

$$W^{A,C} = \frac{1}{f(A, C = 0|L)} = W^A \times W^C$$

where

$$f(A, C = 0|L) = f(A|L) \times P[C = 0|L, A],$$

so it follows that

$$W^C = \frac{1}{P[C = 0|L, A]}.$$

Note that one can also use the stabilized version which is $SW^{A,C} = SW^{A} \times SW^{C}$ where

$$SW^{C} = \frac{P[C=0|A]}{P[C=0|L,A]}.$$

$P[C=0|L,A]$ is estimated the same way we did for $P[A|L]$ e.g., we fit a logistic regression model for the probability of being uncensored. Remember, the above would hold if we have exchangeability; so in the presence of censoring, we must assure exchangeability for the joint "exposure" (A,C) conditional on L i.e., $Y^{a=1,c=0} \perp\!\!\!\perp (A,C)|L$.

3.3.2 Standardization and g-formula

Using IP weights to estimate the average causal effect is one method, but it is not the only one. In the following section, we introduce two alternative methods, standardization and the g-formula. We also ended the last section by introducing censoring in our analysis and specified that we are in fact estimating

$$E[Y^{a=1,c=0}] - E[^{a=0,c=0}]$$

rather than

$$E[Y^{a=1}] - E[^{a=0}].$$

Henceforth, we will assume the presence of censoring, unless specified otherwise. Also, we will still assume the vector of variables, L, is sufficient to adjust for confounding.

Under exchangeability and positivity conditional on L, the standardized mean outcome in the uncensored treated subjects is a consistent estimator of $E[Y^{a=1,c=0}]$ and the same for $E[Y^{a=0,c=0}]$; see proof below. To compute the standardized mean outcome in the uncensored treated, we first need to compute the mean outcomes in the uncensored treated in each stratum l of the confounders L, i.e., the conditional means

$E[Y|A = 1, C = 0, L = l]$. We then weight the latter quantity by $P[L = l]$. The standardized mean for uncensored observations who received treatment level a is

$$\sum_l E[Y \mid A = a, C = 0, L = l] \times P[L = l]. \tag{3.8}$$

We now describe how to estimate the conditional means in 3.8 and an alternative to estimate the standardized means. Note that some of the variables in L may be continuous in which case the sum becomes an integral.

Assuming we do not have censoring:

$$\begin{aligned} E[Y^a] &= E\big[E[Y^a|L]\big], \text{ via } E[X] = E[E[X|L]] \\ &= \sum_l E[Y^a|L = l]P[L = l] \\ &= \sum_l E[Y^a|A = a, L = l]P[L = l], \text{ via conditional exchangeability} \\ &= \sum_l E[Y|A = a, L = l]P[L = l] \text{ , via consistency.} \end{aligned}$$

3.3.2.1 Estimating the Mean Outcome

We want to estimate : $E[Y \mid A = a, C = 0, L = l]$. In the case of IP weights, we fitted logistic models, here to estimate $E[Y \mid A = a, C = 0, L = l]$, we fit a linear regression model for the mean outcome with treatment A and variables in L included as covariates [19]. Note, the model can include n degree terms, or any other functional forms, if needed. Lastly, we weight by $P[L = l]$.

If, for simplicity, all variables in L are discrete, we can calculate the $P[L = l]$ non-parametrically – simply divide the number of subjects in the strata defined by $L = l$ by the total number of subjects in the population. However, this can be very tedious

when L has variables with many levels and if L has higher dimension. What if L includes continuous variable? Hence, we propose an alternative method [19] in estimating the standardized means.

Assume a very simple data of 10 subjects (summarized in Table 3.1 where L is a dichotomous variable, A and Y have two levels and there is no censoring. Recall under censoring, we want to approximate

$$\sum_{l} E[Y \mid A = a, C = 0, L = l] \times P[L = l],$$

under no censoring we are estimating

$$\sum_{l} E[Y \mid A = a, L = l] \times P[L = l].$$

The process of estimating the standardized means is based on 4 steps. First, we create 2 copies of Table 3.1 as shown in Table 3.2.

In simple words, we created one copy where we set the value of $A = 0$ and another where $A = 1$, and in both blocks we set Y to be unknown. Note that Block 0 is just our initial data. As one might guess, we will use Block 1 to estimate the standardized mean in the treated and Block 2 for the standardized mean in the untreated.

Second, we fit a regression model for the mean outcome given treatment A and the confounder L as described above. Observe that only Block 0 will contribute to the estimation as in Block 1 and 2, the outcome values are unknown. Third, we use the parameter estimates to estimate the Y values in Block 1 and Block 2.

Lastly, we compute the average of all predicted values in Block 1 corresponding to the standardized mean in the treated. We do the same for Block 2, obtaining estimate for the standardized mean in the untreated. We then take the difference of the two averages to obtain the estimate of the average causal effect.

	L	A	Y
P_1	0	1	1
P_2	0	1	1
P_3	1	0	0
P_4	1	1	1
P_5	0	0	0
P_6	0	1	0
P_7	0	1	0
P_8	1	0	1
P_9	1	1	0
P_{10}	0	0	1
Block 0			

Table 3.1: Hypothetical Dataset

	L	A	Y	L	A	Y	L	A	Y
P_1	0	1	1	0	1	·	0	0	·
P_2	0	1	1	0	1	·	0	0	·
P_3	1	0	0	1	1	·	1	0	·
P_4	1	1	1	1	1	·	1	0	·
P_5	0	0	0	0	1	·	0	0	·
P_6	0	1	0	0	1	·	0	0	·
P_7	0	1	0	0	1	·	0	0	·
P_8	1	0	1	1	1	·	1	0	·
P_9	1	1	0	1	1	·	1	0	·
P_{10}	0	0	1	0	1	·	0	0	·
Block 0				Block 1			Block 2		

Table 3.2: Data Augmentation

3.3.3 IP Weighting or Standardization?

We have seen two methods to estimate the average causal effect : (1) IP weighting and (2) standardization. One may ask should we use one method over the other? In turns out that these two ways are equivalent under non-parametric model assumptions. From sections 3.3.1 (IP Weighting and Estimating IP Weights) and 3.3.2 (Standardization and g-formula), we know now that $\mathbb{E}\left[\frac{I(A=a)Y}{f[A|L]}\right]$ is the IP weighted mean of Y for treatment level a and $\sum_l \mathbb{E}[Y|A=a, L=l]P[L=l]$ is the standardized mean for treatment level a, respectively. We will assume that $f[a|l]$ is positive $\forall l$ and omit censoring, then

$$\mathbb{E}\left[\frac{I(A=a)Y}{f[A|L]}\right] = \mathbb{E}\left[\frac{I(A=a)Y}{f[A|L]}\middle| A=a, L=l\right] \qquad \text{via the tower property,}$$

$$= \sum_l \frac{1}{f[a|l]} \mathbb{E}[Y|A=a, L=l] f[a|l] P[L=l] \qquad \text{and}$$

$$= \sum_l \mathbb{E}[Y|A=a, L=l] P[L=l].$$

Generally one should use both IP weights and standardization and compare the estimates. A large difference would alarm the analyst of moderate model misspecification [19]. In the example above, we were parametrically estimating conditional mean outcome via standardization. This is a particular case of the *parametric g-formula*.

3.3.4 Required Conditions for IP Weighting and Standardization

We saw two methods of estimating the average causal effect. We then briefly said that, one should not be preferred over the other, but rather both methods should be applied. Once this is done, another natural question is: what is the validity of our estimates? This boils down to, how well were the conditions required by each method met? These conditions are: **(1)** exchangeability, **(2)** positivity, **(3)** well-defined interventions, **(4)** no measurement error, and **(5)** no model misspecification.

The first three were already discussed. The 4^{th} one is saying that no measurement error should be present in the treatment A, the outcome Y, and the confounders L. Otherwise, bias will be introduced. And model misspecification was mentioned in the previous section. Realistically, we can rarely ensure that all these conditions hold; the more we deviate from them, the more our estimate may be biased and less valid.

3.3.5 G-estimation

In this section, we introduce the third and last method – g-estimation – for estimating the average causal effect. G-estimation attempts to estimate

$$\mathbb{E}[Y^{a=1}|L] - \mathbb{E}[Y^{a=0}|L],$$

which is the average causal effect of treatment A on some outcome Y in each strata defined by the covariates L. Models whose parameters are estimated via g-estimation are known as *structural nested models* [19]. In the presence of censoring, the estimate of interest is

$$\mathbb{E}[Y^{a=1,c=0}|L] - \mathbb{E}[Y^{a=0,c=0}|L]$$

Before continuing further, expressing conditional exchangeability in terms of the conditional probability of treatment will be helpful when we describe g-estimation later. Therefore, we propose the following model

$$\text{logit}\, P[A = 1|Y^{a=0}, L] = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L. \tag{3.9}$$

3.3.5.1 Estimating the Mean Outcome

The goal of g-estimation is to estimate

$$\mathbb{E}[Y^{a=1,c=0}|L] - \mathbb{E}[Y^{a=0,c=0}|L].$$

For simplicity, let us omit censoring so that we are estimating

$$\mathbb{E}[Y^{a=1}|L] - \mathbb{E}[Y^{a=0}|L].$$

If L induces no effect modification, then we can assume that

$$\mathbb{E}[Y^{a=1}|L] - \mathbb{E}[Y^{a=0}|L] = \beta_1,$$

where β_1 would be the average causal effect in each strata of L. We can rewrite the above expression as

$$\mathbb{E}[Y^a|L] - \mathbb{E}[Y^{a=0}|L] = \beta_1 a$$

notice that for $a = 1$ we obtain the desired expression while for $a = 0$ we obtain 0. If L induces some effect modification, we need to introduce it in our model so that we have

$$\mathbb{E}[Y^a|L] - E[Y^{a=0}|L] = \beta_1 a + \beta_2 aL,$$

then under exchangeability:

$$\mathbb{E}[Y^a|L] - E[Y^{a=0}|A = a, L] = \beta_1 a + \beta_2 aL. \tag{3.10}$$

The model in 3.10 is referred to as a *structural nested mean model* [19]. In the presence of censoring, one simply has

$$\mathbb{E}[Y^{a=1,c=0}|A, L] - \mathbb{E}[Y^{a=0,c=0}|A, L]$$

Estimating 3.10 requires adjusting for both confounding and bias, but we know that IP weighting and standardization can adjust for both. So what one may do is first use IP weights to create a pseudo-population and apply g-estimation to the pseudo-population [19]. In section 3.3.1.1 (Stabilized IP Weights), we introduced the concept when covariates are present in marginal structural models. Suppose we have the following marginal structural model

$$\mathbb{E}[Y^a|X] = \beta_0 + \beta_1 a + \beta_2 Xa + \beta_3 X \tag{3.11}$$

where X is a covariate in L. Then,

$$\mathbb{E}[Y^{a=1}|X] = \beta_0 + \beta_1 + \beta_2 X + \beta_3 X \tag{3.12}$$

and

$$\mathbb{E}[Y^{a=0}|X] = \beta_0 + \beta_3 X \tag{3.13}$$

so that subtracting 3.13 from 3.12

$$\begin{aligned}
\mathbb{E}[Y^{a=1}|X] - \mathbb{E}[Y^{a=0}|X] &= \beta_1 + \beta_2 X \\
\mathbb{E}[Y^{a=1}|X] &= \mathbb{E}[Y^{a=0}|X] + \beta_1 + \beta_2 X \\
\Rightarrow E[Y^a|X] &= \mathbb{E}[Y^{a=0}|X] + \beta_1 a + \beta_2 X a. \qquad (3.14)
\end{aligned}$$

Hence, $\beta_1 + \beta_2 X$ is the average causal effect in the stratum $X = x$.

Marginalizing over X, 3.11 yields a marginal structural mean model. If we are not interested at estimating $\mathbb{E}[Y^{a=0}|X]$ and leave this term completely unspecified, then the model in 3.14 is referred to as a *semiparametric marginal structural mean model* [19]. Looking at the expression in 3.14 and 3.10, we notice a similarity between them, so we established a connection between marginal structure mean models and structural nested models.

3.3.5.2 Rank Preservation

In general, we can rank our subjects according to their actual outcome, Y. What we mean by this is to order individuals by decreasing order with respect to their outcome. Similarly, we can do the same for the counterfactual outcomes $Y^{a=1}$ and $Y^{a=0}$ if they were known. This will result into two lists; if the two lists are in identical order, that is to say if in the i-th row for $Y^{a=1}$ we find the same individual in the i-th row for $Y^{a=0}$, we then say that there is *rank preservation* [19].

If the effect of treatment A on the outcome Y is additive for all subjects in the population, we say that *additive rank preservation* holds. When we say "additive", we

mean that for individual i we will have some outcome estimate for $Y^{a=0}$ while the estimate of $Y^{a=1}$ for that same individual will be $Y^{a=0}$ plus some constant, m, or simply shifted. If we are interested in the invididual causal effect from a structural nested mean model, we require additive rank preservation within levels of L. In other words, we have *conditional additive rank preservation* which holds if the effect of treatment A on the outcome Y is exactly the same for all individuals with the same values of L [19]. An example of conditional additive rank-preserving structural model is

$$Y_i^a - Y_i^{a=0} = \varphi_1 a + \varphi_2 a L_i \quad \forall i$$

where for $a = 1$, $\varphi_1 + \varphi_2 L_i$ is the constant causal effect for all subjects with covariate values $L = l$. Unfortunately, the additive rank assumption will rarely hold. Yet for simplicity, we will present g-estimation from the perspective of additive rank preservation. We now proceed with formally presenting g-estimation.

3.3.5.3 g-estimation

Let us suppose that we want to estimate the parameter of the model in 3.10 with the second term being zero i.e.,

$$E[Y^a|L] - E[Y^{a=0}|L] = \beta_1 a$$

in simple words, we assume that the average causal effect is constant across stata of L, and we seek an estimate for β_1. We will assume the additive rank-preserving model holds i.e.,

$$Y_i^a - Y_i^{a=0} = \varphi_1 a, \tag{3.15}$$

so that the individual causal effect φ_1 is equal to the average causal effect β_1 of which we are interested at. Omitting the index i since we assume 3.15 is correctly specified and

hence the model is the same for all individuals, we can rewrite 3.15 in the following form

$$Y^{a=0} = Y - \varphi_1 A \tag{3.16}$$

where we used the consistency of Y^a and Y. Consequently, we change the a in the structural model to A as written in 3.16. Why? Remember that by consistency the observed outcome Y is the counterfactual $Y^{a=0}$ if an individual did not receive the treatment $(A = 0)$ or $Y^{a=1}$ if an individual received the the treatment $(A = 1)$. Hence, we need to fix a by the actual value of A – 1 or 0. Remember, our goal is estimating φ_1. We propose the following [19]. Let,

$$H(\varphi^\dagger) = Y - \varphi^\dagger A. \tag{3.17}$$

If we can find $\varphi^\dagger$ such that $\varphi^\dagger = \varphi_1$, then $H(\varphi^\dagger) = Y^{a=0}$. In other words, we want to find the right $H(\varphi^\dagger)$ so that it equals the true counterfactual. So how does one find $H(\varphi^\dagger)$? Under conditional exchangeability, we can fit a model as the one in 3.9 (the reason it was introduced) where α_1 should be 0 [19]. So we can connect $H(\varphi^\dagger)$ to A via a model of the form

$$\text{logit}\, P[A = 1|H(\varphi^\dagger), L] = \alpha_0 + \alpha_1 H(\varphi^\dagger) + \alpha_2 L. \tag{3.18}$$

We need to assess candidates $H(\varphi^\dagger = \varphi^0)$ where φ^0 is some pre-assumed value, we then fit the model in 3.18 and the model for which we fail to reject the null hypothesis for $\alpha_1 = 0$ is the one that returns an estimate, $\varphi^\dagger$, or the true value φ_1. We ideally seek the model that has the weakest evidence against the null i.e., the highest p-value.

What one might ask is: how do we find candidate values for $\varphi^\dagger$? Usually, we might have some idea where the value of $\varphi^\dagger$ should be given the context of the study (say between $[-1, 1]$) and try different values in that interval – $[-1, 1]$ – by increments of 0.01 or 0.001 for more precision.

So far, we assumed that there is no effect modification by L. Generally, such assumption would almost never hold. So we need to include some variables $X \in L$ in our model.

Note that conditioning on X, marginal structural nested models estimate the average causal effect within levels of X. Whereas those that do not condition X, estimate the latter in the whole population. By definition and how structural nested models were defined (above sections), they estimate the average causal effect within levels of L, and not in the whole population. Hence, omitting L (or subset $X \in L$) in the model leads to misspecification and hence bias [19].

Fortunately, we do not need to redefine and re-derive all the above, as it can be simply extended by adding covariates. For instance, suppose we have the structural nested model of the form

$$E[Y^a|A = a, L] - E[Y^{a=0}|A = a, L] = \beta_1 a + \beta_2 aX$$

so that the corresponding rank-preserving model is

$$Y_i^a - Y_i^{a=0} = \varphi_1 a + \varphi_2 aX,$$

then the equivalent to 3.18 is

$$\text{logit}\, P[A = 1|H(\varphi^\dagger), L] = \alpha_0 + \alpha_1 H(\varphi^\dagger) + \alpha_2 H(\varphi^\dagger)X + \alpha_3 L. \tag{3.19}$$

So now, we need to find $\varphi_1^\dagger$ and $\varphi_2^\dagger$ that make both α_1 and α_2 equal to 0. This can be extended to more than two $\varphi^\dagger$'s.

3.4 Summary

In this chapter, we introduced the basics of causal inference; we then established what a causal estimand is. We then saw some conditions that are necessary for the analyses to hold; exchangeability being the main one. Finally, we covered 3 methods – IP weights, standardization, and g-estimation which are subgroups of so called G-methods – that allowed us to estimate the average causal effect. In the next Chapter, we combine Chapters 2 and 3

Chapter 4

Competing Events & Causal Inference

In the two previous chapters, we introduced causal inference and competing events as two independent subjects. In the following chapter, we introduce an approach to analyze a causal inference question in the presence of competing events.

Some common methods that are used to tackle the latter problem are Instrumental Variables (IV) [34] [42], pseudo-observations [2], or both [25]. A less rigorous, yet more intuitive approach was performed by Bolch, Charlotte A. et al. [6] which was to weight the Cumulative Incidence Curve using inverse probability weights.

Another approach discussed by Egleston et al. [11], and Naimi and Tchetgen Tchetgen [30] is to estimate a quantity known as the survivor average causal effect (SACE) which is what we will focus on and estimate in our analysis.

4.1 SACE Estimation

We will follow the notation used by [30] as we discuss the SACE for competing risk data.

Naimi and Tchetgen Tchetgen propose to estimate the SACE using a structural nested survival time model. Suppose we have a failure time outcome, T_δ, that may end in 1 of 2 possible events: death due to an outcome of interest ($\delta = 1$) or death due to a competing risk ($\delta = 2$). Consider further a binary time-varying exposure, $A(j)$, and a set of confounders, W, that can take values for each possible time points until t where t represents the largest integer value just before $T_{\delta=1}$. The Structural Nested Accelerated Failure Time (SNAFT) model can be defined as [30]

$$T^{\bar{0}} = \int_0^{T_{\delta=1}} \exp[\psi A(u)]\, du$$

where $T^{\bar{0}}$ is the outcome that would have been observed under no exposure and ψ is the causal parameter for the time-varying exposure and the cause-specific failure-time, $T_{\delta=1}$. G-estimation of ψ from a structural nested failure time model can be implemented by the following algorithm:

1. Choose a set of candidate values, $\hat{\psi}$ for the true value of the structural nested model parameter.

2. For each candidate from step 1, using the observed failure time, and observed exposure history, we compute

$$H(\hat{\psi}) = \int_0^T e^{\hat{\psi} \cdot A(t)}\, dt.$$

3. Perform a hypothesis test of a null hypothesis that, conditional on all measured confounders, the exposure is independent of $H(\hat{\psi})$.

For a binary exposure, step 3 is usually implemented via a pooled logistic regression model, and the $\hat{\psi}$ value for which the coefficient of $H(\hat{\psi})$ is closest to 0 and its p-value is maximized is used to estimate ψ. To adjust for the bias due to the presence of competing

events, we use standard inverse probability of censoring weights. The following weights are commonly used (weights version 1):

$$w_1(j) = \begin{cases} \prod_{k=j}^{int(t)} \dfrac{P[C(k)=0 \,|\, C(k-1)=Y(k-1)=0, \bar{A}(k-1)]}{P[C(k)=0 \,|\, C(k-1)=Y(k-1)=0, \bar{A}(k-1), \overline{W}(k-1)]} & , C(j)=0 \\ 0 & , C(j)=1 \end{cases} \tag{4.1}$$

The authors suggest to use a different version (weights version 2):

$$w_2(j) = \begin{cases} \prod_{k=j}^{int(t)} \dfrac{P[C(k)=0 \,|\, C(k-1)=Y(k-1)=0, \bar{A}(k-1)=\bar{1}, \overline{W}(k-1)]}{P[C(k)=0 \,|\, C(k-1)=Y(k-1)=0, \bar{A}(k-1), \overline{W}(k-1)]} & , C(j)=0 \\ 0 & , C(j)=1 \end{cases} \tag{4.2}$$

so that the ψ estimate obtained will not have the standard average causal interpretation, but rather an exposure effect among a subgroup of the population i.e., the SACE. In both forms of the weights, $C(k)$ is an indicator that the competing event occurred ($C(k)$=1) at time k or not ($C(k)=0$), $Y(k)$ is an indicator of the event of interest, and overbars denote variable histories.

Once ψ was estimated, what is its interpretation? ψ is estimated from

$$T^{\bar{0}} = \int_0^T e^{\psi \cdot A(t)} \, dt$$

where recall $A(t)$ is the treatment, assumed to be binary, at time t. Suppose that a subject is always treated so $A(t) = 1 \; \forall t$, and let $T^{\bar{a}}$ represent the failure time, then

$$T^{\bar{0}} = \int_0^T e^{\psi \cdot A(t)} \, dt$$

$$T^{\bar{0}} = \int_0^{T^{\bar{a}}} e^{\psi} \, dt$$

$$T^{\bar{0}} = T^{\bar{a}} e^{\psi}$$

or

$$e^{-\psi} = T^{\bar{a}} \Big/ T^{\bar{0}}.$$

If $\psi > 0$, then the right hand side ratio must be less than 1, or $T^{\bar{a}} < T^{\bar{0}}$. So if the patient had not taken the treatment at all, his survival time would have been longer compared to if he had taken the treatment (from time zero until event occurs). In simple words the effect of the treatment is negative.

If $\psi < 0$, then the right hand side ratio must be bigger than 1, or $T^{\bar{a}} > T^{\bar{0}}$. So if the patient had not taken the treatment at all, his survival time would have been shorter compared to if he had taken the treatment (from time zero until event occurs). In simple words the effect of the treatment is positive.

Lastly, if $\psi = 0$, then the ratio of $T^{\bar{a}}$ to $T^{\bar{0}}$ would be 1, or $T^{\bar{a}} = T^{\bar{0}}$; hence, the treatment has no effect. Overall, the sign of ψ determines whether there is a expansion or contraction in the survival time, and its magnitude by how much (once exponentiated) [18].

In the presence of competing events, the interpretation of ψ is the same, but the estimate is biased due to informative censoring. Thus, as mentioned at the beginning of this section we use the weights in 4.1 or 4.2. Depending which weight form is used, as pointed out, the interpretation of ψ changes. Using 4.2 , ψ represents a SACE or "the exposure effect among a subgroup of the population that would not have died from a competing event irrespective of their exposure history" [30]. When we use the weights in 4.1 and view the competing event as "lost to follow-up" (standard censoring), the weights really serve as an inverse probability weight for missing data, so ψ translates to a measure of public health impact. Though if we view the competing event as lost to follow-up, we

are not really solving the problem, but just mask it.

Using the weights in 4.1 and additional assumptions, the interpretation of ψ is a mix of **(1)** the survivor average causal effect, and **(2)** "a function of population average and survivor average causal effects of the exposure through intermediate time-varying covariates" [30].

Chapter 5

Data Simulation and Analysis

In the following chapter, we start by explaining a new data simulation procedure for SNAFT models with competing risks. We then discuss an algorithm that we applied to the real dataset, and lastly, we perform the analysis explained in the previous chapter.

5.1 Data Simulation

A common approach to simulate survival times and consequently survival data is through the hazard function [3]. Namely, using the Cox proportional hazards model introduced section 2.4.1 (Proportional Hazards (PH) Model)

$$h(t \mid \boldsymbol{x}) = h_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{x})$$

From the relations established in Chapter 2, the survival function of the Cox proportional hazards model is

$$S(t \mid x) = \exp(-H_0(t) \exp(\beta' x)) \tag{5.1}$$

where $H_0(t)$ is the cumulative baseline hazard function. Thus, we have that

$$F(t \mid x) = 1 - \exp(-H_0(t)\exp(\beta' x)). \tag{5.2}$$

Appealing to the Inverse Transform Sampling method [3], we can generate random event times. If Y is a random variable with continuous distribution function F, then $U = F(Y) \sim \mathcal{U}(0,1)$, and it is well know that $1-U \sim \mathcal{U}(0,1)$. Hence, if T is the survival time of the Cox model in 5.1, then from 5.2 we have that

$$U = \exp(-H_0(T)\exp(\beta' x)) \sim \mathcal{U}(0,1).$$

Under the assumption that $h_0(t) > 0 \ \forall t$, then $H_0(t)$ can be inverted so that

$$\begin{aligned} \exp(-H_0(T)\exp(\beta' x)) &= U \ ^1 \\ -H_0(T)\exp(\beta' x) &= \ln(U) \\ H_0(T) &= \frac{-\ln(U)}{\exp(\beta' x)} \\ T &= H_0^{-1}\left(\frac{-\ln(U)}{\exp(\beta' x)}\right) \end{aligned}$$

where $U \sim \mathcal{U}(0,1)$ [4].

Although this approach is very appealing, it may not always work. First, it relies on the existence of the inverse of the cumulative baseline hazard, and second, it ignores the possibility of competing events.

To simulate survival times under a competing events setting, instead of using just the hazard and the cumulative hazard, we can use the cause specific hazard and the cumulative all-cause hazard [4] [5]. Recall, the cause specific hazard is defined as:

$$h_{0j}(t) = \frac{P(t \le T < t + \Delta t, \delta_T = j \mid T \ge t)}{\Delta t}, \ j = 1, ..., J$$

[1] Since $U \equiv 1 - U \sim \mathcal{U}(0,1)$, either U or $1-U$ can be used

where δ_j is an indicator specifying which event occurred assuming we have J possible events, and for simplicity, we denoted the hazard of event j by $h_{0j}(t)$. It then follows that the cumulative all-cause hazard is

$$H_{0\cdot}(t) = \int_0^t \sum_{j=1}^{J} h_{0j}(u)\, du \tag{5.3}$$

and we use it to generate the survival times i.e.

$$T = H_{0\cdot}^{-1}\left(-\ln(U)\right). \tag{5.4}$$

Observe that we dropped the effect of covariates; if we have covariates, an extra factor of $\exp\left(-\boldsymbol{\beta}_j'\boldsymbol{x}\right)$ will be present inside the inverse function where $\boldsymbol{\beta}_j$ is a vector of coefficient estimates for event j. We omit covariate effects for introductory purposes. Equation 5.4 will generate a vector of survival times given there are J competing events, but it does not specify the type of event occurring. To determine which event occurs, notice that

$$\begin{aligned} P(\delta_j = 1 \,|\, t \le T < t + \Delta t, T \ge t) &= \frac{P(t \le T < t + \Delta t, \delta_j = 1 \,|\, T \ge t)}{P(t \le T < t + \Delta t \,|\, T \ge t)} \\ &= \frac{h_{0j}(t)}{\sum_{j=1}^{J} h_{0j}(t)}. \end{aligned} \tag{5.5}$$

Hence, we can summarize the simulation steps in a simple algorithm [4] below:

1. Specify the cause specific hazards, and find the cumulative all-cause hazard.

2. Simulate failure times T by generating a random variable $U \sim \mathcal{U}(0,1)$, and then use 5.4.

3. Run a multinomial experiment for a simulated failure time T, which determines with probability, $p_j\big|_{t=T}$, given by 5.5 that cause j occurs, for $j = 1, ..., J$

4. Optionally, if desired, one may generate right-censoring times C and/or left-truncation times.

Note that the censoring and truncation times must be generated as random variables independent of the competing events, if one requires independent censoring. Although we accounted for the possibility of competing events, the equation in 5.4 depends strongly on the existence of the inverse function. To avoid computing an inverse, we consider a different approach [4]. 5.4 is equivalent to

$$H_{0\cdot}(t) = -\ln(u) \Rightarrow H_{0\cdot}(t) + \ln(u) = 0.$$

Hence, for a given u we seek the root, t, to the above equation. The previous algorithm is then equivalent to:

1. Specify the cause specific hazards, and find the cumulative all-cause hazard.

2. Simulate failure times T by generating a random variable $U \sim \mathcal{U}(0,1)$, and then by solving $H_{0\cdot}(t) + \ln(u) = 0$.

3. Run a multinomial experiment for a simulated failure time T, which determines with probability, $p_j\big|_{t=T}$, given by 5.5 that cause j occurs, for $j = 1, ..., J$

4. Optionally, if desired, one may generate right-censoring times C and/or left-truncation times.

5.2 Approximating The Hazard of Gestational Age Data

In the previous section, we established an algorithm via which we can simulate survival times and the event type, in practice, we need to set or predetermine what the cause specific hazards are. For this thesis, we based our choice of hazard function on the observed data.

Recall, our focus is the length of gestational period of live births in the presence of adverse outcomes. In the data set, a conception may result in a live, stillbirth, miscarriage, induced abortion, ectopic pregnancy, or current pregnancy. For simplicity, we subsetted the data to only the four most common events, namely: live births, stillbirths, miscarriages, and ectopic pregnancy.

Figure 5.1 summarizes the count of each possible outcome. From the middle left panel, we notice that the mean gestational age among live births is roughly 39 weeks.

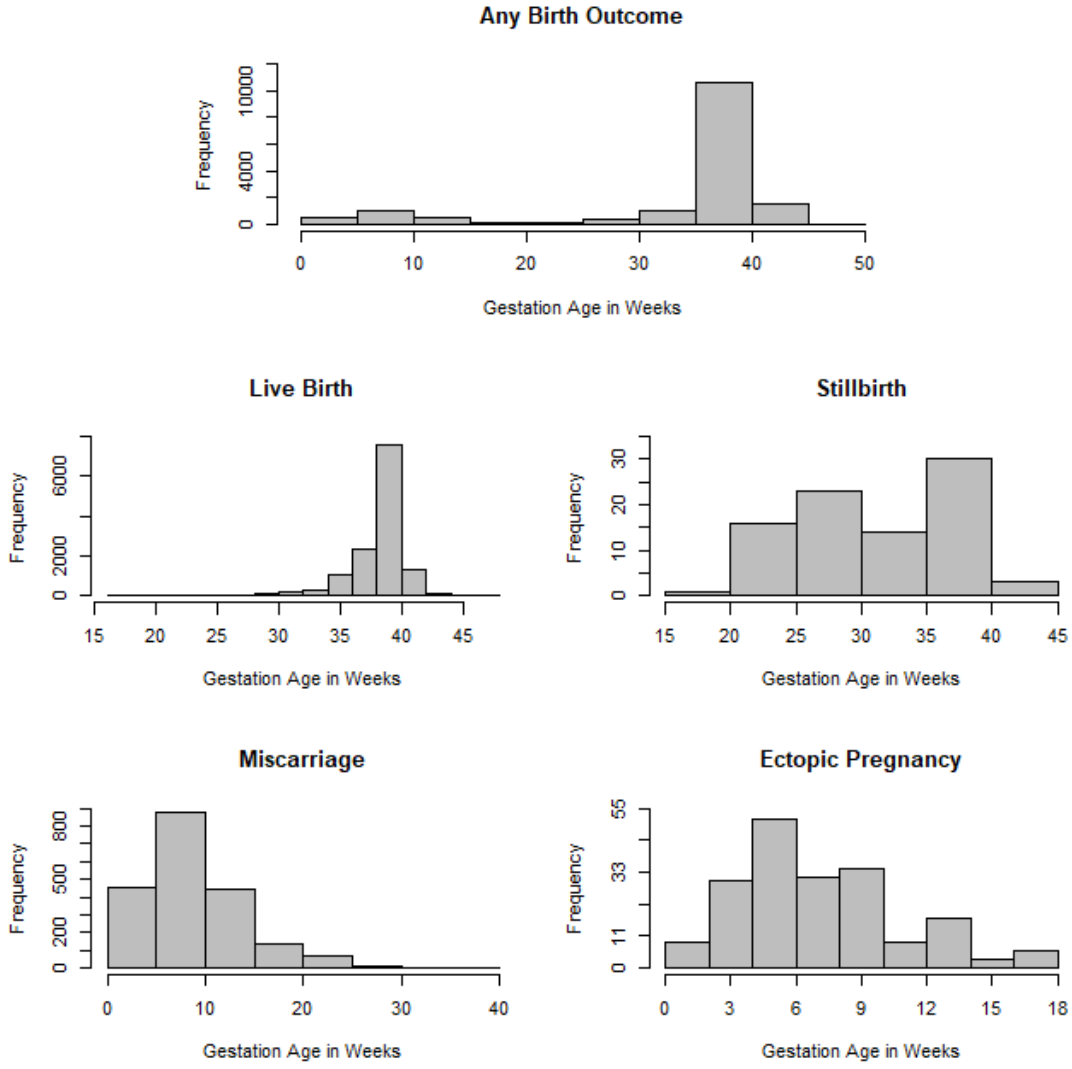


Figure 5.1: Histograms for Birth Outcomes

Miscarriages and ectopic pregnancies are moderately skewed to the right with a rough mean of around 10 weeks while stillbirths are just spread across the span of 15 to about 40 weeks.

The next two figures show estimates for the all-cause hazard, and the cause-specific hazards. The estimates were based on the Nelson-Aalen estimator i.e.,

$$\hat{h}(t_i) = \frac{d_i}{n_i}$$

where recall d_i is the number of events at time t_i and n_i is the total individuals at risk at t_i.

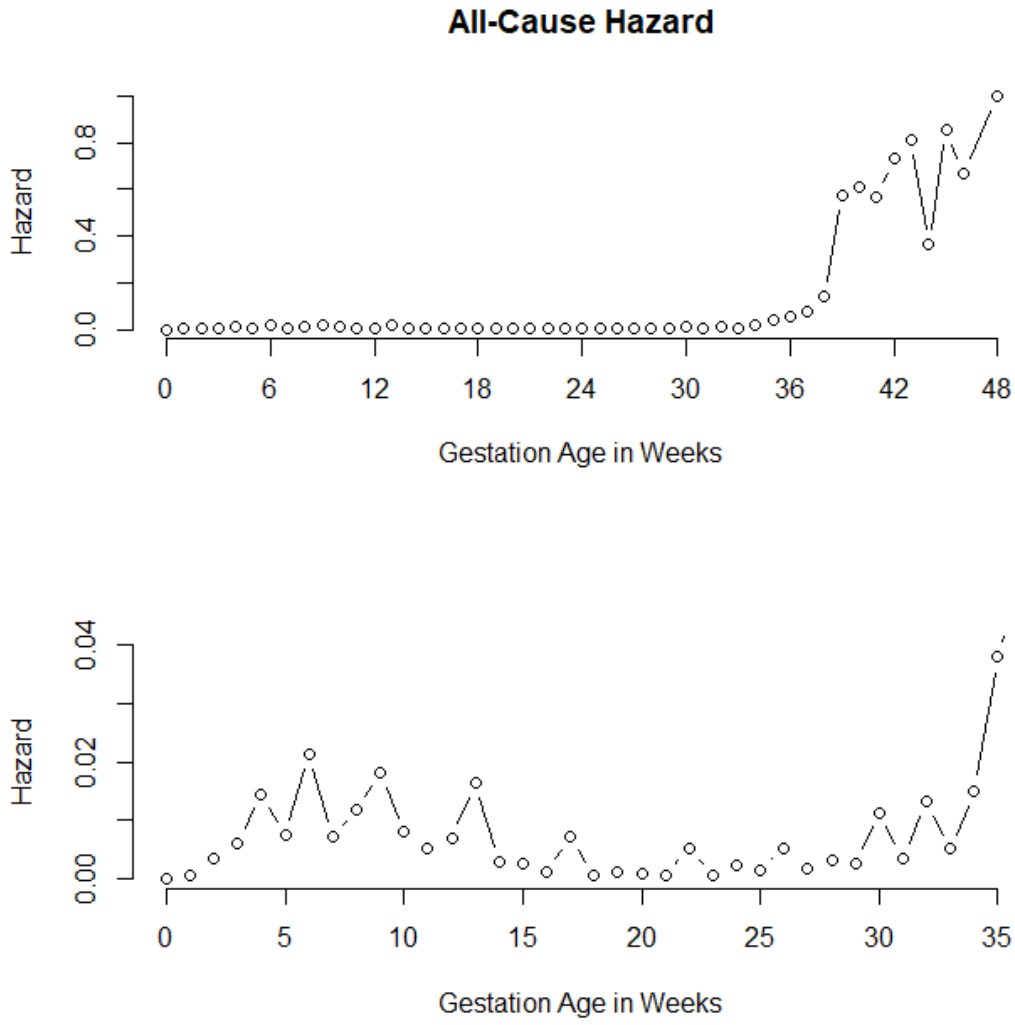


Figure 5.2: All-Cause Hazard For Gestational Ages

The lower panel of Figure 5.2 illustrates the behavior of the all-cause hazard between 0 and 35 weeks as it is not clear from the overall plot (top panel).

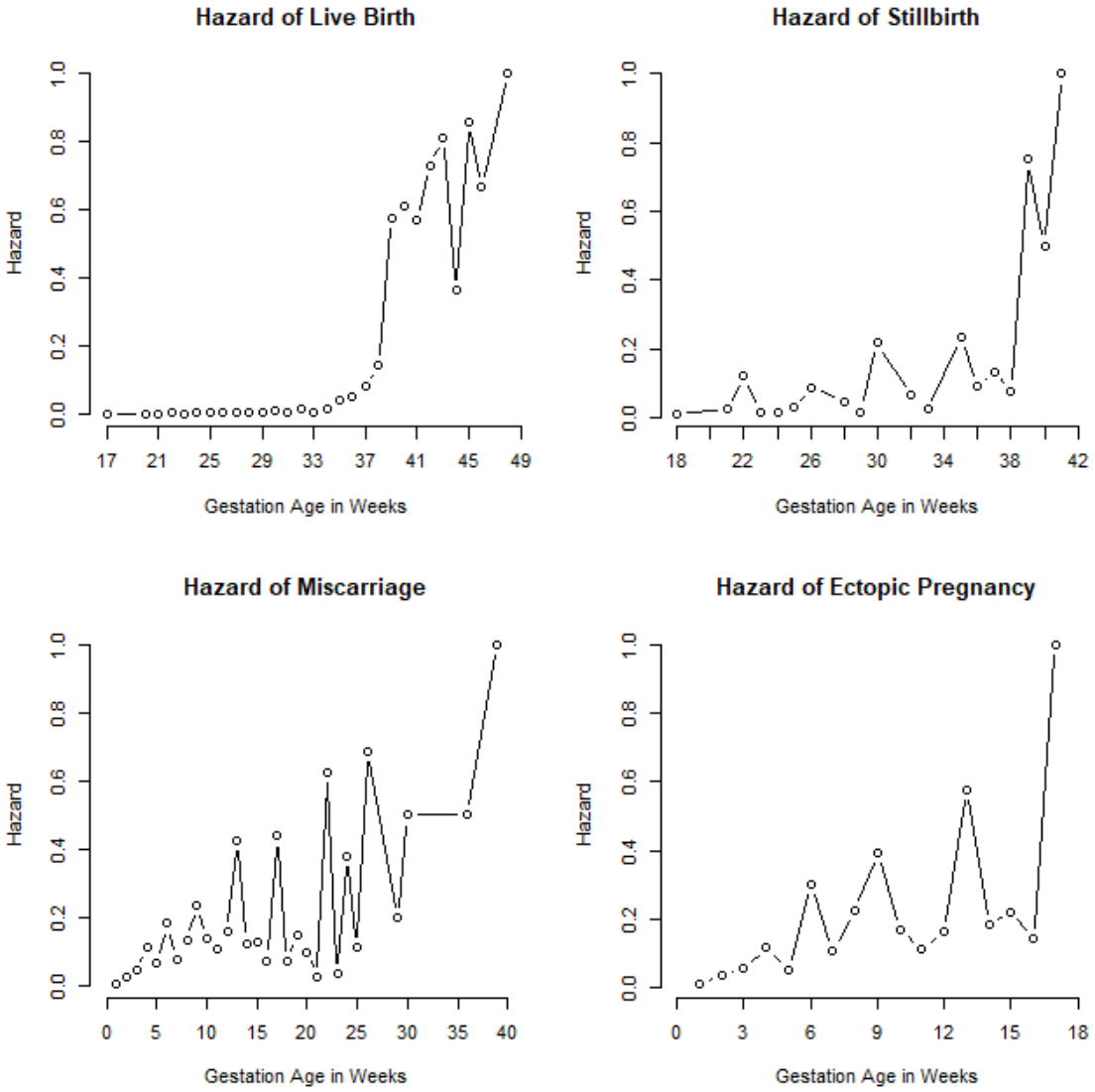


Figure 5.3: Cause-Specific Hazards

Given the 4 figures above, we need to determine an adequate parametric curve that summarizes well enough the behavior of the data. We tackled this problem by determining a probability density function for each birth outcome. Consequently, this will allows us to find the cause-specific hazards and therefore the all-cause hazard from which we can find the cumulative all-cause hazard, see 5.3. For simplicity we assumed that each time

for a particular event arises from the same distributional family, and that the competing events are independent. In other words, the time to event for cause j is given by $f(t\,;\,\boldsymbol{\theta}_j)$ where $\boldsymbol{\theta}_j$ is a parameter vector.

Ideally, we should focus at determining a parametric form for the cause-specific hazards, but recall from Chapter 2 that specifying one function determines all functions.

We chose to describe each time event by a log logistic distribution since its hazard is very flexible depending on the value of the shape and scale parameter; the hazard reaches a peak after some finite period and then slowly declines. The survival, hazard, and cumulative hazard functions have closed form expressions in contrast to other distributions such as the gamma or log normal densities.

A commonly used parameterization of the log-logistic density is

$$f(x\,;\,\mu,s) = \frac{1}{x} \times \frac{1}{s} \times \frac{\exp\left(\frac{\ln(x)-\mu}{s}\right)}{\left(1+\exp\left(\frac{\ln(x)-\mu}{s}\right)\right)^2} \equiv \frac{e^{\,\mu/s} x^{\frac{1}{s}-1}}{s(e^{\,\mu/s}+x^{1/s})^2}. \tag{5.6}$$

The above follows from letting $Y \sim$ Logisitic(μ, s) i.e, Y is distributed according to a logistic distribution with location parameter μ and scale s, then letting $X = \exp(Y)$, and applying the Transformation of Random Variables Theorem. Hence, we assume that each event time is distributed according to

$$f(t\,;\,\mu_j,s_j) = \frac{1}{t} \times \frac{1}{s_j} \times \frac{\exp\left(\frac{\ln(t)-\mu_j}{s_j}\right)}{\left(1+\exp\left(\frac{\ln(t)-\mu_j}{s_j}\right)\right)^2} \tag{5.7}$$

where $(\mu_j,\, s_j) = \boldsymbol{\theta}_j$ are the parameter for cause j, It follows that

$$F(t\,;\,\boldsymbol{\theta}_j) = 1 - \frac{e^{\,\mu_j/s_j}}{e^{\,\mu_j/s_j}+x^{1/s_j}},$$

$$S(t\,;\,\boldsymbol{\theta}_j) = \frac{e^{\,\mu_j/s_j}}{e^{\,\mu_j/s_j}+x^{1/s_j}},$$

$$h(t\,;\,\boldsymbol{\theta}_j) = \frac{x^{\frac{1}{s}-1}}{s_j(e^{\,\mu_j/s_j}+x^{1/s_j})}, \text{ and}$$

$$H(t\,;\boldsymbol{\theta}_j) = \ln(e^{\mu_j/s_j} + x^{1/s_j}) - \frac{\mu_j}{s_j}.$$

Parameter estimates for each time event were obtained via the function `llogisMLE` from the `STAR` package in `R` [32]. Figure 5.4 shows the same histograms from Figure 5.1 with the log logistic densities overlaid.

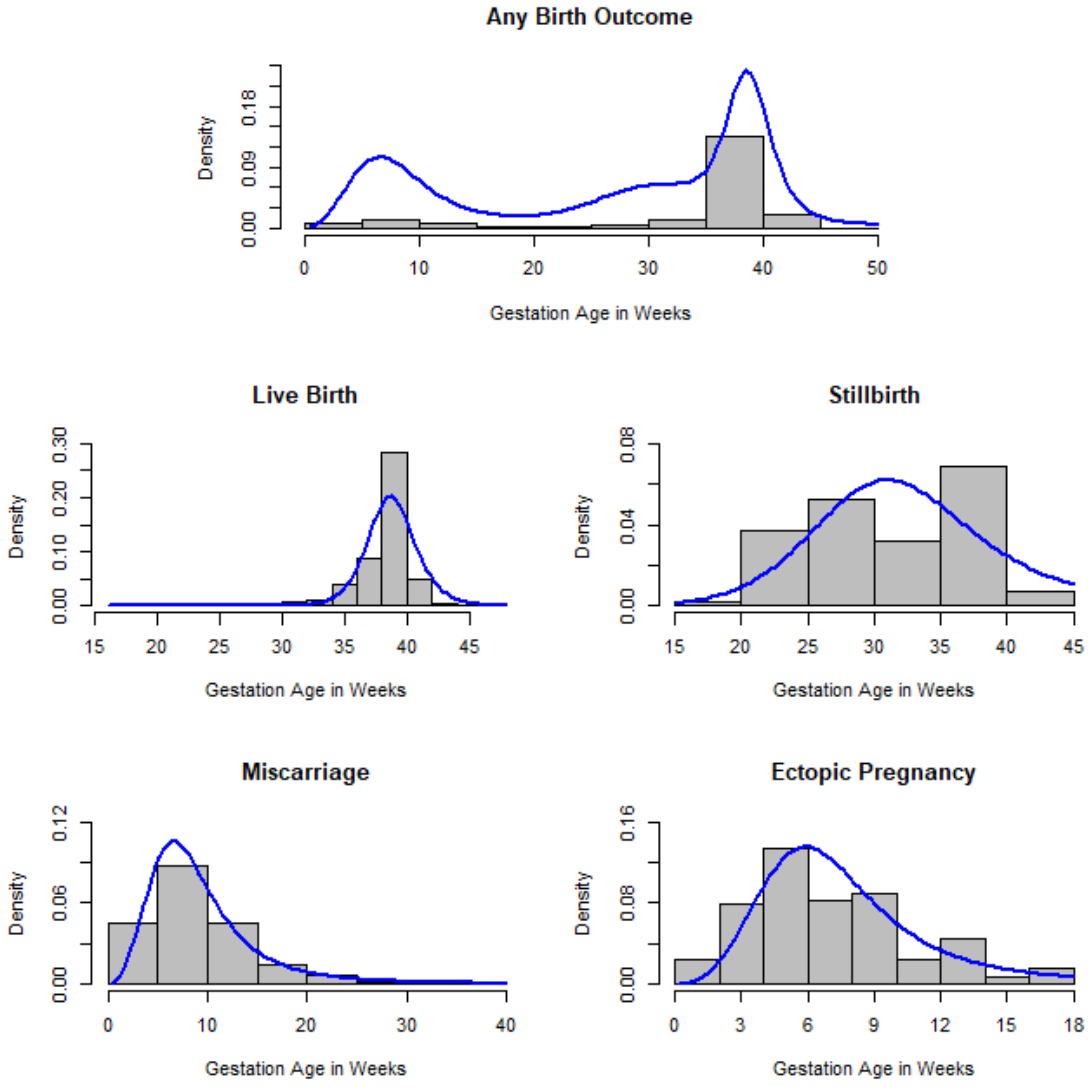


Figure 5.4: Density Curves For Time to Event

The curve in the top panel, or the overall density, was estimated by

$$f(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) = \sum_{j=1}^{4} f(t\,;\,\mu_j, s_j) \tag{5.8}$$

with $f(t\,;\,\mu_j, s_j)$ defined as in 5.7, $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$, and $\boldsymbol{s} = (s_1, s_2, s_3, s_4)$. Note that this is not the true expression for the overall density. From 5.3, we have

$$\begin{aligned}
H_{0\cdot}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) &= \int_0^t \sum_{j=1}^{J} h_{0j}(u\,;\,\mu_j,\,s_j)\,du, \\
H_{0\cdot}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) &= \sum_{j=1}^{J} \int_0^t h_{0j}(u\,;\,\mu_j,\,s_j)\,du, \\
H_{0\cdot}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) &= \sum_{j=1}^{J} H_{0j}(t\,;\,\mu_j,\,s_j), \\
\exp[-H_{0\cdot}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s})] &= \exp\left[-\sum_{j=1}^{J} H_{0j}(t\,;\,\mu_j,\,s_j)\right], \text{ and} \\
S(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) &= \exp\left[-\sum_{j=1}^{J} H_{0j}(t\,;\,\mu_j,\,s_j)\right].
\end{aligned}$$

$S(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s})$ is the overall survival curve and it is also equivalent to $\prod_{j=1}^{J} S_j(t\,;\,\mu_j,\,s_j)$ where $S_j(t\,;\,\mu_j,\,s_j)$ are the cause-specific survival curves.

$$\begin{aligned}
1 - S(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) &= 1 - \exp\left[-\sum_{j=1}^{J} H_{0j}(t\,;\,\mu_j,\,s_j)\right], \\
F(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) &= 1 - \exp\left[-\sum_{j=1}^{J} H_{0j}(t\,;\,\mu_j,\,s_j)\right], \\
\frac{d}{dt}F(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) &= \frac{d}{dt}\left(1 - \exp\left[-\sum_{j=1}^{J} H_{0j}(t\,;\,\mu_j,\,s_j)\right]\right), \\
f(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) &= \exp\left[-\sum_{j=1}^{J} H_{0j}(t\,;\,\mu_j,\,s_j)\right] \times \frac{d}{dt}\sum_{j=1}^{J} H_{0j}(t\,;\,\mu_j,\,s_j), \\
&= \exp\left[-\sum_{j=1}^{J} H_{0j}(t\,;\,\mu_j,\,s_j)\right] \times \sum_{j=1}^{J} \frac{d}{dt} H_{0j}(t\,;\,\mu_j,\,s_j), \\
&= \exp\left[-\sum_{j=1}^{J} H_{0j}(t\,;\,\mu_j,\,s_j)\right] \times \sum_{j=1}^{J} h_{0j}(t\,;\,\mu_j,\,s_j), \text{ and} \\
&\equiv S(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) \times \sum_{j=1}^{J} h_{0j}(t\,;\,\mu_j,\,s_j).
\end{aligned} \tag{5.9}$$

Going back to Figure 5.4, we notice that the overall density curve overestimates the probability densities. In order to stabilize the curve, we weighted each component by the proportion of each birth outcome. The weighted versions of 5.8 and 5.9 are respectively given by:

$$f_w(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) = \sum_{j=1}^{4} w_j f(t\,;\,\mu_j, s_j);$$

$$f_w(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) = \exp\left[-\sum_{j=1}^{4} w_j H_{0j}(t\,;\,\mu_j,\,s_j)\right] \times \sum_{j=1}^{4} w_j h_{0j}(t\,;\,\mu_j,\,s_j).$$

Figure 5.5 and Figure 5.6 show the fit for the density under equations 5.8 and 5.9, respectively.

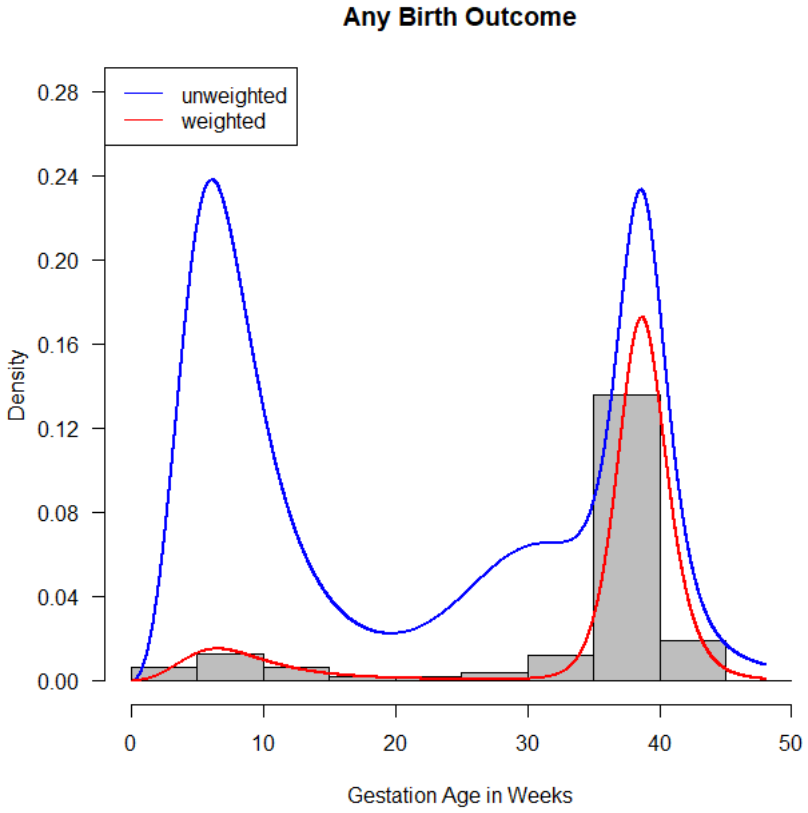


Figure 5.5: Density Curve For Length of Any Birth Outcome given by 5.8

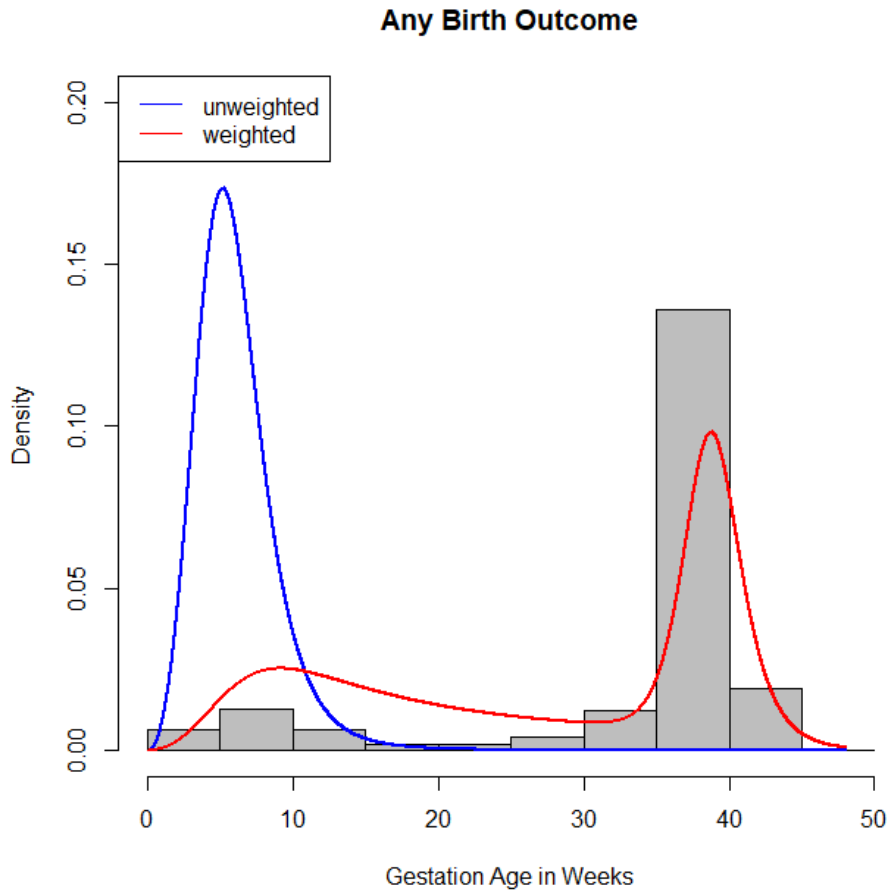


Figure 5.6: Density Curve For Length of Any Birth Outcome given by 5.9

Clearly, the weighted densities are a more accurate fit. Now that we have the density of each outcome, we can obtain the cause-specific hazards and the all-cause hazard. The cause specific hazards are give by

$$h_{0j}(t\,;\,\boldsymbol{\theta}_j) := h(t\,;\,\boldsymbol{\theta}_j)$$

and the all-cause hazard is

$$h_{0\cdot}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) := \sum_{j=1}^{4} h(t\,;\,\boldsymbol{\theta}_j).$$

The weighted versions are respectively given by

$$h^{w}_{0j}(t\,;\,\boldsymbol{\theta}_j) := w_j h(t\,;\,\boldsymbol{\theta}_j)$$

$$h_0^w(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) := \sum_{j=1}^{4} w_j h(t\,;\,\boldsymbol{\theta}_j).$$

The next figure shows the all-cause and the cause-specific hazards. Although the weighted hazard for stillbirths and ectopic pregnancy look flat, they are not; the hazard just spans a very narrow range of values.

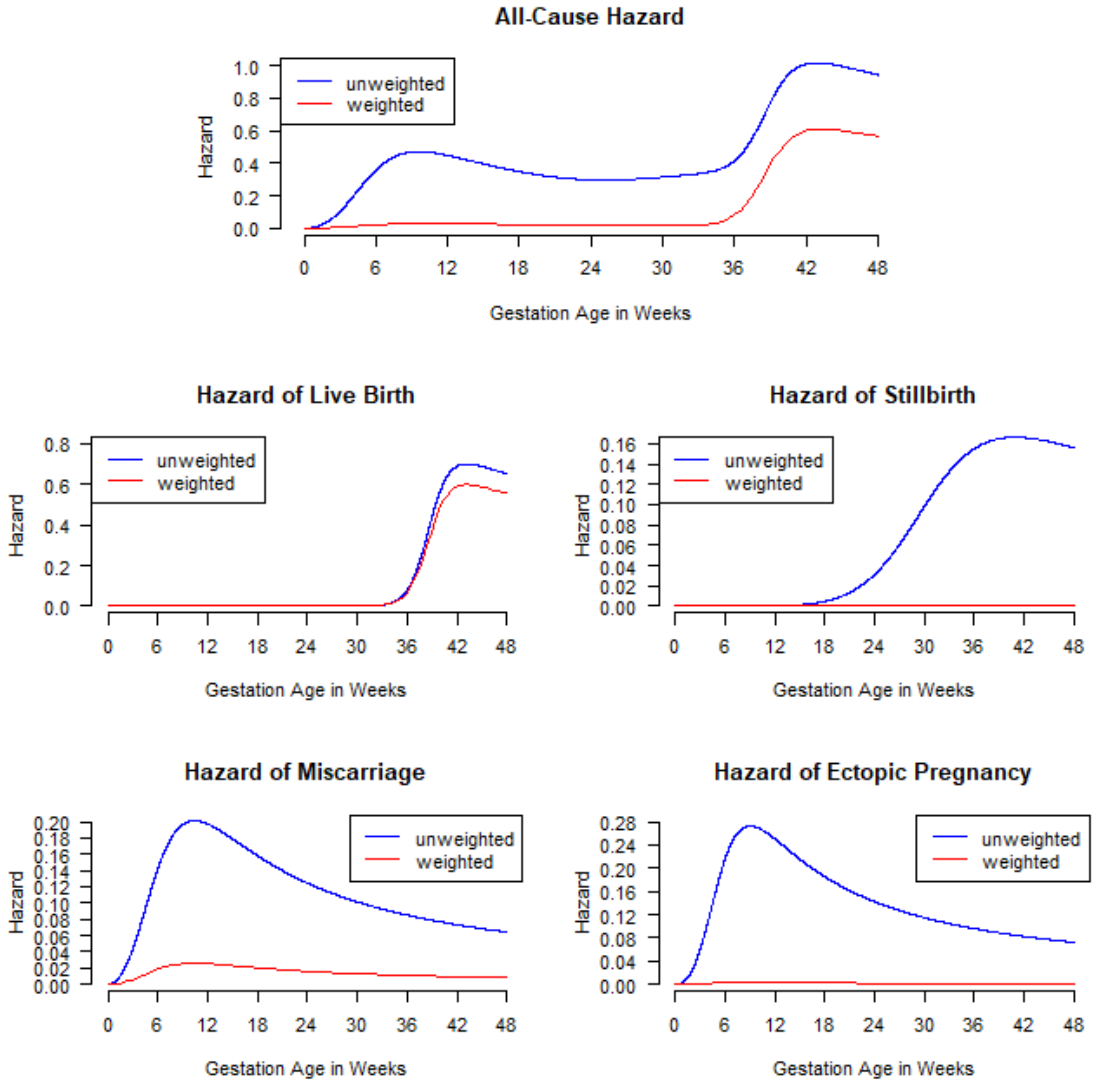


Figure 5.7: Cause-Specific and All-Cause Hazards

The next image shows the weighted and unweighted all-cause hazard superimposed with the top panel of Figure 5.2.

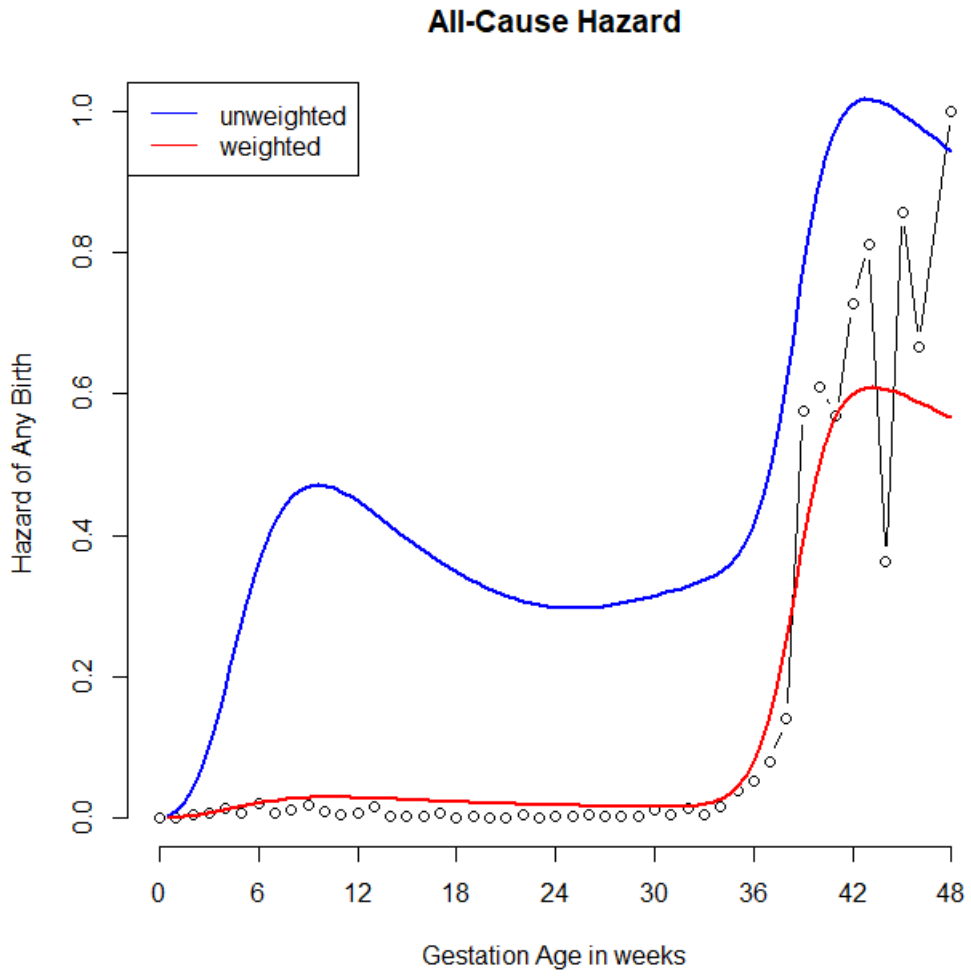


Figure 5.8: Parametric (weighted and unweighted) and Non-parametric Hazard

Now that we have the all-cause hazard, we can find the cumulative all-cause hazard:

$$h_{0.}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) = \sum_{j=1}^{4} w_j h(t\,;\,\boldsymbol{\theta}_j),$$

$$\int_0^t h_{0.}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) = \int_0^t \sum_{j=1}^{4} w_j h(t\,;\,\boldsymbol{\theta}_j),$$

$$H_{0.}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) = \sum_{j=1}^{4} w_j \int_0^t h(t\,;\,\boldsymbol{\theta}_j),$$

$$H_{0.}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) = \sum_{j=1}^{4} w_j H(t\,;\,\boldsymbol{\theta}_j), \text{ and}$$

$$H_{0.}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s}) = \sum_{j=1}^{4} w_j \left(\ln(e^{\,\mu_j/s_j} + t^{1/s_j}) - \frac{\mu_j}{s_j} \right).$$

As we mentioned in the previous section, if we want to account for the effect of covariates, the all-cause hazard will be

$$H_{0\cdot}(t\,;\,\boldsymbol{\mu},\,\boldsymbol{s},\,\boldsymbol{x}) = \sum_{j=1}^{4} w_j \left(\ln(e^{\,\mu_j/s_j} + t^{1/s_j}) - \frac{\mu_j}{s_j} \right) \exp(\boldsymbol{\beta}_j' \boldsymbol{x})$$

where $\boldsymbol{\beta}_j$ is a vector of coefficient estimates for event j. Hence, we can simulate a design matrix and a matrix of coefficients where say column j corresponds to event j, and apply the algorithm described in the previous section to generate a data under a competing events setting.

5.3 Data Simulation With Time Varying Treatment

In section 5.1 we described how to simulate data with baseline covariates using the cause specific hazards. In order to simulate time-varying exposure under a SNAFT model with competing events, we need a simulation approach that captures these three model features.

We used the simulation algorithm described by Young et al. [41] which captures the 3 points mentioned above. The authors assume that past treatment, A, affects present confounder(s), V, and allow for the failure time, T, to arise from a general SNAFTM i.e.,

$$T^{\bar{0}} = \int_0^T \exp\{\gamma(t,\, \bar{A}_t,\, \overline{V}_t,\, \psi)\}\, dt$$

where $\gamma(t,\, \bar{A}_t,\, \bar{V}_t,\, \psi)$ is some function, $T^{\bar{0}}$ is the counterfactual failure time under no treatment history, and ψ is the parameter of interest. The algorithm they propose is briefly as follows. For each subject:

1. Simulate the counterfactual $T^{\bar{0}}$ from a failure time distribution with hazard $h_0(t)$. Then, for each observation time, m, do:

2. Simulate V_m

3. Simulate A_m

4. Given $\gamma(t,\, \bar{A}_t,\, \bar{V}_t,\, \psi)$, let T be the solution to

$$T^{\bar{0}} = \int_0^T \exp\{\gamma(t,\, \bar{A}_t,\, \overline{V}_t,\, \psi)\}\, dt.$$

Assuming that both A_m and V_m are binary, then at time m the authors propose the following models for step 2 and 3, respectively:

$$\text{logit}(V_m) = \beta_0 + \beta_1 T^{\bar{0}} + \beta_2 A_{m-1} + \beta_3 V_{m-1}$$

$$\text{logit}(A_m) = \alpha_0 + \alpha_1 V_m + \alpha_2 V_{m-1} + \alpha_3 A_{m-1}.$$

The model for V_m must include $T^{\bar{0}}$ to guarantee that V_m is a confounder and A_{m-1} so that it is affected by prior treatment. On the other hand, the model for A_m must exclude $T^{\bar{0}}$ to ensure conditional exchangeability. We adapted their algorithm to the case of competing risks. First, we had only a time varying treatment with no time-varying confounders. Second, for the model for A_m we assumed a monotonically increasing intercept; hence,

$$\text{logit}(V) = \beta_0 + \beta_1 T^{\bar{0}}$$

$$\text{logit}(A_m) = \alpha_{0m} + \alpha_1 V + +\alpha_2 A_{m-1}.$$

Notice that so far this simulation approach does not capture the presence of competing events. To our best knowledge, we could not find a reference that explains how to simulate data using SNAFT model and competing events. Therefore, we induced an association between the confounder(s) and type of competing event as follows:

$$\text{logit}(V) = \beta_0 + \beta_1 T^{\bar{0}} + \beta_2 C^{\bar{0}} + \beta_3 \underbrace{\left(T^{\bar{0}} \cdot C^{\bar{0}}\right)}_{\text{interaction}}$$

where $C^{\bar{0}}$ is the counterfactual event type under no treatment history. In our situation we set $C^{\bar{0}}$ to be a binary variable: 1 (live delivery), and 0 (other delivery). $T^{\bar{0}}$ and $C^{\bar{0}}$ are generated as described in section 5.1.

5.4 Simulation Results

We performed a simulation study of 50 replications with 12,000 observations per replication. The true value of ψ was set to be -0.03. For each replication, we fitted three different models as per Chapter 4. More precisely, an unweighted model, a weighted model using weights defined as in 4.1 which we will refer to as Weights Version 1, and a weighted model using weights defined as in 4.2 which we will refer to as Weights Version 2.

We used a grid of ψ values to determine $\hat{\psi}$, so in the end we had 50 estimates of ψ under the unweighted and weighted models. The histograms shown in Figure 5.9 illustrate the distribution of those estimates. We notice that in the first two histograms, the estimates are mostly centered around the truth i.e. -0.03 while in the last histogram they are grouped around -0.02.

Figure 5.10 shows the estimates of ψ from each simulation (black dots) along with the 95% confidence bands under the respective model. To compute the 95% confidence bands for a ψ estimate, we first computed the 95% confidence interval (CI) for $H(\psi)$ using the robust standard errors (given by the `R` output) from the model that returned the best estimate for ψ. We then inverted the 95% CI of $H(\psi)$ to give the 95% CI for ψ. The dashed line cuts at the true value of ψ i.e., -0.03. Under the unweighted model, 90% (or 45 out of the 50) of the confidence intervals contain the true value of -0.03. On

the other hand, under the weighted models, 88% (or 44 out of the 50) of the confidence intervals capture the true value of ψ. Under the model weighted by 4.1, we obtained a slightly less biased estimate compared to unweighted one, but the coverage probability is not as good as the unweighted scenario due to the narrower interval bands. We notice a larger bias when using weights give by 4.2 since we are estimating a different effect than the first two scenarios.

Figure 5.11 shows the box plots for the estimates along with the mean in each group. We observe that the under unweighted and Weights Version 1, we were able to

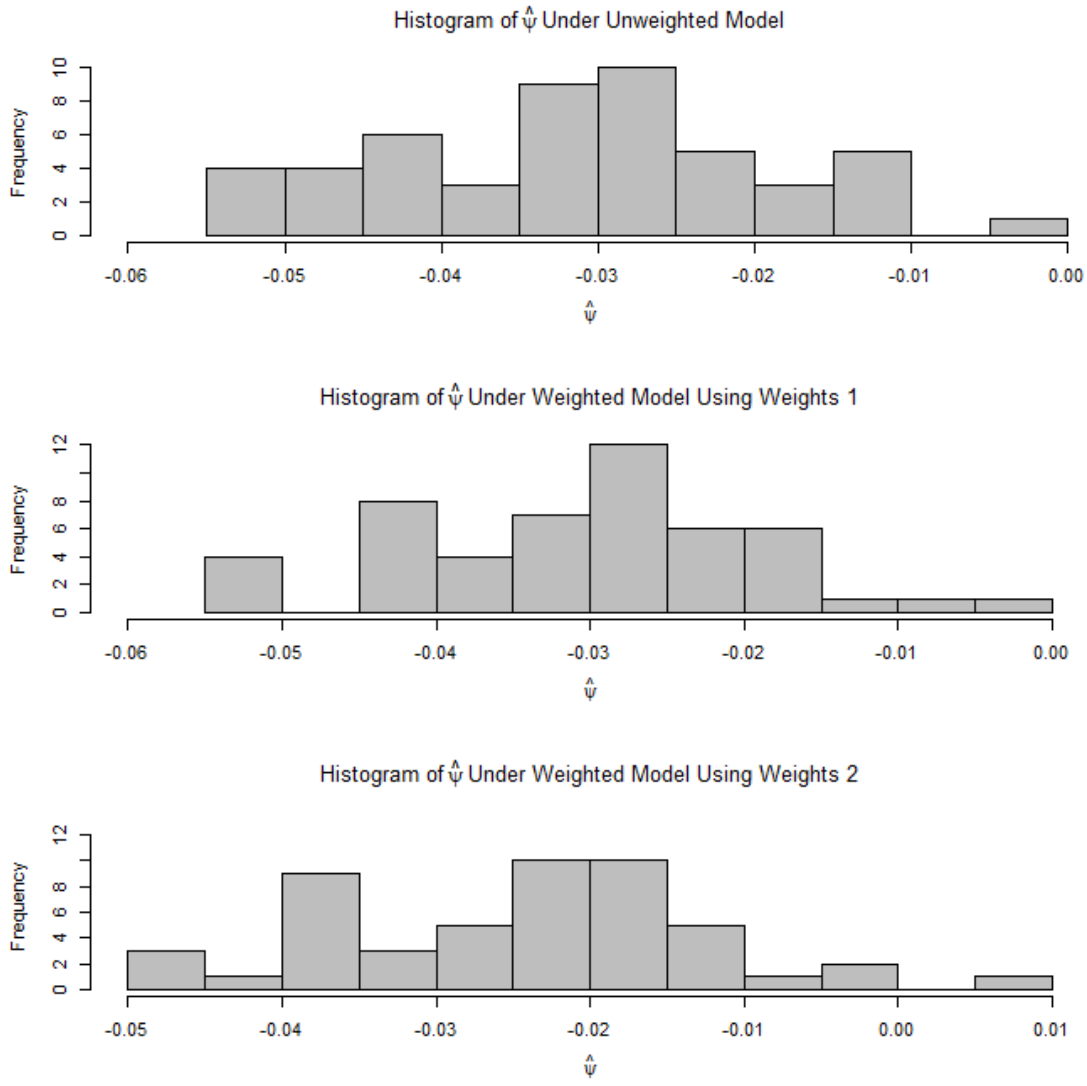
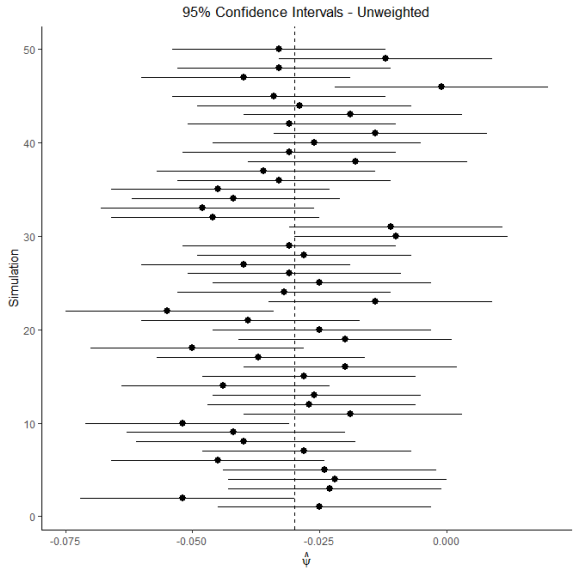
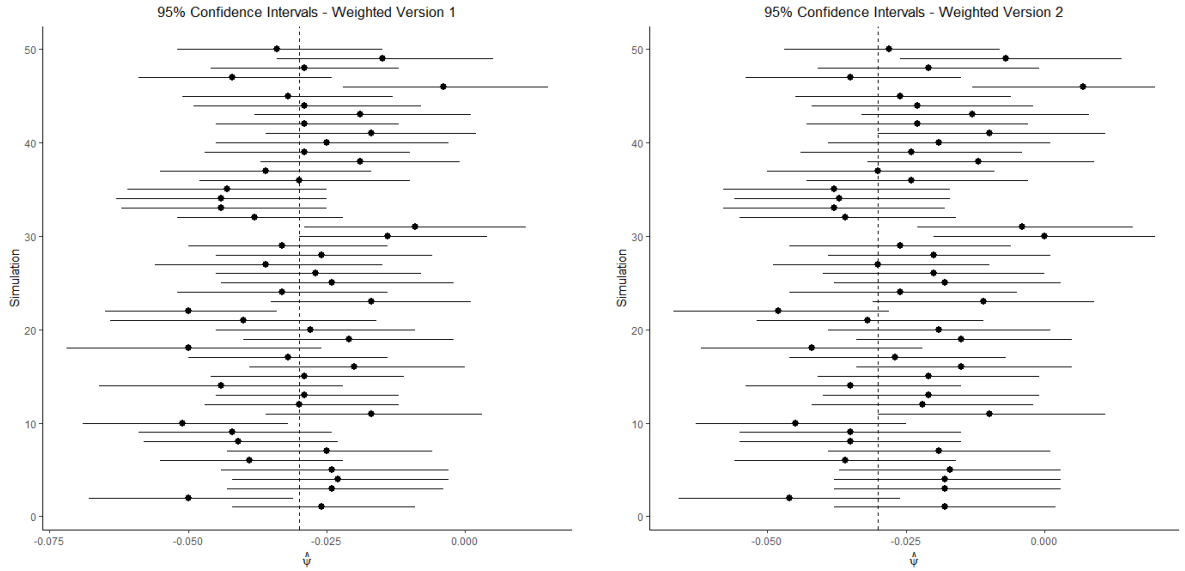


Figure 5.9: ψ Estimates From Each Sample Under The Respective Fitted Model

recover very closely the true value of ψ while under Weights Version 2, we get a mean estimate that differs substantially.



(a)



(b) (c)

Figure 5.10: 95% Confidence Intervals for $\hat{\psi}$ Estimates Under Each Model

Table 5.1 summarizes the mean estimates from the simulations and standard errors, and the average of the standard errors resulting from the constructed confidence intervals

seen in Figure 5.10.

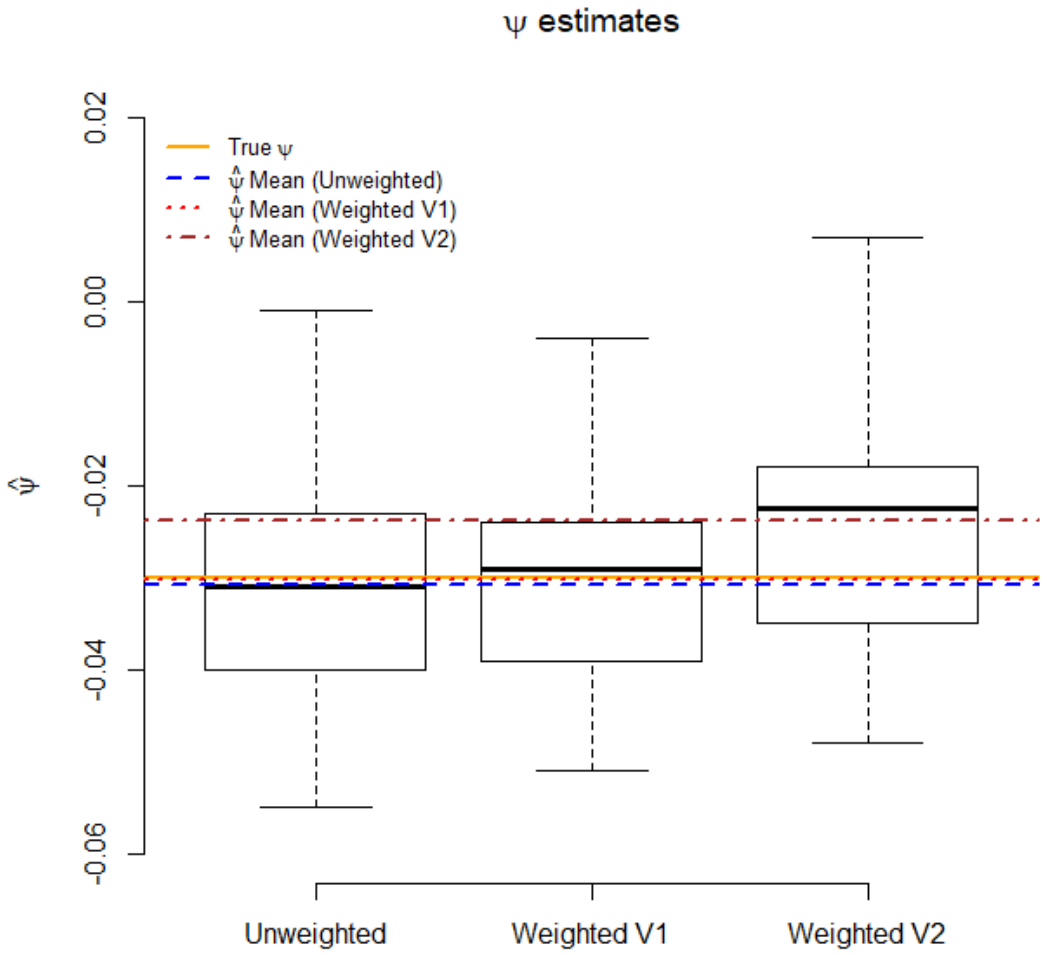


Figure 5.11: Box plots For ψ Estimates Under The Respective Fitted Model

Method	**Mean**$(\hat{\psi})$	**Emp.** $SE(\hat{\psi})$	**Ave** $\widehat{SE}(\hat{\psi})$
Unweighted	-0.03072	0.01220	0.01070
Weighted 1	-0.03024	0.01102	0.00955
Weighted 2	-0.02372	0.01184	0.01015

Table 5.1: Summary Statistics From Simulation Study for 50 Replications Under Each of The Three Weighting Methods: Unweighted, Weighted 1, and Weighted 2

5.5 Clustering Code

In the following section, we discuss an algorithm that we created; its purpose was to cluster the births from a given mother, when there is no unique identifier. We explain the mechanics of the algorithm, and evaluate its efficiency by comparing the output from the algorithm to the truth which was known.

With the increasing usage of computerized record linkage in epidemiological studies [29], and given the nature of our data, we believed that dedicating a section to this topic will be worth. Many papers have analyzed perinatal data, yet some have not considered identifying births from the same mother [33] [20] while others have [40] [39] [16] [21]. Aside from clustering births from the same mother, some authors have focused at hospitalization records linkage such as neonatal readmission [29], congenital anomalies [8], or other type of hospitalization records [15].

5.5.1 Methods

The algorithm was applied to data collected by the 2006-2010 and 2011-2013 National Survey of Family Growth (NSFG) [13]. A total of 17,352 women answered a questionnaire about their birth history and relevant factors during each birth. Hence, the birth history of each mother was known. Some factors, were prior and post smoking status to conception, where was birth given, birth weight, pregnancy outcome, marital status, mother's birth date, pregnancy order, and other.

Prior to applying the algorithm, minor data manipulations were performed. First, for a current birth, we added an extra variable which indicated the date of the previous child's birth. Second, two variables were created called "ID" and "uniqueid;" initially, they take the same values i.e. the observation's position in the dataset, but later they will be used to identify *identical* mothers as per our algorithm. Third, the data was sorted by mother's birth date and each date was assigned to a unique number (variable we called "mom") e.g. all women born on 01/01/1960 were assigned to "mom=1," women born on 01/02/1960 could be assigned to "mom=2" and so on. Lastly, important dates were transformed from century-month format to yyyy/mm format. We then applied the algorithm.

We start by splitting the data into two subsets, one where the date of birth of a mother occurs once (17,330 observations), and another where the date of birth occurs at least twice (22 observations). The first dataset consisted of all the mothers with a plausible birth history. It was further split into mini datasets by using the "mom" variable. In other words, we had smaller datasets where within each dataset we have all women born on the same date. Each dataset was then sorted by descending order of child's date

of birth and pregnancy order. Finally, taking each mini dataset one at the time, the clustering was done within each set and potential matches were identified by using the following variables: ID, uniqueid, date of birth, date of previous birth, and pregnancy order. The procedure is summarized in Figure 5.12.

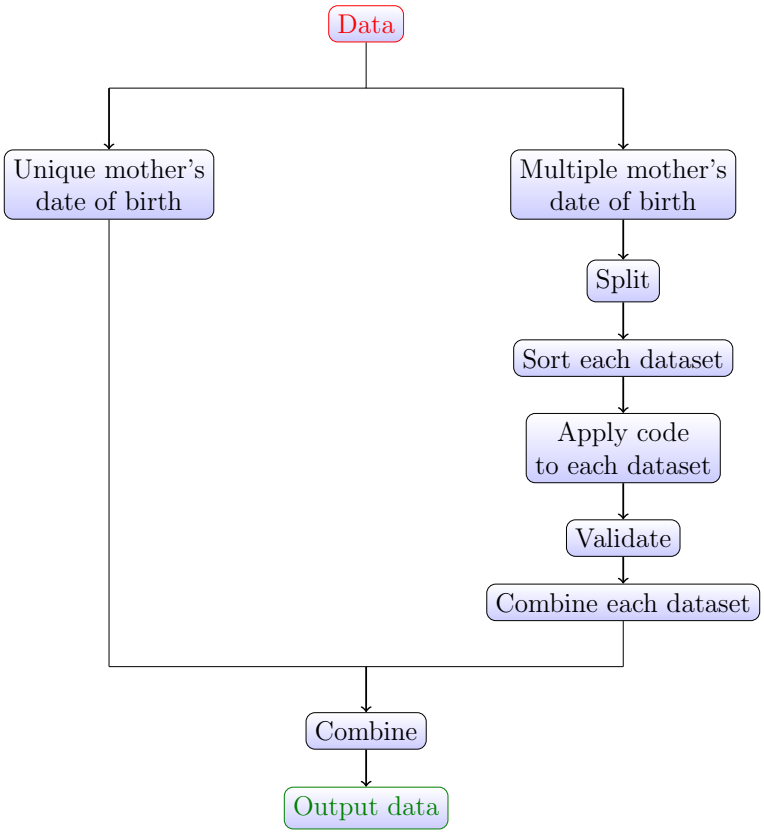


Figure 5.12: Algorithm

5.5.2 Results

Given we knew the true birth history of each mother, we grouped the observations by what we called *sibling sizes*. In other words, observations from women with exactly two, or three, or four, or in general n children were referred to as sibling size 2, 3, 4, and n, respectively. Table 5.2 summarizes the number of observations per sibling size once we removed all the observations where the mother's date of of birth appears once (22

observations).

Sibling size	**Number of Observations**
1	3841
2	5768
3	4032
4	1956
5	920
6	366
7	189
8	120
9	63
10	40
11	11
12	24
Total	17,330

Table 5.2: Number of Observations per Sibling Size n

We knew how many observations belong to each sibling size. So after running our code, their will be a discrepancy between the true and estimated count. Furthermore, children belonging to a specific sibling size could be associated to another sibling size, so we illustrate this via Figure 5.13.

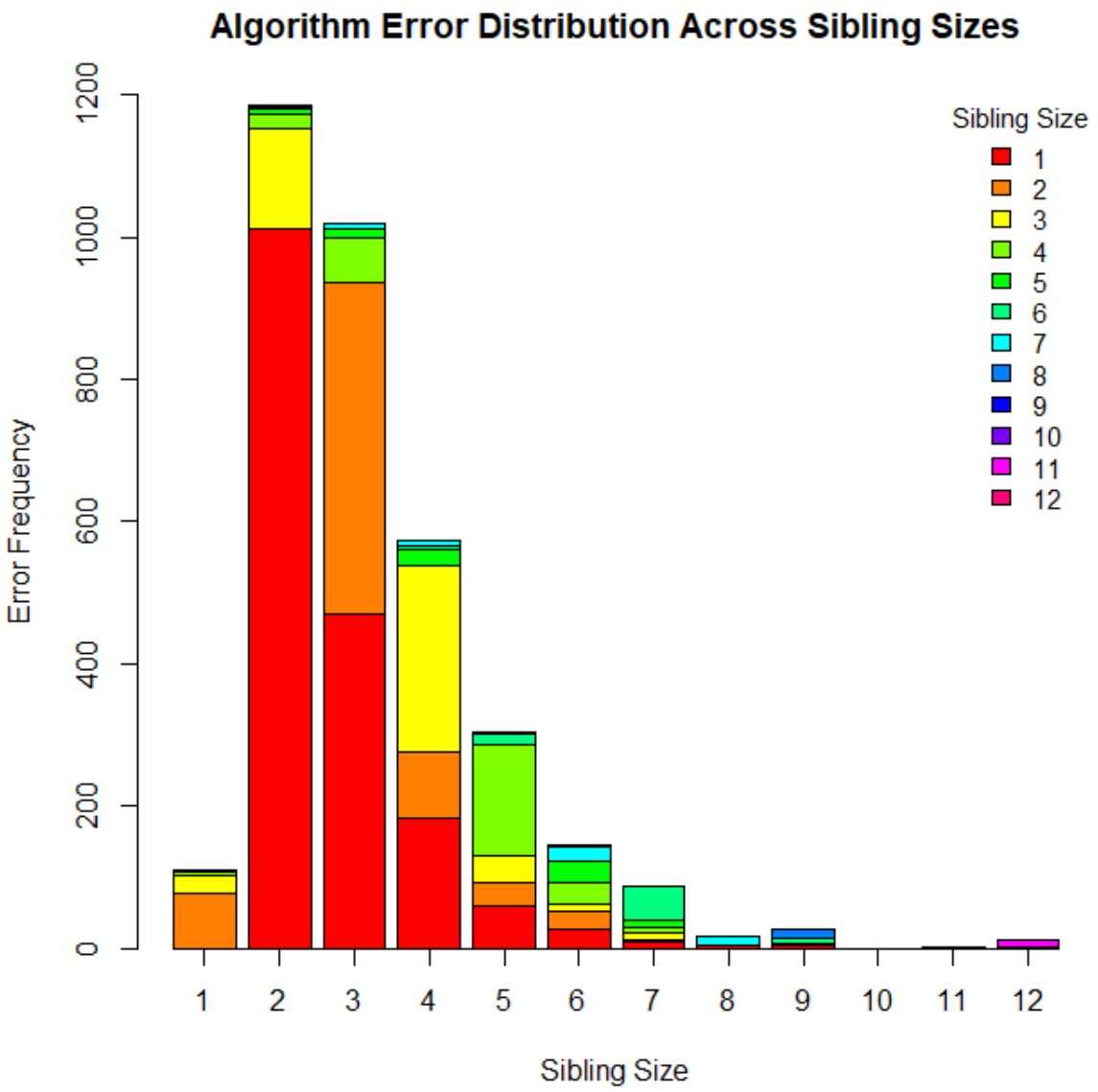


Figure 5.13: Error Allocation Per Sibling Size

An equivalent representation is a heat map which we show in Figure 5.14. Note that we scaled the numbers between 0 and 1.

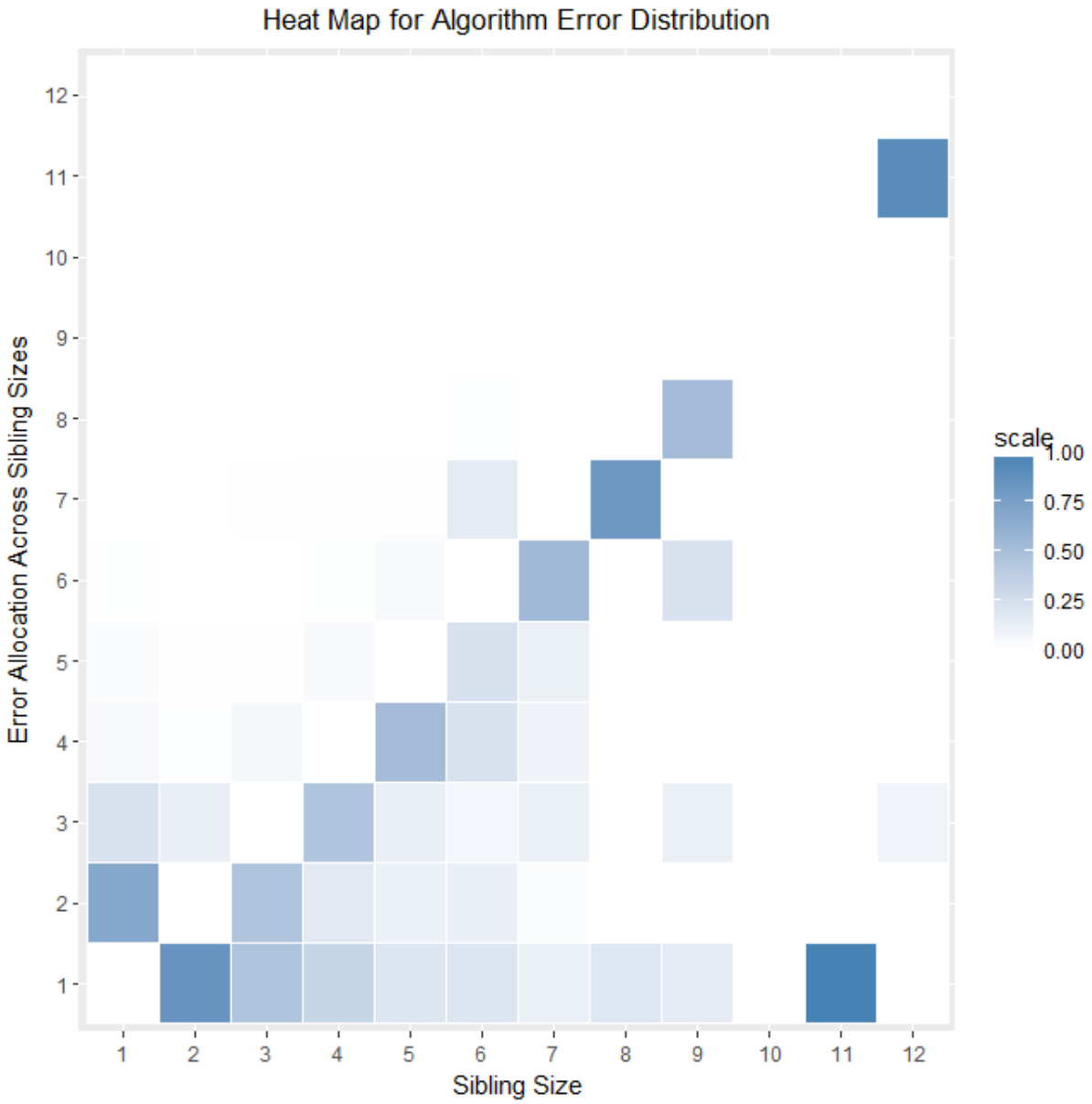


Figure 5.14: Heat Map for Error Allocation Per Sibling Size

To obtain a more accurate measure of efficiency of our algorithm, Table 5.3 shows the true number of observations per sibling size, how many of them were correctly identified, and the percentage.

Sibling Size	True	Estimated	Percentage
1	3841	3730	97.11
2	5768	4414	76.52
3	4032	2835	70.31
4	1956	1300	66.46
5	920	575	62.5
6	366	210	57.38
7	189	98	51.85
8	120	104	86.67
9	63	36	57.14
10	40	40	100
11	11	0	0
12	24	12	0.5
Total	17,330	13,354	77.06

Table 5.3: Number of Correctly Identified Observation Per Sibling Size

5.6 Final Results

Using data from the NSFG, among women with at least two children, we estimated the causal parameter of taking prenatal care during the second birth on the gestational age under no weights, and weights defined by 4.1 and 4.2. Figure 5.15 shows the distribution of the weights under each data type.

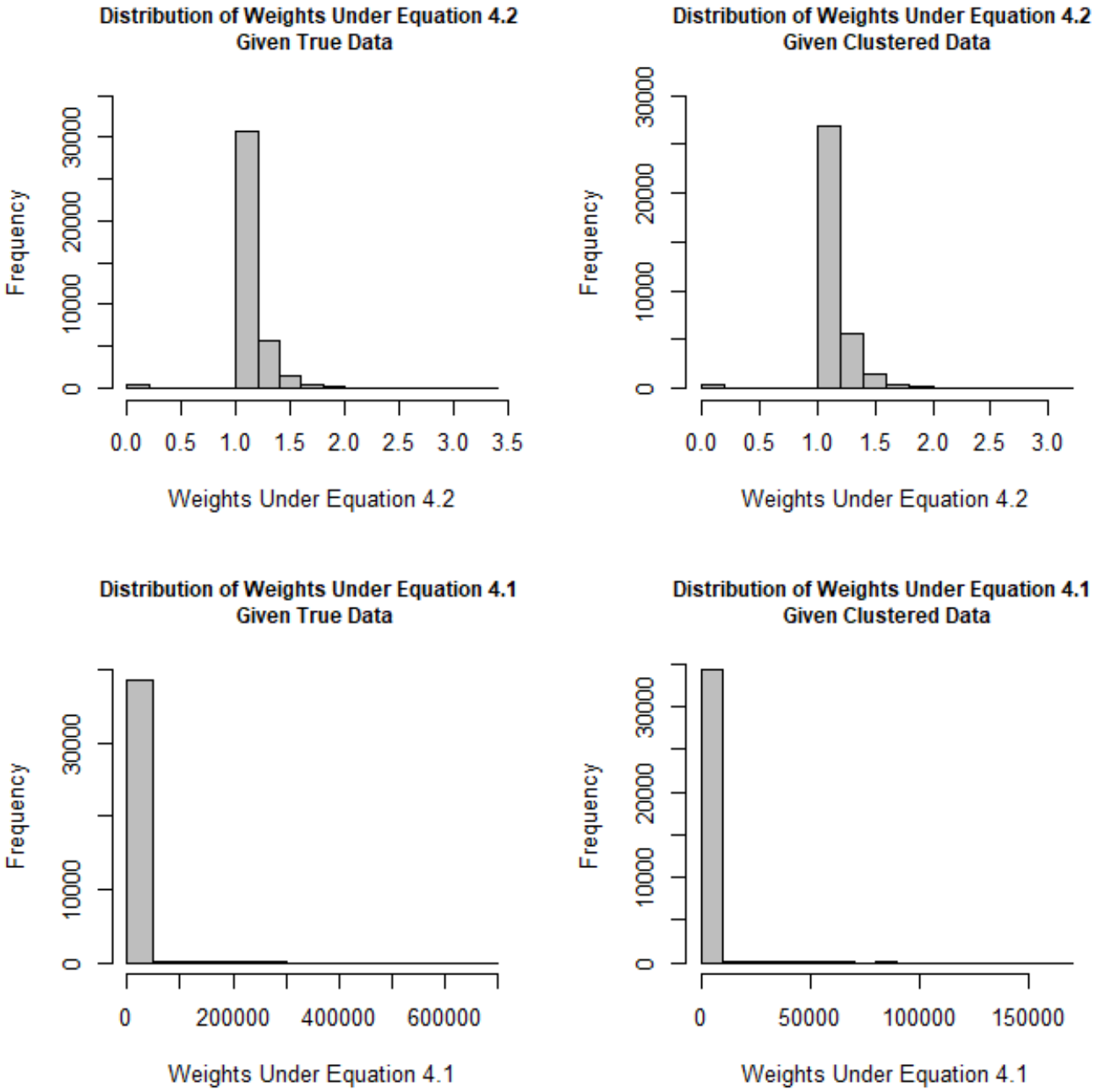


Figure 5.15: Distribution of Weights

We used the mother's age and outcome during the first birth as confounders. Note, mothers already had an identifier variable so the birth history of a woman was known.

We then repeated the analysis, but we first applied the algorithm described in section 5.5.1 (Methods) i.e., as if the data was anonymized. Table 5.4 summarizes the estimates

for ψ along with the 95% confidence intervals, and the grid used in each case.

Data	**Model**	$\hat{\psi}$ (95% CI)	Grid used (From, To, By)
True Data	Unweighted	-0.292 (-0.596, 0.004)	(-0.900, 0.900, 0.004)
	Weighted 1	0.120 (-0.240, 0.992)	(-0.288, 1.040, 0.008)
	Weighted 2	0.088 (-0.244, 0.900)	(-0.900, 0.900, 0.004)
Algorithm	Unweighted	-0.322 (-0.670, 0.050)	(-0.670, 0.902, 0.006)
	Weighted 1	0.264 (-0.088, 1.016)	(-0.288,1.040, 0.008)
	Weighted 2	0.014 (-0.352, 0.902)	(-0.670, 0.902, 0.006)

Table 5.4: Summary of Main Analysis

We see that the unweighted estimate is -0.292 (95% CI: -0.596, 0.004). Thus, a mother that would have started prenatal care earlier would have had a longer gestational age by a factor of 1.339 than if she had not. Given the 95% confidence interval, we conclude that the estimate is not significant. The estimate is incorrect since it was obtained by treating competing events as random censoring. The survivor average causal effect was estimated to be 0.088 (95% CI: -0.244, 0.900), and not to be significant. Thus, we conclude that starting prenatal care earlier does not have a causal effect on the length of the second delivery. More precisely, regardless whether a mother took prenatal care or not, there is no evidence of an effect of the exposure among mothers who would have had a live birth on their second delivery. We arrive at the same conclusion when we estimate the survivor average causal effect by using the date obtained through the algorithm; survivor average

causal effect was estimated to be 0.014 (95% CI: -0.352, 0.902).

Under both data formats, since the weights under equation 4.1 were very unstable (see bottom panel of Figure 5.15), which causes the causal parameter estimate to be unstable and doubtful, we truncated the weights at the 95% quantile (see Figure 5.16). Thus, the estimates under row "Weighted 1" shown in Table 5.4 are obtained using the truncated weights. Under the true data and after applying the algorithm, the estimates were 0.120 (95% CI: -0.240, 0.992) and 0.264 (95% CI: -0.-88, 1.016), respectively. In either case, the estimate is positive so taking prenatal care earlier does not cause a lengthening of the gestational age of the second pregnancy. Notice that from the estimated 95% confidence intervals, both estimates are not significant.

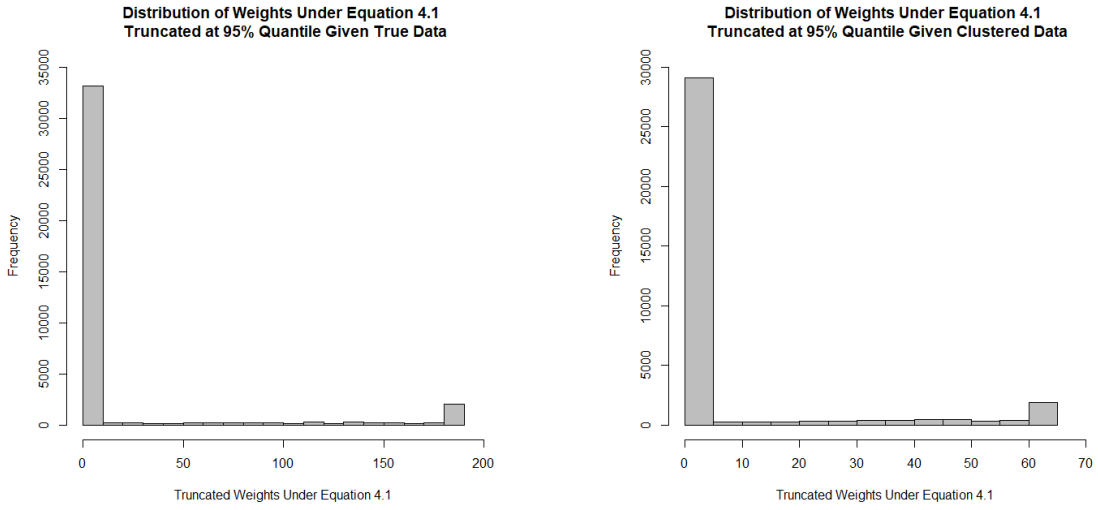


Figure 5.16: Distribution of Truncated Weights Given by Equation 4.1

We see that both weighting methods correct for the presence of competing risks by attenuating the estimate of the casual effect, as well as widening the confidence intervals which indicates that we have more uncertainty about the point estimate. This is an indication that the censoring by competing risks is not random. One possible reason could be that with more advanced prenatal care, babies that otherwise would not have

survived, now will, but those babies have shorter gestational ages, either due to pre-emptive Cesarean sections or earlier induction.

Chapter 6

Discussion and Conclusions

The goal of this thesis was to analyze an anonymized clustered data using causal inference in the presence of competing events.

In Chapter 1, we established the question of main interest, namely the causal effect of prenatal care on the gestational age of live births in the presence adversary deliveries that was answered using data collected by the The National Survey of Family Growth. In Chapter 2, we summarized the basic methodology of survival analysis and common quantities, methods used to analyze a time to event random variable, and briefly discussed competing events. In Chapter 3, we introduced what causal inference is and the class of G-methods – Inverse Probability Weighting, Standardization/g-formula, and g-estimation – that is often used to find the causal effect of certain exposure on some outcome.

In Chapter 4, we reviewed an approach suggested by Naimi and Tchetgen Tchetgen to estimate a causal parameter in the presence of competing events. Instead of estimating a standard average causal effect, using g-estimation, we estimate an average causal effect in a subgroup of the population, in other words, the survivor average causal effect (SACE).

In Chapter 5, we presented the main results of this thesis. section 5.1 explained an approach to simulate survival data in the presence of competing events by using the cause specific hazards. Using an algorithm proposed by G. Young, A. Hernán, Picciotto, and M. Robins [41], section 5.2 expanded the simulation process by allowing for time varying treatment and a causal framework using the model in section 4.1. In section 5.4, we performed a simulation to support the theory established in Chapter 4. In section 5.5, we proposed an algorithm that allows the researcher to find similar observations in a longitudinal data if the latter is anonymized. Lastly, section 5.6 showed the final results and the SACE estimates using data from the NSFG.

Among women with at least two children, we found that at the second birth the unweighted estimate of the causal effect of prenatal care on the gestational age when using the true data was -0.292. The 95% confidence interval from Table 5.4 suggests that the estimate is not significant. The estimate is biased since the presence of competing events was omitted. We then compute the survivor average causal to be 0.088 so that among women who would have had a live birth in the presence of other type of deliveries, regardless of their exposure history, taking prenatal care would be harmful. Similar (or not) results were obtained when the data was assumed to be anonymized; hence, the algorithm in section 5.5 was applied prior to the analysis.

Aside from these results, we used a new simulation approach (inspired by G. Young, A. Hernán, Picciotto, and M. Robins) to generate a longitudinal data using a causal framework in the presence of competing events. Our simulation results agreed with the theory with an exception when the SACE was estimated (see Figure 5.11). We also proposed a clustering algorithm, and to our best knowledge, this was not done before.

The estimates for the effect of prenatal care should not be regarded or reported as

public health factor; they served only to illustrate the methods introduced in this thesis. To improve the results shown, first more confounders should be considered; we only used two. Adjusting for the mother's education level, socioeconomic and smoking status, and other factors should be considered when performing the g-estimation and computing the weights. A finer grid should be used ideally one that has increments of 0.01 or even 0.001, not only to obtain a more accurate estimate, but to also obtain a more precise confidence interval. We tested different grids, and with bigger increments, the 95% CI would sometimes capture 0 and sometimes not. Notice that we did not perform any diagnostics or verified if required assumptions hold. For instance, in order to use 4.2 to estimate the SACE, we need to assume that sequential monotonicity and the concordant survivorship assumption for time-varying exposures hold [30]. We also assumed that the exposure has indeed an effect on the outcome so the SACE bears it interpretation. Unfortunately, if the exposure has no overall effect on the outcome, the SACE can be misleading [22]. Possible future work will be to repeat section 5.4 and use a latent failure time approach to simulate the data, reapply the causal models, and compare the results. Improving the algorithm in section 5.5 by increasing the percentage of correctly matched observations should also be considered. We also ignored possible data entry errors such as the pregnancy orders do not follow increments of 1 or the date of a previous birth does not precede the date of a current birth. Therefore, prior to applying the algorithm, the data should be cleaned. Note that the data was not collected by us, so we did not have access to the original patient records. The data was collected by the National Survey of Family Growth so any validations or modifications done on the data had to be confirmed or approved by them. Therefore, due to time limitations and inaccessibility to the original data, we could not afford to clean it. Lastly, we focused only at second deliveries, so

repeating this analysis by considering the entire birth history of a mother is a question to be investigated.

Bibliography

[1] Definition of term pregnancy. Committee Opinion No. 579. American College of Obstetricians and Gynecologists. Obstet Gynecol 2013;122:1139–40. https://www.acog.org/Clinical-Guidance-and-Publications/Committee-Opinions/Committee-on-Obstetric-Practice/Definition-of-Term-Pregnancy.

[2] Per K Andersen, Elisavet Syriopoulou, and Erik T Parner. Causal inference in survival analysis using pseudo-observations. *Statistics in Medicine*, 2017.

[3] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

[4] Jan Beyersmann, Arthur Allignol, and Martin Schumacher. *Competing risks and multistate models with R.* Springer Science & Business Media, 2011.

[5] Jan Beyersmann, Aurelien Latouche, Anika Buchholz, and Martin Schumacher. Simulating competing risks data in survival analysis. *Statistics in Medicine*, 28(6):956–971, 2009.

[6] Charlotte A Bolch, Haitao Chu, Stephanie Jarosek, Stephen R Cole, Sean Elliott,

and Beth Virnig. Inverse probability of treatment-weighted competing risks analysis: an application on long-term risk of urinary adverse events after prostate cancer treatments. *BMC medical research methodology*, 17(1):93, 2017.

[7] Ebony B Carter, Lorene A Temming, Jennifer Akin, Susan Fowler, George A Macones, Graham A Colditz, and Methodius G Tuuli. Group Prenatal Care Compared With Traditional Prenatal Care: A Systematic Review and Meta-analysis., 2016.

[8] Robert Choinière, Michel Pageau, and Marc Ferland. Prevalence and geographic disparities in certain congenital anomalies in Quebec: comparison of estimation methods [1989-1995 data]. *Chronic Diseases and Injuries in Canada*, 20(2):51, 1999.

[9] David R Cox. Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer, 1992.

[10] David Roxbee Cox and David Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

[11] Brian L Egleston, Daniel O Scharfstein, Ellen E Freeman, and Sheila K West. Causal inference for non-mortality outcomes in the presence of death. *Biostatistics*, 8(3):526–545, 2006.

[12] VT Farewell and DR Cox. A note on multiple time scales in life testing. *Applied Statistics*, pages 73–75, 1979.

[13] National Center for Health Statistics. Public-use data file documentation: 2006-2010 and MD. 2017 2011-2013, National Survey of Family Growth. Hyattsville. `https://www.cdc.gov/nchs/nsfg/index.htm`.

[14] Steven Lawrance Gortmaker. The effects of prenatal care upon the health of the newborn. *American Journal of Public Health*, 69(7):653–660, 1979.

[15] Eric S Hall, Neera K Goyal, Robert T Ammerman, Megan M Miller, David E Jones, Jodie A Short, and Judith B Van Ginkel. Development of a linked perinatal data resource from state administrative and community-based program data. *Maternal and child health journal*, 18(1):316–325, 2014.

[16] Katie Harron, Ruth Gilbert, David Cromwell, and Jan van der Meulen. Linking data for mothers and babies in de-identified electronic health data. *PloS one*, 11(10):e0164667, 2016.

[17] Henry B Hemenway, William H Davis, and Charles V Chapin. Definition of stillbirth. *American Journal of Public Health and the Nations Health*, 18(1):25–32, 1928.

[18] Miguel A Hernán, Stephen R Cole, Joseph Margolick, Mardge Cohen, and James M Robins. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and drug safety*, 14(7):477–491, 2005.

[19] Miguel A Hernán and James M Robins. *Causal inference*. CRC Boca Raton, FL:, 2017. `https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/`.

[20] Lisa Hilder, Kate Costeloe, and Baskaran Thilaganathan. Prolonged pregnancy: evaluating gestation-specific risks of fetal and infant mortality. *BJOG: An International Journal of Obstetrics & Gynaecology*, 105(2):169–173, 1998.

[21] Caroline SE Homer, Charlene Thornton, Vanessa L Scarf, David A Ellwood,

Jeremy JN Oats, Maralyn J Foureur, David Sibbritt, Helen L McLachlan, Della A Forster, and Hannah G Dahlen. Birthplace in New South Wales, Australia: an analysis of perinatal outcomes using routinely collected data. *BMC pregnancy and childbirth*, 14(1):206, 2014.

[22] Marshall Joffe. Principal stratification and attribution prohibition: good ideas taken too far. *The international journal of biostatistics*, 7(1):1–22, 2011.

[23] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[24] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[25] Maiken IS Kjaersgaard and Erik T Parner. Instrumental variable method for time-to-event data using a pseudo-observation approach. *Biometrics*, 72(2):463–472, 2016.

[26] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[27] David G Kleinbaum and Mitchel Klein. *Survival analysis*, volume 3. Springer, 2010.

[28] Paul M Krueger and Theresa O Scholl. Adequacy of prenatal care and pregnancy outcome. *The Journal of the American Osteopathic Association*, 100(8):485–492, 2000.

[29] Shiliang Liu and Shi Wu Wen. Development of record linkage of hospital discharge data for the study of neonatal readmission. *Chronic Diseases and Injuries in Canada*, 20(2):77, 1999.

[30] Ashley I Naimi and Eric J Tchetgen Tchetgen. Invited commentary: estimating population impact in the presence of competing events. *American journal of epidemiology*, 181(8):571–574, 2015.

[31] Tazi Nimi, Sílvia Fraga, Diogo Costa, Paulo Campos, and Henrique Barros. Prenatal care and pregnancy outcomes: A cross-sectional study in Luanda, Angola. *International Journal of Gynecology & Obstetrics*, 135(S1), 2016.

[32] Christophe Pouzat. *STAR: Spike Train Analysis with R*, 2012. R package version 0.3-7.

[33] Joel G Ray, David A Henry, and Marcelo L Urquia. Sex ratios among Canadian liveborn infants of mothers from different countries. *Canadian Medical Association Journal*, pages cmaj–120165, 2012.

[34] Amy Richardson, Michael G Hudgens, Jason P Fine, and M Alan Brookhart. Nonparametric binary instrumental variable analysis of competing risks data. *Biostatistics*, 18(1):48–61, 2017.

[35] Horst Rinne. *The Hazard rate : Theory and inference (with supplementary MATLAB-Programs)*. Justus-Liebig-Universität, 2014. `http://geb.uni-giessen.de/geb/volltexte/2014/10793`.

[36] James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.

[37] James M Robins, Miguel Angel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[38] G. Rodríguez. Lecture Notes on Generalized Linear Models, 2007. http://data.princeton.edu/wws509/notes/.

[39] Malinda Steenkamp. Clustering in Northern Territory Perinatal Data for 2003–2005: Implications for Analysis and Interpretation. *Health Information Management Journal*, 43(1):37–41, 2014.

[40] Marcelo L Urquia, Rahim Moineddin, Prabhat Jha, Patricia J O'campo, Kwame McKenzie, Richard H Glazier, David A Henry, and Joel G Ray. Sex ratios at birth after induced abortion. *Canadian Medical Association Journal*, pages cmaj–151074, 2016.

[41] Jessica G Young, Miguel A Hernán, Sally Picciotto, and James M Robins. Simulation from structural survival models under complex time-varying data structures. *JSM Proceedings, Section on Statistics in Epidemiology, Denver, CO: American Statistical Association*, 2008.

[42] Cheng Zheng, Ran Dai, Parameswaran N Hari, and Mei-Jie Zhang. Instrumental variable with competing risk model. *Statistics in Medicine*, 36(8):1240–1255, 2017.