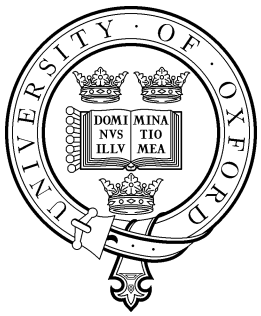
Debiased Machine Learning Causal Inference for Time-Varying Social Variables



Guanghui Pan

Lady Margaret Hall

University of Oxford

A thesis presented for the degree of

Doctor of Philosophy

Michaelmas Term 2024

Acknowledgements

This thesis was completed under the supervision of Professor Richard James Breen, to whom the author is deeply grateful. During the Transfer of Status, Confirmation of Status, and the thesis submission and viva process, Professor Colin Mills, Professor Christiaan Monden, Professor Frank Windmeijer at the University of Oxford, Professor Felix Elwert at the University of Wisconsin–Madison, and Professor Geoffrey Wodtke at the University of Chicago read part or all of the thesis and offered valuable comments and suggestions, for which the author also expresses sincere thanks. The author alone is responsible for any remaining errors.

Abstract

This thesis reviews and develops efficient debiased machine-learning estimators for causal inference and applies the developed methods to empirical research in sociology and demography.

The thesis is composed of five chapters, organized in three parts. Part 1 is the first chapter, which gives a comprehensive methodological review of causal inference and efficient debiased estimators. Part 2 discusses doubly robust/debiased machine-learning techniques for causal inference with survival data; it includes one methodological chapter that develops a twice doubly robust estimator for left-truncated-right-censored survival data and one empirical chapter that addresses the causal effect of widowhood on mortality. Part 3 discusses doubly robust/debiased machine-learning techniques for causal mediation analysis; it includes one methodological chapter that derives debiased nonparametric estimators for both static and time-varying marginal structural models and one empirical chapter that applies these methods to analyze how labor market participation mediates the wage penalties and premiums associated with parenthood and marriage for both genders.

Introduction

Social science fundamentally aims to provide comprehensive explanations of complex social phenomena. With the development in the availability of individual-level survey data and methodological developments in statistics and econometrics, quantitative empirical studies applying the rigorous causal inference framework have appeared in sociological papers since the 1990s. However, for a considerable period, the causal inference estimators employed by social scientists have relied on regression-based parametric models. Ordinary least squares (OLS) based estimators have the advantages of being unbiased, efficient, and consistent, but they also have disadvantages: OLS models (or other parametric models) impose stringent requirements on the distribution of variables, and survey data often have difficulty meeting these requirements; the OLS models are usually incapable of dealing with the high-dimensional data (while the number of predictors exceeds the number of observations); further, models lack flexibility which restricts the relationship between the independent and dependent variables to be linear.

With the use of non-parametric machine learning (ML) techniques in place of parametric methods, the prediction of counterfactual outcomes replaces the interpretation of coefficients. However, with the greater flexibility of ML models, researchers face the so-called variance-bias tradeoff between the risk of overfitting the data and the ability to capture complex relationships. Unbiasedness of the estimator is usually a requirement for predict-

ing counterfactual outcomes, and, as with parametric models, we would also want the ML estimators to be consistent and efficient.

This thesis aims to develop Neyman-orthogonal, efficient-influence-function–based estimators for causal inference under diverse social-science settings. While semiparametric efficiency theory and efficient-influence-function-based causal estimators have a long history (e.g., doubly robust estimators and TMLE), recent Double/Debiased Machine Learning (DML) formalizes the use of orthogonal scores with modern ML and cross-fitting to obtain $\sqrt{n}$-consistent, asymptotically normal estimators in the presence of high-dimensional nuisance functions (Chernozhukov et al. 2018b). When the orthogonal score is chosen to be the efficient influence function, the resulting DML estimator is semiparametrically efficient; otherwise, it is still debiased/orthogonal but not necessarily efficient. DML complements classical GMM and MLE and is increasingly used for treatment effect and other structural parameters in the social sciences.

Indeed, the DML estimator for causal inference requires some prerequisite knowledge on the two sides of the statistical literature: on the one hand, researchers should be familiar with the appropriate methods and prerequisite assumptions for causal inference; on the other hand, researchers should fully understand the efficient estimation theory and how the efficient-based DML estimator for causal inference is produced (van der Vaart 1998; Tsiatis 2006).

Based on this, the thesis attempts to introduce efficient causal estimation into sociology and demography, with empirical studies showing how the estimations can be used. The thesis is composed mainly of three parts: the first part is a thorough review of the two pre-

requisites mentioned in the above paragraph. I try to make the content easier for readers with sociological and demographic backgrounds to digest. Next, I will develop the existing DML framework in two areas. Firstly, I design a twice doubly robust estimation framework for causal inference with left-truncated-right-censored survival data. Demographers often encounter this kind of data when analyzing survey data. After implementing the method, I have a chapter discussing the heterogeneous treatment effect of widowhood on mortality, to which I will apply the DML estimator for causal effects on survival data. Secondly, I will combine the doubly robust estimator with the time-varying causal mediation analysis, with an empirical chapter concerning the classic parenthood and marriage premiums or penalties on wages elaborating on the method. The specific contents for the chapters are below.

I. Part 1: Introduction to Causal Inference under Efficient Theory

This chapter mainly reviews the theoretical and methodological background of causal inference under the efficient estimation theory. I review the basic concepts and assumptions for causal inference, basic ideas of efficient estimators, and estimation methods for the debiased average treatment effect under efficient theory.

II. Part 2: Debiased estimator for heterogeneous treatment effects in survival data

This part discusses the doubly robust (DR) estimator for the average treatment effect and conditional average treatment effect (ATE/CATE) for the survival analysis. The methodological chapter discusses combining the DR causal inference framework with the DR framework for truncated and censored data and develops the twice doubly robust estimator for

causal survival analysis. At the same time, the substantive paper applies the DR causal survival framework, discussing widowhood's effects on mortality.

In the methodological chapter, I review the estimation for the survival outcome and the previous deep learning strategy dealing with truncation and censoring. Based on the previous naïve plug-in augmented inverse probability weighting (AIPW) estimand, I discuss the DR framework for estimating the survival function with truncation and censoring in observational data. I then turn to illustrate the estimation of treatment effects from survival data and again develop the DR estimator for it. Combining the two parts of DR estimation, I introduce the twice doubly robust estimator for causal survival analysis. I compare its estimation with marginal hazard ratio (MHR) estimation, naïve Cox Proportional hazard (Cox-PH) estimation, naïve doubly robust survival estimation, and doubly robust Cox-PH estimation using a simulated dataset.

In the substantive chapter, I discuss how to measure the causal effect of widowhood on mortality. Demographers and sociologists have discussed whether the correlation between widowhood and excess death is due to selection or causality and have had various results on how the causal effects are affected by an individual's education, race, and wealth strata. In this chapter, I point out their methodological shortcomings in estimation and apply the DR causal framework combined with time-varying Cox-PH models to estimate the effects.

III. Part 3: Doubly robust estimator for time-varying causal mediation analysis

In this part, I reexamine the efficient estimator in causal mediation analysis. Statisticians and epidemiologists have developed methods to decompose the causal effects into the di-

rect effects on the outcome and the indirect effects via the mediator on the outcome. Moreover, g-formulas estimate the treatment effects with time-varying treatments, mediators, and confounders. I first derive the efficient estimator for causal mediation analysis based on the controlled response function (CRF) and combined with the g-formula, I discuss the efficient (doubly robust [1]) estimation for the mediation effects (interventional direct and indirect effects; IDE/IIE) for the multiple-period mediation models. I also differentiate the estimands with assumptions for carryover effects and feedback effects. Finally, I applied the dataset that Fearon and Laitin (2003) constructed to test whether political instability is the only mediation path affecting the causal relationship between racial fractionalization and the onset of civil wars in static and dynamic model settings.

In the substantive paper, I apply the DR estimator for causal mediation to investigate the marital and parenthood premiums and penalties for men and women. The causal mediation framework can also be used to measure how the gaps between the two groups are reduced. Using the static causal mediation model, I provide an example of the parenthood and marital effect on young-adult wage ranks between married/not married and between those with and without children. I also test effect heterogeneity by fertility and marital age with the dynamic mediation model. Moreover, in this empirical research, I also test the mediation effect of labor market participation on the marital/parenthood effects. This research provides a different perspective on parenthood and marital inequality, as previous literature focuses more on the treatment effects for the treated switchers (for individuals

[1] Although recent studies attempted to yield even the "multiply robust" estimators, in this thesis, I only discuss the "doubly robust" ones. The difference is not only in the number of nuisance functions: multiple robust suggests that we have multiple nuisance functions ($K > 2$), and any of them correctly specified makes the estimator unbiased, which can be called multiply robust.
An example here is, if we have two distinctive choices, for instance, either the expectation nuisance model μ correctly specified or all probability nuisance model π correctly specified, then the model is doubly robust instead of multiply robust.

before and after the event). Nevertheless, this research focuses on the population-level average treatment effect, which compares the effect between the groups.

Contents

List of Figures

List of Tables

Chapter 1

Methodological Foundations of Modern Causal Inference in Social Science Research

I. Notations, and Fundamental Analytical Framework

This chapter serves as an introduction and literature review on the methods to be discussed and developed in the thesis. The method concerning the efficient-based debiased machine learning for causal inference in social science is mainly based on two veins in statistics. On the causal inference side, it focuses on when and under what conditions a specific value from a statistical model can be interpreted as causal; and on the debiased machine learning side, the technique addresses how estimates derived from observational data can be made unbiased and applied to the statistical model.

We start with the concept of inference, which is the core of statistical analysis. Compared to descriptive analysis of samples, statisticians are more interested in observing a small part of the individuals in a population. Through appropriate inductive reasoning, we can gain knowledge about the population. Therefore, they propose hypotheses and design mathematical models, whether simple or complex, aimed at inferring the characteristics of the population (Neyman 1990).

Mathematically, the population is abstracted as a closed set. Some attributes of the population are **measurable**. We define **measurable sets** Z as a set of non-empty, closed (under complement, countable unions, and countable intersections) subsets of the population [1].If we have a set of functions P **measuring** (assigning numbers to the characteristics of) the measurable set, or mathematically, mapping the measurable set onto the real numbers ($P : Z \to \mathbb{R}$), we call the functions **measures**, or **data-generating process (DGP)**.In other words, suppose in a measurable set Z we have two measurable attributes, $Z = (X, Y)$; the data-generating process gives out the measurable distributions X and Y (Capiński and Kopp 2004; Billingsley 1995).

However, the DGP is, at most times, hard to describe. Consider X to represent the heights and Y to represent the weights of individuals within the British population. It is hard to decipher the data-generating distribution for heights and weights, even if we could possibly obtain them from the census data. An intuitive way to describe it is to set specific **statistical target parameters**, for instance, the mean and variance of the height and weight of the British people. The function to map the DGP to the target parameter is called **estimator** ψ[2], and the number of the target parameter $\psi(P)$ is usually called the **(statistical) estimand** (i.e., $\psi(P) = E_P(Z)$).

A more common condition is that we could never know the true DGP P; we could only

[1] which is called $\sigma-$algebra. Indeed, suppose the population is D and the measurable sets Z, we define a **measurable space** with the pair (D, Z). Further, with measure P, we define the triple (D, Z, P) as the **measure space**.

[2] Rigorously, it should be called "estimandor", and Schuler and van der Laan(2024) do use this term to call it (as it is the generator for the estimand). However, in most literature, it is called the estimator (which should be the generator for value derived from the estimates).

observe the samples $Z_i = (X_i, Y_i)$. **Statistical inference** is the process by which we use the samples to speculate on the characteristics of the population. In this thesis, we assume the samples we used in statistical inference are randomly drawn from the population through the same process, or in other words, [3] (Morgan and Winship 2015; Imbens and Rubin 2015):

Assumption 1.I.1 (Independent and Identically Distributed) *The samples are **independent and identically distributed**.* [4]

With the independent and identically distributed (IID) samples, we have the empirical measure (distribution) based on the observations $\mathbb{P}_n$ (n denotes the sample size). Correspondingly, we have an **estimator** denoted as $\hat{\psi}$ and the empirical **point estimate** $\psi(\mathbb{P}_n)$. [5] If $\psi(\mathbb{P}_n)$ contains the sample mean, or any linear combination of the sample means, under the IID assumption, we have the central limit theorem (CLT):

Theorem 1.I.1 *Central Limit Theorem (CLT): Suppose μ and σ^2 separately denote the expectation and the variance for the IID observations. If σ^2 is finite, as the sample size n approaches infinity, difference between the average sample mean $\hat{\psi}$ and the expectation μ approaches a*

[3] Rigorously, the DGP for random samples Z_i are called, in most textbooks, the random variables. Therefore, the random variable is indeed a function and usually is denoted as $Z = Z_i$.

[4] For instance, using convenience sampling does not draw IID samples. We cannot identify each sample's probability of being sampled from the whole population, and the samples are not connected (consider an extreme scenario in which we sample the heights from a British professional basketball league that can never represent the heights of the British population). Nonetheless, samples from non-IID settings may also perform statistical inference, but I will not address the techniques in this thesis.

[5] We could understand the estimandor and the estimand separately as the function and the quantity to be estimated, while the estimator and the estimate separately as the function and the quantity we perform estimation based on the observed data. However, in most statistical papers and textbooks, estimator and estimate are interchangeable and rarely rigorously defined, and most researchers assume that the estimator and the estimandor have the same function ψ. In this thesis, if the notation is clear, I also use ψ to refer to the estimand $\psi(P)$ and $\hat{\psi}$ to refer to the estimator $\psi(\mathbb{P}_n)$.

normal distribution with mean 0 and variance σ^2, at the rate of $1/\sqrt{n}$:

$$\sqrt{n}\big(\psi(\mathbb{P}_n) - \mu\big) \to N(0, \sigma^2) \tag{1.I.1}$$

If μ is an **unbiased** estimator for $\psi(P)$, Equation 1.I.1 describes the **asymptotic** relationship between $\hat{\psi}$ and ψ, and the difference between $\hat{\psi}$ and ψ is called **statistical error**. I will elaborate more on the analysis of statistical error in Section III.

Indeed, in some statistical analyses, our ultimate goal is not to estimate the statistical estimand. Instead, we use statistical estimand to approach the "real" estimand in our problems. Consider the following common scenarios in sociological or demographic research:

1. Suppose we are interested in the **causal relations** between a treatment and the outcome. Assume that our treatment is binary with only two values: treatment and control. Also, the treatment should be assigned before the outcomes are observed. We use $Y(1)$ to denote the outcome under treatment, while $Y(0)$ denotes the outcome under control (in some textbooks, they are denoted as Y_t and Y_c). Let A denote the treatment variable, and X denote the set of covariates. The causal measurable set can be written as $Z^* = (X, A, Y(1), Y(0))$ under the causal DGP P^*. Suppose our target of interest is the **average treatment effect** defined as the difference between the expectation of $Y(1)$ and $Y(0)$: $\psi^*(P^*) = E_{P^*}\big[E[Y(1)] - E[Y(0)]\big]$. In the real world, due to the fundamental problem in causal inference, we can not observe the outcomes under the treatment and control simultaneously; we could only have the measurable set $Z = (X, A, Y)$ under the statistical DGP P and try to use a reasonable statistical estimand to approach ψ^*, for instance, the conditional expected outcome given the treatment is assigned versus the control is assigned: $\psi(P) = E_P\big[E[Y|A=1] - E[Y|A=0]\big]$.

Only under specific assumptions can we conclude that $\psi(P)$ is equivalent to $\psi^*(P^*)$.

2. Suppose our parameter of interest is the mean survival time for a (sub)population. A simple measurable data frame is $Z^* = (X, T)$, in which X denotes the covariates and T denotes the survival time (under specific survival probability), and therefore, the estimand is $\psi^*(P^*) = E[T]$ [6]. However, we may encounter the problem that the survival time is censored, meaning that we cannot directly know the survival time T but know the censoring time T_C. Therefore, the data structure for us is $Z = (X, T, \delta, T_C)$, where δ signals if we occur censoring and T_C denotes the censoring time. A possible (but uncommon) statistical estimand with the data structure is $\psi(P) = E[T|\delta = 0]$, which ignores the censoring data. Further assumptions are required to make our statistical estimand approach the survival estimand. The detailed techniques for survival inference with truncation and censoring are the main part of Chapter 2 of this thesis.

3. Suppose we are interested in how a mediator interferes with the causal relationship between the treatment and the outcome, for instance, how much the causal effect directly goes from the treatment to the outcome (the direct effect) and how much it goes through the mediator onto the outcome (the indirect effect) (VanderWeele 2015, Chapter 2). We still assume that our treatment is a binary one $A = 1$ denotes the treatment, and $A = 0$ denotes the control. The mediator takes the value of $m(1)$ when $A = 1$ and $m(0)$ when $A = 0$. Thus, there are four combinations of the potential outcome: $Y(1, m(1)), Y(1, m(0)), Y(0, m(1))$, and $Y(0, m(0))$. Our estimands are the two

[6] Of course, in real life analysis, we usually have a more complex data frame like $Z = (X, S(t_1), S(t_2), \ldots, S(t_n))$, where $S(t_i)$ denotes the survival probability at t_i. Thus, we first calculate the survival time under specific survival probability (for instance, half-life survival time, in which $S(t_m) = 0.5$) with S^{-1}, and average the survival time to get $E[T]$.

components of the total average treatment effect $E[Y(1,m(1))] - E[Y(0,m(0))]$: the direct effect, defined as the average effect of treatment in the absence of the mediator $E[Y(1,m(0))] - E[Y(0,m(0))]$; and the indirect effect, defined as the difference between the causal effects with and without the mediator $E[Y(1,m(1))] - E[Y(1,m(0))]$. Indeed, the most important parts, as shown in the decomposition, are the **conditional response function**: $E[Y(a,m)]$ and we let the target mediation estimator $\psi^*(P^*)$ to be that. From our statistical model, in which we have the data formed in a tuple: $Z = (X,A,M,Y)$, we could yield the conditional expectation term, $\psi(P) = E[Y|A = a, M = m]$. Similarly, we need specific assumptions that allow us to equalize $\psi^*(P^*)$ and $\psi(P)$. The details of the technique are discussed in Chapter 4 of this thesis.

Since the content of causal inference (the first scenario above) runs throughout the thesis, we specifically analyze it in this chapter. **Causal inference**, broadly speaking, is a process that determines an independent effect of a particular object (the treatment) on another (the outcome), and it is usually contained in a larger system. For instance, when Galileo experimented on the Leaning Tower of Pisa, he isolated the independent effect of the weights of the two balls, observing that the heavier ball and the lighter ball fell on the ground simultaneously, and therefore inferred that the weights (masses) of the balls (the treatment) have nothing to do with the gravity (the outcome) of the two balls. The strategy to isolate the treatment effect is a **controlled experiment**, in which we are assured that the only difference between the two is the object we deem the treatment. However, in many scientific disciplines, we can not artificially manipulate and completely isolate the treatment[7].

[7]Indeed, in Galileo's experiment, the masses could not be the only difference between the two balls, as either the materials (densities) or the size must be different because the densities and the volumes determine the masses (in Galileo's original experiment he ensured the two balls were made of iron). Therefore, further experiment designs are needed: controlling the size of the two balls and controlling the type of materials of the two balls.

Moreover, due to the requirement of repeatability in modern science, we usually observe the *group-level* differences between treated and untreated instead of the two individuals (balls). Based on this, experiment designers need techniques such as **randomization** or **blind control** for treatment assignment [8].

Experiments are becoming increasingly common in social science studies when researchers are interested in the causal relationship, but most research still relies on **observational data** to infer causal relationships. The observational data do not have the RCT design: they do not randomize the samples to the treatment and control groups. They are not designed to isolate the effects of the treatment variable. Therefore, researchers need to use statistical techniques to transfer the observational cases to approximate the RCT process, and therefore, causal inference with observational data is a **pseudo-RCT**. Our thesis mainly addresses statistical ideas on causal inference with observational data.

In this thesis, causal inference with observational data is the process using the observable estimator from the random samples $\psi(\mathbb{P}_n)$ to estimate the causal estimand $\psi^*(P^*)$ via the statistical estimand $\psi(P)$. In empirical studies, the estimation functions for the estimand (population distribution) (estimandor) and for the observables (sample distribution) are always the same (for instance, the expectation $\psi(P) = E_P[Z]$ and $\hat{\psi}(\mathbb{P}_n) = E_{\mathbb{P}_n}[Z_i]$ have the same functional form), and the only difference between the two functions here is in the measurement choice. Thus, below, we will use $\hat{\psi}$ and $\psi(\mathbb{P}_n)$ interchangeably to represent the estimator from the empirical dataset, we use ψ and $\psi(P)$ interchangeably to represent the estimator (estimandor) from the statistical estimand.

Statisticians prefer $\hat{\psi}$ as an unbiased estimator of ψ^* and regard the divergence between the two terms as error terms. Moreover, we can decompose the error term into the statis-

[8] Randomization and blind control might refer to different techniques. Consider the example of Pavlov's classic conditioning experiment with the dog, which is a blind control instead of a randomized control. Experiments with randomized group assignment are commonly called **Randomized Controlled Trials (RCT)**

tical error, which is the divergence between the estimator and the statistical estimand, and the causal error, which is the divergence between the statistical estimand and the causal estimand:

$$\hat{\psi} - \psi^* = \underbrace{(\hat{\psi} - \psi)}_{\text{statistical error}} + \underbrace{(\psi - \psi^*)}_{\text{causal error}} \tag{1.I.2}$$

For the statistical error, we could further decompose it into the statistical variance and the statistical bias:

$$\hat{\psi} - \psi = \underbrace{(\hat{\psi} - E[\hat{\psi}])}_{\text{statistical variance}} + \underbrace{(E[\hat{\psi}] - \psi)}_{\text{statistical bias}} \tag{1.I.3}$$

Analyzing the error terms is the core of the causal analysis, as in substantive work, especially from observational data, our estimator at most times could not be satisfied as unbiased, yielding the causal estimand (and even in the RCT, further assumptions may be required). In Section II, we discuss the causal error, and in Section III, we discuss the statistical error.

II. Analyzing the Causal Error

A. Unbiased estimators and assumptions

The causal error is defined as the divergence between the statistical estimand $\psi(P)$ (simply ψ) and the causal estimand $\psi^*(P^*)$ (simply ψ^*). We cannot directly apply the statistical estimand as the causal one because of the **fundamental problems of causal inference** that we do not observe the outcome under different treatment conditions simultaneously, as each individual only receives one identifiable treatment. In **Neyman-Rubin's (NR)** causal framework (Neyman 1990; Rubin 1990: 1974), the outcomes under different treatment statuses from the causal DGP ($Y(a)$ are the **potential outcomes** (Holland 1986). In the statistical data frame, the existing outcome is conditioned on the assigned value of the treatment ($Y|A = a$);

nevertheless, we could not directly obtain the outcomes conditioned on the treatment assigned to other values. Therefore, the unobservable outcomes are called **counterfactuals**.

NR's counterfactual framework is not the only way to understand the relationship between the causal and statistical estimands. Computer scientist Judea Pearl (2009) created a framework called **"do-calculus" (DoC)**(Heckman and Pinto 2024). From Pearl's perspective, the outcome is deterministic if the treatment action has been triggered and the mapping rule from the treatment to the outcome is determined: if the treatment is $A = a$, then $a \rightarrow Y(a)$; if $A = a^*$, then $a^* \rightarrow Y(a^*)$[9]. Based on the mapping symbol, Pearl developed the graphical expression for causal analysis called the **Directed Acyclic Graphs (DAG)**, and it is commonly used in structural equation models and causal mediation analysis.

Indeed, both NR's counterfactual and Pearl's DoC framework require additional assumptions to make $\psi^*_a(P^*) = E_{P^*}[Y(a)]$ and $\psi_a(P) = E_P[Y|A = a]$ equivalent [10]. In this whole thesis, we consider the binary treatment. We suppose that the probability of being assigned to treatment and control is a positive number between 0 and 1:

Assumption 1.II.1 (Positivity Assumption) [11] *The probability of being assigned as treatment and control is a positive number between 0 and 1:*

$$P(A = 1) \in (0, 1); P(A = 0) \in (0, 1)$$

[9] In most part of this thesis, the treatment is a dichotomous variable with two groups: the treated group and the control group. In Chapter 5, we demonstrate a multinomial treatment. However, it is worth noting that the method introduced in this thesis (the efficient-influence-function based doubly robust estimator) never restricts the distribution of the treatment, the outcome, or any other relevant variables in the model (for instance, the mediator). This is because we can always model the corresponding nuisance functions for the specific variables, and how they are distributed.

[10] In Pearl's denotation, the conditional probability can be denoted as $E_P[Y \mid do(A = a)]$. In the whole thesis, I randomly adopt the denotation and use the conditional expectation denotation for the statistical estimand.

[11] In some literature, this assumption is also called the overlap assumption (Heckman et al. 1998).

Where A is the treatment.

And suppose that in the "omniscient" causal data frame $Z^* = (X, A, Y(1), Y(0))$, we could observe the following conditional expectations: $E[Y(1)|A = 1]$, $E[Y(1)|A = 0]$,$E[Y(0)|A = 1]$, $E[Y(0)|A = 0]$, while in the statistical data frame $Z = (X, A, Y)$ we could only observe $E[Y|A = 1]$ and $E[Y|A = 0]$. With the following assumption, we could link between $E[Y|A = 1]$ and $E[Y(1)|A = 1]$, and between $E[Y|A = 0]$ and $E[Y(0)|A = 0]$:

Assumption 1.II.2 (Consistency Assumption) [12] *The potential outcome under treatment received is the same as the observed outcome. That is,*

$$Y = Y(A)$$

where Y is the observed outcome, $Y(A)$ is the potential outcome, and A is the treatment.

Indeed, in the DoC framework, consistency has been implied since the mapping function from the treatment execution to the outcome is defined. Also, it is worth noting that when we use observational data to infer the causal estimand, a consistent assumption needs to be held on the individual level: $Y_i = Y_i(A_i)$. Furthermore, since we have Assumption 1.I.1 for the IID samples, we could infer that there's no interference among individuals: the treatment assigned to one observational sample does not affect the outcome of the others. Consistency and no interference assumptions on the individual observational level are collectively known as the **Stable Treatment Unit Value Assumption**, or SUTVA(Morgan and Winship 2015; Imbens and Rubin 2015).

If we have established the relationship between the causal and statistical data frames, that $E_P[Y|A = a] = E_{P^*}[Y(a)|A = a]$. We suppose the proportion assigned to the treatment

[12] This assumption in some literature is also called the faithfulness assumption.

group is ρ, and therefore, $E[Y(1)] = \rho E[Y(1)|A=1] + (1-\rho)E[Y(1)|A=0]$ and $E[Y(0)] = \rho E[Y(0)|A=1] + (1-\rho)E[Y(0)|A=0]$ We may calculate the difference between the causal average treatment effect and the statistical average treatment effect with a simple calculation:

$$\begin{aligned}
&\underbrace{\big(E[Y(1)|A=1] - E[Y(0)|A=0]\big)}_{\text{statistical average treatment effect}} - \underbrace{\big(E[Y(1)] - E[Y(0)]\big)}_{\text{causal average treatment effect}} \\
&= (E[Y(1)|A=1] - E[Y(0)|A=0]) \\
&\quad - \Big[\big(\rho E[Y(1)|A=1] + (1-\rho)E[Y(1)|A=0]\big) \\
&\quad - \big(\rho E[Y(0)|A=1] + (1-\rho)E[Y(0)|A=0]\big)\Big] \\
&= \underbrace{E[Y(0)|A=1] - E[Y(0)|A=0]}_{\text{difference in baseline}} + \underbrace{(1-\rho)(\delta_1 - \delta_0)}_{\text{heterogeneous treatment effect}}
\end{aligned} \tag{1.II.4}$$

where

$$\delta_1 = E[Y(1)|A=1] - E[Y(0)|A=1] \quad \text{and} \quad \delta_0 = E[Y(1)|A=0] - E[Y(0)|A=0].$$

PROOF 1.II.1 *Let* $E[Y(1)|A=1] = \alpha_1$, $E[Y(1)|A=0] = \alpha_2$, $E[Y(0)|A=1] = \alpha_3$, *and* $E[Y(0)|A=0] = \alpha_4$. *Therefore, the left side of the equation is:*

$$(\alpha_1 - \alpha_4) - \big[\rho\alpha_1 + (1-\rho)\alpha_2 - \rho\alpha_3 - (1-\rho)\alpha_4\big].$$

Simplifying this, we have:

$$\begin{aligned}
&(\alpha_1 - \alpha_4) - \big[\rho\alpha_1 + (1-\rho)\alpha_2 - \rho\alpha_3 - (1-\rho)\alpha_4\big] \\
&\qquad = (\alpha_1 - \alpha_4) - \rho\alpha_1 - (1-\rho)\alpha_2 + \rho\alpha_3 + (1-\rho)\alpha_4 \\
&\qquad = \alpha_1 - \alpha_4 - \rho\alpha_1 - (1-\rho)\alpha_2 + \rho\alpha_3 + (1-\rho)\alpha_4 \\
&\qquad = \underbrace{(\alpha_3 - \alpha_4)}_{\textit{difference in baseline}} + (1-\rho)\ \underbrace{(\alpha_1 - \alpha_3) - (\alpha_2 - \alpha_4)}_{\textit{heterogeneous treatment effect}}.
\end{aligned}$$

Rewriting this back in terms of the original expectations, we have the right side of the equation.

Equation 1.II.4 reveals the two origins of bias in causal inference when using the statistical estimator to infer the causal estimand: the baseline difference, or the selection bias, which is the pre-treatment divergence when grouping individuals to the treatment and control groups; and the heterogeneous treatment effect between the treatment and control group, which is the post-treatment divergence between the treatment and the control group. For example, consider evaluating the impact of a training program on workers' productivity. Initially, we measure their productivity levels before the training. Next, we divide the workers into two groups: a treatment group that receives the training and a control group that does not. After a certain period, we measure the change in productivity in both groups to assess the effect of the training program. The bias in this measurement comes from two sources: firstly, a pre-training bias, where workers in the treatment group might have different initial productivity levels compared to those in the control group; and secondly, a post-training bias, where workers in the treatment group might experience a greater improvement in productivity than those in the control group, even if they had all received the training.

Therefore, to eliminate the potential pre- and post-treatment bias, we need further assumptions for identification. Since the pre-treatment selection and post-treatment heterogeneity can be attributed to the non-randomization in the treatment assignment, we have the ignorability/ unconfoundedness assumption:

Assumption 1.II.3 (Ignorability/Unconfoundedness Assumption) *The treatment assignment A is independent of the potential outcomes $Y(1)$ and $Y(0)$:*

$$Y(1), Y(0) \perp\!\!\!\perp A$$[13]

[13]The symbol $\perp\!\!\!\perp$ denotes statistical independence. Independence implies zero covariance when the rele-

Since the potential outcomes are independent of the treatment assignment, we can infer that $E[Y(1)|A=1] = E[Y(1)|A=0]$ and $E[Y(0)|A=1] = E[Y(0)|A=0]$. Therefore, under the unconfoundedness assumption, the average pre-treatment baseline difference is 0. Moreover, the average treatment effect among the treated equals the corresponding effect among the controls, so $\delta_1 = \delta_0$ and the post-treatment heterogeneity term in Equation 1.II.4 is also 0[14]. In this sense, with Assumptions 1.II.1,1.II.2, and 1.II.3, we could finally conclude that the statistical average treatment effect is an unbiased estimand on the causal average treatment effect: $\psi(P) = \psi^*(P^*)$.

In most circumstances, indeed, we may find a set of covariates X in the statistical model that are correlated with both A and Y, violating the ignorability/unconfoundedness assumption. Thus, we may consider the treatment assignment conditioned on X as a **pseudo-randomization**. We update Assumptions 1.II.1 and 1.II.3 to make them include the conditioning on the covariates (the consistency hypothesis remains unchanged):

Assumption 1.II.4 (Causal Inference Assumptions) *Suppose a statistical DGP $Z = (X, A, Y)$, in which X denotes the covariates, A denotes the treatment, and Y denotes the outcome. To make the statistical estimand $\psi(P) = E[E_X[Y|A=1,X]] - E[E_X[Y|A=0,X]]$ equivalent to the causal estimand $\psi^*(P^*) = E[Y(1)] - E[Y(0)]$ from the causal DGP $Z^* = (X, A, Y(1), Y(0))$ (where $Y(1)$, $Y(0)$ denote the potential outcomes under treatment and control, respectively), we need the following hypotheses:*

vant moments exist, but zero covariance alone does not imply independence. Mean independence, written for example as $E[a \mid b] = E[a]$, is also weaker than full independence.

[14] This does not mean that individual-level or covariate-defined treatment-effect heterogeneity disappears. It only means that the between-assignment-group difference in average treatment effects, $\delta_1 - \delta_0$, is removed under this marginal independence assumption. In causal inference, HTE is usually formalized as a conditional contrast such as $E[Y(1) - Y(0) \mid X]$.

1. *Positivity: the probability to be assigned to treatment and control group conditioned on the covariates, is a positive number between 0 and 1:*

$$P(A=1|X) \in (0,1); P(A=0|X) \in (0,1)$$

2. *Consistency: the potential outcome under the treatment received is the same as the observed outcome:*

$$Y = Y(A)$$

3. *Unconfoundedness: conditional on a set of observed covariates X, the potential outcomes $Y(1)$ and $Y(0)$ are independent of the treatment assignment A:*

$$\{Y(1), Y(0)\} \perp\!\!\!\perp A|X$$

Under this framework, Assumption 1.II.4 provides sufficient conditions for the statistical estimand on the average treatment effect to be equivalent to the causal estimand on the average treatment effect (Imbens and Rubin 2015; Schuler and van der Laan 2024; Hernán and Robins 2020; Heckman and Pinto 2024). To simplify, consider our target parameter is the potential outcome $\psi_a^* = E[Y(a)]$, and our statistical estimand is $\psi_a = E[E_X[Y|A=a,X]]$ (as $E_X[Y(a)|X] = \int y f_{Y(a)|X}(y|X)dy$), we have:

$$\begin{aligned}
\psi_a^* &= E[Y(a)] \\
&= E[E_X[Y(a)|X]] \text{ (conditional expectation)} \\
&= E[E_X[Y(a)|A,X]] \text{ (positivity and unconfoundedness)} \\
&= E[E_X[Y|A=a,X]] \text{ (consistency)} \\
&= \psi_a
\end{aligned} \tag{1.II.5}$$

Meanwhile, Equation 1.II.5 can be also written as:

$$
\begin{aligned}
\psi_a^* &= E[Y(a)] \\
&= E[E[Y(a)|X]] \text{ (conditional expectation)} \\
&= E\Big[E[Y(a)|X]\frac{E[\mathbb{1}_A|X]}{E[\mathbb{1}_A|X]}\Big] \text{ (positivity; } \mathbb{1}_A = 1 \text{ if } A = a; 0 \text{ otherwise)} \\
&= E\Big[\frac{E[Y(a)\mathbb{1}_A \mid X]}{E[\mathbb{1}_A|X]}\Big] \text{ (unconfoundedness)} \\
&= E\Big[\frac{E[Y\mathbb{1}_A \mid X]}{P[A = a \mid X]}\Big] \text{ (consistency)} \\
&= E\Big[\frac{E[Y\mathbb{1}_A \mid X]}{\pi_a(X)}\Big] \text{ (define } \pi_a(X) = P(A = a|X)) \\
&= E\Big[\frac{Y\mathbb{1}_A}{\pi_a(X)}\Big] \text{ (reverse conditional expectation)}
\end{aligned}
\tag{1.II.6}
$$

Equations 1.II.5 and 1.II.6 illustrate that we could infer the potential outcome ψ_a^* with either the conditional expectation ψ_a or the propensity score function for the treatment $\pi(a)$ as the unbiased estimators. The results are the foundation of what we refer to as the **doubly robust/debiased estimation** later in this chapter.

B. Methods Addressing Violations on Causal Assumptions

In most social science research scenarios with observational (survey) data, finding an unbiased causal estimator is challenging, as the three conditions in Assumption 1.II.4 are not always satisfied, especially the positivity and the unconfoundedness assumptions. In practical research, violating the positivity assumption is common if our treatment involves policy/reform/law enforcement that affects all our research objects. For instance, suppose our target is to measure how the rules on sports gambling may affect suicide risks for the residents of a state. Since the law affects everyone in the state, it violates the positivity assumption as $P(A) = 1$ (everyone in the state is grouped as treated).

Violating the unconfoundedness assumption is another common matter. The unconfoundedness assumption states that the outcome $(Y(1), Y(0))$ is independent of the treatment A given a set of covariates X.In econometrics terms, we call treatment A the endogenous treatment, while covariates X are exogenous covariates. In practical research, social scientists are concerned that exogenous variables may not be fully captured, and thus, confounding variables may affect the relationship between the treatment and the outcome. For instance, if the treatment of interest is one's occupational mobility and the outcome is fertility, and we set the exogenous variables to include their age, marital status, pre-treatment earnings, and tenure, but we omit to include their education. Since education will affect both occupational attainment and fertility, omitting the variable will lead to the violation of unconfoundedness as occupational mobility A is still correlated with $Y(1)$ and $Y(0)$ through education even after conditioning on the exogenous variables X.

Violations of positivity or unconfoundedness break nonparametric identification of $E[Y(a)]$ from (Y, A, X). With positivity failures, identification might be recovered for a common-support subpopulation or impose extrapolation assumptions (model-dependent and fragile). With unconfoundedness failures, functional-form assumptions alone cannot identify the effect; alternative designs are needed, or we can settle for partial identification and sensitivity analysis.

Below, we introduce some commonly used methods for conducting causal inference with observational data. The methods can be grouped into three layers: identification designs (assumptions and target estimand), modeling devices (representations that make assumptions plausible), and estimation strategies (algorithms to estimate the identified estimand if violations on the assumptions are known).

B.1 Identification Designs

Methods concerning identification designs usually make substantive *assumptions* that enable a causal effect to be learned from observational data, and set the target estimand for the assumptions to identify. We introduce three methods here: the difference in differences (DID) method, which sets the parallel trends assumption and identifies the average treatment effect for the treated group over time as the causal estimand; the regression discontinuity (RD) design, which set the continuity assumption of potential outcomes at the threshold and identify the causal estimand as the treatment effect at the cutoff point; and the instrumental variable (IV) method, which relies on the relevance, exogeneity, and monotonicity assumption and identify the causal estimand as the local average treatment effect for the compliers.

- **Difference in Differences**

  A common technique to address causal inference when standard cross-sectional assumptions of positivity and unconfoundedness are strained is the **difference in difference (DID)** method, which leverages time and comparison groups to identify causal effects (Card and Krueger 1994; Abadie 2005; Angrist and Pischke 2009). When a policy applies to everyone within a treatment unit (e.g., a state-level law that $A = 1$ for all after the reform), cross-sectional positivity fails. DID circumvents the problem and creates variation by comparing treated units to units that did not adopt the policy, creating treatment variation across group and time cells. Meanwhile, DID also relaxes cross-sectional unconfoundedness by permitting selection on time-invariant unobservables, which removes treated-control differences by differencing, so that identification does not require $A \perp\!\!\!\perp (Y(1), Y(0)) \mid X$ at each time point. Therefore, identifica-

tion then relies on the parallel-trends restriction, conditioned on observed covariates X:

Assumption 1.II.5 (Parallel Trends for DiD) *In the absence of treatment, the average outcome change for treated and control groups would be equal:*

$$E\big[Y_{t_1}(0) - Y_{t_0}(0) \mid D = 1\big] = E\big[Y_{t_1}(0) - Y_{t_0}(0) \mid D = 0\big],$$

where D indicates the treated group[15]*, and t_0 and t_1 are pre- and post-periods. A conditional version requires equality within X:*

$$E\big[Y_{t_1}(0) - Y_{t_0}(0) \mid D = 1, X\big] = E\big[Y_{t_1}(0) - Y_{t_0}(0) \mid D = 0, X\big].$$

With Assumption 1.II.5 (together with the causal inference assumptions), the causal estimand, in this regard, should be the two-way difference between the observable expected post-treatment outcome and the counterfactual post-treatment outcome (assuming the treatment was not received):

$$\big\{E[Y_{t_1} \mid D = 1] - E[Y_{t_0} \mid D = 1]\big\} - \big\{E[Y_{t_1} \mid D = 0] - E[Y_{t_0} \mid D = 0]\big\},$$

It is worth noting that the causal estimand here is, by default, the average treatment effect for the treated (ATT) because the parallel assumption is used only to impute counterfactual $Y(0)$ for the treated group, and we never impute the treated counterfactual $Y(1)$ for the control group. If we would like to recover the population-level

[15] Indeed, we use D to label group membership (treated vs. control, which is time-invariant, while A usually denotes the actual treatment status at a given time (changes over time in DID). So in a two-period DID, everyone in the treated group is untreated before and treated after. Hence, we could not condition on A: at t_0, $A = 1$ is for nobody and at t_1, $A = 0$ is true for the control group but not for the treated group. This is why $E[Y_{t1}(0) - Y_{t0}(0) \mid A = 1]$ is not defined due to the positivity violation.

average treatment effect (ATE), further homogeneity assumptions are required.

Further, we can also address the conditioned ATT by averaging group-specific contrasts over X. By doing so, we address both cross-sectional positivity and mitigate unconfoundedness by differencing out time-invariant confounding. However, DID is still sensitive to time-varying confounders that affect treated and control groups differently, which would lead to the violation of Assumption 1.II.5.

DID framework can be extended to the **Difference in difference in differences (DDD)** (Gruber 1994; Wolfers 2006; Meyer et al. 1995), if some institutional features suggest heterogeneous trends across the third dimension. We could apply the DDD framework to net out group-specific trend heterogeneity. For instance, consider the possible unparalleled trends in urban and rural areas, and we might derive the DDD:

$$\delta_{\text{DDD}} = \left(\Delta Y_{\text{treated, urban}} - \Delta Y_{\text{control, urban}}\right) - \left(\Delta Y_{\text{treated, rural}} - \Delta Y_{\text{control, rural}}\right),$$

which relaxes reliance on a single parallel assumption.

- **Regression Discontinuity**

Like DID, **Regression Discontinuity (RD)** can address violations of both positivity and unconfoundedness under additional design assumptions. RD exploits a known assignment rule based on a running variable with a cutoff, generating quasi-experimental variation at the threshold (Thistlethwaite and Campbell 1960; Hahn et al. 2001; Imbens and Lemieux 2008; Lee and Lemieux 2010). When a policy deterministically treats everyone above a threshold (so cross-sectional positivity fails), RD recovers

identification by contrasting outcomes just below and just above the cutoff.

In practice, there are two types of RD design: the first is the sharp RD, which suggests the rule is enforced perfectly at the cutoff. For instance, everyone above the threshold gets the treatment, while everyone below does not. Meanwhile, we may also encounter the fuzzy RD, which suggests that the actual treatment is not perfectly enforced, and it only changes people's eligibility at the cutoff. For instance, on the population level, we may see a change in the probability distributions, but not everyone is strictly treated above the threshold.

Suppose the target of our research is to evaluate how a tax change affects consumption for those with annual income near \$100,000. If the policy increases the tax rate from 25% to 30% for $X \geq 100{,}000$ (with X = income), everyone above the threshold is assigned to the new regime, violating the positivity assumption. RD uses observations *just* below and *just* above \$100,000 to form a local contrast that is as-if randomized at the cutoff. Hence, we could have the continuity assumption:

Assumption 1.II.6 (Continuity Assumption for Regression Discontinuity) *The expected potential outcomes* $E[Y(0) \mid X = x]$ *and* $E[Y(1) \mid X = x]$ *are continuous at the threshold* c*:*

$$\lim_{x \to c^-} E[Y(0) \mid X = x] = \lim_{x \to c^+} E[Y(0) \mid X = x], \qquad \lim_{x \to c^-} E[Y(1) \mid X = x] = \lim_{x \to c^+} E[Y(1) \mid X = x].$$

Here $Y(1)$ *and* $Y(0)$ *denote potential outcomes under treatment and control, and* X *is the running variable.*

Under Assumption 1.II.6 and no manipulation of X at c, the limits of observed outcomes identify the local potential-outcome means at the cutoff:

$$\lim_{x \to c^-} E[Y \mid X = x] = \lim_{x \to c} E[Y(0) \mid X = x], \qquad \lim_{x \to c^+} E[Y \mid X = x] = \lim_{x \to c} E[Y(1) \mid X = x],$$

Under the sharp RD scenario, the causal estimand can be identified as:

$$\tau = \lim_{x \to c^+} E[Y \mid X = x] - \lim_{x \to c^-} E[Y \mid X = x].$$

The causal estimand can be viewed as a local average treatment effect at the cutoff c (see the section below for the details). Moreover, under the fuzzy scenario in which the cutoff is not strictly enforced, the effect is given by the ratio between the jump in the outcome and the jump in the treatment probability at the cutoff, which identifies the local average treatment effect at c for compliers (similar to the IV-identification below):

$$\tau = \frac{\lim_{x \to c^+} E[Y \mid X = x] - \lim_{x \to c^-} E[Y \mid X = x]}{\lim_{x \to c^+} P(A = 1 \mid X = x) - \lim_{x \to c^-} P(A = 1 \mid X = x)},$$

which leverages the discontinuity in treatment probability to address positivity locally and replaces cross-sectional unconfoundedness with continuity and no-manipulation at c.

In practice, researchers might choose a bandwidth (e.g., \$5,000 around the \$100,000 cutoff) and compare average consumption for $X \in [95{,}000, 100{,}000)$ versus $X \in [100{,}000, 105{,}000]$. The identifying assumptions further require that the density of X and the distribution of covariates vary smoothly at c (ruling out sorting/bunching at the threshold). Under these conditions, any jump in $E[Y \mid X]$ at c is attributed to the treatment, yielding

a credible local causal effect even when global overlap fails and unobservables differ across units. [16]

- **Instrument Variables (IV)**

  Instrumental variables (IV) methods address violations of causal assumptions by exploiting exogenous variation to handle endogeneity, identifying the **local average treatment effect (LATE)** for compliers under additional exogeneity, exclusion, and monotonicity assumptions (Imbens and Angrist 1994).

  A comprehensive review from the biostatistical and clinical perspective of IV methods can be found in Baker et al. 2016. Specifically, we illustrate a binary IV scenario for *simplification*. In an empirical study, suppose we have assigned individuals to treatment and control groups to measure the post-treatment divergence as the treatment effect. We have discussed that the main concern in this setting is the action of "assignment"—whether it is random. With random assignment, as we have discussed,

$$\begin{aligned}\hat{\psi} &= E[Y_i \mid A_i = 1] - E[Y_i \mid A_i = 0] \\ &= \psi \;=\; E[Y \mid A = 1] - E[Y \mid A = 0] \\ &= \psi^* \;=\; E[Y(1)] - E[Y(0)],\end{aligned}$$

  as the fundamental link between statistical estimands and causal effects.

  Now consider that some "naughty" individuals do not follow the group assignment, creating noncompliance. Let X denote the binary *assignment* (instrument), and A

[16] For fuzzy RD, one also requires a nonzero first stage and monotonicity, which we will discuss later in the instrumental variable part.

denote whether they *receive* treatment. We then have four observed groups: assigned treatment, received treatment ($X_i = 1, A_i = 1$); assigned treatment, received control ($X_i = 1, A_i = 0$); assigned control, received treatment ($X_i = 0, A_i = 1$); and assigned control, received control ($X_i = 0, A_i = 0$). Covariates may influence the "naughty" decisions, but the *instrument* must affect the outcome only through treatment.[17]

Assumption 1.II.7 (Instrumental Variable Assumptions) *Let X be the instrument, and it should satisfy:*

(a) Instrument relevance: the instrument is correlated with the treatment,

$$\text{cov}(X, A) \neq 0.$$

(b) Instrument exogeneity/independence: the instrument is uncorrelated with the outcome error term (as-good-as-random with respect to potential outcomes) [18]*,*

$$\text{cov}(X, \epsilon) = 0.$$

(c) Exclusion restriction: the instrument affects the outcome only through the treatment (no direct effect of X on Y).

For each of the four observed cells, we can classify underlying compliance types. Individuals who would take treatment regardless of assignment are **always-takers** (they appear in $(1,1)$ and $(0,1)$); those who would never take treatment are **never-takers**

[17] This is not to say the "naughty decisions" are unaffected by other covariates—for instance, past medical history. What matters for IV is that any effect of assignment X on Y operates only through A (the exclusion restriction) and that X is otherwise as-good-as-random with respect to the outcome error.

[18] A common formulation of (b) is that the instrument is independent of potential outcomes (possibly conditional on covariates).

(they appear in $(1,0)$ and $(0,0)$); those who follow assignment are **compliers** (they appear in $(1,1)$ and $(0,0)$); and those who do the opposite are **defiers** (they appear in $(1,0)$ and $(0,1)$).

We cannot identify causal effects for always-takers and never-takers with the instrument, because their treatment does not vary with X. Further, we need to assume that there are no defiers in the experiment, as we need the relationship between the treatment and the instrument to be monotonic:

Assumption 1.II.8 (Monotonicity Assumption) *There are no defiers: the instrument moves treatment in one direction (it never decreases treatment for some while increasing it for others).*

Under Assumptions 1.II.7–1.II.8, thus, the treatment effect on the outcome for the compliers can be reliably estimated using the instrument, provided the exogeneity and monotonicity assumptions hold. Since adding the instrument meets all the assumptions necessary for causal inference, particularly unconfoundedness, the outcome is now independent of the treatment given the instrument. The Local Average Treatment Effect (LATE) for the compliers is determined by dividing the covariance between the outcome and the instrument by the covariance between the treatment and the instrument.

$$\text{LATE} = \frac{\text{cov}(X,Y)}{\text{cov}(X,A)} = \frac{E[Y \mid X=1] - E[Y \mid X=0]}{E[A \mid X=1] - E[A \mid X=0]}. \tag{1.II.7}$$

We refer to the covariance ratio more generally as the IV estimand.[19] The pseudo-randomization is thus achieved via the instrument for the compliers.

[19] IV first arose in econometrics to address identification in **simultaneous equations** where outcomes are

A classic example in social science is the causal effect of military service on future earnings (Angrist 1990). Because voluntary enlistment is endogenous, individuals who choose to enlist may differ from those who do not, so a direct comparison of enlistees and non-enlistees is biased. To address this, Angrist uses the Vietnam War draft lottery as an instrument: lottery numbers are randomly assigned, and draft eligibility shifts the likelihood of military service. Individuals whose service status is determined by eligibility (serve if eligible, do not if ineligible) are compliers. Among compliers, the study estimates that military service lowers subsequent earnings, i.e., a negative local average treatment effect (LATE).

Does this study have a good research design? First, we need to point out the external validity of the causality: the analysis focuses on men born 1950–1953 and draft years 1970–1972, so the estimate pertains most directly to that complier population. Second, while draft evasion creates noncompliance, IV accommodates this provided the standard assumptions hold—relevance, independence (exogeneity), exclusion, and monotonicity (no defiers). Note that exclusion could be threatened if draft eligibility affects earnings through channels other than military service (e.g., schooling deferments or legal consequences). Under these assumptions, the design identifies the complier LATE via the Wald estimand, rather than a simple difference between those

equilibria of structural relations. For example, with price P and quantity Q, the reduced forms

$$\begin{cases} P = \pi_1 X_s + \pi_2 X_d + \nu_p, \\ Q = \theta_1 X_s + \theta_2 X_d + \nu_q, \end{cases}$$

are functions of exogenous shifters X_s (supply) and X_d (demand). One can recover structural slopes (e.g., the demand slope $\alpha_d = \theta_2/\pi_2$) if valid instruments shift one side but not the other (Haavelmo 1944; Stock and Trebbi 2003; Pearl 2015). Here, "exogeneity" means that the instruments are uncorrelated with the structural error.

drafted-and-served and those not-drafted-and-not-served.

In today's social science, "finding an appropriate instrument is rather an art than a science." Researchers often exploit naturally random assignments (Angrist and Krueger 2001), geographic/spatial variation (Card 1999), weather (Dell et al. 2009), policy changes (Angrist and Lavy 1999), economic shocks (Autor et al. 2013), and even demographic, biological, or health variables (Oreopoulos 2006; Fletcher and Wolfe 2009) as instruments.[20]

It is worth noting that candidate instruments may be **weak** (Stock et al. 2002). According to Assumption 1.II.7, a valid instrument must satisfy relevance and exogeneity. A weak instrument is one with weak relevance, as it barely shifts treatment. The consequence is **weak identification**: the IV estimand becomes unstable, and conventional two-stage least squares (2SLS) inference can be severely *biased* and nonnormal.

To see the intuition, consider the binary-IV (Wald) estimand for the complier effect:

$$\widehat{\text{LATE}} = \frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[A \mid Z = 1] - E[A \mid Z = 0]}.$$

When the first-stage difference $E[A \mid Z = 1] - E[A \mid Z = 0]$ is near zero, the denominator is small, so the variance of the ratio explodes, and finite-sample distributions become highly nonnormal. In linear models with one endogenous regressor, 2SLS equals the sample analog of $\beta_{\text{IV}} = \frac{\text{cov}(Z,Y)}{\text{cov}(Z,A)}$, so as $\text{cov}(Z, A) \to 0$ the estimator becomes

[20] When using demographic/biological/health instruments, questions often concern policy effects on socioeconomic outcomes; with policy/economic shocks, questions often concern relationships among socioeconomic variables.

unstable. In finite samples, 2SLS estimates drift toward OLS, and standard t-tests can over-reject (Bound et al. 1995; Staiger and Stock 1997; Stock and Yogo 2005).

Practically, researchers often (i) report first-stage strength (e.g., first-stage F); as a rule of thumb, $F > 10$ indicates acceptable strength in many single-endogenous-regressor settings; (ii) use heteroskedasticity-robust diagnostics (e.g., Kleibergen–Paap rk F) and, when multiple instruments are used, Stock–Yogo critical values; and (iii) rely on weak-IV-robust inference such as Anderson–Rubin or conditional likelihood ratio confidence sets, or use less biased estimators such as LIML/Fuller.

Formally, in a linear setup with one endogenous regressor,

$$Y = \beta A + u, \qquad A = \pi Z + v,$$

The IV/2SLS estimator is the Wald ratio we elaborated above. With weak π (small first-stage), sampling distributions are nonnormal, and the finite-sample bias of 2SLS toward OLS increases roughly with the inverse of the first-stage F (Bound et al. 1995; Staiger and Stock 1997; Stock and Yogo 2005). Thus, weak instruments primarily create weak identification and distorted inference, rather than reintroducing endogeneity.

B.2 Modeling Devices

Methods in this category do not identify effects on their own; rather, they are representational choices about the **data structure** that help make identification assumptions more plausible and reduce misspecification. We therefore discuss fixed-effects (FE) models as a modeling device: by imposing additive, time-invariant unit effects (and often common time effects), FE absorbs unobserved heterogeneity that is constant over time. On their

own, FE models do not identify causal effects; identification requires additional assumptions (e.g., strict exogeneity conditional on FE and controls, or a DID-style parallel trends condition). Unlike instrumental variables (IV), which is an identification strategy based on relevance and exclusion to handle time-varying unobservables and simultaneity, FE primarily addresses endogeneity arising from time-invariant omitted variables with a parametric modeling.

- **Fixed Effects (FE)**

  As discussed above, the essence of IV/LATE is to isolate exogenous variation in the endogenous treatment and obtain an unbiased, consistent estimator of a causal effect. Fixed effects (FE) can also help by removing time-invariant confounding through within-entity variation. Consider the panel-data model

  $$y_{ij} = \alpha_i + f(a_{ij}) + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2),$$

  where i indexes units (individuals) and j indexes time (or repeated observations). Here, α_i is a unit-specific term capturing all time-invariant heterogeneity, $f(a_{ij})$ is the effect of treatment a_{ij} on the outcome y_{ij}, and σ_y^2 is the outcome variance. We do not need a distributional assumption for α_i under FE; a multilevel (random-effects) alternative would instead posit

  $$\alpha_i \sim \mathcal{N}(\mu, \sigma_\alpha^2),$$

  cf. (Gelman and Hill 2006, ch. 12, p. 257).[21]

  Under FE, we assume mean independence and serial uncorrelated errors:

  $$E[\epsilon_{ij} \mid a_{i1}, \ldots, a_{iJ}, \alpha_i] = 0 \quad \text{and} \quad E[\epsilon_{it}\epsilon_{is}] = 0 \text{ for } t \neq s.$$

[21] In a random-intercept specification, one may write $y_{ij} = \mu + f(a_{ij}) + u_i + \epsilon_{ij}$ with $u_i \sim \mathcal{N}(0, \sigma_\alpha^2)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2)$. FE does not require a distribution for α_i.

With these assumptions, $f(a_{ij})$ is interpreted as a within-unit (over time) causal relationship under the absence of time-varying confounding. A convenient estimator is the within (demeaned) transformation:

$$y_{ij} - \bar{y}_i = f(a_{ij}) - \overline{f(a)}_i + \epsilon_{ij} - \bar{\epsilon}_i,$$

and, if $f(a) = \beta a$,

$$y_{ij} - \bar{y}_i = \beta(a_{ij} - \bar{a}_i) + (\epsilon_{ij} - \bar{\epsilon}_i),$$

where $\bar{y}_i$ and $\bar{a}_i$ are unit-specific means over j.

As can be seen from the above example, when j indexes time, FE implementations coincide with difference-in-differences (DID) in the canonical two-period, two-group setup; more generally, DID can be implemented via FE under a parallel-trends-type assumption. FE addresses time-invariant unobserved heterogeneity but does not by itself guarantee positivity, nor does it address time-varying confounding without additional structure or controls.

B.3 Estimation Strategies

Compared to the previous methods, researchers apply these methods with a clear idea of how the unconfoundedness bias could be eliminated with auxiliary variables (covariates). Under Equation 1.II.6, the methods can be categorized as propensity weights based (propensity-based matching and weighting) or conditional expectation based (outcome regression) (Rubin 1974; Rosenbaum and Rubin 1983).

- **Propensity Function based Matching and Weighting**

With covariates affecting group assignment known, matching or weighting is the intuitive choice to eliminate the bias due to unconfoundedness. If all confounders X

are observed and the positivity (overlap) condition holds, then as Equation 1.II.3 indicates, with the nuisance function $\pi_a(X) = P(A = a \mid X)$ we can identify the causal estimand $E[Y(a)]$. Traditionally, $\pi_a(X)$ is called the **propensity score**, as it measures the likelihood of treatment assignment given covariates X. The expression using the indicator divided by the propensity score, $\frac{\mathbb{1}(A=a)}{\pi_a(X)}$, yields the **inverse probability weighting (IPW)** estimator (Horvitz and Thompson 1952; Hernán et al. 2002). Therefore, when unconfoundedness and positivity hold (and regularity conditions apply), IPW delivers unbiased (and regular, asymptotically linear; see Section III) estimation of the target effect.

The IPW method achieves pseudo-randomization because reweighting creates a synthetic sample in which treatment is independent of X. A key design requirement is that X be measured **ex ante** the treatment assignment in longitudinal settings. Suppose we have a set of variables M that occur **ex post** treatment and affect the outcome. Then M should not be conditioned on when the target is the **total** effect of A on Y; including M would block part of the treatment effect. Instead, M can be used to decompose the total effect into **direct** and **indirect** components, which we discuss in the mediation section.[22]

Based on the propensity score function, weighting with inverse probability (propensity) provides a way to manipulate the treatment assignment to attain pseudo-randomization.

[22]The unconfoundedness assumption $(Y(1), Y(0)) \perp\!\!\!\perp A \mid X$ implies $\text{cov}(Y, A \mid X) = 0$ but does **not** imply $\text{cov}(Y, A) = 0$, $\text{cov}(X, A) = 0$, or $\text{cov}(Y, X) = 0$. Indeed, $\text{cov}(Y, A) \neq 0$ and $\text{cov}(X, A) \neq 0$ are typical in observational studies and motivate adjustment. Also, X are **confounders**, not instrumental variables: an instrument Z shifts A while affecting Y only through A and is independent of the potential outcomes.

Similarly, a matching method based on the propensity score function [23] can also achieve the effect of pseudo-randomization. This is the classic **propensity score matching (PSM)** method for causal inference (Rosenbaum and Rubin 1983; Hirano et al. 2003; Robins 1986; Rubin 1974). Suppose we could match the cases assigned to the treatment group and the control group with exactly the same propensity score and calculate the difference between the matched cases along the propensity score spectrum. We could then calculate the average, yielding the causal estimand for the average treatment effect based on the matching method.

It is worth noting that the propensity score method seems plausible theoretically, but when dealing with real observational data, researchers have to make a tradeoff between the quality of the matching algorithm and the selection of cases. The real problem is that we rely on the observational data to generate the estimator $\hat{\pi}_a$ to estimate π_a. We can imagine that when using the observational data to estimate the propensity function, for the individuals in the treatment group, the distribution is likely to be dense at the end towards 1 (if 1 indicates being assigned to the treatment group) and relatively sparse at the end towards 0, while for the control group individuals tend to be distributed more densely on the side of 0 and more sparse on the side of 1. Therefore, it is infeasible to have a one-on-one match between the individuals from the treatment group and the control group, with the exact same propensity value, and get everyone matched (see Figure 1.1 for the illustration). Researchers have to adopt methods either to allow the divergence (caliper) in propensity scores between the matched cases, to drop the unmatched cases, or a method that combines the two

[23] Propensity score function is not a score function that will be introduced in Section III. Therefore, to avoid confusion, in the following part of this chapter, we only call it "propensity function."

(for instance, set a threshold for nearest neighborhood matching and drop the cases beyond the threshold).

The problem of PSM, as King and Nielsen (2019) advocate, is not on the process of causal inference (from π_a to ψ_a^*, as Equation 1.II.3 suggests the unbiased process), nor on the process of statistical inference $\hat{\pi}_a^k$ to π_a (where k denotes the $k-$th method for propensity estimation, and all $\hat{\pi}_a^k$ can be an unbiased estimator for π_a if causal Assumptions 1.II.4 hold), it lies in the choice of $\hat{\pi}_a^k$, or the problem they call "model dependence": we rely on empirical observational data to simulate the DGP for propensity function, and further use the simulated function to predict the propensities for individuals from treatment and control groups (the dots in Figure 1.1), leaving cases unmatched due to uneven densities for the treatment and control groups. The bias, as King and Nielsen (2019) suggest, is a *subjective* bias originating from model choices, and the subjective choice of model somehow increases the imbalance [24], model dependence, and bias for the causal estimation.

Intrinsically, the problem with PSM, if any, arises from the failure of the specific propensity function to satisfy the unconfoundedness assumption. Suppose the unconfoundedness assumption holds, which specifies the independence between the outcome and the treatment under the covariates. In that case, we can imagine that the expected outcomes for the treated and the control should have a constant distance along the propensity score (the slope does not necessarily have to be zero though). Thus, as long as we have a balanced match along the propensity score, the choices of $\hat{\pi}_a^k$ would

[24] Imbalance refers to the deviations from the exact match.

be the same and an unbiased estimator for π_a (see the illustration of the middle and lower panel of Figure 1.1). However, as in empirical studies, there are remaining covariates which still influence the distribution of outcomes for treatment and control groups based on the propensity scores (the unconfoundedness assumption is not satisfied, for example, in the lower panel of Figure 1.1, we have different slopes for treatment and control groups along the x-axis), hence, the matching based on propensity scores results in bias.

In summary, weighting and matching methods achieve causal inference by identifying the propensity function: $\pi_a(X) = P[A = a|X]$ and expect the estimated $\hat{\pi}_a(X)$ to be an unbiased estimator for $\pi_a(X)$. In other words, the propensity function only matters for X and A[25], and it is almost unavoidable that we have statistical error between the estimated propensity score and the true value (similar to the omitted variable bias in regression analysis).

Thus, different models that researchers adopt will unavoidably generate different sample estimators. So, how do researchers claim that the causal effect they captured makes sense by adopting a specific weighting or matching method? We believe two things researchers need to claim before they describe their causal findings: one is the preconditions the models rely on: for instance, what covariates they have included and how they contribute to address or reduce the bias from confounding effects; the other is the theoretical guidance for them to choose the preconditions: the choice of the specific causal identification with variables included would be better theory-

[25]Strictly speaking, the unconfoundedness assumption in the propensity function based method turns to: the outcome is independent to the treatment conditioned on the propensity score, $(Y(1), Y(0)) \perp\!\!\!\perp A \mid \pi_a(X)$.

driven than pure data-driven.

Methods based on propensity functions with weighting and matching have multiple variant forms other than IPW and PSM. For instance, researchers could adopt an evolutionary search algorithm (Diamond and Sekhon 2013) or a Hungarian algorithm (Rosenbaum 1989) to optimize the matching process, or stratify the samples and match within different stratification (Rosenbaum and Rubin 1984), or match cases based on Mahalanobis distance (MDM, see Rubin 1980; King and Nielsen 2019).

- **Outcome Regression**

Like propensity-based approaches, researchers use **outcome regression** to obtain counterfactual means and average treatment effects when all confounders X are observed and the standard identification conditions (consistency/SUTVA, unconfoundedness given X, and positivity) hold. Outcome regression estimates the conditional outcome function and then averages predicted outcomes over the covariate distribution to obtain the counterfactual mean. With the nuisance function $m(a,x) = E[Y \mid A = a, X = x]$, the causal estimand is identified via the law of iterated expectations as $E[Y(a)] = E_X\{m(a,X)\}$.

Therefore, for outcome regression, researchers first fit the model between the outcome, the treatment and covariates, then manipulate the treatment into the treated and control values and substitute it into the fitted model to predict the counterfactual outcomes based on the manipulated treatment. With the counterfactual outcomes

under treated and control, the treatment effect can be yield.

In some literature, outcome regression is also called g-computation (VanderWeele et al. 2014), reflecting this integration over X to identify $E[Y(a)]$. As we elaborate later, the outcome-regression term and the IPW term are the two canonical components that compose doubly robust estimators.

III. Analyzing the Statistical Error

In the section above, we discussed methods that address causal error, in other words, how a statistical estimand approximates a causal estimand, $\psi^* - \psi$. We now turn to how to estimate the statistical estimand from sample data, yielding the statistical error $\hat{\psi} - \psi$[26].

According to Equation 1.I.3, it is standard to analyze statistical error via its mean squared error (MSE), which decomposes into a variance term and a squared-bias term:

$$\mathrm{MSE}(\hat{\psi}) = \mathrm{Var}(\hat{\psi}) + \left\{E[\hat{\psi}] - \psi\right\}^2.$$

In the machine-learning literature this motivates the **bias–variance trade-off** (James et al. 2013; Hastie et al. 2009)[27]. Although biased estimators can sometimes achieve lower MSE[28],

[26] Some procedures described earlier (e.g., in the section of estimation strategies) also act to reduce statistical error.

[27] A quick derivation: $\mathrm{MSE}(\hat{\psi}) = E\left[(\hat{\psi} - \psi)^2\right] = E\left[(\hat{\psi} - E[\hat{\psi}])^2\right] + \left(E[\hat{\psi}] - \psi\right)^2 = \mathrm{Var}(\hat{\psi}) + \mathrm{Bias}(\hat{\psi})^2$.

[28] A standard example is the maximum-likelihood estimator (MLE) of the variance in a normal model with unknown mean:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{is unbiased with } E[S^2] = \sigma^2,$$

whereas

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is biased, with $E[\hat{\sigma}^2_{\mathrm{MLE}}] = \frac{n-1}{n}\sigma^2$, yet can have smaller MSE due to reduced variance (useful in some industrial settings).

in the causal-inference methods used throughout this thesis we require estimators to be **unbiased for the statistical estimand** (at least asymptotically), i.e. $E[\hat{\psi}] = \psi$.[29]

Our target in this chapter is to derive the efficient-influence-function based debiased machine learning estimator for causal inference. To obtain such estimators, we rely on **asymptotic** analysis (van der Vaart 1998). We view $\hat{\psi} = \psi(\mathbb{P}_n)$ as a functional of the empirical distribution $\mathbb{P}_n$ for the true law P, and study smooth perturbations along a regular parametric submodel $\{P_\varepsilon : \varepsilon \in \mathbb{R}\}$ with $P_0 = P$. Two objects organize this analysis: the **score** (which gives the local direction of departure of P_ε from P) and the **influence function** (which quantifies the first-order effect of that departure on $\psi(P_\varepsilon)$). As we will show, the optimal (semiparametric) **efficient** estimator within the class of regular, asymptotically linear estimators is asymptotically unbiased and attains the smallest possible asymptotic variance (the efficiency bound). Moreover, its estimating equation is Neyman-orthogonal to first-order perturbations in the nuisance parameters, providing robustness to small misspecification.

Because these ideas may be unfamiliar, this section proceeds as follows. We first introduce regular and asymptotically linear estimators, then develop the concepts of score and influence function and their connection to efficiency. Then we derive the efficient influence function for the average treatment effect. Finally, we discuss how efficient influence functions can lead to the efficient doubly robust/debiased machine learning estimators. In subsequent chapters, we reuse this machinery to construct efficient estimators tailored to a range of social-science and demographic applications.

[29] In causal inference the target ψ^* is defined by a hypothetical intervention. Once identification assumptions equate ψ^* with a statistical target ψ, we want $E[\hat{\psi}] = \psi$ so the estimator is centered at the true effect. Estimators derived from unbiased estimating equations (mean-zero influence functions) are regular and asymptotically linear, enabling valid Wald inference and efficiency comparisons. Persistent bias can be indistinguishable from a true effect and risks misleading conclusions; therefore we prioritize (asymptotic) unbiasedness and then pursue variance reduction.

A. Introduction to RAL Estimators

For all the **regular and asymptotically linear (RAL)** estimators (Newey 1990: 1994)[30], it would be best to find the one with the lowest variance so that the MSE for the estimator with respect to the statistical estimand would be lowest. In regular parametric models, the **Cramer-Rao bound** provides a benchmark lower bound for the variance of unbiased estimators (Greene 2012):

Lemma 1.III.1 (Cramer-Rao Bound) *The Cramer-Rao Bound (CRB) states that for an unbiased estimator $\hat{\psi} = \psi(\mathbb{P}_n)$ of $\psi = \psi(P)$, the variance of $\hat{\psi}$ is at least as large as the inverse of the Fisher information:*

$$Var(\hat{\psi}) \geq \frac{1}{I(\psi)}$$

where $I(\psi)$ is the Fisher information given by:

$$I(\psi) = E\left[\left(\frac{\partial}{\partial \psi} \log f(Z;\psi)\right)^2\right]$$

[30]Not all unbiased estimators are RAL. For instance, consider the median estimator for all symmetric distributions. It is an unbiased estimator for the population median, but it is not asymptotically linear as $\phi(\psi;Z)$ for the median is not a smooth function. Another classic example here is the Hodges' estimator (van der Vaart 1998): in the normal mean model with known variance, define

$$\hat{\theta}_H = \begin{cases} 0, & \text{if } |\bar{X}_n| \leq n^{-1/4}, \\ \bar{X}_n, & \text{otherwise,} \end{cases}$$

where $\bar{X}_n$ is the sample mean. At $\theta_0 = 0$ it is super-efficient so its asymptotic variance at the $\sqrt{n}$ scale is 0. But this gain is non-uniform: in $n^{-1/4}$-neighborhoods of 0 the risk inflates and the estimator fails to be regular: under local alternatives $\theta_n = h/\sqrt{n}$ it does not have a stable $\sqrt{n}$-normal limit. Consequently, there is no influence function ϕ such that

$$\sqrt{n}(\hat{\theta}_H - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(Z_i) + o_p(1)$$

uniformly over a neighborhood of P_{θ_0}. Hodges' estimator is therefore not regular asymptotically linear (RAL).

The proof will be given in Appendix A.3. In simple terms, the Cramer–Rao Bound is a fundamental result that says: no unbiased estimator can have a variance smaller than a certain threshold determined by the information in the data (Bickel et al. 1993; van der Vaart 1998).

The threshold, as Lemma 1.III.1 suggested, is the reciprocal of the **Fisher information** of the estimand ψ. Fisher information is a measure of how much information the data distribution carries about the quantity we are estimating. Intuitively, if the data are very sensitive to changes in the parameter (for example, a slight change in the parameter makes the probability of observations change a lot), then the Fisher information is high. With high information, we can estimate the parameter more precisely.

In this parametric setting, if an unbiased estimator attains the CRB, we call it an **efficient estimator** (Kay 1993; Lehmann and Casella 1998; Rao 1973), because it uses the information in the data as efficiently as possible for the parameter of interest. However, deriving an efficient estimator is often challenging, especially in complex models, because it requires carefully using the data without waste. The concepts of regularity, asymptotic linearity, score functions, and influence functions introduced below are the tools that help us move towards the goal of efficient estimation. With all these concepts introduced, they provide a framework for developing estimators that are not only unbiased, but also have variance as low as possible and follow convenient large-sample behavior. Before delving into those concepts, it's important to note that we usually restrict our attention to estimators that behave well as the sample size grows.

In practice, to apply the above efficiency theory, we firstly require our estimator to satisfy certain **regularity** conditions. Regularity means that as the sample size increases, the estimator's behavior stabilizes in a nice, smooth way. In other words, a regular estimator is one that doesn't react erratically to tiny changes in the underlying data-generating process when we have lots of data. Instead, its distribution settles down and converges to something well-behaved as n (the sample size) becomes large.

One way to express regularity formally is through a convergence condition. Generally speaking, an estimator $\hat{\psi}_n$ for a parameter ψ is regular if, when we scale its error by $\sqrt{n}$, it converges in distribution to a fixed distribution as $n \to \infty$. In notation, this is often written as:

$$\sqrt{n}\big(\psi(\mathbb{P}_n) - \psi(P)\big) \overset{P}{\rightsquigarrow} D, \tag{1.III.8}$$

Where P denotes an empirical measure which we will discuss shortly, and D denotes a fixed distribution [31]. This condition means that the fluctuations of the estimator around the true value shrink at the rate $1/\sqrt{n}$ and eventually follow a stable distribution. The $1/\sqrt{n}$ rate is the classic parametric convergence rate, as it's the same rate at which the sample mean converges to the true mean by the Central Limit Theorem. Converging in distribution to some D means that the shape of the estimator's probability distribution approaches the shape of D as we get more data. Essentially, regularity rules out estimators that behave irregularly

[31] In this thesis, the denotation on the convergence uses the expressions in van der Vaart (1998). In short, convergence in distribution (weak convergence) $X_n \xrightarrow{D} X \iff X_n \rightsquigarrow X$ suggests that a sequence of random variables $\{X_n\}$ converges in distribution to a random variable X if for all points t at which $F_X(t)$ is continuous:$\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$, where $F_{X_n}(t)$ and $F_X(t)$ are the cumulative distribution functions of X_n and X; convergence in probability $X_n \xrightarrow{P} X$ (or $X_n \xrightarrow{prob.} X$, to differentiate with the measure P) suggests the relationship between X_n and X, for every $\epsilon > 0$, is: $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$. Almost sure convergence $X_n \xrightarrow{a.s.} X$ suggests X_n almost surely converges to X: $P(\lim_{n\to\infty} X_n = X) = 1$ (we sometimes also write it as $a.s.$). Meanwhile, we also use big-O probability and small-o probability to denote convergence: big $O_p : X_n = O_p(a_n)$ indicates that the sequence of random variable X_n is bounded by a_n in probability: for every $\epsilon > 0$, there exist constants $M > 0$ and $N > 0$ such that$P(|X_n| \leq M a_n) \geq 1 - \epsilon$ for all $n \geq N$. Small o_p suggests that X_n is asymptotically smaller than a_n as n increases, or the difference between X_n and a_n is negligible: for every $\epsilon > 0$ and $\delta > 0$,$P(|X_n| \geq \delta a_n) \to 0$ as $n \to \infty$.

(e.g. jumpy or discontinuous estimators) which might not settle into a clear pattern as the sample size grows. In summary, regularity essentially guarantees the stability and smoothness of the estimand: with more data, the distribution change of the estimand should be in a controlled, smooth manner (often around the true value). With the smoothness of convergence, we may use **score function** to identify the directions of local change so that regularity lets us differentiate along the directions.

While regularity cares about the estimator's overall convergence behavior, **asymptotic linearity** suggests how the estimator's error can be represented when the sample is large. An estimator is asymptotically linear if, for large n, the difference between the estimator and the true parameter can be well-approximated by a linear term which is the average of some function of the individual observations, plus a smaller remainder term. Suppose the function is $\phi(\psi; P; Z_i)$, for the estimator ψ, measure P and dataset $Z_i = (X_i, Y_i)$[32],therefore,

$$\sqrt{n}\big((\psi(\mathbb{P}_n) - \psi(P)) - \frac{1}{n}\sum_{i=1}^{n}\phi(\psi; \mathbb{P}_n; Z_i)\big) \xrightarrow{prob.} 0, \tag{1.III.9}$$

We call $\phi(\psi; \mathbb{P}_n; Z_i)$ as the **influence function** for the empirical data Z_i with respect to the empirical measure $\mathbb{P}_n$ (Schuler and van der Laan 2024; Kennedy 2016; Ichimura and Newey 2022). As we will show later in this chapter, the influence function of an estimator tells us how each individual data point influences the estimator's value and thus it breaks down the estimator's error into contributions from each observation. A more rigorous way to write Equation 1.III.9 is:

$$\big(\psi(\mathbb{P}_n) - \psi(P)\big) - \frac{1}{n}\sum_{i=1}^{n}\phi(\psi; \mathbb{P}_n; Z_i) = o_p(n^{-1/2}) \tag{1.III.10}$$

[32]The influence function has three entries, ϕ refers to the estimation functional form, P refers to the measure. Z_i refers to the measurable set (dataset). Due to our assumptions in Section I that the estimation function and the measure (DGP) will not change simultaneously, thus, if we specify the influence function for an estimator, we omit the measure and the dataset. We use Z_i for discrete elements in the measurable set and z if the elements are continuous.

which suggests that the difference between the estimator and the true value equals the average influence $\frac{1}{n}\sum(\psi;\mathbb{P}_n;Z_i)$ plus a remainder term that is negligible compared to $n^{-1/2}$ in probability. The $o_p(n^{-1/2})$ term means that when you multiply the remainder by $\sqrt{n}$, it goes to zero in probability as $n \to \infty$. By the Central Limit Theorem (Theorem 1.I.1), we know that the average $\frac{1}{n}\sum(\psi;\mathbb{P}_n;Z_i)$ will be approximately normal for large n. Thus, asymptotic linearity could also be expressed as:

$$\sqrt{n}(\psi(\mathbb{P}_n) - \psi(P)) \xrightarrow{d} N(0,\sigma^2). \tag{1.III.11}$$

where $\sigma^2 = \mathrm{Var}(\phi(\psi;\mathbb{P}_n;Z_i))$ and we call it the **asymptotic variance**. While regularity describes the convergence of a distribution, asymptotic linearity gives us asymptotic normality and thus the ability to do inference using familiar Gaussian approximations. More importantly, it provides a blueprint for how to improve an estimator: if we know the influence function, we can often reduce bias or adjust the estimator by subtracting out the average influence. For the details of regularity and the score function, asymptotic linearity, and the influence function, we will discuss them in the following sections.

B. Regularity and Score Function

As mentioned above, regular estimators correspond to functionals P that change smoothly when P is perturbed, and the score function is the tool that identifies the direction of perturbation. The score function acts like a directional derivative or a compass pointing in the direction of a small change in the distribution or model.

Suppose the true data-generating process is described by some probability distribution P over the data $Z = (A, X, Y)$. Now imagine a slightly perturbed version of that distribution, call it $\tilde{P}$, which is "close" to P. We can consider a smooth path of distributions from P to $\tilde{P}$,

indexed by a small parameter $\epsilon \in [0,1]$. At $\epsilon = 0$ we're at the original distribution ($P_{\epsilon=0} = P$), and at $\epsilon = 1$ we're at the new distribution ($P_{\epsilon=1} = \tilde{P}$). One convenient way to form such a path is by mixing the two distributions: $P_\epsilon = (1-\epsilon)P + \epsilon\tilde{P}$. This defines a smooth trajectory in the space of possible distributions.

The score function $s(z)$ associated with this path is defined as the rate of change of the log-likelihood (the log of the probability density) at $\epsilon = 0$, in the direction of the new distribution. In simpler terms, $s(z)$ tells us how the log probability of observing a data point z would change if we nudge the distribution from P toward $\tilde{P}$ at an infinitesimal rate (Bickel et al. 1993; Kennedy 2016; Schuler and van der Laan 2024). We can write:

$$s_{\epsilon_0}(z) = \frac{\partial log\, \tilde{p}_\epsilon(z)}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} \tag{1.III.12}$$

where $p_\epsilon(z)$ is the density of P_ϵ. This derivative essentially compares the new density $\tilde{p}(z)$ to the original $p(z)$. In fact, if we do the calculus (see Appendix A.1), it turns out:

$$s(z) = \frac{\tilde{p}(z)}{p(z)} - 1 \tag{1.III.13}$$

when evaluated at $\epsilon = 0$. This formula says the score function is, at first order, proportional to the fractional change in the probability of z under the perturbation. Another way to write it is $\tilde{p}(z) = (1+\epsilon s(z))\,p(z)$ for small ϵ. So, $s(z)$ indicates in which direction the probability of each point z is being pulled by the new distribution relative to the old one and how strongly it is.

An important characteristic of the score function is that the score function at the original distribution has mean zero:

$$E_P[s(Z)] = 0$$

Because $\tilde{P}$ is just a reweighting of P, at $\epsilon = 0$ there is no change, so the expected score should be zero. Intuitively, scores form a family of "compasses," each pointing to a distinct local

perturbation direction of the data-generating process around P. We can imagine these directions spanning a space [33]. The collection of these scores is called the tangent space of the model; it consists of all directions along which we can locally perturb the distribution while remaining within the model. [34]:

$$T(P_Z) = \overline{\left\{ \sum_i \alpha_i s_{\eta_i} : \alpha_i \in \mathbb{R},\ s_{\eta_i} \text{ are scores of smooth submodels through } P_Z \right\}}.$$

If the tangent space contains all the mean-zero functions, the model that specifies the DGP is a **saturated model** (Schuler and van der Laan 2024). For saturated models, the tangent space exists for all directions, meaning that if we move towards any direction, we are still in the model. Suppose our models are fully nonparametric; then, our model is saturated since no restrictions impede us. Otherwise, the model is not saturated. For instance, a general causal model is saturated, but an RCT causal model is not because the propensity in the randomized trial is fixed.

The mean-zero property of the tangent space yields a useful decomposition in practice: suppose we have a joint distribution: $P(X, Y)$, we can decompose any joint score functions as:

$$s_{X,Y}(x, y) = s_X(x) + s_{Y|X}(x, y) \quad \text{with} \quad E\left[s_{Y|X}(X, Y) \mid X\right] = 0, E[s_X(X)] = 0,$$

and any pathwise derivative along the joint score can also be split into marginal and conditional components:

$$\nabla_{s_{X,Y}} \psi = \nabla_{s_{Y|X}} \psi + \nabla_{s_X} \psi$$

[33] Strictly speaking, they are mean-zero, square integrable functions. Since the score functions are defined on the L_2^0 space, we could define the tangent space as the set of the square-integrable functions with respect to $P(Z)$ whose means are 0:

$$T(P(Z)) = \left\{ h \in L_2(P(Z)) : E_{P(Z)} h(x) = 0 \right\}.$$

[34] In a fully nonparametric/saturated model, that tangent space coincides with all mean-zero, square-integrable functions; in restricted models, it's a proper subset.

As we noted above, the score function is the tool to identify the direction of perturbation, and we use it to calculate how an estimand $\psi(P)$ would change if the distribution P is moved in a certain direction. Hence, the pathwise derivative of an estimand along a score direction ∇_s leads us to the influence function, which we will discuss next.

In the above steps, we are actually **factorizing** the score function (Lehmann and Casella 1998; Cox and Hinkley 1974). Moreover, since any score function belonging to the tangent space of the specific measure can be used during factorization, therefore, we are actually factorizing the tangent space. In the above bivariate example, we could write the factorization as: $T_{X,Y} = T_{Y|X} \oplus T_X$, where $T_{Y|X}$ is defined as the tangent space associated with the conditional distribution $P(Y|X)$: $T_{Y|X} = \left\{h_{Y|X}(x,y) : E[h_{Y|X}(x,y) \mid X = x] = 0 \text{ for all } x\right\}$. Correspondingly, $T(X)$ denotes the tangent space associated with the marginal distribution $P(X)$: $T(X) = \{h_X(x) : E[h_X(X)] = 0 \text{ for all } x\}$. The symbol $\oplus$ denotes the direct sum, indicating that these spaces are orthogonal.

An advantage of factorization is that, after our appropriate factorization, if the estimator is perturbed only on one dimension, we need only consider the change in the score function on that dimension and keep the others unchanged. The proof of the factorization transformations can be seen in Appendix A.1. We will use this advantage when we yield the efficient estimator later.

Finally, we may also notice the relationship between the score function and Fisher information: the denominator in the Fisher information is the average of the squared score function, and this reveals the importance of the score function in the Cramer-Rao bound. In regular parametric models, an estimator that attains the Cramer-Rao bound has fluctu-

ations that align with the score function $s(Z;\psi)$ that captures information about ψ. This intuition carries into the influence-function framework below, where efficiency is characterized by the efficient influence function and the corresponding efficiency bound. We will once again discuss this when we reach the efficient influence functions.

C. Asymptotic Linearity and the Influence Function

As we introduced above, asymptotic linearity cares about how the estimator's error can be represented when the sample size is large. For large n, an estimator would be asymptotically linear if the difference between the estimator and the true parameter can be approximated by a linear term, which is an average of some functional form of the individual observations and a smaller remainder term. The "some functional form", as we elaborated before, is called the **influence function** (Kennedy 2016).

Based on the definition, the influence function of an estimator shows how each individual data point influences the estimator's value. Indeed, the influence function can also be regarded as the (Gateaux) pathwise derivative of the estimated functional, the pathwise gradient, the first-order term of the Taylor expansion of the estimand around the true distribution, and the Neyman orthogonal score (Chernozhukov et al. 2018b). Due to the importance of the influence function, we will demonstrate the equivalence.

C.1 Influence Function as Individual Contribution to Estimation Error

Firstly, the influence function can serve as the description of the contribution of each data point to the estimation error, which we can infer directly from the asymptotic linear representation. As

$$\hat{\psi}_n - \psi(P) \approx \frac{1}{n}\sum_i \phi(\psi;\mathbb{P}_n;Z_i)$$

So we can say $\phi(\psi;\mathbb{P}_n;Z_i)$ is the contribution of the i-th observation to the estimation error. For an unbiased estimator, we have mean-zero error terms (the positive and negative influences cancel out on average), suggesting that if an estimator is unbiased, a small random change in data should raise the estimate in some cases and lower it in others, with no systematic bias. The influence function formalizes this idea.

C.2 Influence function as Gateaux/ Pathwise Derivative of the Estimand Functional

In the second definition, the influence function is essentially the **Gateaux derivative** of the functional $\psi(P)$ in the direction of a **point mass** at z.

To understand the above statement, we need to define what the Gateaux derivative is. In the discussion of the score functions, we have illustrated the process of the distribution P nudging toward $\tilde{P}$. The Gateaux derivative is the slope for the target estimand $\psi(P)$ under that nudge, or in other words, how fast ψ would change per unit of that reweighting in that direction. [35].

Here, we use the point mass at a specific data point z to replace $\tilde{P}$, so the nudge becomes the increase of a very tiny weight (an infinitesimal bit of probability) on a single point z and the removal of the same tiny weight from the rest accordingly. It is exactly the same as the contribution of the single point z to the estimation error: The influence function $\phi(\psi;\mathbb{P}_n;Z_i)$ is defined so that $\psi(\tilde{P}) \approx \psi(P) + \epsilon, \phi(\psi;\mathbb{P}_n;Z_i)$ for a small perturbation that adds

[35] Mathematical definition: Let P be a class of probability measures on a measurable space (Z, A), and let $\psi : P \to \mathbb{R}$ be a (possibly nonlinear) functional. For $P \in P$ and any signed finite measure $\tilde{P}$ with total mass $\tilde{P}(Z) = 0$ (a direction), define the path $P_\epsilon := P + \epsilon\tilde{P}$ for small ϵ such that $P_\epsilon \in P$. The Gateaux derivative of ψ at P in the direction $\tilde{P}$ is

$$D\psi(P;\tilde{P}) := \lim_{\epsilon\to 0} \frac{\psi(P+\epsilon\tilde{P}) - \psi(P)}{\epsilon},$$

Whenever the limit exists. Equivalently, for any $\tilde{P} \in P$ one may use the mixture path $P_\epsilon = (1-\epsilon)P + \epsilon\tilde{P}$ and write $D\psi(P;\tilde{P}-P)$.

an ϵ amount of probability at z. Intuitively, it's answering: "if we tweak the distribution to include a little more of point z, how does my target quantity ψ change to first order?" If $\phi(\psi;\mathbb{P}_n;Z_i)$ is large, it means that point z has a big impact on the estimator; if $\phi(\psi;\mathbb{P}_n;Z_i)$ is small or zero, z doesn't influence the estimator much (in the first-order term).

The fact that the influence function can be regarded as a kind of derivative is an important and widely used property because, in practical computations, all rules of differentiation can apply to the influence function, and the most useful one is the chain rule for differentiation, as we will show below in the derivation of the efficient influence function.

C.3 Influence Function as Pathwise Gradient

Now, following the Gateaux derivative definition, recall that the score function describes the direction of the moves. So all the tiny, smooth moves we can move away from the true DGP have a direction described by a mean-zero score function $s(Z)$ and the path $P_\epsilon = (1+\epsilon s)P$ (Equation 1.III.13). The pathwise derivative is just the slope of $\psi(P)$ along such a tiny move. Indeed, under linearity and certain (Riesz) representation, this slope equals an inner product:

$$\nabla_s \psi(P) = \frac{d}{d\epsilon}\psi(P_\epsilon)\Big|_{\epsilon=0} = E\big[\phi(\psi;\mathbb{P}_n;Z_i)\, s(Z)\big]. \tag{1.III.14}$$

Equation 1.III.14 is called the **central identity for influence functions** and we append its proof in Appendix A.2 for interested readers (Schuler and van der Laan 2024; Tsiatis 2006; Bickel et al. 1993; Ichimura and Newey 2022). So the influence function ϕ is the object that, for any chosen direction s, gives the correct directional rate of change of ψ by this simple average. That's why we call ϕ the pathwise gradient: it plays the same role as a gradient does in ordinary calculus, but now for small changes of the whole distribution.

The central identity for influence functions also reveals the relationship between the influence function and the score function: the direction of the gradient is aligned with the interaction between the influence function and the score function (as the expectation of the score and the influence functions under RAL are both zero). This equation is the most helpful tool for extracting the influence function or validating if the influence function is correct, especially for efficient influence functions for the saturated models. We will give a specific example later in this chapter.

C.4 Influence Function as the First-Order (Bias Correction) Term

As the influence function serves as the Gateaux derivative and pathwise gradient, it is also the first-order term (the linear part) of the functional/ Taylor expansion of the target ψ around the true distribution P. Recall Equation 1.III.10, when we nudge along path $P_\epsilon = (1-\epsilon)P + \epsilon\tilde{P}$, as the influence function can be expressed as the Gateaux derivative (let $D\psi(.)$ denote the derivative), we have

$$\psi(P_\epsilon) = \psi(P) + \epsilon\, D\psi(P; \tilde{P} - P) + o(\epsilon) = \psi(P) + \epsilon \int \phi(\psi; \mathbb{P}_n; Z_i)\, d(\tilde{P} - P)(z) + o(\epsilon).$$

As the influence function is also the pathwise gradient, for the point-mass direction $\tilde{P} = \delta_z$,

$$\psi\big((1-\epsilon)P + \epsilon\,\delta_z\big) = \psi(P) + \epsilon\,\phi(\psi; \mathbb{P}_n; Z_i) + o(\epsilon),$$

The Taylor expansion of a function can be written as:

$$f(X_0) \approx f(X_1) + \nabla f(X_1)(X_0 - X_1) + \frac{1}{2}(X_0 - X_1)\nabla^2 f(X_1)(X_0 - X_1)^T + \ldots \tag{1.III.15}$$

or,

$$f(X_0) \approx f(X_1) + \frac{\partial f(x)}{\partial x}\Big|_{x=X_1}(X_0 - X_1) + \frac{1}{2}\frac{\partial^2 f(x)}{\partial x^2}\Big|_{x=X_1}(X_0 - X_1)^2 + \ldots$$

Thus, for the function $\psi(P_\epsilon)$ in the domain $\epsilon \in [0,1]$, we have [36]:

$$\psi(P_\epsilon) = \psi(P) + \underbrace{\epsilon \left.\frac{\partial}{\partial \epsilon}\psi(P_\epsilon)\right|_{\epsilon=0}}_{\text{first-order bias correction}} + \underbrace{\frac{\epsilon^2}{2}\left.\frac{\partial^2}{\partial \epsilon^2}\psi(P_\epsilon)\right|_{\epsilon=0}}_{\text{second-order remainder}} + o(\epsilon^2). \tag{1.III.16}$$

The first-order "bias correction" (we will revisit it later) term is exactly the pathwise gradient expression of the influence function (Fisher and Kennedy 2021). To visualize the idea, we illustrate the ideas in Figure 1.2, which shows how to use results of $\psi(P)$ to approach $\psi(P_\epsilon)$.

C.5 Influence Function as Neyman Orthogonal Score

In semiparametric models, we separate the target (ψ) and the nuisance parts that we don't care about (η). A **Neyman Orthogonal score** is a mean-zero target (moment) whose first-order sensitivity to the nuisance part is zero at the truth. In other words, small errors in η do not move the moment to first order. As the influence function can be seen as the first-order (bias correction) term, has zero-mean, and is orthogonal to all nuisance functions (the score functions), due to the central identity of the influence function $E[\phi(\psi; \mathbb{P}_n; Z_i)s_\eta(z)] = 0$ (Chernozhukov et al. 2018b), thus, building an estimation equation from the influence function yields moments that are insensitive to small misspecification of nuisance parts, enabling valid inference with flexible machine learning techniques for nuisance functions. This is the mechanism behind the debiased/ doubly robust machine learning estimators we discuss throughout the thesis.

In summary, we derive the four expressions of the influence function, which can serve as the individual contribution to estimation error, Gateaux derivative to the estimand, pathwise gradient, first-order bias correction term, and Neyman orthogonal score. As we dis-

[36] This is also called **"von Mier Expansion"** in some literature.

cussed the necessity of asymptotic linearity, it gives us asymptotic normality and thus the ability to do inference and provide a method to improve the estimator: if we know the influence function, we can reduce bias with the idea of a one-step estimator, which precisely takes an initial (possibly biased) estimator and then adds the average influence function to correct it. By doing so, we cancel out the first-order bias and often dramatically improve accuracy. This is how we construct the efficient estimators, with the derivation of the efficient influence functions (EIF).

D. Efficient Influence Functions

We have introduced the basic concepts of regularity, score functions, asymptotic linearity, and influence functions. These concepts are a necessary toolbox to identify and construct efficient estimators, which attain the relevant parametric or semiparametric efficiency bound under the corresponding regularity conditions.

In large samples, according to Equation 1.III.11, the variance of an estimator is given by the asymptotic variance, which is the variance of the influence function. Therefore, to minimize the variance is the same as finding the influence function with the smallest possible variance. The optimal influence function is called the **efficient influence function** for the estimand. And if the efficient influence function derives an estimator, we call the estimator an **efficient estimator**.

So the target to derive an efficient estimator is intrinsically the same as finding the efficient influence function (EIF) first and constructing the estimator whose influence function is that EIF. An intuitive idea is to consider the set of all influence functions for all regular,

asymptotically linear estimators of ψ, and then pick the one with the smallest variance.

We know that the tangent space is spanned by the score functions—that is, the set of all directions the model lets us move the data-generating distribution. Therefore, if we want the target estimand to change in a well-defined, first-order way, the push direction must lie in the tangent space; conversely, any component pointing outside these allowable directions cannot change $\psi(P)$ to first order and is just noise.

We can split a candidate influence function into two orthogonal parts: a target-of-interest part, which lies in the tangent space and moves ψ; and a nuisance part, which also lies in the tangent space but is tied to features that do not identify ψ. Any remaining component outside the tangent space is discarded. We then **project** within the tangent space to remove the nuisance part, keeping only the part that legitimately moves the estimand. This projection yields the smallest-variance valid choice, which is exactly how the EIF is defined. Hence, the influence function should lie on the tangent space to be efficient (Schuler and van der Laan 2024; Newey 1994; Bickel et al. 1993). The detailed mathematical proof can be seen in Appendix A.3.

The necessary condition that the EIF lies in the tangent space reveals the relationship between the score function and the EIF: the EIF is always aligned with the score function for the target parameter. In classical parametric models, the score function for ψ and the EIF are proportional to each other. Intuitively, this means the estimator is nudging in the same direction as the true parameter itself.

From the relationship between the EIF and the tangent space, we can derive an important corollary: for fully nonparametric saturated models, the efficient influence function is

unique. This is somehow obvious: because the tangent space for the saturated model is the whole space, meaning all directions of the small nudge to the distribution are allowed. The influence functions are the first-order change of ψ for every such nudge. And if two candidates both worked, then their difference would have zero correlation with every allowed nudge. The only function that's uncorrelated with all nudges is the zero function, so the two candidates must be the same and the EIF is unique [37].

In non-saturated models (with constraints, for instance, semiparametric models with nuisance parameters), the efficient influence function still exists but must lie within the allowed tangent space of that model.

Below, we provide an example of how we derive the EIF for the average treatment effect (ATE), which enables us to yield the efficient doubly robust/debiased machine learning estimator for the ATE, the main target of this thesis.

D.1 Deriving EIF for the ATE in the Saturated Model

We start with the saturated model: the ATE in the observational study, where there is no restriction on the statistical estimand to infer the causal estimand, and there's only one influence function, which is the EIF for the estimator. Similar to the operation in Equation 1.II.5, we use $\psi_a = E[E_X[Y|A=a,X]]$ from the observational study to infer $\psi_a^* = E[Y(a)]$ (therefore, the ATE is $\psi_1 - \psi_0$). From the perspective of the efficient theory, the estimators we applied in Section II, for instance, the IPW estimator, as we have shown, is a RAL estimator but not the efficient one, as it is only the "naive plug-in estimator" part in the Taylor distributional expansion decomposition (Equation 1.III.16). Our goal is the efficient

[37] Mathematically, we can write: suppose we have two influence functions ϕ_1 and ϕ_2, based on the central identity of the influence functions we have: $E[\phi_1 s] = E[\phi_2 s] = E[(\phi_1 - \phi_2)s] = 0$, then obviously $\phi_1 = \phi_2$.

estimator for ψ_a:

$$\psi_a = E[E_X[Y|A=a,X]] = \sum_x \underbrace{E[Y|A=a,X]}_{:=\mu_a(X)} p(x) = \sum_x \mu_a(x)p(x)$$

As we discussed above, to derive the efficient estimator, we first capture the EIF and then derive the efficient estimator based on the EIF. Based on the chain rule,

$$\phi(\psi_a) = \sum \phi(\mu_a(x)p(x)) = \sum \Big(\phi(\mu_a(x))p(x) + \mu_a(x)\phi(p(x))\Big). \tag{1.III.17}$$

As the equation shows, we need the EIFs for the expectation $\phi(E[X])$ (to yield $\phi(p(x))$)and the conditional expectation $\phi(E[Y|X=x])$ (to yield $\phi(\mu_a(x))$.

EIF for Unconditional Expectation Because of space limitations, we relegate the step-by-step derivations to the Appendix A.3 for interested readers. We first derive the EIF for $\psi(P) = E_P[X] = \int x dP(x)$. From the Gateaux derivative definition of the influence function, its expression for the unconditional mean $\phi(\psi = E_P[X]; P, x)$ is (Kennedy 2023):

$$\phi(E_P[X]) = \frac{\partial}{\partial \epsilon}\psi(\tilde{P}_\epsilon)\Big|_{\epsilon=0} = x - \psi(P) = x - E_P[X] \tag{1.III.18}$$

EIF for Conditional Expectation We then derive the EIF for $\psi(P) = E_P[Y|X=x] = \int_y y dP(y|x)$. Still, $\tilde{P}_\epsilon = (1-\epsilon)P + \epsilon\delta_{y|x}$. Recall the Bayesian rule $P(y|x) = \frac{P(y,x)}{P(x)}$. Therefore, we can get:

$$\phi(\psi = E_P[Y|A=a,X=x], P_X(y,a,x)) = \frac{\mathbb{1}_{(a,x)}}{p[A=a,x]}\Big[y - E_P[Y|A=a,X]\Big] \tag{1.III.19}$$

Equation 1.III.19 is the EIF for the conditional expectation.

EIF for ψ_a With Equations 1.III.18 and 1.III.19, we have the EIF for $\phi(\psi_a) = E[E_X[Y|A=a,X]]$. Based on the chain rule in Equation 1.III.17, we have:

$$\phi(\psi_a) = \sum_x \Big[\Big(\frac{\mathbb{1}_{a,x}}{p(a,x)}[y-\mu_a(x)]p(x)\Big) + \Big(\mu_a(x)(\mathbb{1}_x - p(x))\Big)\Big]$$

$$= \frac{\mathbb{1}(a)}{\pi_a(x)}[y - \mu_a(x)] + \mu_a(x) - \psi_a$$
$$= \frac{\mathbb{1}(A = a)}{p(A = a|X)}(Y - E[Y|A = a, X]) + E[Y|A = a, X] - E[Y|A = a]$$

Thus, the EIF for ψ_a with discrete measurable elements $\phi(\psi_a)$ is:

$$\phi(\psi_a) = \frac{A_i}{\pi_a(X_i)}[(Y_i - \mu_a(X_i)] + \mu_a(X_i) - \psi_a \tag{1.III.20}$$

In other words,

$$\phi(\psi_1) = \frac{\mathbb{1}(A_i = 1)}{\pi_1(X_i)}(Y_i - \mu_1(X_i)) + \mu_1(X_i) - \psi_1$$

and

$$\phi(\psi_0) = \frac{\mathbb{1}(A_i = 0)}{\pi_0(X_i)}(Y_i - \mu_0(X_i)) + \mu_0(X_i) - \psi_0$$

The central identity of the influence function for the EIFs from the saturated models can validate the correctness of the EIFs. We provide the validation in Appendix A.3 for interested readers.

D.2 Deriving EIF for the ATE in the Non-Saturated Model

Further, we consider the non-saturated model scenario. The average treatment effect under the randomized controlled trial setting is a non-saturated model since we have placed restrictions on the treatment and control cases. Therefore, unlike the saturated models, we cannot derive the efficient influence function with an arbitrary score function, as the influence function may not be efficient. However, we could start with a known RAL estimator and derive its influence function through the definition of asymptotic linearity, and then project it (find its minimized square error) onto the tangent space. As elaborated before, the EIF should be the projection of any influence functions on the tangent space. However, it might be hard to find the projection on the tangent space directly if the underlying estimator is complex. If so, we could still use the factorization technique that first projects the

influence function onto tangent subspaces and then sums the factorized influence functions together.

For the ATE under the RCT setting, we start with the IPW estimator. Obviously, as shown in Equation 1.II.6, the IPW estimator is a regular and asymptotically linear one. Due to the symmetry in expressions for ψ_0 and ψ_1[38], to reduce the space in this thesis, we use the estimation on the EIF for ψ_0 as an example below:

$$\hat{\psi}_0^{IPW} = E\left[\frac{\mathbb{1}(A_i=0)}{\pi_0(X_i)}Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}(A_i=0)}{\pi_0(X_i)}Y_i\right].$$

Therefore, based on Equation 1.III.9, we could derive the influence function as:

$$\phi_0^{IPW} = \frac{\mathbb{1}(A_i=0)}{\pi_0(X_i)}Y_i - \psi_0.$$

Then we project the influence function from the IPW estimator into the tangent subspace of $T_{Y|A,X}$, $T_{A|X}$, and T_X in which $s_{Y|A,X}$, $s_{A|X}$, and s_X forms (and $T_{Y,A,X} = T_{Y|A,X} \oplus T_{A|X} \oplus T_X$). Again, we will not show the detailed projection derivations here; the step-by-step processes are in Appendix A.3 for interested readers.

Project the Influence Function on T_X First, we try to project the influence function of the IPW estimator on the tangent space T_X. The projection function is defined to find the score function on the tangent space for which its mean squared error with the influence function from the IPW is minimal[39]; therefore, the projection of the influence function for the IPW estimator on the tangent space T_X as its conditional expectation on the X axis:

$$\phi^{\dagger}_{0\langle T_X\rangle} = E\left[Y_i|A_i=0,X\right] - \psi_0 = \mu_0(X_i) - \psi_0$$

[38]The derivation is also applicable when the treatment is multinomial, or even continuous.

[39]We have a sketch Figure A.1 illustrating the projection process for the readers' reference for understanding the algebraic process here.

Project the Influence Function on $T_{A_0|X}$ Now we derive the projection of the influence function to the tangent space $T_{A|X}$. Indeed, the parameter $\psi_0 = E\{\mu_0(X)\}$ does not depend on the treatment mechanism $P(A \mid X)$; hence the canonical gradient has no $T_{A|X}$ component:

$$\phi^{\dagger}_{0\langle T_{A|X}\rangle} = 0.$$

Project the Influence Function on $T_{Y|A,X}$. For the projection of ϕ_0^{IPW} onto the tangent space $T_{Y|A_0,X}$, we have:

$$\begin{aligned}\phi^{\dagger}_{0\langle T_{Y|A_0,X}\rangle} &= E[\phi_0^{IPW} \mid Y, A_0, X] - E[\phi_0^{IPW} \mid A_0, X] \\ &= \left(\frac{\mathbb{1}(A_i = 0)}{\pi_0(X_i)} Y_i - \psi_0\right) - \left(\frac{\mathbb{1}(A_i) = 0}{\pi_0(X_i)}\mu_0(X_i) - \psi_0\right)\end{aligned}$$

Sum up the three sub-EIFs on the three tangent subspaces and we can get the EIF for ψ_0 under the RCT settings:

$$\begin{aligned}&\phi^{\dagger}(\hat{\psi}_0^{IPW}) \\ &\quad= \phi^{\dagger}_{0\langle T_X\rangle} + \phi^{\dagger}_{0\langle T_{A_0|X}\rangle} + \phi^{\dagger}_{0\langle T_{Y|A_0,X}\rangle} \\ &\quad= \big(\mu_0(X_i) - \psi_0\big) + 0 + \left(\frac{\mathbb{1}(A_i = 0)}{\pi_0(X_i)} Y_i - \psi_0\right) - \left(\frac{\mathbb{1}(A_i) = 0}{\pi_0(X_i)}\mu_0(X_i) - \psi_0\right) \\ &\quad= \frac{\mathbb{1}(A_i = 0)}{\pi_0(X_i)}(Y_i - \mu_0(X_i)) + \mu_0(X_i) - \psi_0\end{aligned}$$

Which is, unsurprisingly, exactly the EIF for ψ_0 we obtained from the saturated model. Similarly, the EIF for ψ_1 from the non-saturated model of the RCT will also be the same as the EIF from the saturated model of the observational study.

E. Efficient Estimators

E.1 One-step Estimator

With the EIF in hand, we can build the efficient estimator with the approach called **one-step estimator**. The one-step estimator is actually what Equation 1.III.10 and the first-order bias-correction definition for the influence function describe: we start with the consistent estimator $\psi(\mathbb{P}_n)$ from the sample, then add the correction term based on the EIF to nudge the consistent estimator towards efficiency (Fisher and Kennedy 2021):

$$\hat{\psi}_{\text{1-step}} \approx \hat{\psi}(\mathbb{P}_n) + \sum_{i=1}^{n} \phi(\hat{\psi}; \mathbb{P}_n; Z_i) \tag{1.III.21}$$

We call $\hat{\psi}(\mathbb{P}_n)$ the **naive plug-in estimator**, and $\phi(\hat{\psi}; \mathbb{P}_n; Z_i)$ is the EIF evaluated at each observation in $\mathbb{P}_n$. Under regularity conditions, this one-step updated estimator $\hat{\psi}_{\text{1-step}}$ will be consistent and asymptotically efficient, and thus it is the efficient estimator.

With the EIFs for ψ_1 and ψ_0, we can get the EIF for the ATE:

$$\begin{aligned}
\phi(\psi) &= \phi(\psi_1) - \phi(\psi_0) \\
&= \left[\frac{\mathbb{1}(A_i = 1)}{\pi_1(X_i)}(Y_i - \mu_1(X_i)) + \mu_1(X_i) - \psi_1\right] - \left[\frac{\mathbb{1}(A_i = 0)}{\pi_0(X_i)}(Y_i - \mu_0(X_i)) + \mu_0(X_i) - \psi_0\right] \\
&= \frac{\mathbb{1}(A_i = 1)}{\pi(X_i)}\big(Y_i - \mu_1(X_i)\big) - \frac{1 - \mathbb{1}(A_i = 1)}{1 - \pi(X_i)}\big(Y_i - \mu_0(X_i)\big) + \big(\mu_1(X_i) - \mu_0(X_i)\big) - \psi
\end{aligned} \tag{1.III.22}$$

As we let $\pi(X_i) = \pi_1(X_i) = P(A_i = 1|X_i)$ and $\psi_1 - \psi_0 = \psi$. With Equation 1.III.9 at the start of this section, we could derive the efficient estimator for the average treatment effect:

$$\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}(A_i = 1)}{\hat{\pi}(X_i)}\big(Y_i - \hat{\mu}_1(X_i)\big) - \frac{1 - \mathbb{1}(A_i = 1)}{1 - \hat{\pi}(X_i)}\big(Y_i - \hat{\mu}_0(X_i)\big) + \big(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\big)\right] \tag{1.III.23}$$

Equation 1.III.23 is the core equation in the whole thesis. The estimator in Equation 1.III.23 has several different names in different literature. It is, as we elaborated, the efficient causal

estimator, and it is also the **doubly robust (DR)** causal estimator since the estimator will be consistent if either our specification of $\pi(X_i)$ or our specification of $\mu_a(X_i)$ is correct (Chernozhukov et al. 2018b). Further, it is also called the **debiased machine learning (DML) estimator** as it eliminates the bias between the estimator and the true estimand with the double robustness details below. Meanwhile, some literature also calls this estimator the **Neyman-orthogonal estimator** of the ATE, as the EIF satisfies Neyman orthogonality, yielding this result. Finally, it is also an **augmented inverse probability weighting (AIPW)** estimator for the ATE: it starts with an outcome regression estimate (plug-in) and adds a correction term involving inverse propensity weights and outcome residuals. That correction term is precisely derived from the EIF for the treatment effect. We would like to call it **EIF-based doubly robust/debiased machine learning (DML) estimator**, showing that we derive the DML estimator from the EIF derivation.

E.2 Cross-fitting/ Sample-splitting for Nuisance Components

In practice, researchers use **cross-fitting** or **sample-splitting** techniques with one-step estimators [40]to help ensure the one-step estimator converges no slower than $1/\sqrt{n}$ rate with asymptotic normality so that the second-order and higher-order terms are negligible, especially when using complex ML models for nuisance components. With sample-splitting involved, the method is also called **double machine learning**. In other words, sample splitting helps maintain regularity and asymptotic linearity even when using flexible, high-dimensional

[40]Indeed, we use sample-splitting to ensure the convergence rate is no slower than $1/\sqrt{n}$. A replacement for the sample splitting technique is that we could assume that the empirical influence function falls into the **Donsker class**. A class of functions Φ is a Donsker class if the empirical process indexed by Φ converges in distribution to a Gaussian process: $\mathbb{G}_n(\Phi) = \sqrt{n}(\mathbb{P}_n - eP)\Phi$, as $\mathbb{P}_n$ denotes the empirical measure and eP is the true underlying probability measure. The Donsker class has the property $\mathbb{G}_n \rightsquigarrow \mathbb{G}$, converging to a Gaussian Process. In some ways, we could rewrite the empirical process as $\sqrt{n}(\mathbb{P}_n - eP)(\hat{\Phi} - \Phi)$, if we regard the convergence of the empirical measure $\mathbb{P}_n$ and the convergence of the influence function $\hat{\Phi}$ together at the rate of $1/\sqrt{n}$. Obviously, the Donsker class property ensures that the empirical process converges uniformly and at a controlled rate. When the influence function of an estimator belongs to a Donsker class, this guarantees that the empirical process does not exhibit erratic behavior and converges smoothly.

models for the nuisance parts. The doubly robust estimators with sample-splitting are a practical way to construct efficient estimators in complex settings, and they inherently rely on the theory of influence functions. Once again, we put the theoretical proof of how sample splitting achieved negligible higher-order terms in Appendix A.4 for interested readers.

Now we turn to focus on why Equation 1.III.23 yields the doubly robust estimator for the ATE. We know that if either the propensity score function or the conditional expectation function is correctly specified, the estimator will be consistent. When the propensity score function $\pi(X_i)$ is correctly specified, we have:

$$E\left[\frac{A_i}{\pi(X_i)}(Y_i - \mu_1(X_i)) \mid X_i\right] = 0$$

and

$$E\left[\frac{1-A_i}{1-\pi(X_i)}(Y_i - \mu_0(X_i)) \mid X_i\right] = 0.$$

Therefore, the first two terms of the influence function become mean-zero conditional on X, leaving:

$$\hat{\psi}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mu_1(X_i) - \mu_0(X_i)\right].$$

So even if $\mu_1(X_i)$ and $\mu_0(X_i)$ are misspecified, the terms involving the propensity score correct the bias introduced by the misspecified outcome models, resulting in a consistent estimator for the ATE.

Similarly, if the outcome regression models $\mu_1(X)$ and $\mu_0(X)$ are correctly specified, we have:

$$E[Y_i \mid A_i = 1, X_i] = \mu_1(X_i) \quad \text{and} \quad E[Y_i \mid A_i = 0, X_i] = \mu_0(X_i).$$

In this case, the terms $Y_i - \mu_1(X_i)$ and $Y_i - \mu_0(X_i)$ are mean-zero conditional on A_i and X_i. Thus, the first two terms of the influence function average out to zero:

$$\hat{\psi}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i}{\pi(X_i)}(Y_i - \mu_1(X_i)) - \frac{1-A_i}{1-\pi(X_i)}(Y_i - \mu_0(X_i)) + \mu_1(X_i) - \mu_0(X_i)\right].$$

Here, even if $\pi(X_i)$ is misspecified, the correctly specified outcome regression models ensure that the estimator is consistent for the ATE.

We also prove that with sample-splitting, the second and higher-order remainders are negligible in Appendix A.4.

F. Summary

The foregoing narrative outlines the full workflow for deriving and validating an EIF-based DML estimator for the ATE. In Chapter 2, we apply the same method to the ATE with survival data. In Chapter 4, we extend it to several estimands in causal mediation analysis that are widely used in social science applications. To make the approach portable, we conclude with a practical, step-by-step summary so that interested social scientists can derive EIF-based DML estimators tailored to their own research questions.

The target of our workflow is to derive the most efficient estimator, with the requirements of unbiasedness and as low variance as possible. Under regularity conditions, the efficient estimator reaches the relevant efficiency bound and is a regular and asymptotically linear estimator.

Regularity means that the estimator nudges smoothly as data grows, and under regularity, we define score functions to describe directions of change in the distribution.

With the restrictions of regularity, we look for asymptotic linearity so that we can capture the asymptotic variance with the expression of the influence function. We express the estimator's error as an average of influence functions, and the influence function tells us the impact of each observation. We also demonstrated that the influence function can be regarded as the derivative, the gradient, the first-order term, and the Neyman orthogonal scores.

We focus most on the influence function because we would like the most efficient influence function (EIF), which minimizes the variance. The efficient influence function is aligned with the score and resides in the allowed space of perturbations. We elaborated on the process of deriving the EIF in both saturated and non-saturated models and gave the EIF for the ATE.

With the EIF, we can construct an efficient estimator with one-step estimation and sample splitting to achieve the goals of robustness and efficiency. We demonstrated how the efficient estimator for the ATE is constructed, and we refer to it as the EIF-based double/debiased machine learning (DML) estimator for the ATE, which is our initial target.

Practically, we have the algorithm deriving the EIF-based DR causal estimator:

1. **Set up Input:**
    - Dataset $\{(X_i, A_i, Y_i)\}_{i=1}^{n}$, where X_i represents covariates, A_i represents treatment assignment (0 or 1), and Y_i represents outcomes.
    - Number of folds for cross-validation k.
2. **Split Dataset for Cross-Validation**

- Randomly split the dataset into k approximately equal-sized folds. Each fold will be used as an estimation sample while the remaining $k-1$ folds will be used for training.
- Label these folds as $\{D_1, D_2, \ldots, D_k\}$.

3. **Cross-Validation Loop**
   - Initialize lists to store the fold-specific estimates of $\hat{\psi}_1$ and $\hat{\psi}_0$.
   - For each fold j (from 1 to k):
     - **Training Set:** Combine all folds except D_j to create the training set $\{(X_i, A_i, Y_i)\}_{i \in \text{Training}}$.
     - **estimation samples:** Use fold D_j as the estimation samples $\{(X_i, A_i, Y_i)\}_{i \in \text{Estimation}}$.
4. **Estimate Propensity Scores in the Training Set**
   - Fit a propensity score model $\hat{\pi}(X)$ using the training set.
   - Calculate the estimated propensity scores $\hat{\pi}(X_i)$ for all i in the estimation samples.
5. **Estimate Outcome Regressions in the Training Set**
   - Fit outcome regression models $\hat{\mu}_0(X)$ and $\hat{\mu}_1(X)$ using the training set.
   - Calculate the predicted outcomes $\hat{\mu}_0(X_i)$ and $\hat{\mu}_1(X_i)$ for all i in the estimation samples.
6. **Calculate the Doubly Robust Estimator in the estimation samples**
   - Initialize two variables to accumulate the contributions from the treated and control groups in the estimation samples: $\hat{\psi}_1^j$ and $\hat{\psi}_0^j$.

- For each observation i in the estimation samples:
  - Compute the contribution for the treated group:

$$\hat{\psi}_{1i} = \frac{A_i}{\hat{\pi}(X_i)} \left(Y_i - \hat{\mu}_1(X_i)\right) + \hat{\mu}_1(X_i)$$

  - Compute the contribution for the control group:

$$\hat{\psi}_{0i} = \frac{1 - A_i}{1 - \hat{\pi}(X_i)} \left(Y_i - \hat{\mu}_0(X_i)\right) + \hat{\mu}_0(X_i)$$

  - Accumulate the contributions:

$$\hat{\psi}_1^j \leftarrow \hat{\psi}_1^j + \hat{\psi}_{1i}$$

$$\hat{\psi}_0^j \leftarrow \hat{\psi}_0^j + \hat{\psi}_{0i}$$

- Calculate the averages for the estimation samples:

$$\hat{\psi}_1^j = \frac{1}{n_j} \sum_{i \in \text{estimation}} \hat{\psi}_{1i}$$

$$\hat{\psi}_0^j = \frac{1}{n_j} \sum_{i \in \text{estimations}} \hat{\psi}_{0i}$$

- Store the fold-specific estimates.

7. **Aggregate Results Across Folds**

   - Calculate the overall estimates by averaging the fold-specific estimates:

$$\hat{\psi}_1 = \frac{1}{k} \sum_{j=1}^{k} \hat{\psi}_1^j$$

$$\hat{\psi}_0 = \frac{1}{k} \sum_{j=1}^{k} \hat{\psi}_0^j$$

8. **Compute the Average Treatment Effect (ATE)**

- Estimate the ATE:

$$\hat{\psi} = \hat{\psi}_1 - \hat{\psi}_0$$

9. **Output**

- The doubly robust estimator of the average treatment effect $\hat{\psi}$.

G. Further Discussions: Comparisons with MLE and GMM

So far, we have given out the process and algorithms for deriving the EIF-based, doubly robust machine learning estimator for causal inference. To make social scientists understand more about the advantages and disadvantages of the estimand, we here compare the method with two other very commonly used methods: the maximum likelihood estimation (MLE) and the Generalized Methods of Moments (GMM), so that researchers could understand the benefits and limitations of the doubly robust estimators.

Maximum likelihood estimation (MLE) is the most commonly used approach for estimation in social science due to its merits in interpretability: it has a specified model for the data and, based on maximizing the likelihood function, it yields the estimated parameters, which intuitively present the size of effects. However, it has the highest requirement (assumption) among the three methods: the model it fits *must* be correctly specified. If the model is correctly specified, the MLE model is consistent, asymptotically unbiased, and asymptotically efficient among regular parametric estimators. As for validity, MLE's validity is only good as the model and robust (sandwich) standard errors can fix some variance misspecification, but not bias from a wrong mean structure.

Generalized Method of Moments (GMM) is comparatively more flexible than the MLE model: it does not require a fully specified probability distribution of the data (the full model); instead, it focuses on the moment condition: the population expectations that equal zero at the true parameter. If the moment conditions are correct and the parameters are identified, GMM is consistent and asymptotically normal. However, GMM is usually less statistically efficient than MLE if the MLE assumptions are correct because GMM uses less information, but within the class of moment-based estimators, an optimally weighted GMM is the most efficient. As for validity, GMM relies on the credibility of the moments and often uses the sandwich covariance estimator. Due to the difference in efficiency, GMM estimates often have larger standard errors than MLE for the same problem. In practice, most studies using the GMM estimator (especially in causal inference) have the instrumental variable identification designs, as the moment condition can be regarded as instrumental exogeneity conditions, and researchers could obtain a consistent treatment effect without specifying the distribution of the outcome. If the instrument is valid, or if the endogeneity or likelihood functions are complex, under a large sample size, the GMM works better than the MLE. However, with weak instruments, finite-sample bias and size distortions can be severe, and even asymptotically problematic, so diagnostics and stronger instruments are essential.

The EIF-based double robust (DR) machine learning estimator, which we introduce in this thesis, "hedges" the modeling risk from the MLE and the GMM estimators by combining the nuisance function through the influence function updates: if one nuisance function (the propensity model or the outcome model) is mis-specified but the other is well specified, the target estimate remains consistent. However, the correct identification of the DR methods in causal inference relies on the causal inference assumptions (consistency,

positivity, and unconfoundedness), and DR methods cannot circumvent the assumptions. Because the DR estimators are all EIF-based in this thesis, they are regular and asymptotically linear (RAL) estimators, and when both models are correctly specified, the DR estimators are efficient for the target parameter. As for validity, compared to the MLE estimator, which purely relies on the probability distribution, the DR estimator protects against misspecifying the form of one of the models. Hence, its tolerance to the risk of model misspecification is relatively higher than that of the MLE and GMM models, as consistency can still be achieved if one model performs well. So if one (propensity/outcome) model is wrong, DR estimation will still give the unbiased estimates, and if both are correct, DR estimation will give near-MLE level efficiency for the causal parameters– and its ability to balance robustness and efficiency is the crucial merit adopted by many researchers.

IV. Conclusion

In this chapter, we introduce the basic ideas and mathematical tools to perform the causal inference in an efficient/doubly robust way from the observational/ survey data in social science. Intrinsically, to correctly identify the causal estimand from the observational data (via the statistical estimand), for all the methods, no matter whether it is a parametric (regression) model, a nonparametric (machine learning) model, or a semi-parametric (bias reduction) model, the key is the same: to correctly specify/identify the two models: the propensity score model which allocates cases into the treatment and control group and the outcome result model which specifies the factual or counterfactual outcomes.

The doubly robust estimator provides a toolbox that, compared to the previous models which does not have the correct specification of both the propensity score model and

the outcome result model, we only need to correctly specify one of them. This makes social science research, especially under theoretical-driven studies, much more convenient in yielding an unbiased and efficient estimator for the treatment effect. In empirical studies, researchers are always afraid of omitted variable bias when specifying the outcome model, but if the researchers' theories and hypothesis could make the propensities allocating to the treatment and control groups deterministic, then the results are robust (this is pretty like the idea behind the local average treatment effect estimation).

In the chapters below, we will again use asymptotic analysis methods (with the score and influence function) to yield efficient estimators under different data structures and model settings. In this chapter, we just give readers from social science backgrounds a preliminary introduction (with necessary mathematical transforms) to this area. The method for semiparametric doubly robust target double machine learning (Kennedy 2022a) is definitely one of the fastest developing areas in statistics, econometrics, data science, and relevant discipline's methodological discussions. Like other disciplines, social scientists and demographers need the toolbox to have better causal estimations of their research interests.

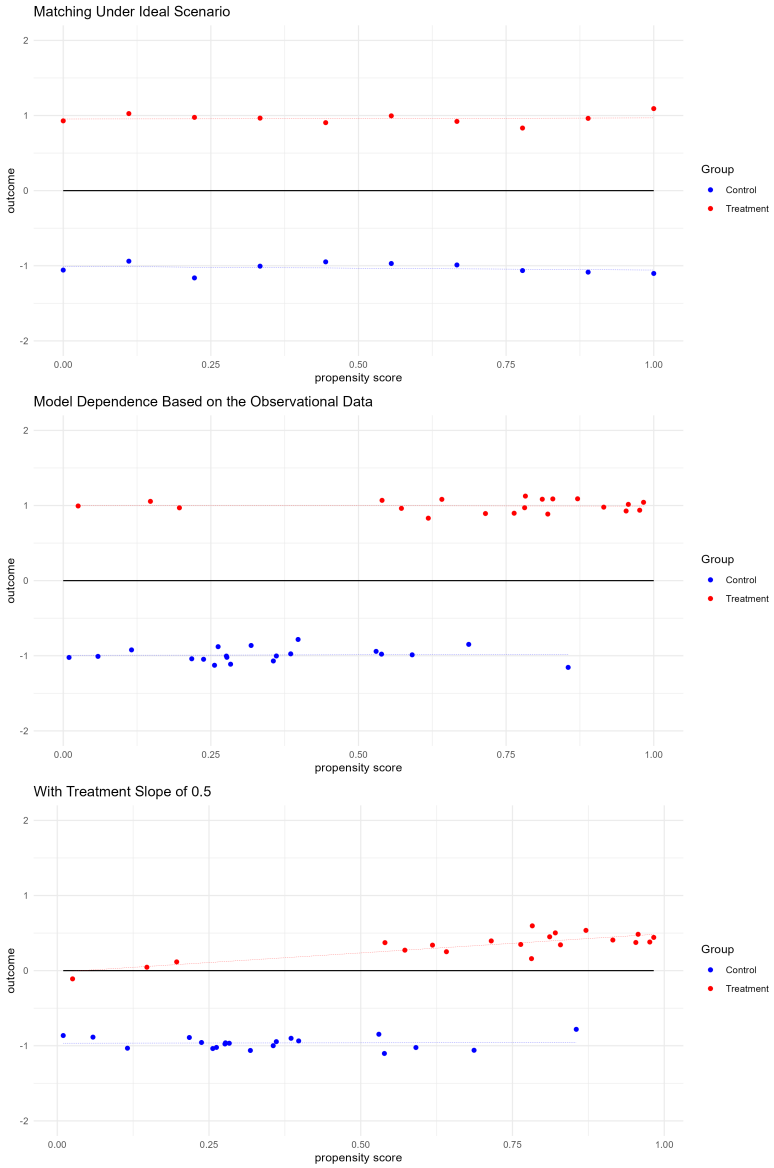


Figure 1.1: Illustrations on propensity score matching

Note: The upper panel shows the ideal scenario for propensity score matching, where for each individual in the treatment group, we could find a corresponding individual in the control group with the same propensity value and achieve one-on-one matching. However, as the lower two panels show, with observational data to train and predict the propensity function, the distributions of the treatment and control individuals are unequal along the propensity score, making the one-on-one ideal matching infeasible. But if for the treatment group and for the control group, the outcome is irrelevant to the propensity scores (the middle panel), using any method to get $\hat{\pi}_a$ yields the same unbiased estimation for π_a. Otherwise, as the lowest panel indicates, the choice of $\hat{\pi}_a$ yields bias.

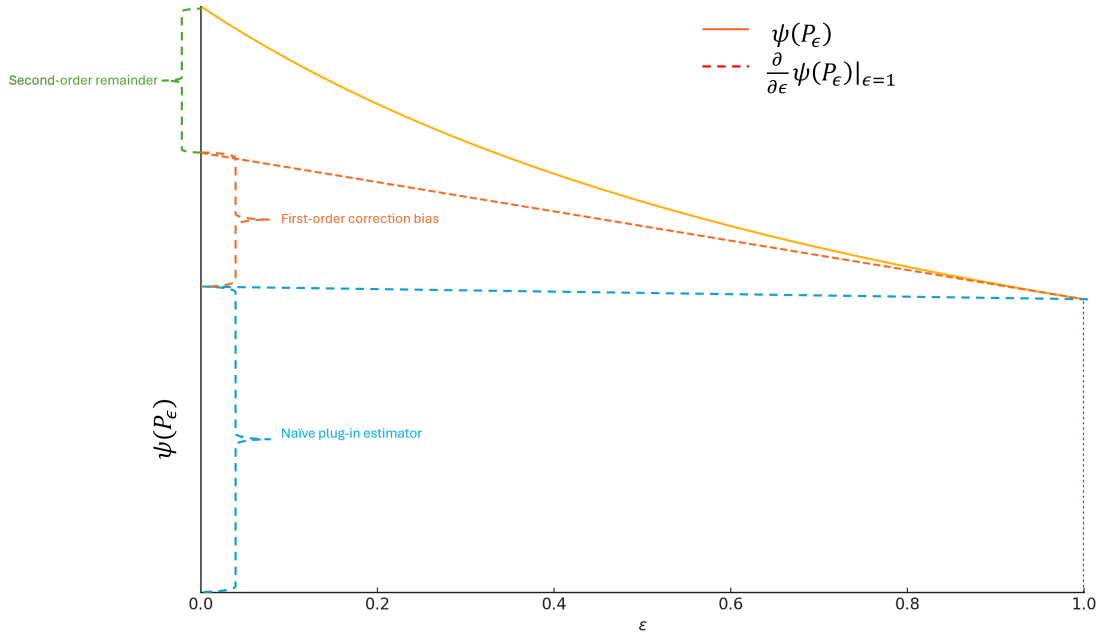


Figure 1.2: Illustration on Distributional Taylor Expansion

Note: This is a simplified illustration of the Distributional Taylor expansion. Indeed, P_ϵ and P are two measures, but we simplify them as two values. If so, the direction of the score function should be horizontal, and the direction of the gradient of the score function on the estimator should be the tangent line of ψ at $\psi(P)$, which has the same direction as the influence function. Thus, the angle of the score and the influence function points in the same direction as the gradient.

Chapter 2

Estimating Heterogeneous Treatment Effects for Survival Data with Twice Doubly Robust Estimator

I. Causal Inference for Survival Data

In Chapter 1, we discussed the general concepts of causal inference and the efficient (doubly robust) estimator for the average treatment effect (ATE). In this chapter, we turn to a very specific extension of the framework to analyze the application of the efficient/debiased/doubly robust causal estimator in survival analysis.

Causal inference for survival outcomes is ubiquitous in public health and medical research, as researchers are interested in how the trial changes patients' potential longevity or disease progression. It is also applicable in social science areas like demography and public policy when studies are interested in how a social variable may change the length of time spent in the initial state before transitioning to another state (for instance, from unemployment to employment).

In medical studies, researchers may have fully observed data with **randomized controlled trials (RCT)**. For example, they test the effectiveness of a drug on the survival of

mice. Under such a scenario, researchers implement the treatment for one of the two randomized groups (making the other group without implementation the control), set the time of the implementation as $t = T_0$, and record the time interval between the start of T_0 and the demise time for both the treatment and control groups. Then, researchers could compare the expectation of the survival time between the treatment and control groups, using statistical methods to make inferences from the sample parameters to the population and derive the causal estimand for the drug.

Sometimes, the data structure that the researchers encounter is more complex. On the one hand, researchers often use observational data for causal inference, which does not specify the assignment of treatment and control with randomization. We have discussed the techniques of causal inference with observational data in Chapter 1[1]. Specifically for the observational survival data, researchers can encounter the missing data scenarios respectively called **truncation** and **censoring**. Truncation is a sampling/inclusion restriction: some units never enter the dataset because their event time may not be included in our observational window. Specifically, we have left-truncation, in which some of the subjects are not observed because their events occurred before our observational window started. **Censoring** refers to partial observation: we can observe the subject, but the exact event time is not in our observational window. Specifically, we have the case for right-censoring, in which some of the subjects are in our observation window but their exact event time is after our observation ends. We will give a detailed definition of truncation and censoring in Section II. In short, in the above example exploring the drug's treatment effect on the mice, for observational data, we may observe that for some mice the treatment had been implemented

[1]Also, as we have discussed in the introductory chapter, the efficient estimators for the average treatment effect given by the difference between the treatment and control expectations are the same for the RCT and the observational data.

before we started our observation window and they enter the dataset if they survived up to our entry time (left truncation), and for some mice, they remain alive at the end of follow-up and thus their exact demise times are unknown (right censoring).

Therefore, like the missing counterfactuals in causal inferences, for the survival data containing truncation and censoring, we need an unbiased estimator to estimate the survival time (van der Laan and Robins 2003). This is the twice doubly robust estimator for the causal inference on the truncated and censored survival data we introduced in this chapter: we apply a first doubly robust estimation on the causal effect estimation and a second doubly robust estimation on the survival curve estimation.

The organization of this chapter is as follows. Section II reviews the assumptions we mentioned in the introduction chapter for causal inference and the efficient causal inference estimator. Section III reviews the notations, basic concepts, and basic assumptions for survival data analysis and machine learning methods. We will elaborate on the derivation of the mean survival time and the loss function based on different parametric and non-parametric survival models. Section IV derives the efficient/doubly-robust estimator for the survival analysis. Section V summarises the algorithm to apply the twice doubly robust estimator to infer the average treatment effects and heterogeneous treatment effects with the survival outcomes. In Section VI, we run simulations comparing our doubly robust estimation model with other model settings.

II. Assumptions and Doubly Robust Estimator for Causal Inference

This part serves as a very simplified review of Chapter 1. From that chapter, we have known that the causal inference with observational data indeed uses the observational estimator

for a specific treatment status $E[Y_i|A_i = a, X_i]$ as the unbiased estimator for the statistical estimand $E[Y|A = a, X]$ and makes inference on the causal estimand of the treatment status $E[Y(a)]$ (notations here: the observational dataset: $Z_i = (X_i, A_i, Y_i)$, and the statistical measurable set: $Z = (Y, A, X)$; Y denotes the dependent variable/ predictor/ outcome, X denotes the covariates/independent variable/features, and A denotes the treatment assignment status), solving the fundamental problem in causal inference that we could not observe $Y(1)$ and $Y(0)$ (as the treatment is dichotomous) simultaneously. To make the statistical estimand inferable on the causal estimand, we have the assumptions for causal inference, which we elaborated as Assumption 1.II.4 in our last chapter:

Assumption 2.II.1 (Causal Inference Assumptions) *Suppose a statistical DGP $Z = (X, A, Y)$, in which X denotes the covariates, A denotes the treatment, and Y denotes the outcome. To make the statistical estimand $\psi(\mathrm{P}) = E[E_X[Y|A = 1, X]] - E[E_X[Y|A = 0, X]]$ equivalent to the causal estimand $\psi^*(\mathrm{P}^*) = E[Y(1)] - E[Y(0)]$ from the causal DGP $Z^* = (X, A, Y(1), Y(0))$ (where $Y(1)$, $Y(0)$ denote the potential outcomes under treatment and control, respectively), we need the following hypotheses:*

1. *Positivity: the probability to be assigned to treatment and control group conditioned on the covariates, is a positive number between 0 and 1:*

$$P(A = 1|X) \in (0, 1) \quad P(A = 0|X) \in (0, 1)$$

2. *Consistency: the potential outcome under the treatment received is the same as the observed outcome:*

$$Y = Y(A)$$

3. *Unconfoundedness: conditional on a set of observed covariates X, the potential outcomes $Y(1)$ and $Y(0)$ are independent of the treatment assignment A:*

$$\{Y(1), Y(0)\} \perp\!\!\!\perp A|X$$

Meanwhile, we use the estimator from the observational data to yield an unbiased estimation on the statistical estimand. Under regularity and asymptotic linearity conditions, an efficient estimator attains the relevant efficiency bound within the class of regular asymptotically linear estimators. The EIF-based estimator for the average treatment effect (ATE) is also doubly robust, debiased, and Neyman-orthogonal:

$$\hat{\psi}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i}{\pi(X_i)}(Y_i - \mu_1(X_i)) - \frac{1-A_i}{1-\pi(X_i)}(Y_i - \mu_0(X_i)) + \mu_1(X_i) - \mu_0(X_i)\right]. \tag{2.II.1}$$

The nuisance functions are $\pi(X_i) = P(A = 1|X)$ and $\mu_a(X_i) = E[Y_i|X_i, A_i = a]$. Equation 2.II.1 estimates the causal estimand for the ATE ($E[Y(1) - Y(0)]$) via the statistical estimand ($E[Y|A = 1, X] - E[Y|A = 0, X]$). Suppose the outcome variable Y_i in the observational data indicates the survival (time-to-event) of the individual; this is the data structure we are discussing in this chapter.

III. Notations and Basic Concepts of Survival Data Analysis

A. Discrete and Continuous Survival Outcomes

Survival analysis deals with the time-to-event outcome: the death of an animal/ species, the survival time of an unstable atom, and the failure time of a system/ a machine are all time-to-event outcomes. Observing the process of the survival event and making statistical inferences about the population is the key to survival analysis. Hence, there are two ways to observe the survival outcome: for the first method, and usually for the short-life objectives, we could calculate the accurate **failure time**, in which case we treat the survival outcome

as the **continuous** survival outcome; or we could observe whether the event has occurred within specific intervals like hours, days, weeks, and years. We have a dichotomous indicator on whether the object failed, and another indicator showing the number of the interval. We call this the **discrete survival outcome** (Allison 1982; Willett and Singer 1995; Jenkins 1995; Singer and Willett 1993).

For discrete survival outcomes, the outcome Y in the statistical estimand can be inferred as the probability of survival grouped by the specific time interval using the frequencies of survived cases in the dataset. Hence, for causal inference, we can identify the causal estimand as $E[Y_t(1) - Y_t(0)]$ (where $t \in [0, +\infty)$ and $t \in \mathbb{Z}^+$; thus t stands for the specific time interval), which is the difference in the probability of survival for the treatment and control groups (can also be denoted as $P_t(1) - P_t(0)$, where P is the probability of failure or survival). Therefore, for every time interval, we may have a doubly robust/efficient estimator on the ATE, and the ATE will change with the change in the time interval. Social scientists applied this data structure to conduct policy analysis (Andersen et al. 1993; Box-Steffensmeier and Jones 2004). For instance, in Wolfers (2006), he initially used a fixed-effect discrete-year setting to study the change of divorce laws on the change in divorce rates. Indeed, the paper measured the probability difference of not being divorced at each time interval between the policy-affected and not-affected groups.

Sometimes, in social science, even when the survey is conducted at fixed intervals, and we observe discrete survival outcomes, we treat the survival outcomes as continuous and use continuous survival models (Allison 2014; Singer and Willett 2003; Box-Steffensmeier and Jones 2004; Kalbfleisch and Prentice 2002). This usually requires further assumptions on the data distribution. In the following chapter, we mainly discuss the estimator for the

continuous survival outcomes. The survival outcome, under the continuous scenario, can be written as a function of t. Suppose T stands for the time the event of research interest occurs, and we can define the survival function $S(t)$ on the domain $t \in [0, T]$ and $t \in \mathbb{R}$ simply as the survival probability (rate) at time t:

Definition 2.III.1 (Survival Function)

$$S(t) = P(T \geq t)$$

For the conditional survival function, we have:

$$S(t|X) = P(T \geq t \mid X)$$

Since the cumulative distribution function (CDF) for the random variable T is defined as: $F(t \mid X) = P(T \leq t \mid X)$, we could rewrite the conditional survival function $S(t|X)$ as:

$$S(t|X) = 1 - F(t|X) = 1 - \int f(t|X)\,dt^{2}.$$

The term $f(t)$ refers to the probability density function of t, which has a relationship with the survival function, deriving from the equation above:

$$f(t) = -\frac{dS(t)}{dt}. \tag{2.III.2}$$

Estimating the survival function to yield the ATE that satisfies the theoretical requirements of the data-generating process is the key to the causal inference with survival outcomes (Klein and Moeschberger 2003). Obviously, like the discrete survival outcomes, we

[2] As the mathematical transformation of the survival functions is irrelevant with the covariates X, thus, if we are not estimating specific parameters with the covariates, we only discuss the unconditional survival function $S(t)$, the CDF $F(t)$, the PDF $f(t)$, the mean survival time $E[T]$, and the hazard function $h(t)$. When we are estimating parameters from the covariates, for instance, when we need the covariates X to estimate the rate parameter λ in the exponential distribution, we will give the conditional expression, for example, the PDF $f(t \mid X)$.

could still set a specific time point $t = T$ and compare the survival probability for the treatment group $S^1(T)$ and the control group $S^0(T)$ for the ATE [3]. For instance, Wu and Wen (2022) also examined the change of divorce laws on the longevity of marriage. They construct the continuous survival function based on the Cox-Proportional Hazard Assumption (see below), set the observational year as the end of year seven, and applied the difference-in-difference (DID) framework to measure the causal effect [4].

Our goal for survival analysis with observational data is to have the survival function estimated from the observational data $\hat{S}(t \mid A = a, X)$ to be the estimator for the true survival function with specific treatment status $S^{A=a}(t)$. For the independent and identically distributed (IID) samples, the group-level survival function (for either the treatment or the control group) is the product of all the single survival curves for the individuals in that group, namely,

$$\hat{S}(t|A = a, X) = \prod_{i:A_i=a} \hat{S}_i(t|X_i)$$

In statistics (and machine learning), we use the **loss function** to quantify the discrepancy between the observational data and the predictions on the survival function, which takes the form of the negative log-likelihood of the survival function:

$$\mathfrak{L} := -\sum_{i=1} \log L(\theta) = -\sum_{i=1} \log(f(t;\theta)).$$

In which θ refers to the parameters we estimate. Below, we will introduce several common parametric and nonparametric methods for estimating the survival function using ob-

[3] For the efficient estimators for the ATE in Equation 2.II.1, indeed, we are using the estimated $\hat{S}^0(T|X_i)$ and $\hat{S}^1(T|X_i)$ to simply be $\hat{\mu}_0(X_i)$ and $\hat{\mu}_1(X_i)$ for the estimations. We here put the indicator for treatment status $A = a$ in superscript to differentiate $S^0(T)$ as the survival function for the control group at time T and $S_0(T)$ which is the baseline survival function $S_0(T) = \exp(-H_0(T))$.

[4] As mentioned in the Chapter 1, the DID framework solves the violation of the positivity assumption: in the states where the divorce laws were enforced, there were no counterfactual cases in the control group, and the cases in the neighborhood states have to be included in the research as the control.

servational data.

B. Estimating the Complete-Case Loss and Mean Survival Time

Besides comparing the difference in survival rates, we could also construct the **mean survival time** separately for the treatment and control groups [5]. Usually, for parametric and nonparametric models, without any transformation, they are defined on the domain of $(-\infty, +\infty)$ (for instance, if the outcome is defined on $[0,1]$, we will use sigmoid (logistic) or probit models to transform the original predicted value from the covariates). The time-to-event data (survival time T) are defined on the domain of $[0, +\infty)$. Therefore, we need appropriate model specifications for the non-negative nature of the survival time and derive the mean (expected) survival time. According to the definition of the expectation, we could derive the relationship between the mean survival time and the survival function as:

$$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty t \left(-\frac{dS(t)}{dt} \right) dt = \int_0^\infty S(t) dt. \tag{2.III.3}$$

Thus [6], $\hat{E}[T|X] = \int_0^\infty \hat{S}(t \mid X) dt$.

B.1 Parametric Models and the Hazard Function

If we apply a parametric model to model the survival data, a prerequisite assumption is that the survival data follow specific types of distribution:

[5] In many studies, researchers traditionally estimate the median survival time (Mao et al. 2018), or the half-life during the survival process to represent the survival function. According to the definition, the half-life time is $T_m : S(T_m) = 1/2$, which is more intuitive and more revealing of the central tendency than the expectation (as the survival time data are usually right-skewed: some cases have longer survival time). However, when we use the median survival time to capture the treatment effect, we could still have the individual treatment effect (ITE) for individuals, but on the group level (the divergence between the treatment group and the control group), we are capturing the survival median treatment effect (see Hu et al. 2021). This is still feasible as long as we can yield the doubly robust/efficient estimator for the median treatment effect (for instance, we could use the heuristic method mentioned in the introduction chapter to accomplish this). However, this is beyond our discussion here as we focus on the treatment effect based on the expectation.

[6] Because $\int_0^\infty t \left(-\frac{dS(t)}{dt} \right) dt = \int_0^\infty -t dS(t) = -[tS(t) \mid_0^\infty - \int_0^\infty S(t) dt] = \int_0^\infty S(t) dt$.

Assumption 2.III.1 (Parametric Modeling Assumption) *If we choose to model survival data with the parametric models, we assume that the survival data follow a particular (known) distribution.*

We introduce three common parametric models for survival data: log-normal (Meeker 1998), exponential (Klein and Moeschberger 2003), and Weibull (Lawless 2003). The log-normal distribution is somewhat like the probit modeling for the data on $[0,1]$: which originally spreads over the entire real number line to map it onto a different range. Specifically, for the log-normal distribution, the transformation shifts the range to $[0,+\infty)$. Therefore, the log-normal distribution suggests

$$\log(T \mid X) \sim \mathcal{N}(\mu(X), \sigma^2(X)).$$

This model setting is similar to the logistic transformation when the outcome is distributed in $[0,1]$. If $\log T \mid X$ is normally distributed, the mean survival time is

$$\mathbb{E}[T \mid X] = \exp\left(\mathbb{E}[\log T \mid X] + \frac{1}{2}\sigma^2(X)\right) = \exp\left(\mu(X) + \frac{\sigma^2(X)}{2}\right).$$

If we assume that the time-to-event data is distributed log-normally, we thus have the survival function as:

$$S(t|X) = P(T > t|X) = 1 - \Phi\left(\frac{\log(t) - \mu(X)}{\sigma(X)}\right)$$

in which Φ refers to the CDF for the standard normal distribution. Thus, the PDF is:

$$f(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\log(t)-\mu)^2}{2\sigma^2}}$$

The loss function, therefore, is the negative log-likelihood:

$$\mathfrak{L} = -L(\mu,\sigma) = n\log(\sigma) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log(t_i) - \mu)^2 + \frac{n}{2}\log(2\pi)$$

Also, in some literature, $T \mid X$ can be approximated as the **exponential distribution** with scale parameter λ: $T \mid X \sim \text{Exponential}(\lambda(X))$, where $\lambda(X)$ is usually the exponential of the linear combination of the covariates $\mu(X)$ (Klein and Moeschberger 2003). The exponential distribution clearly is defined on the domain of $[0, +\infty)$ and could also address the skewed cases with adjustments on the scale parameter λ. According to the definition of the exponential distribution, the survival function is expressed as $S(t|X) = \exp(-\lambda(X)\,t)$ [7], and the mean for exponential distribution is the inverse of $\lambda(X)$:

$$E[T|X] = \frac{1}{\lambda(X)} = \frac{1}{\exp(\mu(X))}$$ [8]

The third parametric model for the survival time usually takes the survival time T as a **Weibull distribution**. The Weibull distribution is an extension of the exponential distribution. In the Weibull distribution, we have the scale parameter λ and the shape parameter κ, which controls the change of probability of event occurrence over time. The survival function for the Weibull distribution is: $S(t|X) = \exp\left(-\left(\frac{t}{\lambda(X)}\right)^{\kappa(X)}\right)$. Thus, an exponential distribution is a special Weibull whose $\kappa = 1$ and λ is the inverse of the scale parameter in the Weibull. To make it more clear why including parameter κ is crucial, we first introduce the concept of the **hazard function**.

[7] The probability density function for the exponential distribution is

$$f(t) = \lambda \exp(-\lambda t); \quad t \geq 0$$

Therefore, $f(t|X) = \lambda(X)\exp(-\lambda(X)\,t)$. With the relationship between the PDF and the survival function, we have $f(t|X) = -\frac{dS(t|X)}{dt}$, thus,

$$S(t|X) = \int_0^\infty f(t|X)\,dt = \int_0^\infty \lambda(X)\exp(-\lambda(X)\,t)\,dt = \exp(-\lambda(X)\,t).$$

[8] Because $S(t|X) = \exp(-\mu(X)\,t)$,

$$E[T|X] = \int_0^\infty S(t|X)\,dt = \int_0^\infty \exp(-\lambda(X)\,t)\,dt = \left[-\frac{1}{\lambda(X)}\exp(-\lambda(X)\,t)\right]_0^\infty = \frac{1}{\lambda(X)}.$$

As the name indicates, the hazard function indicates the instantaneous possibility (rate) at which the event occurs at time t (of course, with the condition that the case has survived until time t). Therefore, the hazard function is defined as a derivative (Box-Steffensmeier and Jones 2004):

Definition 2.III.2 (Hazard Function) *The hazard function is defined as the limit of the probability that an event occurs in a small time interval, divided by the length of that interval, given that the event has not occurred before time t:*

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Based on the definition of the hazard function, we can also derive the relationship between it and the survival function. According to the rule of conditional probability, the numerator of $h(t)$ can be transformed as: $P(t \leq T < t + \Delta t \mid T \geq t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} = \frac{f(t)\Delta t}{S(t)}$. Therefore,

$$h(t) = \lim_{\Delta t \to 0} \frac{\frac{f(t)\Delta t}{S(t)}}{\Delta t} = \frac{f(t)}{S(t)}. \tag{2.III.4}$$

We can regard that the survival probability is the opposite **cumulative hazard**, or regard the hazard as the derivative of the demise (opposite of survival), as the hazard function describes the instantaneous case of demise at the specific time. This can be also shown from the mathematical transformation, combining Equations 2.III.2 and 2.III.4,

$$h(t) = \frac{f(t)}{S(t)} = \frac{-\frac{dS(t)}{dt}}{S(t)} = -\frac{dS(t)}{S(t)}\frac{1}{dt} = -\frac{d\log S(t)}{dt}; \tag{2.III.5}$$

$$S(t) = \exp\left(-\int_0^t h(u)\,du\right) = \exp\left(-H(t)\right). \tag{2.III.6}$$

With the relationship between the survival and the hazard function, we now can illustrate that for the Weibull distribution, if $\kappa = 1$, we indeed assume a constant hazard: the

hazard function does not change over time for the exponential distribution since $h(t|X) = -\frac{d\log S(t)}{dt} = -\frac{d\cancel{\log}\cancel{\exp}(-\frac{t}{\lambda(X)})}{dt} = \frac{1}{\lambda(X)}$ (constant over time). Correspondingly, $\kappa > 1$ indicates that the hazard increases over time, and hence, the event is more likely to occur as time progresses, while $\kappa < 1$ suggests that the hazard decreases over time. The mean survival time is [9]:

$$E[T|X] = \lambda(X)\Gamma\left(1 + \frac{1}{\kappa(X)}\right).$$

in which $\Gamma(\cdot)$ denotes the Gamma function. Usually, we use two sets of functions to fit $\lambda(X)$ and $\kappa(X)$ separately. For example, we use the exponential transformation of the two sets of linear combinations of X: $\lambda(X) = 1/\exp(X\beta)$ and $\kappa(X) = \exp(X\gamma)$, or nonparametric machine learning models to obtain the results.

For the Weibull-class distribution (including exponential), the loss for the survival function is:

$$\mathfrak{L} = -\sum_{i=1} \log(f(t_i; \lambda, \kappa))$$

and as the survival function for the Weibull distribution is $S(t|X) = P(T > t|X) = e^{-\left(\frac{t}{\lambda(X)}\right)^{\kappa(X)}}$, we have its PDF as:

$$f(t|X) = \frac{\kappa(X)}{\lambda(X)}\left(\frac{t}{\lambda(X)}\right)^{\kappa(X)-1} e^{-\left(\frac{t}{\lambda(X)}\right)^{\kappa(X)}}$$

Thus, the empirical loss is given as follows:

$$\mathfrak{L} = -L(\lambda, \kappa) = n\log(\lambda) + n\log(\kappa) + (\kappa - 1)\sum_{i=1}^{n} \log(t_i) + \sum_{i=1}^{n}\left(\frac{t_i}{\lambda}\right)^{\kappa}.$$

[9] Since $E[T \mid X] = \int_0^\infty S(t \mid X)dt = \int_0^\infty \exp(-(\frac{t}{\lambda(X)})^{\kappa(X)})dt$. Let $u = (\frac{t}{\lambda(X)})^{\kappa}(X)$, thus, $t = \lambda(X)u^{1/\kappa(X)}$ and $dt = \frac{\lambda(X)}{\kappa(X)}u^{(1/\kappa(X))-1}du$. Substituting the terms in the integral term, we have $E[T \mid X] = \int_0^\infty \exp(-u) \cdot \frac{\lambda(X)}{\kappa(X)}u^{(1/\kappa(X))-1}du$. The Gamma function $\Gamma(\cdot)$ is defined as: $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$. Therefore, we have:

$$E[T \mid X] = \frac{\lambda(X)}{\kappa(X)}\Gamma\left(\frac{1}{\kappa(X)} + 1\right) = \lambda(X)\Gamma\left(1 + \frac{1}{\kappa(X)}\right).$$

B.2 Nonparametric and Semiparametric Model Specifications

Usually, researchers cannot approximate the distribution of the survival data into any known distribution, or the parametric methods are not applicable. A very common non-parametric modeling for the survival data is called the **Kaplan-Meier method**, which is quite similar to the discrete method but yields the survival function. Even though many studies on survival data do not apply the Kaplan-Meier method, researchers often draw the **Kaplan-Meier curve** to show the trend of the survival probability over time Kaplan and Meier (1958). The idea for the Kaplan-Meier method is quite simple: it just counts, stepwise, the number of cases **at risk** (meaning that they have survived up to the observation time) n_j and the number of cases the event occurs d_j at time t_j. So, the survival function from the Kaplan-Meier method is:

$$S(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

With the condition that the survival at each time point t_j is independent. Since we observe the conditions of survival at every t_j, the mean survival time can be approximated by the sum of the survival probability at t_j times the interval length from the last observation point t_{j-1} to t_j:

$$E[T] = \sum_{T_0}^{T_{\max}} S(t_j)(t_j - t_{j-1})$$

Geometrically, the mean survival time is represented by the area under the survival curve. This applies equally to both parametric models and the Kaplan-Meier model (Efron 1988). However, the Kaplan-Meier method does not involve a loss function, unlike the parametric models or the semiparametric Cox Proportional Hazard model to be mentioned below, but there are several loss function analogies with the Kaplan-Meier method to compare the predicted outcome and the true statistical estimand. For instance, we can use the concordance index or the log-rank tests for nonparametric models to analyze the divergence between the

Kaplan-Meier estimates and the true statistical estimand.

We may also assume some flexible model that does not specify the distribution of the survival data but has some restrictions that could incorporate the covariates, and even better, it can estimate the effect of a specific covariate on the survival data. This is the most-used model in survival analysis: the **Cox Proportional-Hazard (Cox-PH)** method(Cox 1972: 1997), which is, a semiparametric model as it specifies parameters partially in the model. As the name suggests, although it does not assume any distributions for the survival and hazard functions, it restricts the hazard function to be proportional over time across covariates:

Assumption 2.III.2 (Proportional Hazard Assumption for Cox-PH Model) *Cox Proportional Hazard models assume the effect of a covariate on the hazard function is multiplicative and does not change with time. Mathematically, the baseline hazard form, $h_0(t)$ and the conditional hazard $h(t \mid X)$, both at t, have the following relationship:*

$$h(t \mid X) = h_0(t) \exp(X\beta)$$

and the term $\exp(X\beta)$*, the* ***risk score****, is constant over time.*

For instance, suppose we want to analyze how gender affects survival time. We could derive the hazard ratio between men and women. The assumption here is that the hazard ratio between men and women is constant at any time or interval. For instance, if we suppose men have higher risks than women with the hazard ratio of $2:1$, then at any given time, the hazard ratio between men and women is $2:1$.

A very convenient way to capture the causal effect of survival data in the previous sociological and demographic studies is to yield the hazard ratio between the treatment and

the control group. The estimator is the **marginal hazard ratio (MHR)**. However, most of the time, MHR is a biased estimator for causal estimations as it violates the assumption of unconfoundedness (the treatment variable without restricting exogeneity is not independent of the observed outcomes).

A method to circumvent this is similar to the Inverse Probability Weighting (IPW) or two-step regression: we first predict the propensity for the treatment given the exogenous covariates and use the predicted propensity for the treatment from the first step in the Cox-PH model to yield the coefficient β. If the unconfoundedness assumption holds, this will yield an unbiased causal estimand, as the IPW estimator itself is unbiased. However, this method has two shortcomings: first, as we noted in the introduction chapter, the IPW estimator is not an efficient/doubly robust estimator. Thus, the standard error of the estimator will be larger than the most efficient one. More importantly, the causal effect derived from this method is quite hard to interpret. The coefficient suggests the ratio between the hazard functions, which is a relative risk. If we want to show the direction of the treatment (increases or decreases the risks), we could use the estimation of the hazard ratio; however, if we need to interpret quantitatively how much difference the treatment changes for survival, the hazard ratio is not enough.

However, the Cox-PH model could be the foundation for us to calculate the survival function $S^{A=a}(t|X)$ and mean survival time $E_{A=a}[T|X]$. Cox-PH method provides a toolbox to capture the hazard function $h(t \mid X$, and using Equation 2.III.6, we could get the hazard function for $S(t \mid X)$, and with Equation 2.III.3 we can get the estimated mean survival time. Compared to the estimation of the hazard ratio, if the survival function is required, we need to figure out the baseline hazard $h_0(t)$. Here, we use the **Breslow method** (Breslow 1975)

for the estimation on $\hat{h}_0(t)$ with the observational data. In short, the Breslow estimator suggests that the estimated cumulative baseline hazard $\hat{H}_0(t)$ is the sum of the inverse of the risk scores for cases at risk at the particular time t_j. Mathematically, it is expressed as :

$$\hat{H}_0(t) = \sum_{t_j \leq t} \frac{1}{\sum_{i \in R_j} R_i}$$

In which R_i denotes the risk score for individual i. As previously mentioned, if with linear estimation, the risk score is the exponentiated linear predictor $\exp(X_i\hat{\beta})$. Therefore, the baseline hazard function at t_j is estimated as:

$$\hat{h}_0(t_j) = \hat{H}_0(t_j) - \hat{H}_0(t_{j-1})$$

Thus, the estimated survival function is:

$$\hat{S}(t \mid X_i) = \hat{S}_0(t)^{\exp(X_i\hat{\beta})} = \exp\left(-\hat{H}_0(t)\right)^{\exp(X_i\hat{\beta})}.$$

For Cox-PH models, we usually derive the loss for the hazard function. Since $h(t \mid X) = h_0(t)\exp(X\beta)$, we may derive the **partial likelihood function** (ignoring the baseline hazard) to yield the estimation on $\hat{\beta}$:

$$\hat{\beta} = \arg\max_{\beta} L(\beta) = \arg\max_{\beta} \left[\prod_{t} \frac{\exp(X_i\beta)}{\sum_{j \in R(t_i)} \exp(X_j\beta)} \right]$$

Here $R(t_i)$ is the risk set at t_i (individuals who are still at risk of experiencing the event before t_i [10]. Thus, the log-likelihood is:

$$L(\beta) = \sum_{i=1}^{n} \log\left(\frac{\exp(X_i\beta)}{\sum_{j \in R(t_i)} \exp(X_j\beta)} \right) = \sum_{i=1}^{n} \left[\beta^T X_i - \log\left(\sum_{j \in R(t_i)} \exp(\beta^T X_j) \right) \right]$$

As the loss function is the negative log-likelihood, it is:

$$\mathfrak{L} = -\log L(\beta) = -\sum_{i=1}^{n} \left[\beta^T X_i - \log\left(\sum_{j \in R(t_i)} \exp(\beta^T X_j) \right) \right]$$

[10] Cox(1997) gives the calculation for partial survival function and its corresponding Hessian matrix for readers who are interested as a reference.

An advantage of the Kaplan-Meier and Cox-PH methods is they are convenient to use with censoring data (see below). For these two models, they calculate the survival function stepwise at each time point, and therefore, for any given time t_j, we can capture the **restricted mean survival time**, which accumulates the survival from the start point:

$$\text{RMST}(t_j) = \int_0^{t_j} S(t)\,dt.$$

C. Machine Learning Framework for Survival Outcomes

As we derive the loss functions for the survival outcome from both the parametric models (log-normal and Weibull) and the semiparametric Cox-PH model, we can implement the loss in a machine learning algorithm to better reduce the divergence between the predicted survival outcome and the true value. In the twice doubly robust estimator for the time-to-event data, we use the architecture of the neural network to minimize the loss. For readers who are unfamiliar with the neural network, we briefly introduce its architecture here.

In general, the neural network architecture transforms the input data through a series of interconnected layers consisting of multiple neurons applying specific transformations (linear combinations). A basic neural network has three layers: the input layer, which puts the covariates predicting the survival outcome (the features) into the model[11]; the hidden layer, which executes specific transformations for the input; and the output layer, which outputs the predicted value and compares it with the observational data.

[11] In some literature, the input features are not counted as a layer as they will directly be involved in the mathematical calculations of the hidden layers. Therefore, the first $K-1$ layers are all hidden layers.

Suppose our neural network has $K \geq 3$ layers. We start with the input layer. In this layer, every covariate is regarded as a node and takes the form of a vector. The input nodes are transferred into the $K-2$ hidden layers, and mathematical transformations are executed here. There are two forms of transformations in each layer. The first is a linear transformation. We set Γ_k as the weight/slope parameter of the layer k and ρ_k as the bias/intercept parameter. The linear transformation is thus $X\Gamma_k + \rho_k$. The second transformation is the activation transformation. In order to learn more complex patterns, non-linear functions are applied. For instance, we may use the rectified linear unit (ReLU) function to achieve non-linearity: $\mathfrak{f}_{\eta_{k:2\leq k\leq K-1}} = \max(X\Gamma_k + \rho_k, 0)$ ($\eta_k = (\Gamma_k^T, \rho_k)^T$ on the $k-$th layer). Therefore, after the ReLU transformation, the results are in the range of $[0, +\infty)$. In sum, the hidden layer transformation is indeed $\hat{F}(X_i) = \prod_{k=1}^{K} \mathfrak{f}_{\eta_k}$ ($\eta = (\eta_1^T, \ldots, \eta_K^T)^T$). Finally, we feed our result from the hidden layer $\hat{F}(X_i)$ into the output layer and with some transformation to compare it with the target outcome to capture the empirical loss. The target outcome could be the hazard or survival rate without functional form, the mean survival time, or the parameters set in our parametric or semiparametric models. In other words, we could directly use neural networks to predict the survival time $\hat{T}$ and compare it with observed survival time T_i, or we could predict the parameters of the Cox-PH model with the neural network $\hat{\beta}$, optimize it and use the Cox-PH transformation discussed in the last subsection to generate the mean survival time. In this part, researchers need to pay attention to the domain of the data. For instance, for the parameters in the Cox-PH model β, as it can be valued in the whole real number set, we directly use linear activation in the output layer, and it could produce any real number. Likewise, if our outcome is the survival rate in the range of $[0, 1]$, we may use a sigmoid transformation in the output layer. If the outcome is the hazard function and the survival time is defined on $[0, +\infty)$, we can use the exponential transformation so that our predicted outcome is in that range.

The process described above (from the input through the hidden layer to the output layer) is called **forward propagation**. After calculating the empirical loss between the predicted outcomes from the output layer and the target outcome, we further perform the **backward propagation** to update every parameter in the model η_k in order to get the minimized loss (Rumelhart et al. 1986). To achieve this, we compute the gradient of the loss function with respect to each parameter using the chain rule:

$$\Gamma_k \leftarrow \Gamma_k - \iota \frac{\partial \mathfrak{L}}{\partial \Gamma_k}; \quad \rho_k \leftarrow \rho_k - \iota \frac{\partial \mathfrak{L}}{\partial \rho_k};$$

We set parameter ι as the **learning rate**. With multiple training iterations, the neural network could minimize the loss and adjust the parameters to predict the outcome better. Optimization algorithms like stochastic gradient descent (SGD) or Adam are deployed in this process.

As mentioned above, we could set the outcome in the loss function directly for the survival time, or the parameters in our parametric and semiparametric models. If we directly set the survival time as the outcome, we call the model a deep survival learning model or simply a neural network (NN) model for survival outcome (Faraggi and Simon 1995; Steingrimsson and Morrison 2020). If our predictors are the parameters for a log-normal survival distribution, we call the model the NN-log-normal model. Correspondingly, we have the NN-Exponential, NN-Weibull, and NN-Cox-PH models for the predicted outcomes, respectively, as the parameters for the exponential, Weibull distributions, and Cox-PH models. After obtaining the corresponding survival or hazard function, we could transfer them to capture the mean survival time and apply it to our doubly robust/efficient estimator for the

ATE to get the ATE for the survival outcomes.

However, in social science and medical research, the observed survival outcome usually contains missing data, and thus, we cannot have the complete loss function in our modeling. Dealing with the missing data for the survival outcome is a crucial subject in our method.

D. Truncation and Censoring

In general, survival data are missing because the time-to-event is not in our observation window. In some cases, if they are excluded from the analysis because of the occurrence of the event relative to our observational window, this type of missing is called **truncation**. In some cases, we could only observe their survival status partially, but the time they experience the event is beyond our observation window, and this missingness is **censoring**. More specifically, if individuals who experience the event before the initial time point are never included in the dataset (for example, we start observing one year after treatment, and individuals who died in the first year are not in our sample at all), we say the data are **left truncated**. Conversely, if individuals are included only when their event occurs before a fixed upper time point (for example, a retrospective study that samples only those who have already experienced the event before the study start), the data are **right-truncated**. Similarly, for censoring, if an individual would experience the event after our observation period ends (so we only know that the event time is greater than their last follow-up time), this is **right censoring**. If an individual has already experienced the event before the start of our observation, but the exact event time is unknown (we only know it occurred before the first observation), this is **left censoring**. Furthermore, if we observe subjects at discrete time points and only know that the event occurs within a certain interval between two ob-

servation times, we call this **interval censoring** (van der Laan and Robins 2003; Kalbfleisch and Prentice 2002).

For data in social science and demography, as we use **coarsened data**, we do not consider the issue of interval censoring as if the event occurs in our observational window; by default, we will approximate its survival time as we first observe the event. We are mainly concerned with the missing left truncation and right censoring. For left truncation, if we conduct observations after a period of time following the start of the experiment, it means that the group-level survival rate at our observation start time is lower than 1 (in contrast, without left-truncation, the start survival rate for both the treatment and the control groups are 1). For right-censoring, as we are unaware of when the real time-to-event is, our estimation based on the complete survival data will doubtlessly be biased. The method in this paper mainly deals with the **left-truncated-right-censored (LTRC)** survival data (Klein and Moeschberger 1992). LTRC data are common in social science research. For instance, the empirical case in the next chapter aims to explore how widowhood affects the survival outcome of mortality (as a survival outcome, life expectancy) among the elderly. We start our observation age at 50 so that individuals widowed before 45 are left-truncated, and individuals who are still alive in our last survey (observation) are right-censored.

As in other analyses of missing data, assumptions are required about the missing data patterns. In this paper, for the LTRC data, we assume that truncation is **missing at random (MAR)**, and censoring is also MAR, after controlling the covariates:

Assumption 2.III.3 (Assumption for Truncation and Censoring) *Suppose* $\tau > 0$ *denotes the time of truncation,* $C_i > 0$ *denotes the censoring time,* T_i *is the failure/survival time,* δ *is the*

indicator for censoring: $\delta_i = 1 \iff T_i < C_i$ *(*$\delta = 0$ *means censoring and* $\delta = 1$ *means observed), we assume the following conditions are satisfied:*

- *The truncation time is independent of the survival time:*

$$\tau \perp\!\!\!\perp T_i \iff P(\tau > t \mid T_i) = P(\tau > t)$$

- *The censoring time is independent of the survival time, given the observed covariates:*

$$C_i \perp\!\!\!\perp T_i \mid X_i \iff P(C_i > t \mid T_i, X_i) = P(C_i > t \mid X_i)$$

- *The truncation and censoring mechanisms are conditionally independent of each other given the covariates:*

$$C_i \perp\!\!\!\perp \tau \iff P(\tau, C_i \mid T_i, X_i) = P(\tau) \cdot P(C_i \mid X_i)$$

With the independence of truncation and censoring, we may further denote $\delta_i(\tau) = 1 \iff (T_i > \tau) \vee (T_i < C_i)$[12], which is the complete observed cases. Obviously, the set $\delta_i(\tau) = 1$ is a subset for $\delta_i = 1$. Like for the survival function $S(t \mid X_i) = P(T_i > t \mid X_i)$ for the prediction of the survival time, we have the censoring function $G(t \mid X_i) = P(C_i > t \mid X_i)$ to predict the time for censoring. Correspondingly, we could have the hazard function for censoring, denoted as $h_G(t \mid X_i) = P(C_i = t \mid C_i \geq t, T_i \geq t, X_i) = \frac{d \log G(t|X)}{dt}$ and the cumulative hazard $H_G(t \mid X) = -\log G(t \mid X)$ Based on Assumption 2.III.3, the survival function and the censoring function are independent.

Due to the missing values on time-to-event, the loss function used in the models is no longer obvious, as we couldn't obtain the likelihood function directly from the dataset.

[12] We could further infer that $T_i < C_i \iff \delta_i T_i \leq \tau$.

Therefore, we must apply efficient/doubly robust estimation techniques to the empirical loss so that we can yield the estimator for the mean survival time and apply it once again to the efficient/doubly robust estimator for the treatment effect. In the process, we used the technique of an efficient/doubly robust estimator twice, once for the loss function estimation of the survival outcome and once for the causal effect estimation. Therefore, we call our method a twice doubly robust estimator.

IV. Doubly Robust Loss for the LTRC Survival Outcomes

In this section, we will elaborate on how we derive the doubly robust/efficient loss for the LTRC survival outcome. Since truncation and censoring are independent processes, and truncation does not affect the relative distribution for survival time, we start with the scenario that only contains the censoring case.

As we did in Chapter 1 for the efficient estimator on non-saturated models, we begin our estimation from a regular and asymptotically linear (RAL) estimator. The IPW estimator of the survival loss is an RAL estimator, with the form of the inverse of the censoring probability. Let $\tilde{T}_i$ denote the observed time $\tilde{T}_i = min(T_i, C_i)$ and $\mathfrak{L}(\hat{F}(X_i), \theta)$ denotes the loss obtained from the complete data for the neural network in Subsection C between the predicted value $\hat{F}(X_i)$ for the target θ (as noted above, could be the survival time, the survival function, the hazard function, etc.), we have:

$$\mathfrak{L}^{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\hat{G}(\tilde{T}_i \mid X_i)} \mathfrak{L}(\hat{F}(X_i), \theta) \tag{2.IV.7}$$

As by definition $G(\tilde{T}_i \mid X_i) > 0$. Since censoring is MAR given the covariates, the estimator from Equation 2.IV.7 should have the same expectation as $\frac{1}{n} \sum_{i=1}^{N} \mathfrak{L}(\hat{F}(X_i), \theta)$.

In the introductory chapter, we have shown the efficient influence function (EIF) for the unconditioned expectation $\psi = E[Y]$ is $\phi^\dagger = Y - E[Y]$, and the EIF for conditional expectation $\psi = E[Y|X = x]$ is $\phi^\dagger = \frac{\mathbb{1}(X=x)}{P(X=x)}(Y - E[Y \mid X]) + (E[Y \mid X] - E[Y])$. Therefore, for the fully-observed data (without censoring and truncation), the EIF for the loss function on θ is:

$$\phi^\dagger_{\text{Full-data}} = \mathfrak{L}(\hat{F}(X_i), \theta) - E[\mathfrak{L}(\hat{F}(X_i), \theta)]$$

Now, we take censoring into consideration. Notice Equation 2.IV.7 has almost the same structure as we derive the EIF for the conditioned treatment effect $\phi^\dagger(\psi(a)) = \phi^\dagger(E[E_X[Y|A = a, X]] = \frac{\mathbb{1}(A=a)}{P(A=a|X)}[Y - E[Y \mid A = a, X]] + (E[Y \mid A = a, X] - E[Y \mid A = a])$ in our causal analysis (Equation III.23 in the introduction chapter), Thus, we could similarly derive the EIF and the efficient estimator for the loss as (Steingrimsson and Morrison 2020):

$$\phi^\dagger_{\text{Censoring}} = \frac{\delta}{G(\tilde{T} \mid X)}\left(\mathfrak{L}(\hat{F}(X), \theta) - E[\mathfrak{L}(\hat{F}(X), \theta) \mid \tilde{T} \geq t]\right) + E[\mathfrak{L}(\hat{F}(X), \theta) \mid \tilde{T} \geq t] - E[\mathfrak{L}(\hat{F}(X), \theta)]; \tag{2.IV.8}$$

$$\mathfrak{L}^{DR}_{\text{Censoring}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i}{\hat{G}(\tilde{T}_i \mid X_i)}\left(\mathfrak{L}(\hat{F}(X_i), \theta) - \hat{E}[\mathfrak{L}(\hat{F}(X), \theta) \mid \tilde{T} \geq t]\right) + \hat{E}[\mathfrak{L}(\hat{F}(X), \theta) \mid \tilde{T} \geq t]\right] \tag{2.IV.9}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\underbrace{\frac{\delta_i \mathfrak{L}(\hat{F}(X_i), \theta)}{\hat{G}(\tilde{T}_i \mid X_i)}}_{:a} + \underbrace{\left(1 - \frac{\delta_i}{\hat{G}(\tilde{T}_i \mid X_i)}\right)\hat{E}[\mathfrak{L}(\hat{F}(X), \theta) \mid \tilde{T} \geq t]}_{:b}\right]$$

In survival analysis literature, Equation 2.IV.9 is usually called the **augmented inverse probability weighted complete-case (AIPWCC)** estimator (Tsiatis 2006, chapter 9, pp. 199-220), as part a is the IPW for the complete case loss and part b is the augmented term which more efficiently uses information from the censoring individuals (Steingrimsson et al. 2016). Let

$\hat{U}(X_i,t) = \hat{E}[\mathfrak{L}(\hat{F}(X_i),\theta) \mid \tilde{T} \geq t]$, the estimation will be doubly robust if we correctly specified either $\hat{G}(t \mid X_i)$ or $\hat{U}(X_i,t)$.

However, Equation 2.IV.9 is uncommonly seen in the survival literature [13]. In most literature, they will simplify part b as:

$$\left(1 - \frac{\delta_i}{\hat{G}(\tilde{T}_i \mid X_i)}\right)\hat{U}(X_i,t) = \int_0^{\infty} \frac{\hat{U}(X_i,t)}{\hat{G}(t \mid X_i)} dM_G(t \mid X_i)$$

As function $M_G(t \mid X)$ is the censoring martingale at t given the covariates X. To understand this, we first introduce the definition of the martingale:

Definition 2.IV.1 *Martingale is a process describing the difference between a counting process ($N(t)$) and the "compensator" generated by the intensity function ($\Lambda(t) = \int_0^t h(u)du$):*

$$M(t) = N(t) - \int_0^t h(u)du$$

Let's say we have counted the number of items censoring at time t: $N(t) = \mathbb{1}_{\delta=0,\tilde{T}\leq t}$ (the actual number of case censoring from the observational data). Besides, we could also calculate the predicted counts for censoring from the model relevant to censoring, for example, with the censoring hazard function: $\Lambda(t) = \int_0^t \mathbb{1}_{\tilde{T}\geq u} h_G(u \mid X)$. Therefore, the **censoring martingale** represents the cumulative excessive amount of censoring from time 0 to t:

$$M_G(t \mid X) = \mathbb{1}_{\delta=0,\tilde{T}\leq t} - \int_0^t \mathbb{1}_{\tilde{T}\geq u} h_G(u \mid X)$$

With the definition, we may infer that (Strawderman 2000; Robins and Rotnitzky 1992):

$$1 - \frac{\delta_i}{G(\tilde{T} \mid X)} = \int_0^{\infty} \frac{dM_G(t \mid X)}{G(t \mid X)}$$

[13] This is because the way we derive the doubly robust estimator for the survival loss is not strictly mathematical derivation. Instead, for readers with social science and demographic backgrounds, we simplified the derivation of the doubly robust estimator by analogy with deriving the conditioned treatment effect, and then we will prove that the derived doubly robust estimator for the survival loss is intrinsically the same as the results from the survival analysis literature. For the rigorous proof, see textbooks like Bickel et al.(1993) and Tsiatis (2006).

Simply because $\int_0^\infty dM_G(t \mid X) = (1-\delta_i) - (1 - G(\tilde{T} \mid X)) = G(\tilde{T} \mid X) - \delta_i$. Therefore, we may rewrite Equation 2.IV.9 into the form with the martingale:

$$\mathfrak{L}^{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i \mathfrak{L}(\hat{F}(X_i),\theta)}{\hat{G}(\tilde{T}_i \mid X_i)} + \int_0^\infty \frac{\hat{U}(X_i,t)}{\hat{G}(t \mid X_i)} dM_G(t \mid X_i)\right] \tag{2.IV.10}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i \mathfrak{L}(\hat{F}(X_i),\theta)}{\hat{G}(\tilde{T}_i \mid X_i)} + \left(\frac{(1-\delta_i)\hat{U}(\tilde{T}_i,X_i)}{\hat{G}(\tilde{T}_i,X_i)} - \int_0^{\tilde{T}_i} \frac{\hat{U}(t,X_i)}{\hat{G}(t \mid X_i)} dH_G(t \mid X_i)\right)\right] \tag{2.IV.11}$$

While Equation 2.IV.11 just expands Equation 2.IV.10 with the definition of martingale.

Finally, we consider the restrictions of left truncation. Since the truncation process is independent of the censoring process, we may just rewrite the expressions in Equation 2.IV.10 and Equation 2.IV.11, changing δ_i into $\delta_i(\tau)$ and $\tilde{T}_i$ into $\tilde{T}_i(\tau)$ (indicating that we only observe those who survive beyond the truncation time, τ):

$$\hat{\mathfrak{L}}^{DR}_{\text{LTRC}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i(\tau)\,\mathfrak{L}(\hat{F}(X_i),\theta)}{\hat{G}(\tilde{T}_i(\tau) \mid X_i)} + \left(1 - \frac{\delta_i(\tau)}{\hat{G}(\tilde{T}_i(\tau) \mid X_i)}\right)\hat{U}\big(X_i, \tilde{T}_i(\tau)\big)\right] \tag{2.IV.12}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i(\tau)\,\mathfrak{L}(\hat{F}(X_i),\theta)}{\widehat{G}(\tilde{T}_i(\tau) \mid X_i)} + \int_0^\infty \frac{\widehat{U}(t,X_i)}{\widehat{G}(t \mid X_i)} d\widehat{M}_G(t \mid X_i)\right] \tag{2.IV.13}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i(\tau)\,\mathfrak{L}(\hat{F}(X_i),\theta)}{\widehat{G}(\tilde{T}_i(\tau) \mid X_i)} + \left\{\frac{(1-\delta_i(\tau))\,\widehat{U}\big(\tilde{T}_i(\tau),X_i\big)}{\widehat{G}(\tilde{T}_i(\tau) \mid X_i)} - \int_0^{\tilde{T}_i(\tau)} \frac{\widehat{U}(t,X_i)}{\widehat{G}(t \mid X_i)} dH_G(t \mid X_i)\right\}\right] \tag{2.IV.14}$$

In summary, the procedures to apply the doubly robust estimator for the LTRC survival outcomes are as follows: first, based on the observations, we estimate the survival of censoring $\widehat{G}(\tilde{T}_i(\tau) \mid X_i)$, and construct the architecture of the neural network for the complete case to yield the loss function of the target (survival function, hazard function, mean survival time, log-normal/exponential/Weibull/Cox-PH model parameters) under $T(\tau) \geq t$: $\hat{U}(X_i, \tilde{T}(\tau))$; then we use the doubly robust characteristics for the loss function in either Equation 2.IV.12, 2.IV.13, or 2.IV.14 to solve the parameters ($\theta = \arg\min_\theta \mathfrak{L}^{DR}_{LTRC}$); and finally, based on the model we trained, we predict the mean survival time and apply it to our causal

inference framework.

Sometimes, researchers might find it hard to solve $\theta = \arg\min_\theta \mathfrak{L}^{DR}_{LTRC}$ even with iterations. In such cases, approximate methods may be required. Optimal restriction assumptions may be further needed. For the technical details, see Tsiatis (2006: Chapter 12).

V. Twice doubly Robust Estimation Algorithm

In sum, based on the two doubly robust estimators in Section II and IV, we develop a twice doubly robust estimator to capture the average treatment effects and the heterogeneous treatment effect in left-truncated-right-censored survival outcome: we use the first doubly robust estimator to estimate the loss function for the survival outcome and yield the estimation for mean survival time; we then apply a doubly robust estimator to have the efficient/doubly robust estimator for the causal outcome.

Beyond that, we can capture the **heterogeneous treatment effect (HTE)** by conditioning the treatment effect on particular covariate values. This quantity is the conditional average treatment effect (CATE):

$$\text{CATE}(x) = E[Y(1) - Y(0) \mid X = x] = E[Y(1) \mid X = x] - E[Y(0) \mid X = x]$$

Therefore, we can have the procedure to estimate the CATE with survival outcomes as follows:

- First, identify our measurable set $D_i = \{X_i, T_i, Z_i, A_i, C_i, R_i, \tau, \delta_i\}$. In the dataset, X_i denotes the covariate for the CATE, T_i denotes the observable survival time, Z_i are the covariates controlling randomization, and R_i are the covariates controlling censoring, A_i is the treatment assignment, C_i denotes the censoring time, τ denotes the

restriction horizon, and δ_i is the event indicator. Further, based on the variables that appeared in the dataset, we can generate $\tilde{T}_i = \min\{T_i, C_i\}$, $\tilde{T}_i(\tau) = \min\{\tilde{T}_i, \tau\}$, and $\delta_i(\tau) = \delta_i \mathbb{1}(\tilde{T}_i \leq \tau) + \mathbb{1}(\tilde{T}_i > \tau)$. Then, we divide the dataset D into three subsets D_1, D_2, and D_3.

- We use D_1 to fit the nuisance parameter π in causal inference. Namely, we have $\hat{\pi}(Z_1)$: $\hat{\pi}(Z_1) = P(A_{1i} = 1 | Z_{1i})$ from D_1.

- We use D_2 to fit the nuisance parameter μ_a in causal inference. As we noted above, in our study, μ_a is the mean survival time under the treatment status $A = a$.

- Regress the pseudo-outcome using the third dataset and yield the doubly robust estimation for the conditional average treatment effects: $\widehat{CATE}^{DR} = E\Big[\big[\frac{A_{3i} - \hat{\pi}(Z_{3i})}{\hat{\pi}(Z_{3i})(1-\hat{\pi}(Z_{3i}))}(Y_{3i} - \widehat{\mu}_{A_{3i}}(X_{3i})) + (\widehat{\mu}_1(X_{3i}) - \widehat{\mu}_0(X_{3i}))\big] \mid X = X_{3i}\Big]$.

- Rotate the three datasets in the previous steps for cross-validation. In other words, separately use D_2 and D_3 to fit the nuisance function for propensity; use D_3 and D_1 to fit the nuisance function for survival function, and use D_1 and D_2 to generate the doubly robust estimation for the CATE. Finally, average the three results.

The algorithm is indeed exactly the same as the algorithm for the efficient/doubly robust/debiased ATE/CATE. The specific part that needs to be addressed is in Step 3, in which we generate the mean survival time with neural networks with the doubly robust loss:

- Split D_2 into M subsets: $D_2^{(1)}, \ldots, D_2^{(M)}$.

- For each subset $D_2^{(m)}$, decide the target parameter for the loss function. The target parameter can be the survival rate, instantaneous or cumulative hazard, mean survival time, and log-normal/exponential/Weibull/Cox-PH model parameters. Calcu-

late the loss function from the complete case $\mathfrak{L}(\hat{F}(R_i),\theta)$ and its conditional expectation $\hat{U}(R_i,\tilde{T}_i(\tau)) = E[\mathfrak{L}(\hat{F}(R_i),\theta) \mid \tilde{T}_i(\tau),R_i]$. Calculate the survival function for censoring $\hat{G}(t \mid R_i) = P(C_i \geq t \mid R_i)$.

- Derive the empirical average estimated loss function based on either Equation 2.IV.12, 2.IV.13, or 2.IV.14. Predict the parameter $\theta = \sum_{m=1}^{M} \eta_m \mathbb{1}_{R \in D_2^{(m)}}$.

- Based on the optimized parameter from the last step, correspondingly derive the correct estimation for the mean survival time.

VI. Simulation Work

A. Model Settings

This section uses a simulation to illustrate our method. In total, we examine the estimation from five models: first, we estimate the ATE and the CATE without an RCT framework; we use the effects via the marginal hazard ratio (MHR) directly from the Cox Proportional Hazard Model with specific baseline settings. Then we turn to the RCT settings. We initially use the separated Cox-PH models to fit the survival functions for the treatment and the control, and then predict the expected outcome under the treatment and control for the whole dataset with the treatment and control survival functions. Then we use a neural network with a doubly robust loss function to fit the treatment and survival function and to calculate the average and heterogeneous treatment effects in the same way as the Cox-PH models. Since the two models give out naïve plug-in estimations for the treatment effects (for both the ATE and the HTE), we call the naïve Cox-PH model and the naïve "NN-DR loss" models separately. Finally, we apply the Cox-PH and NN-DR loss models under the doubly robust/debiased treatment effects framework. We refer to them as the debiased Cox-PH and debiased NN-DR loss models, respectively. The debiased NN-DR loss model

yields a twice-doubly robust estimator for the average and heterogeneous treatment effects in survival data. For the naïve NN-DR loss model, the debiased Cox-PH model, and the twice-doubly robust model (debiased NN-DR loss model), we further execute cross-fitting due to the requirements of doubly robust estimation.

The settings of our simulated data are as follows. We generate $N = 1,800$ independent units with covariates $(Z_1, Z_2) \sim N(0,1)$ that drive treatment assignment and $(X_1, X_2) \sim N(0,1)$ that drive survival. We assign the treatment via a logistic propensity model:

$$\Pr(A = 1 \mid Z_1, Z_2) = \text{logit}^{-1}\{-0.3 + 1.0\, Z_1 + 1.2\, Z_2 + 0.6\, Z_1 Z_2\},$$

which induces broad overlap and meaningful heterogeneity along the propensity score. We set failure times to follow a Weibull law with arm-specific parameters: shape $k_1 = 1.8$ and baseline scale $\lambda_1 = 2.6$ for the treated, and shape $k_0 = 1.2$ and baseline scale $\lambda_0 = 1.1$ for controls. To induce outcome heterogeneity and confounding, the individual scale is modulated as

$$\lambda(a \mid Z_2, X_1, X_2) = \lambda_a \exp\{-0.35\, Z_2 + 0.25\, X_1 + 0.10\, X_2\}, \qquad a \in \{0,1\}.$$

We set the event times sampled as $T = \lambda(A \mid Z_2, X_1, X_2)\,(-\log U)^{1/k_A}$ with $U \sim \text{Unif}(0,1)$. Further, We introduce left truncation by an entry time $U^{\text{entry}} \sim \text{Unif}(0, 0.5)$ and retain only subjects with $Y = \min(T, C) > U^{\text{entry}}$. Right censoring is generated as $C \sim \text{Exp}\{\rho(Z_1, X_2)\}$ with

$$\rho(Z_1, X_2) = 0.25 \exp\big(0.15\, Z_1 - 0.15\, X_2\big),$$

yielding covariate-dependent censoring. Because treatment alters both shape and scale, hazards are not strictly proportional, resulting in genuine heterogeneous treatment effects (larger benefits at lower propensities). We summarize performance using the restricted

mean survival time (RMST) up to the horizon $\tau = 5$.

Under one simulation replicate with these settings, 1,547 subjects remain after left truncation (out of 1,800 generated; 253 truncated), with 708 treated and 839 controls; 437 observations are right-censored. The mean observed time is 1.634 in the treated arm and 1.200 in the control arm (difference = 0.434). Propensity scores are estimated by logistic regression in $(Z_1, Z_2, Z_1 Z_2)$. We report Kaplan–Meier curves in Figure 2.1. The overlap of estimated propensities by treatment group, and the heterogeneous treatment effect (RMST difference) across ten propensity-score bins or deciles, visualize effect variation with treatment likelihood. The propensity score overlap for the treatment and control groups can be seen in Figure 2.2. The simulation truth has an average treatment effect on restricted mean survival of approximately 0.86 time units, with heterogeneous treatment effects across propensity strata.

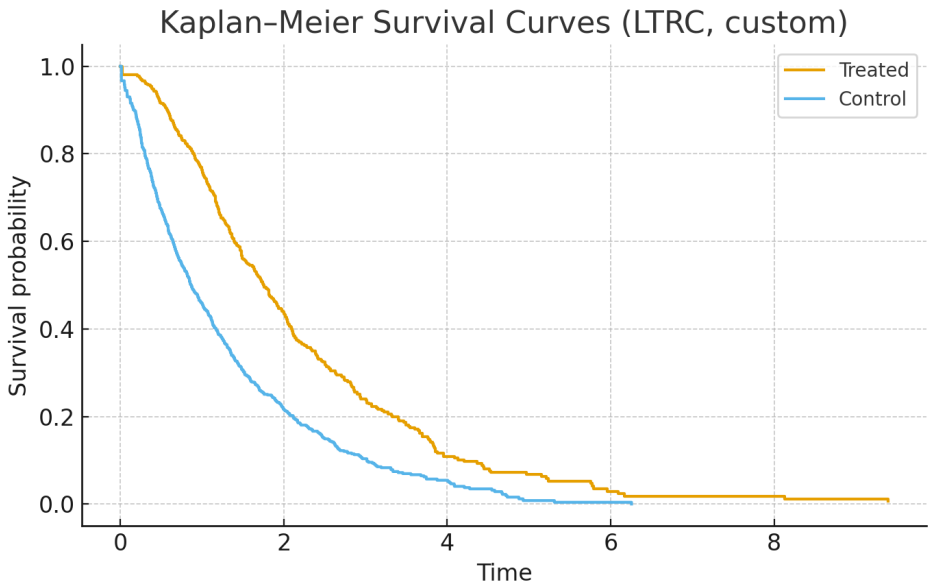


Figure 2.1: Kaplan–Meier survival curves for treated and control groups with left truncation (delayed entry) accounted for.

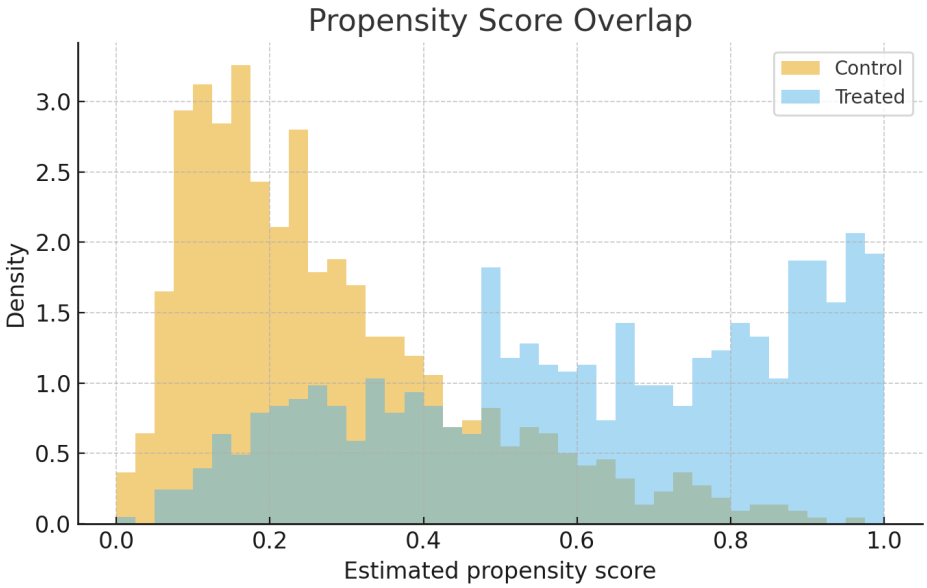


Figure 2.2: Overlap of estimated propensity scores by treatment group. Histograms (or density estimates) of $\hat{\pi}(X)$ for treated and control units show common support across $[0,1]$.

B. Results from Simulation

Table 2.1: Table 1: Simulation performance (RMST estimand at $\tau = 5$). RMSE = $\sqrt{\text{MSE}}$.

Estimator	Bias	Variance	MSE	RMSE
MHR	-0.4892	0.0088	0.2465	0.4965
Naïve Cox plug-in	0.0404	0.0315	0.0269	0.1640
debiased (DR) Cox-PH	-0.6297	0.0081	0.4020	0.6340
naïve NN-DR loss	-0.4979	0.0042	0.2513	0.5013
twice doubly robust	-0.2199	0.0304	0.0713	0.2670

In Table 2.1, we present the simulation results for the five estimators: Marginal Hazard Ratio (MHR), a naïve Cox-PH plug-in estimator, a DR (debiased) Cox-PH model, a naïve neural network doubly robust loss model (naïve NN-DR loss), and the proposed twice doubly robust estimator in terms of their bias, variance, mean squared error (MSE), and root MSE (RMSE).

As expected, the results of the MHR are heavily biased: fitting a single Cox-PH on the full sample without accounting for confounding underestimates the average treatment effect (ATE) (ATE ≈ 0.63 vs. true $= 0.86$ in one simulation), and it misleadingly suggests a declining treatment benefit for higher-propensity patients. The bias arises as MHR estimators violate the unconfoundedness assumption and fail to adjust for the selective treatment assignment. Therefore, the MHR has a high MSE despite its moderate variance.

We also report the HTE results for the MHR model in Figure 2.3. The MHR estimator produces an HTE curve that is flat across propensity bins (mean RMST difference is approximately 0.70 with comparable uncertainty in each bin). Indeed, the trend is by construction: as the A-only Cox model imposes a single proportional-hazards effect for all subjects, it cannot express heterogeneity by covariates or propensity. Thus, MHR's HTE display is stable but uninformative about effect modification and biased at every bin. Therefore, we could not use the MHR model to detect any heterogeneous effect.

We then turn to the results on the Naïve Cox-PH plug-in model. The naïve Cox plug-in (separate Cox models for treated and control to predict counterfactual outcomes) reduces bias substantially: its ATE estimate is closer to the truth (e.g., approximately 0.73 in one run) and its variance is low (bootstrapped SD is 0.21), yielding a surprisingly small MSE. In fact, the naïve Cox has the lowest RMSE in our simulation despite a slight residual bias. However, this strong performance should be interpreted cautiously: the arm-specific Cox plug-in model approximates the simulated survival curves well in this setting, but the data-generating process includes non-proportional hazards and heterogeneous RMST effects. In empirical studies, when the underlying distribution of the survival outcome is unknown, we should be more cautious in the model selection.

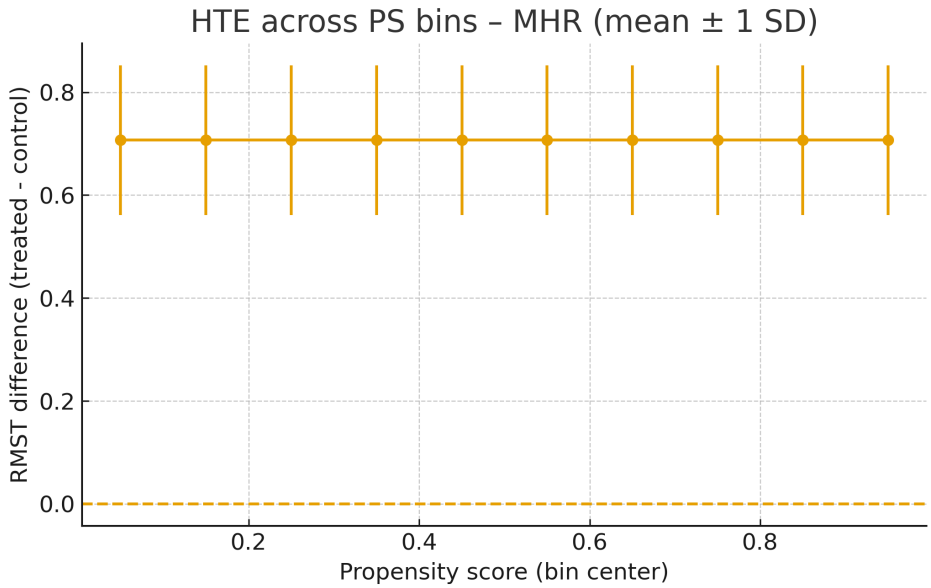


Figure 2.3: HTE across 10 propensity-score bins (mean ± 1 SD over R=100, N=1200, τ=5): **MHR** (A-only Cox).

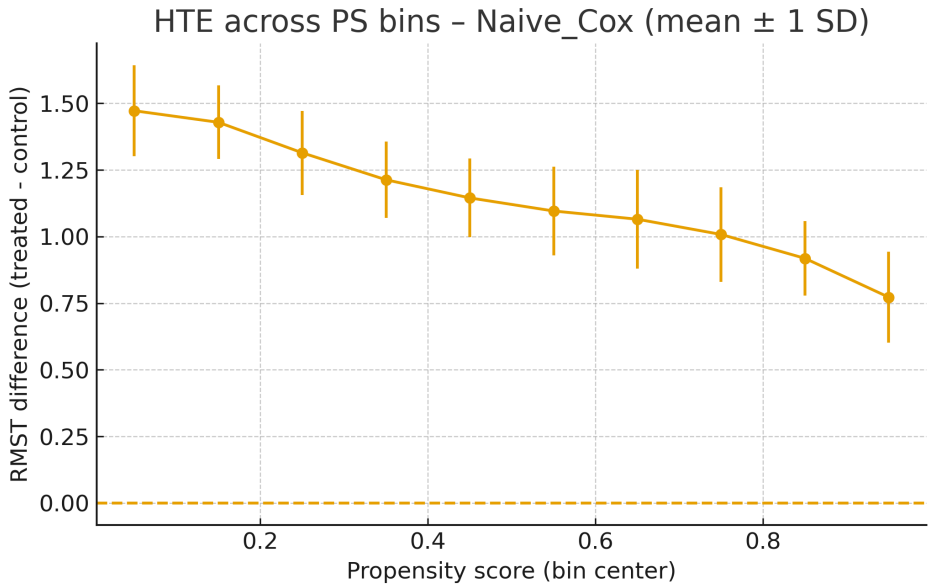


Figure 2.4: HTE across 10 propensity-score bins (mean ± 1 SD over R=100, N=1200, τ=5): **Naive Cox plug-in**.

The HTE graph for the naive Cox-PH model can be seen in Figure 2.4. As can be inferred from the graph, it has the narrowest uncertainty, and the trend is downward. Because it fits separate Cox models by treatment arm with covariates and then performs g-computation for counterfactual RMST, it removes most confounding while retaining high stability. Again, as we warned above, this performance depends on how well the working Cox specification approximates the survival process; if the Cox model is materially mis-specified, for instance, with stronger non-PH hazards or nonlinear covariate effects, the bias will be large.

We then discuss the three models involving the doubly-robust techniques. First, we look at the results from the DR (debiased) Cox-PH estimator. In fact, as shown in Table 2.1, its variance is most affected by weight noise in the sample (with the largest bias and the largest RMSE among the models). We have the HTE graph for the DR (debiased) Cox-PH estimator in Figure 2.5. The HTE function is broadly flat across propensity score bins, with larger uncertainty (and occasional non-monotonic fluctuations) at the extremes. This is because the AIPW correction introduces treatment and censoring weights, and when overlap is limited on the very low and high propensity score ends, those weights become variable, inflating bin-level variance and producing the jagged appearance. In summary, the debiased Cox-PH models will give more noise from weighting, which is consistent with their high RMSE.

The next model is the naive NN-DR loss model. As shown in Table 2.1, the model is also biased downwards (-0.498) with moderate variance. The HTE graph (Figure 2.6) shows a clear downward slope, with larger effects in lower propensity-score bins and smaller effects in the higher bins. Broadly speaking, the shape of the HTE graph for the naive NN-DR loss model is quite similar to that of the naive Cox-PH model, and this is because the naive NN plug-in actually inherits a Cox-like inductive bias under the DGP and does not contain the

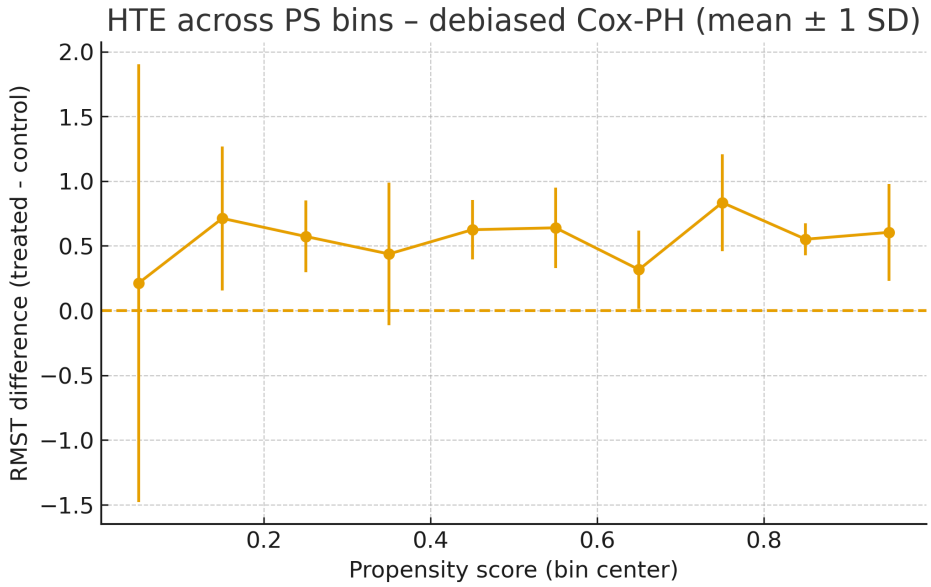


Figure 2.5: HTE across 10 propensity-score bins (mean ± 1 SD over R=100, N=1200, τ=5): **debiased Cox-PH**.

augmentation part for variance augmentation, so the shapes are quite similar, while the magnitude of the slope and uncertainty are different.

Finally, we visit the results from our proposed twice doubly robust model. Theoretically, the estimator is consistent if the required nuisance components are correctly specified in the doubly robust sense, and it can approach the efficiency bound when the EIF components are well estimated under the regularity conditions. In practice, it is still the best model behind the naive Cox-PH model in terms of RMSE, confirming that it substantially reduces bias. However, its variance is larger than that of most competing models. As the MSE is the sum of squared bias and the variance, since the model has a smaller bias, its MSE and RMSE are smaller compared to other doubly robust models.

We then turn to the HTE graph for the twice doubly robust model, and the results are

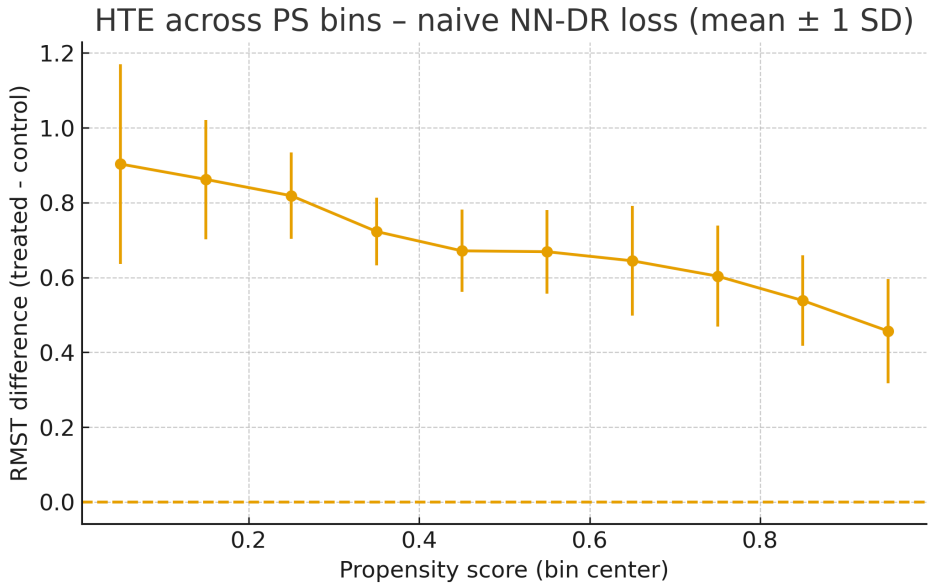


Figure 2.6: HTE across 10 propensity-score bins (mean ± 1 SD over R=100, N=1200, τ=5): **naive NN-DR loss**.

in Figure 2.7. As can be seen from the graph, the estimated CATE pattern is less dominated by a single plug-in model specification, but it carries more variance, especially when the propensity is sparse at the lowest end. This illustrates the strengths and weaknesses of the twice doubly robust estimator: on the positive side, it is robust to some nuisance-model misspecification and well-targeted for estimating HTEs, and it controls bias as long as the required nuisance components are correctly specified in the doubly robust sense. However, because of the requirements on the nuisance functions, it requires more tuning than simple plug-in models. Additionally, in practice, larger samples and careful stabilization are essential for the estimator to achieve the theoretical efficiency.

VII. Conclusion and Further Discussions

In this chapter, we introduce a twice doubly robust estimator to estimate the average and heterogeneous treatment effects for the left-truncated-right-censored survival data: we con-

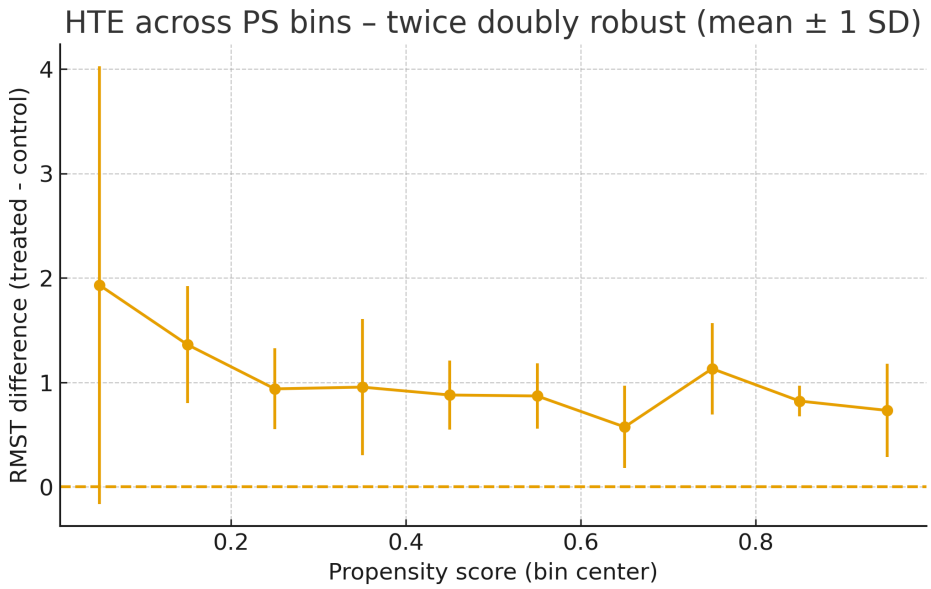


Figure 2.7: HTE across 10 propensity-score bins (mean ± 1 SD over R=100, N=1200, τ=5): **twice doubly robust**.

duct doubly robust estimation both for the causal estimand and for the survival functions and combine them. It is a regular, asymptotically linear, and efficient estimator used to identify causal effects.

Some readers may have noticed that in this chapter, we assume that the treatment is an instantaneous variable and do not discuss the scenario of the time-varying treatment. In fact, we believe that for time-varying treatment, calculation on the average and heterogeneous treatment effect is not a statistical/methodological problem; instead, it is a theoretical problem as we need to use a theoretical framework to identify what the "treatment effect" stands for. Hence, for time-varying treatments, the main challenge may not be statistical but definitional: researchers may better specify the causal estimand, not only what the intervention is, but also its timing and duration.

For instance, with treatment history $\bar{A}_t = (A_0, \ldots, A_t)$ and information history H_t, ef-

fects are indexed by a static regimen $\bar{a}$ or a dynamic rule $d(H_t)$, so the relevant targets are $E[Y^{\bar{a}}]$ or $E[Y^d]$, not a single undifferentiated "ATE." Week-specific contrasts (e.g., initiating therapy at week t given no prior use; continuing vs. stopping at week t; sustaining therapy through week T) condition on different histories and populations, so computing ATEs at weeks $-1, 0, 1, 2, \ldots$ and averaging them generally fails to define the overall effect. Once the substantive question fixes the intervention, we can use standard longitudinal methods like the g-formula, marginal structural models, or structural nested models to estimate $E[Y^{\bar{a}}]$ or $E[Y^d]$ and their heterogeneity. We will see a very similar problem in Chapter 3 when we apply the model to empirical studies of the causal effect of widowhood on mortality.

Chapter 3

Widowhood and Mortality: Delineating Heterogeneous Effects Using Doubly Robust Estimation

I. Introduction

The widowhood effect, or the increased mortality due to bereavement, has been well documented in sociological and demographic literature. In earlier studies, researchers have noticed that bereaved people have higher mortality rates than married couples and have found some causal mechanisms between widowhood and death (Hu and Goldman 1990; Lillard and Waite 1995; Subramanian et al. 2008; Liu et al. 2020).

Recent research on the widowhood effect has shifted the focus to its causal interpretations. Widowhood may have a causal effect on the increase in mortality; however, it could also result from the similarities between the couples or the couple's environmental exposures. With ingenious designs, some research responds to the question without sophisticated statistical techniques. For example, Elwert and Christakis (2008a) compare the widowhood effects of the current spouse and ex-spouses on widowed individuals' mortality. They found no widowhood effects of the ex-spouses, indicating that the widowhood effect

of the current spouse is a causal mechanism for the mortality increase among widowed individuals.

Meanwhile, recent research on the widowhood effect also draws attention to heterogeneous treatment effects (HTE) on two dimensions: first, the heterogeneous effect of pre-loss preparedness on the post-loss mortality outcome: Prior evidence shows that low preparedness and unexpected bereavement are associated with poorer post-loss outcomes and elevated risks among surviving spouses (Hebert et al. 2006; Treml et al. 2021; Tang et al. 2021; Wen et al. 2021; Shah et al. 2013; Hauksdóttir et al. 2010). Second, different characteristics/ social attributes moderate the heterogeneous treatment effect at different treatment levels. Studies have analyzed effects by educational levels, marital types (homogamy vs. heterogamy), and asset levels, and found that the magnitude of the widowhood effect correlates with socioeconomic status across subgroups. Substantively, we therefore estimate HTE across preparedness strata (and their intersections with education and income) and assess whether the socioeconomic factors lead to differences in the heterogeneous treatment effect of preparedness on mortality after widowhood.

Although studies have elaborated on the importance of the causal effects of widowhood on mortality and their varieties among subgroups, sociologists need to be more explicit in addressing the sociological meaning of widowhood for widowed individuals' lives and be more specific in acknowledging the causal mechanisms of the widowhood effect. In sociological literature, it is crucial to understand why social relationships may affect the life chances of individuals (House et al. 1988). Based on statistical analyses, researchers established the causal relationship between widowhood and health status (Stroebe et al. 2007) and between widowhood and death (Elwert and Christakis 2008b), and we may assume that

widowhood increases the risks of mortality by exacerbating widowed individuals' health status. We still need to understand how the variations in the widowhood effects among different subgroups formed.

On the methodological side, although researchers could circumvent the problem of causal identification with innovative research designs, the "causal estimand" identified in the previous studies differs from the "treatment effect" of widowhood in rigorous statistical terms because the treatment variable of widowhood is not randomly assigned. Previous literature applied a survival model and compared the hazard ratio between the widowed and non-widowed based on the coefficients (see Shor et al. 2012 for the review). However, they fail to randomize the "treatment" of widowhood by controlling the confounders that affect its likelihood. Therefore, the effects might be overstated due to the endogeneity of widowhood itself.

This paper examines the varying impact of widowhood on mortality rates in the United States, utilizing a doubly robust causal estimator for the analysis. Substantively, we define a latent variable–preparedness for widowhood–to analyze the causally heterogeneous widowhood effect. In confronting the experience of widowhood, we posit that elderly individuals exhibit varying degrees of preparedness for the loss, influenced by their lifelong familiarity with their spouse. This leads to differing levels of grief experienced among them. Indeed, previous studies acknowledged that unexpected widowhood is likely to bring a higher mortality risk (Kristensen et al. 2012; Shah et al. 2013), and widowhood due to acute conditions leads to higher hazard ratios than chronic diseases (Elwert and Christakis 2008b). Furthermore, the preparedness for widowhood could also interfere with different socioeconomic traits, mainly education and wealth. In other words, we elaborate on the results that even

for people with the same level of preparedness for losing their spouse and the same level of grief, the causal effect of widowhood still varies among different social groups.

Methodologically, this paper applies the doubly robust estimator in causal inference to analyze the treatment effect of widowhood. Constructing baseline preparedness for widowhood from propensity scores, we establish the counterfactual mean survival time (life expectancy) based on time-varying Cox Proportional Hazard (Cox-PH) models for the widowed and non-widowed groups and apply the doubly robust analytical framework to estimate the average treatment effects (ATE) and the heterogeneous treatment effects (HTE) among different preparedness levels.

In this paper, using longitudinal data from the US Health and Retirement Study (HRS) between 1998 and 2018, we aim to illustrate how widowhood for elders causally contributes to their mortality and how the causal effects vary across different preparedness levels and educational and wealth levels. This paper is structured as follows: Section II reviews the previous literature concerning the causal identification of widowhood and heterogeneous effects among different social traits, including education and wealth. Section III reviews the prior literature's analytical strategy and introduces our analytical framework under Rubin's causal model. Section IV describes the dataset we use, the measurements we adopt, and the descriptive statistics for the variables. The results are presented in Section V. Section VI concludes the contribution of this paper and discusses the possible methodological advance based on the insufficiency of the current version.

II. Literature Review

A. Causal Widowhood Effect

A classic sociological vein is understanding how social relationships affect an individual's life chances since Durkheim's analysis of suicide (Durkheim 1951). Demographers noticed the relationship between marital status and mortality in the early years. In Farr's (1858) short essay on the marital effects on the mortality of the French people, he documented the positive correlation between widowhood and mortality among widowed people. Later studies applied survival analysis methods on census or survey data and confirmed that the higher risks of mortality for the widowed group (with reference group as the married or the general population) existed ubiquitously in the US (Shurtleff 1955: 1956; Goldman et al. 1995; Lillard and Panis 1996; Espinosa and Evans 2008), the UK (Parkes et al. 1969; Jones and Goldblatt 1987; Hart et al. 2007), Europe (Grundy and Kravdal 2008; Joung et al. 1996; Kolip 2005; Kravdal 2003; Malyutina et al. 2004; Manor et al. 1999; Samuelsson and Dehlin 1993; Shkolnikov et al. 2007; Thierry 2000), and Asian countries (Goldman and Hu 1993; Iwasaki et al. 2002; Rahman et al. 1992).

Specifically, recent studies focused on the relationship between widowhood and mortality seek to substantiate the "widowhood effect." This theory posits that the experience of widowhood has a causal impact on the increased likelihood of death for the surviving spouse (Lillard and Waite 1995; Lillard and Panis 1996; Elwert and Christakis 2006: 2008b). The causal mechanisms underpinning the widowhood effect are intuitively understandable: The onset of widowhood can precipitate a cascade of adverse psychological and physiological responses in the surviving spouse. These may encompass amplified emotional distress, such as heightened feelings of grief and sorrow, increased health-related anxieties,

and an intensified sense of isolation owing to the absence of intimate companionship. Although several studies point out the negative correlation between widowhood experience and health status, it is hard to justify the pathways with quantitative methods, as the level of grief, vexation, and loneliness may change rapidly, so finding valid measures of these variables is difficult.

On the other hand, a body of research posits that the observed positive correlation between widowhood and increased mortality risk should not be automatically equated with a causal relationship. They indicate that the similarities between a couple's deaths are due to a selection process instead of the causal mechanism (Sullivan and Fenelon 2014). The homogamy assumption suggests that they may have similar lifespans due to the similarities from assortative mating (Kalmijn 1998). The family resources assumption suggests that couples living together share socioeconomic resources that directly and indirectly (via healthcare conditions) affect their healthcare outcomes (Boyle et al. 2011). The environmental exposure assumption suggests their common exposure to similar environmental variables leads to the related death between a couple (Schaefer et al. 1995). In summary, some scholars believe that the selection process is the reason for the spurious relationship between widowhood time and death time.

To determine whether widowhood has a causal effect on mortality or if they are correlated because of selection, Elwert and Christakis (2008a) had an ingenious research design: they compared the widowhood effects of losing the current spouse and ex-spouse. An ex-spouse may also exhibit homogamy, share access to family resources, and be subject to the same environmental factors as the widower. Researchers found that the ex-spouse's death does not lead to a significant increase in mortality, while they observed the increased ef-

fect of the death of the current spouse. Therefore, the authors concluded that the effect of widowhood on mortality is causal. However, rigorously speaking, their paper had limitations in its conclusions: a possible alternative explanation for the result might be that the divorce effect offsets the spurious relationship between widowhood and mortality due to selection. Further, when the widower lives with the ex-spouse, they may experience different resource restraints and environmental exposures compared to when the widower lives with the current partner. Intrinsically, the researchers couldn't trace and compare the interactions between the widower and their spouse/ex-spouse, and thus could not decipher the specific social meaning of the widowhood event on the widower.

From the statistical point of view, even if we observe that the current spouse's death significantly increases mortality among widowed individuals, the effect cannot be identified as causal. In the context of causal inference, the assumption is that a randomly assigned treatment will have a significantly different impact on the group that receives the treatment than on the group that does not. In this case, widowhood should be a random event; given this prerequisite, the statistically significant difference in mortality risks between widowed and non-widowed individuals can be regarded as causal. For widowed individuals, the spouse's death is not a random event, not only because they have similarities and share the resource restraints and environmental exposures, but also because the widowed individual may have expectations of their spouse's death *before* the event. Therefore, to address the causal widowhood effects, relying on simple survival analysis and using the hazard ratio for widowhood is insufficient.

Assumption 3.II.1 *Widowhood causes widowed individuals to have higher mortality risks.*

B. Heterogeneous Effects among Preparedness for Widowhood

In fact, a large number of studies have noticed that preparedness for a loved one's death before the actual loss can affect post-loss outcomes. Widowed and non-widowed individuals may differ in their preparedness level *before* the widowhood event occurs. [1] Research from Vable et al. (2015) found that the worsening of widowed individuals' health condition began before the actual death of the spouse. In this paper, we want to elaborate that the pre-loss preparedness level not only has a causal effect on mortality, but also the effect is heterogeneous. If widowed individuals have better preparedness for the loss of their partners, they suffer lower risks.

Indeed, previous literature has introduced the concept of preparedness (for loved one's death) and documented the case of widowhood (Wortman et al. 1993; Treml et al. 2021; Sullivan and Fenelon 2014; Shah et al. 2013). As Kristensen et al. (2012) reviewed, sudden and violent losses of significant others dramatically increase mental health risks, which is a crucial channel linking the widowhood effect on mortality. Barry et al. (2002) pointed out that preparedness for death mediates mental health risks, but they measure the preparedness *after* the loss. Moreover, Shah et al. (2013) further used a UK dataset and found that the mortality risk after the first year of bereavement between unexpected bereavement and bereavement preceded by chronic diseases is significantly higher than 1, suggesting that a higher level of preparedness could protect widowed individuals' health.

[1] For the non-widowed group, widowhood events are a little trickier: the treatment never occurs in the observational period. As widowhood is a time-varying treatment, we can only assume that the occurrence time is when an individual who "cloned" all traits from an original individual assigned to the control group is assigned to the treatment group (i.e., being widowed). When the "cloned" experienced the event, we assume it is the comparable treatment time for the original individual in the control group. We apply the manipulation imputation method described in the analytical strategy section.

Research tracing the spouse's cause of death reaffirms the role of preparedness in the widowhood effect. Elwert and Christakis (2008b) reported that a spouse's death due to acute conditions (infections/sepsis/ influenza/accidents/severe fractures) raises the risks of mortality for widowed individuals, while bereavement due to chronic diseases (diabetes/Alzheimer/Parkins has more minor or no increase in risks for them[2]. On the contrary, Boyle et al. (2011) used Scottish data. They found little evidence supporting that the hazard ratios vary for bereavement causes, while the interactions between the causes and preexisting risks for widowed individuals are significant. Indeed, they differentiated the causes into informative, avoidable, and risky causes, which are not intuitive when directly classifying by the causes of diseases.

In summary, preparedness can be constructed by the health status (preexisting diseases of the spouse) and the socioeconomic status of the couple. Moreover, if the causal widowhood effect exists, we could capture the significant difference in mortality risks after balancing the preparedness for the widowed and non-widowed groups. Given the construction of the preparedness, we presume that the causal widowhood effect is *heterogeneous* along the preparedness:

Assumption 3.II.2 *Widowed individuals with better preparedness for bereavement have lower mortality risks than less-prepared widowed individuals*[3].

[2]However, an exception among the causes of bereavement is chronic obstructive pulmonary disease (COPD), which is a chronic disease but significantly raises the risks for mortality.

[3]Here I take preparedness as a pre-loss attribute that moderates the effect of widowhood, not as a different "version" of the treatment. Under this interpretation, SUTVA's no-interference part is unchanged, and consistency holds in the usual way: for someone who is widowed (or not) the observed outcome equals the corresponding potential outcome, with preparedness simply indexing strata of heterogeneity.

C. Heterogeneous Effects Among Socioeconomic Statuses

So far, we have discussed the role of preparedness *before* the widowhood event, namely the spouse's death, in a heterogeneous way. Social welfare and public policy literature also pointed out that interventions *after* widowhood could reduce the mortality hazard. For instance, high-quality informal care and the coordination between formal and informal care reduced the widowhood effect on mortality risks (Leggett et al. 2020; Jin and Christakis 2009). As social science researchers, we focus more on how social traits heterogeneously affect the widowhood effect (Elwert and Christakis 2006) and how they differentiate the results for widowed individuals with various preparedness levels after bereavement.

In general, two competing theories on social traits and the widowhood effect are the **protective theory** and the **compensation theory**. According to the protective theory, having a high socioeconomic status (SES) can act as a buffer, mitigating the damages associated with widowhood and reducing the negative impacts traditionally observed in the widowhood effect. On the other hand, the compensation theory posits that losing a high-SES partner results in higher compensatory costs, thus making individuals with higher SES more susceptible to the adverse effects of widowhood, thereby associating a high SES with a more significant widowhood penalty. The dichotomy between these theories presents a nuanced examination of how social stratification can differentially influence individuals' experiences of widowhood and its subsequent effects on their well-being. To dialogue with previous literature, we consider educational levels and wealth.

C.1 Protective Effects

As discussed above, ideas from protective theory suggest that socioeconomic status is a buffer for the effect of widowhood on mortality risk. The idea is intuitive from the discussions of the positive relationship between socioeconomic status and health: People with higher socioeconomic status maintain better health status because they have better access to healthcare services (Braveman et al. 2011), health insurance (Adler and Newman 2002), and health-related information (Berkman et al. 2011); furthermore, they usually live in better environments, have more nutritional food, and adopt healthier behaviors (Stringhini et al. 2010). Since health status is highly correlated with mortality risks, we expect lower mortality risks for widowed individuals with higher socioeconomic status.

On the dimension of education, usually, individuals with higher educational levels are healthier than their less-educated counterparts (Kitagawa and Hauser 1973; Montez et al. 2011; Sasson 2016). Similarly, researchers expect an "educational gradient" in mortality: less-educated people will likely have higher mortality risks (Fan and Qian 2019). Thus, the widowhood effect on mortality will also follow the gradient: less-educated people suffer higher penalties because they have limited resources to help them recover from grief, compared to the more educated. For instance, highly educated people are better at health literacy, enabling them to make better-informed health decisions after the trauma.

Assumption 3.II.3 (Education Protective Assumption) *Under the protective theory, more highly educated people suffer less from the widowhood penalty on mortality than the less educated.*

Another dimension concerning education and the widowhood effect is educational homogamy. Previous literature has indicated that educational homogamous families have

better health conditions (Brown et al. 2014; Monden et al. 2003) since couples with similar educational backgrounds have economic stability, which positively impacts access to healthcare and environments (Schwartz and Mare 2005), share comparable health behaviors (Makela et al. 1997), social support (Kiecolt-Glaser and Wilson 2017), and reduce stressed situations better (Luo and Klohnen 2005). Ostergren et al. (2022) reported that educational homogamy level attenuates the mortality risks with Swedish data. Therefore, if the protective theory holds, we expect educationally homogamous widowed individuals to suffer lower widowhood penalties on mortality risks than heterogamous widowed individuals, whether upward or downward married.

Assumption 3.II.4 (Educational Assortative Mating Protective Assumption) *Under the protective theory, educationally homogamous widowed individuals suffer less from the widowhood penalty on mortality than upward and downward heterogamous widowed individuals.*

Finally, the family's financial condition is widely regarded as a protective mechanism against health and mortality risks (Kitagawa and Hauser 1973; Christenson and Johnson 1995; Elo 2009). Wealthy individuals could better afford high-quality caretaking services for their spouses and themselves (Sullivan and Fenelon 2014), especially for lengthy and costly illnesses (Sullivan and Fenelon 2014), contributing to a lower mortality rate (Jin and Christakis 2009). Therefore, we expect widowed individuals with higher family assets to have lower widowhood effects on mortality rates.

Assumption 3.II.5 (Family Asset Protective Assumption) *Under the protective theory, widowed individuals with more family assets suffer less from widowhood penalty on mortality.*

However, some previous empirical studies have found evidence contradictory to the predictions from the protective theory. Although higher socioeconomic status may protect individuals from higher mortality rates, it may not protect against mortality directly from losing a spouse.

C.2 Compensation Effects

Compensation and specialization theory suggests that the heterogeneity in socioeconomic status for widowhood affects mortality differently from the mechanism by which socioeconomic status contributes to health status and mortality risks. Instead, they focus on the meaning of losing a spouse. For the compensation or specialization theory, they suggest that people in higher socioeconomic status families are associated with more specialized gender roles in marriage. Thus, if they lose their spouse, the compensation cost is higher, exposing them to higher mortality risks (Manor and Eisenbach 2003).

For education levels, several empirical results suggest that mortality risks increased as widowed individuals' education levels improved. Lusyne et al. (2001) used the dataset from Belgium in 1991 – 1996 and found that higher educational levels, although they generally alleviate mortality risks, increase the number of excess deaths shortly after bereavement. Meanwhile, Ostergren et al. (2022) used Swedish register data from 2007 – 2016 and reached the same conclusion that the relative risk of mortality for losing a spouse is higher for higher-educated individuals.

Assumption 3.II.6 (Education Compensation Assumption) *Under the compensation theory, more highly educated people suffer more from the widowhood penalty on mortality than the less educated.*

Indeed, there are two limitations of the literature above: first, almost all studies supporting the compensation theories were conducted in high-income European welfare states (besides the studies mentioned above in Belgium and Sweden, similar results are shown in Scotland (Boyle et al. 2011), Israel (Manor and Eisenbach 2003), and Finland (Martikainen and Valkonen 1998)), while the educational system (especially secondary and tertiary education) in the US is slightly different. The compensation cost of losing one's spouse might be higher in these countries than in the US context. Secondly, the studies above measure the widowhood effects on mortality with short-term excess deaths after bereavement, and in our studies, we mainly focus on longevity instead of the short-term impact.

Compensation theory also explains the relationship between educational homogamy and widowhood effects on mortality. The compensation cost would rise as the spouse's educational level increases; meanwhile, we expect that if the spouse's educational level is higher than the widowed individual's, then the compensation cost would be higher, and the widowhood penalty should be more significant. This corresponds to the conclusions from Fan and Qian (2019), who documented that educational homogamy has little contribution to widowhood effects. With compensation theory, the mortality risk associated with widowhood should be higher in cases of upward educational matching, followed by educational homogamy, and is lowest in scenarios of downward matching.

Assumption 3.II.7 (Educational Assortative Mating Compensation Assumption) *Under compensation theory, widowhood effects on mortality risks should be higher in educational upward matching cases, followed by educational homogamy, and lowest in scenarios of downward matching.*

Finally, we discuss family assets. Literature with specialization and compensation theory has documented that wealthy people may be more vulnerable to grief and depression from losing their spouse (Bowling 1987: 1989; Martikainen and Valkonen 1998). Wealthier couples may have a higher degree of role specialization, where each partner takes on distinct roles. In addition, more affluent individuals might incur higher compensatory costs following the loss of a spouse due to a lifestyle built around a higher level of financial resources. This lifestyle becomes unsustainable or emotionally taxing in the absence of their partner. Consequently, losing a spouse can disrupt this balance, potentially leading to a heightened experience of grief and depression as the surviving spouse grapples with newfound responsibilities or a sense of loss concerning their partner's unique contributions to their shared life.

Assumption 3.II.8 (Family Asset Compensation Assumption) *Under compensation theory, widowed individuals with higher family assets suffer more widowhood penalty on mortality than widowed individuals with lower family assets.*

III. Analytical Strategy

A. Marginal Hazard Ratios

To our knowledge, most recent literature analyzes the widowhood effect with the parametric discrete-time hazard models, which treat the survival time (mortality time) as outcomes observed in discrete periods, or the semiparametric continuous-time Cox Proportional Hazard (Cox-PH) models, which assume the survival time follows a continuous distribution. Therefore, suppose the occurrence of death is S_j at time j : $S_j = 1$ if alive, $S_j = 0$ otherwise. D denotes the dummies for periods, A_j indicates whether widowhood occurred at

j, P_j means the preparedness score, C_j represents the time-varying covariates, and X denotes the time-invariant confounders. $\alpha, \beta, \gamma, \delta, \eta, \zeta$ are all parameters to be solved. If we use the logit function as the linking function for the discrete-time survival model (Suresh et al. 2022), it is:

$$logit(S_j = 1) = \log\left(\frac{S_j}{1 - S_j}\right) = \alpha + \beta \sum_{j=1}^{J} D_j + \gamma A_j + \delta P_j + \eta C_j + \zeta X$$

Meanwhile, we need to specify the proportional hazard assumption for the continuous-time Cox-PH model: variables on the right side of the equation only change the failure chance but not the timing. Suppose $h(t)$ denotes the hazard for individuals at time t, $h_0(t)$ is the reference baseline hazard, based on the proportional hazard assumption, for any given time t, we have (Cox 1972; Therneau and Grambsch 2000):

$$\log\left(\frac{h(t)}{h_0(t)}\right) = \gamma A(t) + \delta P(t) + \eta C(t) + \zeta X$$

This model above is an extended Cox-PH model because A, P, and C are time-varying variables. To make it more intuitive, we assume their covariates γ, δ and η are consistent through time (for technical details, see Thomas and Reyes 2014, Therneau and Grambsch 2000,Zhang et al., 2018). For both models, we can directly calculate the hazard ratio between widowed and nonwidowed by calculating e^{γ}. The estimand is called the marginal hazard ratio (MHR) (Hernán 2010; Aalen et al. 2015). However, the e^{γ} obtained from a standard (extended) Cox model is an *associational* hazard ratio; it is not, by itself, a causal estimand in the Rubin–Neyman potential–outcomes framework [4] because it is not defined as a contrast of coun-

[4] Formally, a causal hazard-ratio estimand would specify $h^{a(\cdot)}(t)$, the hazard under intervention (or regime) $a(\cdot)$—e.g., a path that sets widowhood vs. nonwidowhood—and contrast $h^{a(\cdot)}(t)/h^{a'(\cdot)}(t)$. Identification requires consistency, positivity, and (sequential) exchangeability. With time-varying confounders affected by prior exposure, standard Cox adjustment can be biased, whereas a marginal structural Cox model can identify the *marginal causal* hazard ratio under those assumptions (Hernán 2010; Aalen et al. 2015). Hazard ratios are also non-collapsible, so conditional and marginal HRs generally differ even without unmeasured confounding.

terfactual hazards under a well-specified intervention on widowhood (Mao et al. 2018). [5].

B. Average Treatment Effect Estimation

B.1 Mean Survival Time Differences

In Rubin's causal framework, under unconfoundedness, positivity, and consistency assumptions, the average treatment effect (ATE) refers to the difference in outcome between the treated and the control groups, and the conditional average treatment effect (CATE) refers to the difference between treated and control outcomes conditional on covariates. In our analysis above, the heterogeneous treatment effect (HTE) is intrinsically the CATE evaluated at fixed levels of preparedness. If the outcome is a survival variable, we may generate the estimand of the ATE by measuring the treatment and control group difference in probability of survival before a specific time, or we could compare the average survival time difference between the treatment and the control group (Mao et al. 2018):

$$\int_0^\infty S_{A=1,C,X}(t)dt - \int_0^\infty S_{A=0,C,X}(t)dt$$

Where $S(t)$ is defined as the survival function $S(t) = P(T \geq t) = \exp(-\int_0^t h(u)du)$ (Therneau and Grambsch 2000). In most cases, we could not observe the entire survival curve in the empirical analysis due to right censoring. Therefore, we rely on the model assumption to yield the mean survival time. For instance, in this paper, as we use the Cox-PH model, we first use the Breslow method (Breslow 1975) to calculate the baseline hazard and survival function at t: $\hat{h}_0(t)$ and $\hat{S}_0(t) = \exp(-\int_0^t \hat{h}_0(u)du)$. We then calculate the individual survival function as $\hat{S}_i(t) = \hat{S}_0(t)^{\exp(\hat{\gamma}A_i(t)+\hat{\delta}P_i(t)+\hat{\eta}C_i(t)+\hat{\zeta}X_i)}$. For a counterfactual group-level survival curve under $A = a$, we average individual counterfactual survival curves, $\hat{S}_{A=a}(t) =$

[5]To compare the quality of the data we cleaned with previous studies, we reported our results with the MHRs from the continuous-time Cox-PH models for men and women, and the results could be referred to in Appendix B.1 for interested readers.

$n^{-1}\sum_i \hat{S}_i(t \mid A_i(t) = a)$, rather than multiplying the individual survival functions. Finally, we integrate the survival function to get the (restricted) mean survival time: $\int_0^{\tilde{T}} \hat{S}_{A=a,C,X}(t)dt$ (in which $\tilde{T}$ refers to the latest observation time). It is worth noting that unlike most treatment variables in survival analysis, which are assigned before the observation time starts, the treatment (widowhood) in this research is a time-dependent variable that occurs during our observation time; we use the extended Cox-PH model discussed above to predict $\hat{h}(t)$. We will discuss the technical details of the hazard function estimation in Appendix B.2 for interested readers.

B.2 Nuisance Function for Propensity Scores and Preparedness Scores

We calculate the propensity score of the treatment effect (widowhood). Let π represent the propensity score. In general, π is a function of a set of covariates predicting the probability of widowhood. We use Z to denote the set of covariates here, and hence,

$$\pi(Z) = P(A = 1 \mid Z = z)$$

Which is in the range between 0 and 1 and $\hat{\pi}(Z)$ is our target. In the main paper, we present the predictions on the propensity score from the logistic regression.

It is worth noting that we also use the same set of covariates to construct a baseline "preparedness" measure, denoted $S = g(Z)$, which we treat as an effect modifier. We consider the operationalization on expectedness-as-risk: where $S \equiv \hat{\pi}(Z)$. $\hat{\pi}(Z)$ is used for confounding adjustment (via weighting, matching, or stratification), while S enters the outcome/MSM stage as the moderator. We then estimate preparedness-specific effects

$$\text{CATE}(s) \;=\; E[Y(1) - Y(0) \mid S = s]\,,$$

under the usual identification conditions (consistency, positivity, and conditional exchangeability given Z). When $S \equiv \hat{\pi}(Z)$, heterogeneity is interpreted as variation by baseline risk of widowhood [6].

C. Doubly Robust Estimator for Treatment Effect

Finally, we will employ the doubly robust method to debias the ATE and HTE estimations. The pure-imputation or IPW estimands are naïve plug-in estimators: with asymptotic analysis, we know they will yield plug-in bias in estimation (Tsiatis 2006; Kennedy 2022b). We adopt the sample splitting and cross-fitting method and construct the doubly-robust learner (DR-Learner) estimator to make our ATE and HTE estimators regular and asymptotically efficient under the stated regularity and nuisance-estimation conditions (van der Laan and Robins 2003). We split the dataset into two subsets, D_1 and D_2, let $\hat{\mu}_1 = \int \hat{S}(t \mid A = 1)dt$ and $\hat{\mu}_0 = \int \hat{S}(t \mid A = 0)dt$, Y denotes the observed survival/censoring time. For the general model and every socioeconomic trait (education, marital type, and asset levels), the steps are as follows:

- Use D_1 to estimate the nuisance functions for the propensity score $\hat{\pi}$, the mean survival time for the treatment group $\hat{\mu}_1$ and for the control group $\hat{\mu}_0$.
- Construct the pseudo-outcome based on the efficient influence function (EIF) of Rubin's ATE:

$$\hat{\lambda}_i = \frac{\mathbb{1}(A_i = 1)}{\hat{\pi}(Z_i)}\left[Y_i - \hat{\mu}_1(X_i)\right] - \frac{1 - \mathbb{1}(A_i = 1)}{1 - \hat{\pi}(Z_i)}\left[Y_i - \hat{\mu}_0(X_i)\right] + \left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right)$$

[6] Using the same covariates Z for both $\pi(Z)$ and $S = g(Z)$ does not make the two quantities identical. To avoid redundancy, we estimate $\hat{\pi}(Z)$ from Z and then include S only as the effect modifier in the outcome stage (not alongside Z inside the propensity model). We also assess overlap within S strata and, if necessary, trim extreme weights.

- Regress $\hat{\lambda}_i$ on the preparedness scores S in D_2 to yield the doubly robust estimator for the heterogeneous treatment effect (the conditional average treatment effect on preparedness scores).

- Swap D_1 and D_2 and repeat the previous steps for cross-validation and use the average outcome as the final doubly robust estimator.

IV. Data and Measurement of Variables

The data used in this research are from the Health and Retirement Study (HRS) 1998 – 2018 waves. HRS is a biennial longitudinal household survey designed to understand aspects of population ageing in the US. Although the HRS began before 1998, our analytic sample uses the 1998 wave as the baseline for this study. There are, in total, 40,957 individuals included in the study. In our analysis, as we mainly investigate the widowhood effect among older adults, we select respondents over 50 and married in their first interview, and set the first interview year as the start time (27,913 cases dropped). If the respondents exited the interview in the following waves or their deaths were not observed until the 2018 survey, we treat them as right-censored cases. Moreover, we drop the cases who were same-sex married or experienced remarriage after widowhood (232 cases) as we believe the patterns between widowhood and mortality for them could differ from those of heterosexuals who only experienced bereavement. Thus, we have 12,812 individuals. We establish models separately for men and women, with 6,667 men and 6,145 women in our study. Table 3.1 describes their distributions stratified by education and wealth. Because of the time-varying covariates, we transfer the wide data (person) into the long form (person-year) for analysis.

The treatment variable is whether the respondent experienced widowhood. We use a set of time-varying and time-constant variables to predict the propensity of widowhood, which

Table 3.1: **Sample Size By Gender (HRS 1998 - 2018)**

Characteristics	Total	%	Not Widowed Dead	Not Widowed Alive	Widowed Dead	Widowed Alive
Men (N = 6,667)						
Lower than College	3917	58.75%	1983	1276	369	289
College and above	2750	41.25%	1112	1300	166	172
Husband College, Wife College	1713	25.69%	643	871	99	100
Husband College, Wife Non-college	1004	15.06%	464	404	66	70
Husband Non-college, Wife College	799	11.98%	385	301	61	52
Husband Non-college, Wife Non-college*	3151	47.26%	1603	1000	309	239
Lowest Strata Asset	2250	33.70%	1151	721	209	169
Middle Strata Asset	2221	33.26%	1031	889	162	139
Upper Strata Asset	2206	33.04%	913	966	174	153
Women (N = 6,145)						
Lower than College	3888	63.27%	1012	1197	717	962
College and above	2257	36.73%	444	995	287	531
Wife College, Husband College	1506	24.51%	282	713	172	339
Wife College, Husband Non-college	730	11.88%	161	270	113	186
Wife Non-college, Husband College	930	15.13%	212	346	134	238
Wife Non-college, Husband Non-college*	2979	48.48%	801	863	585	730
Lowest Strata Asset	2022	32.90%	548	566	388	520
Middle Strata Asset	2062	33.56%	460	770	340	492
Upper Strata Asset	2061	33.54%	448	856	276	481

is also the individual's preparedness score for losing a spouse. Sociodemographic variables include the spouse's birth year (age), educational years, age difference with the partner, race, religion, migration status, number of shared children (in the current household), and number of living children. The socioeconomic variables include the household's past-three-year average log assets, spouse's past-three-year average log income (pension), retirement status, years since retirement, and years since children died, if ever. Spouse's health conditions are a set of variables measuring the mean or maximum values in the last three years: the average number of drinking days and quantity, average hospital time per year, average BMI, average CESD scores, the maximum time visiting the hospital per year, the mean spouse's medical expenditures, whether the spouse was living in the nursing home, and whether the spouse was diagnosed with high blood pressure, diabetes, any cancer, lung diseases, heart diseases, stroke, and psychological illness (with each diagnosis as a dummy). Once the spouse dies, the propensity of widowhood for the individual in the following waves remains the same as the propensity score predicted in the last alive wave until the individual's death or censoring.

The model predicting the survival time for the widower has a similar structure to predicting the propensity of widowhood, except we replace the spouse's traits with her own characteristics. Moreover, we include the predicted propensity from the last stage and years since widowhood into the model. The detailed descriptive statistics are listed in Table 3.2.

Table 3.2: Descriptive Table for Variables (HRS 1998-2018)

	Value for Respondent	Value before Death	Value for Spouse	Value before Widowhood
Men, widowed (N = 996)				
Observed Death	0.53 (0.50)			
Death Age	85.81 (7.74)		79.01 (8.46)	
Years between Death and Widowhood	5.17 (3.92)			
Years between Death and Retirement	23.10 (8.67)		20.16 (11.48)	
Number of Symptoms[b]	2.24 (1.24)	2.16 (1.23)	2.19 (1.29)	2.20 (1.28)
Mean Household Log Asset	10.17 (2.18)	10.87 (2.73)		
Mean Log Income	8.21 (1.71)	9.13 (1.97)	5.73 (1.77)	7.99 (1.98)
Men, not widowed (N = 5,671)				
Observed Death	0.55 (0.50)			
Death Age	79.13 (8.72)			
Years between Death and Retirement	18.15 (8.53)			
Number of Symptoms[b]	2.04 (1.28)	1.99 (1.24)	1.38 (1.14)	
Mean Household Log Asset	9.55 (2.50)	10.65 (3.09)		
Mean Log Income	7.21 (2.27)	8.47 (2.74)	6.23 (2.12)	
Women, widowed (N = 2,497)				
Observed Death	0.39 (0.49)			
Death Age	84.93 (8.00)		76.33 (8.38)	
Years between Death and Widowhood	6.18 (4.11)			
Years between Death and Retirement	25.93 (12.39)		19.03 (8.31)	
Number of Symptoms[b]	1.97 (1.21)	1.92 (1.18)	2.28 (1.30)	2.29 (1.28)
Mean Household Log Asset	10.46 (2.09)	10.71 (2.73)		
Mean Log Income	7.76 (1.59)	8.86 (1.94)	6.36 (1.96)	9.03 (1.77)
Women, not widowed (N = 3,648)				
Observed Death	0.40 (0.49)			
Death Age	77.21 (8.87)			
Years between Death and Retirement	20.16 (12.01)			
Number of Symptoms[b]	1.74 (1.25)	1.79 (1.21)	1.75 (1.21)	
Mean Household Log Asset	9.88 (2.55)	10.39 (3.33)		
Mean Log Income	6.49 (2.15)	7.46 (2.78)	7.42 (2.33)	

Note: [a] Standard deviation is in parentheses. [b] We include six symptoms: high blood pressure, diabetes, cancer, lung problems, heart problems, and stroke. In models, we separate the six symptoms into six dummies, adding them as covariates. Source: Author's calculation based on the HRS data, 1998—2018.

V. Results

In this section, we present the ATE and the HTE results for widowhood effects on mortality for men and women. As noted above, we compare the difference in mean life expectancy between the widowed and non-widowed groups.

A. General Results

Figure 3.1 demonstrates the general results of the widowhood effects for men and women under different levels of preparedness. Results from Figure 3.1 validate Assumptions 3.II.1 and 3.II.2. For both genders, we verify the causal effect of widowhood on mortality from the HRS dataset, while the widowhood effect is more prominent for men than for women. For men, widowed individuals are, on average, 2.6 years shorter in life expectancy than non-widowed individuals, while for women, widowed individuals live 0.4 years shorter than non-widowed peers.

Moreover, as seen in Figure 3.1, for both men and women, the worst-prepared individuals bear a higher widowhood penalty than those well-prepared for their spouse's passing. The worst-prepared men suffer a widowhood penalty for 1.5 more years than the best-prepared men, while the gap between the worst-prepared women and best-prepared women is only 0.10 years.

B. College Education and Educational Homogamy

We then discuss the heterogeneity in educational levels and educational homogamy patterns for widowhood effects. Figure 3.2 demonstrates the average and heterogeneous treatment effects for widowhood effects separately for college-educated and non-college-educated men and women.

We first look at the difference between college-educated men and non-college-educated men. As the upper panel of Figure 3.2 shows, on average, men with college degrees and above have an average of 0.6 years shorter life expectancy if they ever experience widowhood. In comparison, men without college degrees average 4.3 years shorter if ever wid-

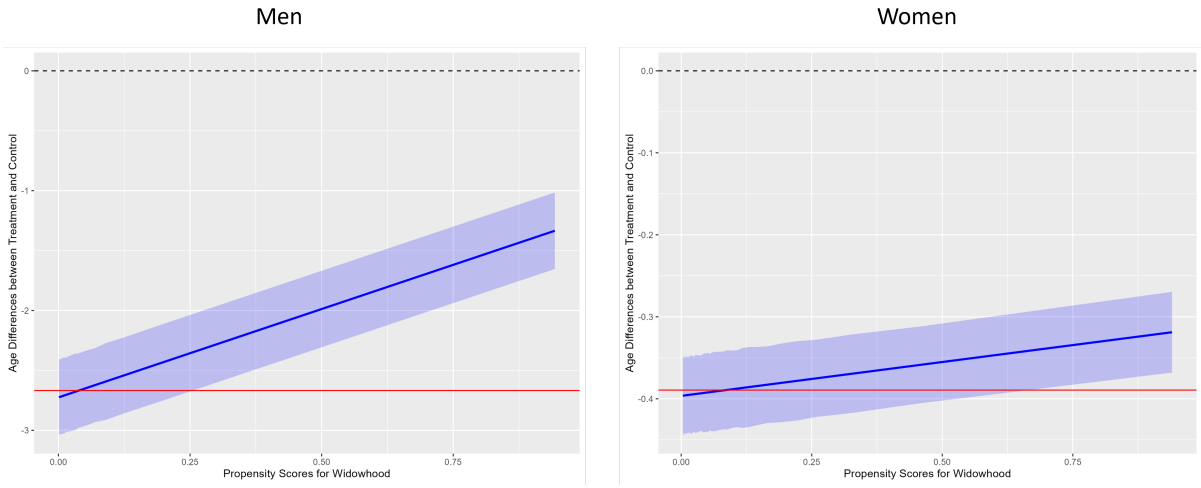


Figure 3.1: Average Treatment Effect and Heterogeneous Treatment Effect of Widowhood Effects

Note: Red and blue lines separately indicate the linear approximation of the ATE and the HTE, while the black dash indicates the horizontal line of difference is 0. The blue ribbon denotes the 95% confidence interval for the HTE estimation from 100 bootstraps.

owed. The results suggest that a college degree reduces the widowhood penalty for men. However, the average widowhood effect for college-and-above and non-college women is almost the same. Widowhood reduces college-and-above women's life expectancy by an average of 0.45 years, while non-college women's life expectancy by an average of 0.35 years.

We then discuss the heterogeneity in the widowhood effect. The upward slopes in the four graphs show that the best-prepared widowers suffer less than the worst-prepared widowers in life expectancy reductions, although the magnitudes differ. For college-and-above-educated men, although the average treatment effect indicates that widowhood in general still negatively affects life expectancy, the heterogeneous analysis shows that at the well-prepared end, widowed men survive longer than the non-widowed men for 0.3 years, though the effect is not significant. For non-college-educated men, we find that the increase in preparedness level reduces the difference in life expectancy gap between wid-

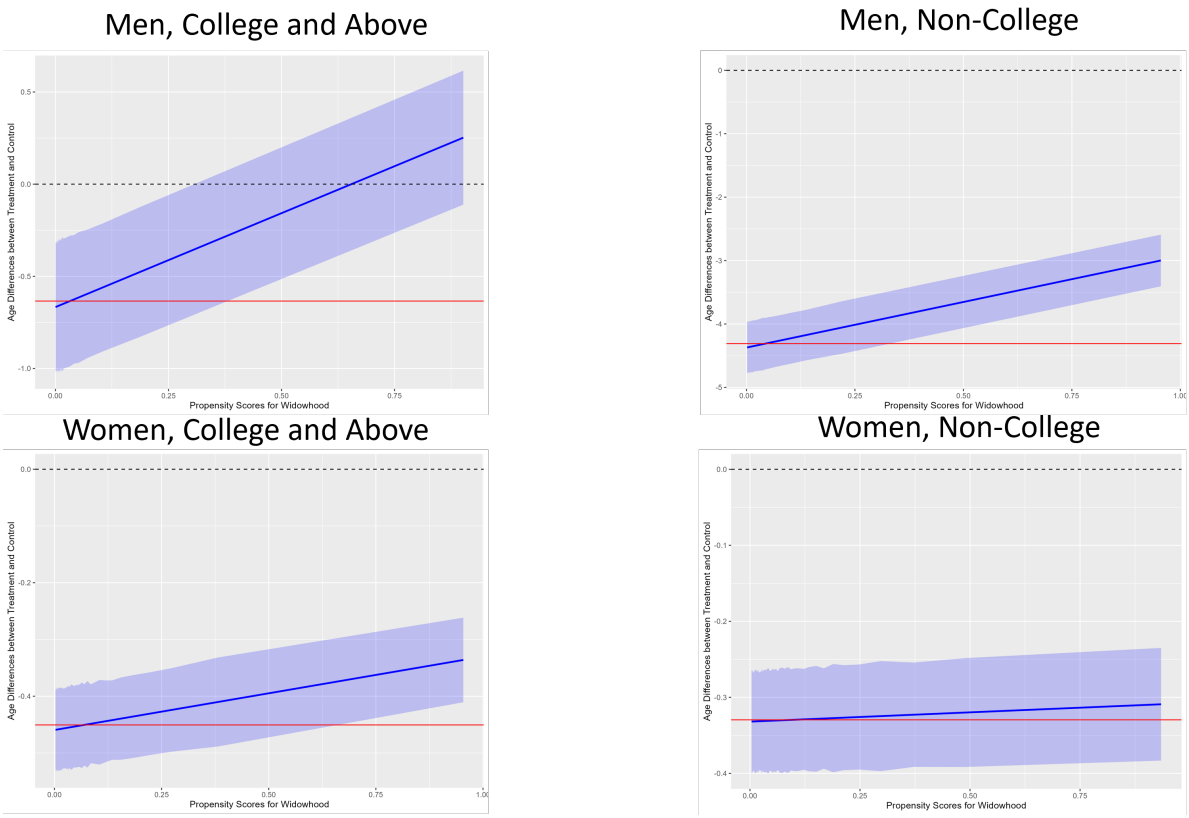


Figure 3.2: College Education and Widowhood Effects

Note: Red and blue lines separately indicate the linear approximation of the ATE and the HTE, while the black dash indicates the horizontal line of difference is 0. The blue ribbon denotes the 95% confidence interval for the HTE estimation from 100 bootstraps.

owed and non-widowed most rapidly, with 4.3 years fewer for the widowed group at the worst-prepared end but three years fewer at the best-prepared end. For both college-and-above-educated and non-college-educated women, preparedness could only reduce the life expectancy gap between the widowed and non-widowed by 0.1 years. The effect for non-college-educated women is not significantly different from the average treatment effect.

Then, we turn to the heterogeneity in the widowhood effect for homogamy status. We regard college–college and non-college–non-college matches as homogamous marriages,

while college–non-college matches are considered heterogamous. The results are shown in Figure 3.3. The results show that homogamy reduces the widowhood penalty on mortality for women but not for men. The average treatment effect of widowhood shortens educational homogamous men's life expectancy by 2.5 years, while shortening educational heterogamous men's life expectancy by 1.8 years. However, heterogamous women, on average, suffer one year less in life expectancy if they experience widowhood, while homogamous women, on average, only suffer 0.1 years less in life expectancy if widowed. The results indicate that homogamous men suffer more grief from losing their partners than heterogamous men. In contrast, homogamous women cope with losing their partner better than heterogamous women.

We then discuss the impact of preparedness on widowhood for homogamous and heterogamous groups. The results also show a gender disparity. For men, the increase in preparedness significantly reduces the widowhood penalties. The widowhood penalty shrinks from -2.5 years to -1.5 years for homogamous men and from -1.9 years to -0.1 years for heterogamous men. Also, at the best-prepared end, the difference in life expectancy between the widowed and non-widowed men is not statistically significant. On the women's side, there's almost no difference between the widowed and non-widowed for the best-prepared homogamous women, while on the worst-prepared end, the widowed have a 0.1-year gap in life expectancy with the non-widowed. For heterogamous women, there's almost no difference between the best and the worst-prepared ends.

In summary, as for education, we find that higher education reduces the widowhood penalty, but the reduction only exists for men. Furthermore, we find educational homogamy asymmetrically impacts the widowhood effect for men and women: educational homogamous men suffer a larger widowhood penalty than heterogamous men. In contrast, the

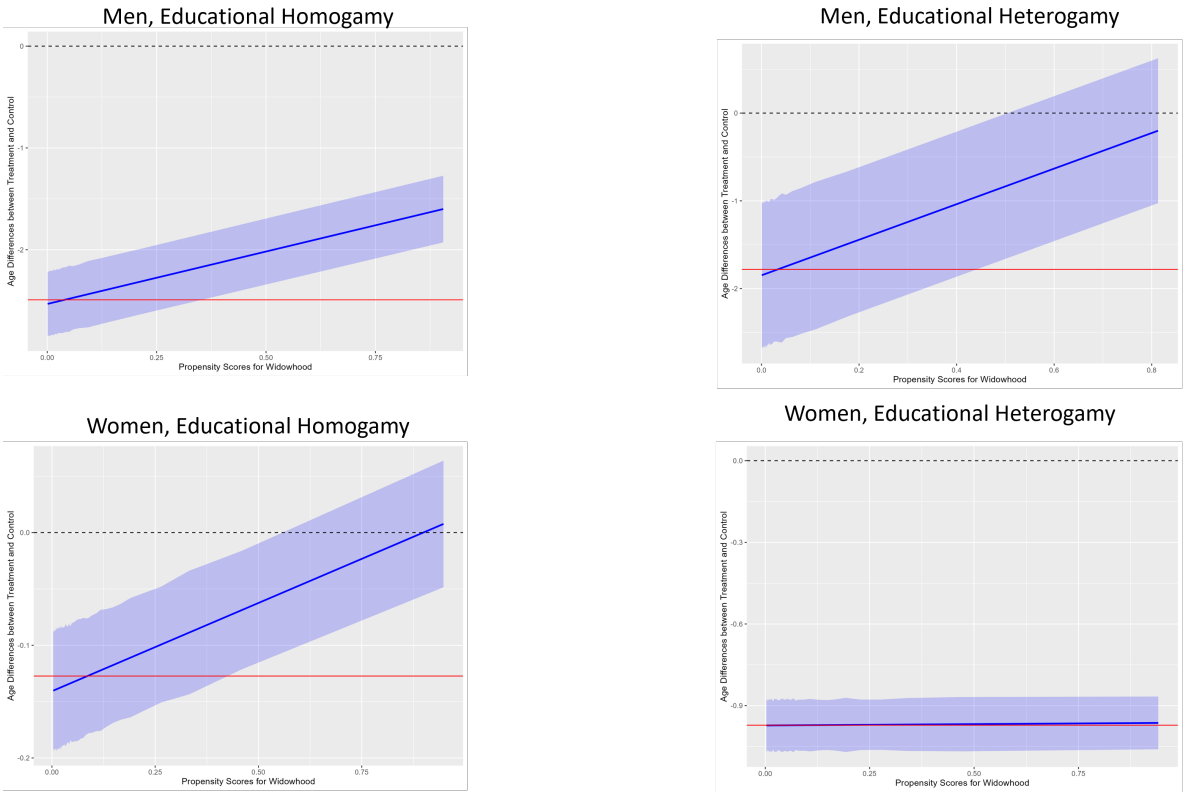


Figure 3.3: Educational Homogamy and Widowhood Effects

Note: Red and blue lines separately indicate the linear approximation of the ATE and the HTE, while the black dash indicates the horizontal line of difference is 0. The blue ribbon denotes the 95% confidence interval for the HTE estimation from 100 bootstraps.

heterogamous women suffer more. We believe that the results indicate different sociological meanings for education and educational homogamy for men and women before Generation X (the research population in this study) and reflect the educational gradient in the widowhood effect. Before Generation X, a large proportion of women (both college and non-college) in those generations worked as housewives. Therefore, education does not stratify the effect of their husband's death. In this regard, education has a protective effect, but only for men. Besides, since most college entrants were men (though the gap reduced and flipped in the later decades), the heterogamous men are mostly married downward (college men with non-college women) while the heterogamous women are more likely

to be upwardly matched. If we bring in the compensation and specialization theory, we may expect that the compensation cost for widowhood is highest for upward matching and lowest for downward matching. Thus, educational heterogamy augmented the widowhood penalty for women while alleviating it for men.

C. Wealth

Finally, we discuss how widowhood effects differ among different asset groups for men and women. Since assets may change dramatically due to medical expenditures, when we classify the families into the three terciles, we use their family assets around the age of 50 (the start of the observation time). The results are shown in Figure 3.4. Intuitively, we could find that all graphs indicating better preparedness scores reduce widowhood penalties. However, only men at the lowest tercile, men at the highest tercile, and women at the highest tercile show statistically significant improvement. The detailed description of the figure is as follows.

We start with the ATE for men. The average widowhood penalties in life expectancy for men at the lowest, middle, and highest terciles are 2.0 years, 0.5 years, and 2.0 years, respectively. The results indicate that men at the lowest and the highest terciles suffer almost the same widowhood penalties in life expectancy. In contrast, the middle tercile men suffer a much smaller effect than the other groups. We may attribute the results to the existence of both the protection effect and the compensation effect: the lowest tercile men are more vulnerable to widowhood than the middle tercile men because their assets cannot protect them from the grief; meanwhile, the highest tercile men are also more vulnerable because of the more specialized gender roles within the family and higher compensation costs.

On the women's side, widowed women in the lowest asset terciles have an average of

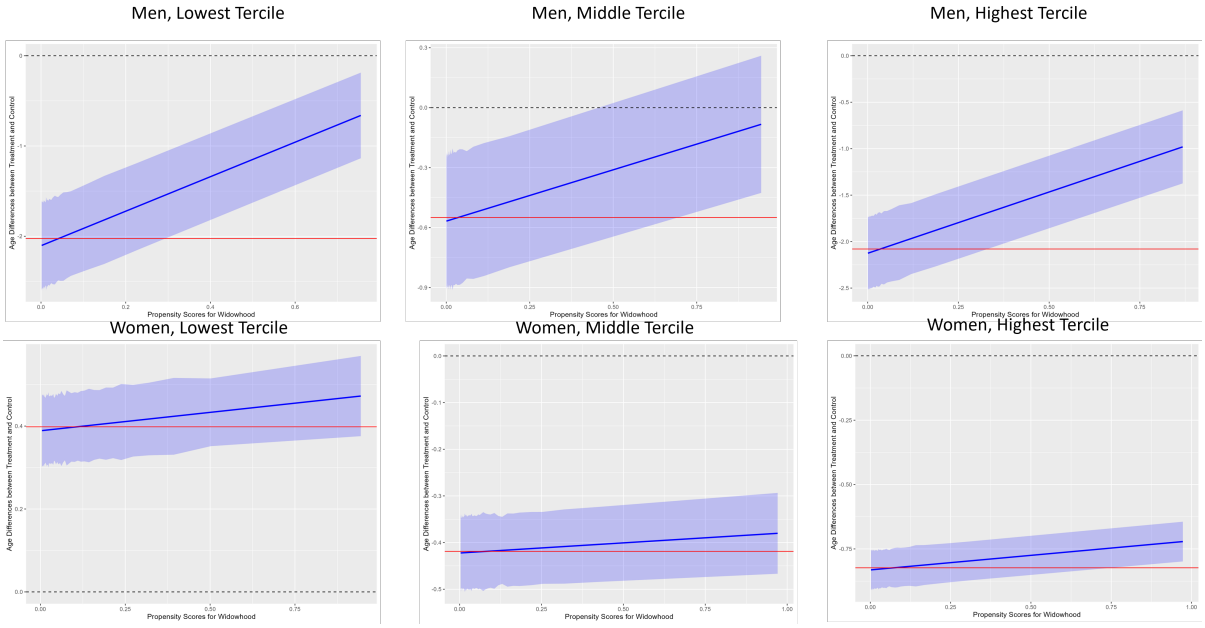


Figure 3.4: Wealth and Widowhood Effects

Note: Red and blue lines separately indicate the linear approximation of the ATE and the HTE, while the black dash indicates the horizontal line of difference is 0. The blue ribbon denotes the 95% confidence interval for the HTE estimation from 100 bootstraps.

0.4 years longer in life expectancy than non-widowed women. For the middle terciles, the effect turns to a widowhood penalty for widowed women with a shorter life expectancy of 0.42 years, and the penalty increases for the highest tercile at an average of 0.83 years. The result shows that the widowhood penalty increases with family assets for women. Therefore, for women, the effect of widowhood is mainly on the compensation side; they suffer higher risks in high SES families when they lose their valuable spouse.

We then look at the heterogeneous treatment effects along with the preparedness scores for widowhood. For men at the lowest tercile, the increase in preparedness scores mitigates the widowhood penalties from 2.1 years to 0.6 years for widowers compared to non-widowers. For the middle tercile men, the widowhood penalties mitigate the worst-prepared

0.8 years gap between widowers and non-widowers to almost no gap. However, the heterogeneous effect is not statistically significant compared to the average effect. Finally, for the men with the highest tercile assets, the increase in preparedness scores reduces the widowhood penalty for the widowers from nearly 2 years to 1 year at the best-prepared end.

For women, better preparedness scores do not reduce the widowhood penalties on life expectancy, as the bottom panel of Figure 3.4 shows. For women with the lowest tercile, better preparedness enlarges widowed women's lead in longevity over non-widowed women from 0.39 years to 0.47 years, while for the middle tercile and the highest tercile women, better preparedness respectively shrinks the penalty of widowhood from 0.43 years to 0.38 years, and from 0.83 years to 0.72 years. The results indicate that for women, the increase in preparedness scores has little impact on the causal effects of widowhood.

In summary, our analysis of both average and heterogeneous treatment effects reveals distinct gender-based mechanisms through which assets affect the impact of widowhood on mortality for American elders before Generation X. For men, we find that both compensation and protection effects coexist, rendering the middle tercile the most resilient group against the adverse consequences of widowhood on mortality. Conversely, higher asset levels are associated with a magnified widowhood effect for women, implying that the relationship between assets and the widowhood effect is best understood through the lens of compensation cost effects. Moreover, we discover that preparedness scores can more effectively mitigate the negative impact of widowhood among vulnerable groups, specifically men in the lowest and highest asset terciles and women in the highest asset tercile, compared to other segments.

D. Further Discussions

In the HRS data, since most respondents are white, we further restrict our sample to white men and women to test the robustness of the results. We present the results in Appendix B.3. The results show that, except for white women in the lowest strata of assets, patterns in other subgroups are consistent with the results presented in the main paper. We suspect the abnormal results in the lowest strata for white women are due to the reduced sample size, as white women are more likely to be included in the middle and higher asset strata than minorities. Thus, we believe our results are generally robust to this subsample restriction.

VI. Conclusion and Further Discussions

This study uses the HRS data to present the heterogeneous treatment effect of widowhood on mortality in the US. Empirically, under the stated identification assumptions, we find evidence consistent with a causal widowhood effect on longevity for both men and women born before Generation X, although the impact is more significant for men than women. The empirical results also support our assumptions on the heterogeneous effects of preparedness. For both the general population and sub-populations filtered by different socioeconomic strata, we find that better preparedness will improve widowed individuals' mean life expectancy compared with people on the side of lower preparedness scores.

We also analyzed the heterogeneous treatment effects of education and family wealth. We find that men and women have different mechanisms for the effects of widowhood. For men, higher socioeconomic status, on the one hand, protects them from the widowhood penalty while, on the other hand, increases their compensation cost. For women, we only

find evidence supporting the compensation theory from the analysis of both education and wealth. We believe the result is caused by gender segregation in education and the labor market before Generation X in the US, as women in that generation socioeconomically rely more on their husbands than the later generations. Men, as household heads, could find social resources based on their socioeconomic traits, while on the other hand, relying on a specialized gendered role of the wife in domestic work. Therefore, the meaning of widowhood differs for widowed men and women. With later cohort data available, it would be intriguing to analyze whether the patterns continue or change when more women receive higher education, participate in the labor market, and become more independent.

Methodologically, this paper adopts the doubly robust estimation to estimate the heterogeneous treatment effect of widowhood on mortality. Compared to the previous studies, we tried our best to set the pseudo-randomized scenario for widowhood treatment and then test the causal effects. We estimate the mean life expectancy based on time-varying Cox Proportional Hazard (Cox-PH) models to grab the average and heterogeneous treatment effects. The results should be more intuitive than previously presented with hazard ratios.

However, a shortcoming in our model is that for the covariates we selected for fitting the likelihood of widowhood and mortality, we still could not avoid the endogenous selection problem: unobserved dyad- or household-level factors (e.g., shared lifestyle, caregiving burden, genetic risk, financial shocks) may jointly increase the partner's death risk and the respondent's mortality. If these variables really affect the conditional unconfoundedness assumptions, our ATE and HTE estimates can only be interpreted as the associational contrasts conditional on the covariates, not fully identified causal effects. Indeed, we use the

traits of the spouse to model widowhood and the traits of the respondents to model mortality, and future research may use stronger identification designs to reduce confounding. For instance, researchers could seek quasi-exogenous variation in partner mortality risk to construct instruments and identify local effects.

Another shortcoming of the methodological side is that we still apply a Cox-PH model to estimate the survival function. The Cox-PH model is biased if the proportional hazard assumption is violated. However, in our research, our primary interest is the treatment of widowhood, and thus, in both the full-population model and the sub-population models on education and wealth, we add the interactive term between the treatment and years since widowhood into the models so the problem is addressed [7]. Ideally, we feel it would be best if our model could be doubly robust in estimating the causal effects (dealing with the missing counterfactuals in the treatment or control group) and also doubly robust in estimating the survival function (dealing with the missing counterfactual for censoring data). Future methodological improvements could overcome the shortcomings and deepen our understanding of the sociological meaning of bereavement.

[7] Meanwhile, we also test the proportional hazard assumptions with the Schoenfeld test. The results are presented in Appendix B.5.

Chapter 4

Doubly Robust Estimation for Static and Dynamic Causal Mediation Analysis

I. Introduction

In this chapter, we continue discussing the efficient/doubly robust causal estimator in the context of **causal mediation**. Previously, researchers commonly used mediation analysis in social science and epidemiology to understand the mechanism of causality and the formation of post-treatment heterogeneity. The treatment effect can be decomposed into the direct treatment effect on the outcome and the indirect treatment effect on the outcome via the mediator. In summary, this chapter aims to explain the efficient/doubly robust estimand for the decomposed terms [1].

Previous literature analyzing the causal mediation effect used Pearl's (2009) *"do-calculus" (DoC)* framework (see Chapter 1), which treats the treatment variable and the mediator vari-

[1] In this chapter, we only cover doubly robust estimators, although some relevant literature has mentioned the "multiply robust" estimators in causal mediation analysis (for instance, see Zhou 2021). In general, it is not the number of nuisance functions that determines if the estimator is doubly robust or multiply robust. Instead, if and only if we have multiple nuisance functions, and any one of them is correctly specified, that makes the estimator unbiased, we can call the estimator multiply robust. As we will elaborate in the chapter, the estimators given here are all doubly robust. Although we have multiple nuisance functions, only one of the two conditions is satisfied, and we can still yield unbiased estimation, which is clearly the doubly robust estimator.

able as actions and derives the flow chart called the **directed acyclic graph (DAG)** to assist the analysis. VanderWeele's (2015) book gives the basic concepts and analytical framework for causal mediation, which is the starting point of the discussions in this chapter.

Recently, terms and analytical frameworks in causal mediation have been further applied to the causal disparities between groups. In epidemiological and demographic research, researchers put variables that could not be regarded as treatments in the DoC framework (for instance, gender, race, birthplace, etc.) into groups and measure how contextual variables amplify or alleviate the gaps between the groups. The analytical framework is called the **"reduced disparities"** in epidemiology (Jackson and VanderWeele 2018) and **"gap-closing estimand"** (Lundberg 2024) in demography and social science.

Furthermore, researchers may apply the causal mediation analytical framework to the time-varying models, as they may encounter time-varying treatments, mediators, and confounders in their empirical research. For instance, they may try to decompose the total treatment effects into direct and indirect effects through the time-varying mediators; they may assume heterogeneity in treatment effects if the treatments occurred at different times, or they may differentiate the short-term causal and mediation impact from the long-term ones. We call the models with time-varying confounders the "dynamic models" (to distinguish them from the static models assuming no time-varying effects). $G-$**formula** is the common toolbox to address the time-varying confounding (Robins 1986; 1994). In this paper, we give an efficient influence function (EIF) based doubly robust estimator of the components in dynamic/time-varying causal mediation models.

In this chapter, we will present the EIF-based DR estimator for the natural, controlled, and interventional direct and indirect effects for the static and dynamic causal mediation analysis. The sections in this chapter are as follows: in SectionII, we will review the DoC framework for causal inference and the terms and assumptions for causal mediation analysis; Section III gives a review of the efficient/doubly robust estimation for causal/treatment effects (which has been covered in the introduction chapter); Section IV gives the efficient/doubly robust estimation for the components in static causal mediation models. In Section V, we turn to the concepts and the assumptions for the dynamic model. Then, in Section VI, we derive the doubly robust/efficient estimator for the components in the dynamic causal mediation model. Finally, in Section VII, we set out our proposed algorithm in the empirical research: we replicate classic causal mediation analysis—Fearon and Laitin's (2003) analysis of the mediation effects of political instability on the causal mechanisms between racial fractionalization and civil wars—with our new estimators.

II. DoC Framework and Static Causal Mediation Model Assumptions

A. Mutual Causality, Confounders, Colliders, and Mediators

The causal inference we discussed in the previous chapters is based on the Neyman-Rubin (NR) framework (Neyman 1990; Rubin 1974), in which we assume the existence of the counterfactual *status* of the outcome. We aim to apply statistical methods to quantitatively identify the value of the counterfactual status. By comparing the observed/factual status values, we could evaluate the effects of the treatment on the outcome.

The **Do-Calculus (DoC) framework**, developed by Judea Pearl (Pearl 2009), tells another story. Due to the self-evident message conveyed in the causal relationship that the cause precedes the effect, we can use a unidirectional arrow to represent the relationship

between the treatment A[2] and the outcome Y: $A \rightarrow Y$. In the dyadic relationship between A and Y, the status of Y will not change unless we manipulate the status of A (for instance, manipulate $A = 1$ to $A = 0$ in the binary treatment-control setting). The manipulation, in Pearl's (2009) term, is a *do*. The right arrow denotes the causal effect of A on Y, and statistical models aim to identify the value of the causal effect, in the DoC notations, $E[Y \mid do(A = 1)] - E[Y \mid do(A = 0)]$.

Based on the dyadic relationship between A and Y: $A \rightarrow Y$, consider a third variable M interfering in the effect of A on Y. With the unidirectional arrow notation, the way M could intervene $A \rightarrow Y$ can only fall into four types: $A \leftarrow M \leftarrow Y$; $A \leftarrow M \rightarrow Y$; $A \rightarrow M \leftarrow Y$; and $A \rightarrow M \rightarrow Y$. In the introduction chapter, we have discussed the two sources of bias for causal effect estimation: the pre-treatment selection and the post-treatment heterogeneity. Here, we discuss how the four diagrams contribute to the bias in causal identification.

We start with the notation $A \leftarrow M \leftarrow Y$. Obviously, if we simultaneously have the relationships $A \rightarrow Y$ and $A \leftarrow M \leftarrow Y$, then A and Y are in a bidirectional relationship/ feedback loop, and we call A and Y **mutually causal** as we cannot identify the order in which A and Y occur. Thus, there's no causal effect between A and Y if $A \leftarrow M \leftarrow Y$ holds [3].

Then we simultaneously have $A \rightarrow Y$ and $A \leftarrow M \rightarrow Y$. In this case, M can be regarded as a **confounder**, which affects both the pre-treatment selection and the post-treatment heterogeneity. In this case, the order of occurrence is first the confounders M, then A, and

[2] In many DoC framework and causal mediation literature, the treatment is also called the exposure.

[3] In Pearl's notation, circulation is actually not permitted because the graphs are "directed acyclic graphs". However, in this part, we just introduce the scenarios for the relationships among A, M, and Y and hence we include the scenario for mutual causality here.

finally Y. To make a valid causal inference of A on Y, we must ensure the outcome is independent of the treatment conditioned on the confounders, which is the unconfoundedness assumption for causal inference we have discussed. In this chapter, we use C to denote the confounders affecting both the treatment and the outcome.

Next, if $A \rightarrow M \leftarrow Y$ and $A \rightarrow Y$ are simultaneously satisfied, then the M variable here is called a **collider**. The collider will not affect our identification on $A \rightarrow Y$, as in the triangle relationship Y precedes M, so the order of occurrence is first A, then Y, and finally M. However, some researchers may encounter **collider bias** if they over-complicate the identification of the causal relationship between A and Y by stratifying on the levels of M or manually fixing M at specific values. To illustrate this, consider a prevalent scenario in social science where we set A as the education level, Y as the income, and M (the collider) here as the job performance. We assume the colliding structure, which implies that theoretically, we assume that people with higher educational levels will have better job performances, and people with higher income will perform better in their jobs as they have greater access to resources and experiences. However, if we select our samples from specific job performance groups, we may find within the performance level, the correlation between education and income may be negative, leading to biased causal estimation between education and income. Hence, the collider bias is also a form of pre-treatment selection.

The final type of the triangle relationship is we simultaneously have $A \rightarrow Y$ and $A \rightarrow M \rightarrow Y$. In this situation, M stands for the **mediator**. The time order of occurrence is first A, then M, and finally Y. Therefore, adding the mediator M is indeed decomposing the causal effect estimation into two parts: the **direct effect** $A \rightarrow Y$ and the **indirect** effect $A \rightarrow M \rightarrow Y$. If the mediator is omitted, the estimation of the **total causal effect** of A on Y is still un-

biased, but we overestimate the direct effect (the specific path $A \rightarrow Y$) as we ignore the indirect paths between the treatment and the outcome.

In the DoC framework, unidirectional arrows are crucial as the **fairness** rule of $A \rightarrow Y$ is *"do A, then Y"*: fixing $A = a$, then we have $Y(a)$ (Recall this is called the consistency assumption in the NR framework). If, in the diagram, we select any variable as the starting point and follow any unidirectional arrows and will not return to the starting variable, we call the diagram a **directed acyclic graph (DAG)**, and we can infer the causal relationship between any variables in a DAG. In the four types above, since we can turn back to A if we start at A in the mutual causality graph ($A \rightarrow Y \rightarrow M \rightarrow A$), we cannot infer any causal relationship among the variables. On the other hand, for the confounding, colliding, and mediation graphs, we can all infer the causal relationship between any of the two variables. The essence of confounding, colliding, and mediation is almost identical, with only the time sequence of variables exchanged. Therefore, in this chapter, we will only discuss the mediation relation.

B. G-formula, and Sequential Ignorability Assumption

In the mediation analysis, as the fairness rule suggests, due to the two paths $A \rightarrow M$ and $A \rightarrow M \rightarrow Y$, we should have the potential mediators $M(a)$ when we do the treatment $A = a$ if the treatment is the only intervention, and the potential outcomes $Y(a, m)$ when we do $A = a, M = m$ if the treatment and the mediator are the only interventions. Like in the introductory chapter, we still assume that we have a binary treatment: $A = 1$ or $A = 0$. Now we discuss what assumptions are required for the statistical estimand (conditional expectation) to estimate the potential outcomes for the mediator $M(a)$ from the conditional expectation

of the observational data. [4]

Under the DoC framework, $E[M(a)]$ is equivalent to $E[M \mid do(A = a)]$ if the treatment is the only intervention without any **back-door** paths. In social science research with observational survey data, the condition is too ideal because researchers could not artificially manipulate the treatment; however, theoretical framework may guide researchers to set a group of covariates C_{AM} which blocks all back-door paths from A to M and contains no descendant of A (C_{AM} is ignorable). In potential outcome terms, as we discussed in the introduction chapter, this corresponds to the familiar assumptions: consistency (fairness), positivity (overlap), and unconfoundedness (ignorability) assumptions. The consistency assumption is satisfied if the fairness rule holds; the positivity assumption says that the probability of the treatment should be between $(0,1)$ for any treatment value; and the unconfoundedness assumption requires that all the confounding variables are controlled between the mediator and the treatment. In the DoC framework, the unconfoundedness assumption is equivalent to suggesting that there is no (backdoor) path between the treatment and the mediator (conditional ignorability of A for M given C_{AM}). Formally, we have:

Assumption 4.II.1 (Assumptions for Mediator Fairness) *To identify $E[M(a)]$ from $E[M \mid A]$* [5]*, we have the following assumptions:*

- *Consistency: the observed mediator is the same as the potential mediator under the*

[4] Following the terms in the introduction chapter, we still call the function/expression from the statistical models to the potential outcomes the estimand, and the function from the observational/empirical data to estimate the statistical function the estimator.

[5] In the following two chapters, we use $E[Y \mid do(A = a)]$ to denote the artificial manipulation of the treatment in the do-calculus term, $E[Y(a)]$ to denote the potential outcomes, and use $E[Y \mid A = a]$ to denote the conditional expectation with the observational data. As we explained above, in social science research with observational survey data, we rarely have the pure do manipulation, and thus our focus is still on how to use the observational conditional expectation to identify the potential outcomes.

treatment:

$$M = M(a) \text{ if } A = a;$$

- *Positivity: the probability for the mediator to be assigned to treatment and control group should be a positive number between* 0 *and* 1*, for every covariates* c *in the support of the back-door path* C_{AM}*:*

$$P(A = a \mid C_{AM} = c) \in (0,1); \quad a = \{0;1\}, f(c) > 0;$$

- *Unconfoundedness: conditioned on our controlled treatment-mediator confounders* C_{AM}*, the potential outcomes for the mediator is independent of the treatment assignment* A*:*

$$M(a) \perp\!\!\!\perp A \mid C_{AM}.$$

With the assumptions held, we have :

$$\begin{aligned} E[M(a)] &= E[E_{C_{AM}}[M(a) \mid C_{AM}]] \quad \text{(conditional expectation)} \\ &= E[E_{C_{AM}}[M(a) \mid A, C_{AM}]] \quad \text{(positivity and unconfoundedness)} \\ &= E[E_{C_{AM}}[M \mid A = a, C_{AM}]] \quad \text{(consistency)} \\ &= \int_{c_{AM}} E[M \mid A = a, c_{AM}] f_{C_{AM}} dC_{AM} \end{aligned} \tag{4.II.1}$$

The DAG for Assumption 4.II.1 is quite straightforward:

Similarly, if we only consider the binary relationship between A and Y and would like to use $E[Y \mid A = a]$ to infer $E[Y(a)]$, with observational data and in the DoC terms, we require a set of covariates C_{AY} to separate all back-door paths from A to Y and contains no descendants of A (to make the effects ignorable). As we discussed in Chapter 1, in potential

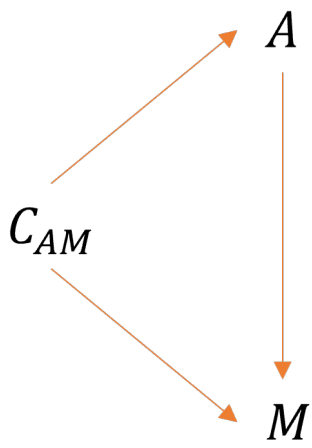


Figure 4.1: Directed acyclic graph (DAG) for Assumption 4.II.1.

outcome terms, this aligns with the trio of assumptions– consistency, positivity, and unconfoundedness (ignorability) [6]:

Assumption 4.II.2 (Assumptions for Outcome Fairness) *To identify $E[Y(a)]$ from observed data, it suffices that:*

- ***Consistency:*** *the observed outcome equals the potential outcome under the realized treatment:*

$$Y = Y(a) \;\; \text{if } A = a.$$

- ***Positivity:*** *for every c in the support of C_{AY}, treatment assignment has nonzero probability:*

$$P(A = a \mid C_{AY} = c) \in (0,1), \quad a \in \{0,1\}; f(c) > 0$$

- ***Unconfoundedness:*** *given C_{AY}, the potential outcome is independent of treatment assignment:*

$$Y(a) \perp\!\!\!\perp A \mid C_{AY}.$$

[6] See also in Chapter 1 Assumption 1.II.4.

Under Assumption 4.II.2, the standardization (the g-formula) identifies the interventional mean:

$$\begin{aligned} E[Y(a)] &= E\{E[Y(a) \mid C_{AY}]\} \\ &= E\{E[Y(a) \mid A, C_{AY}]\} \quad \text{(by exchangeability and positivity)} \\ &= E\{E[Y \mid A = a, C_{AY}]\} \quad \text{(by consistency)} \\ &= \int E[Y \mid A = a, C_{AY} = c]\, f_{C_{AY}}(c)\, dc \end{aligned} \tag{4.II.2}$$

The transition in Equation 4.II.2 is called the $g-$**formula** (which is a more common term in mediation analysis), as "g" stands for the generalized (Robins 1986; Hernán and Robins 2020; Naimi et al. 2017; Taubman et al. 2009; Snowden et al. 2011):

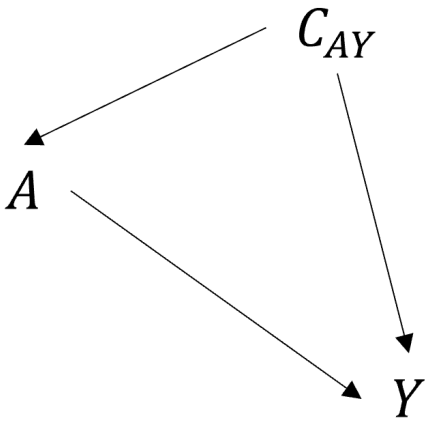
Definition 4.II.1 ($g-$formula) *The $g-$formula, or the generalized formula, states the relationship between the causal estimand and the statistical estimand. Specifically, if we have the treatment A, the outcome Y, and the covariates C which suffice for ignorability, under the consistency, positivity, unconfoundedness assumptions ($Y(a) \perp\!\!\!\perp A \mid C$), we have the $g-$formula linking the causal estimand to a function of the observed data:*

$$E[Y(a)] = E\{E[Y \mid A = a, C]\} = \int E[Y \mid A = a, C = c]\, f_C(c)\, dc.$$

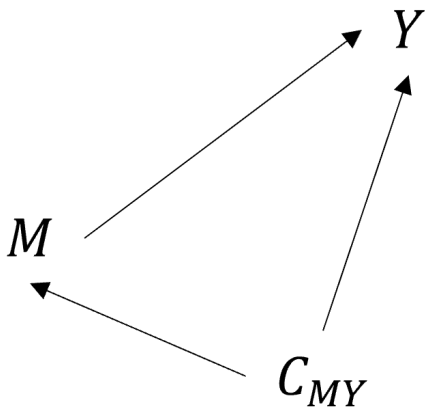
where f_C is the density of C when it exists and F_C is the distribution function of C.

In words, the g-formula standardizes the conditional outcome regression $E[Y \mid A = a, C]$ over the marginal distribution of C, thereby identifying $E[Y \mid A = a] = E[Y(a)]$ under the stated assumptions. Similarly, if we consider identifying the binary relationship between M and Y for $E[Y(m)]$ with the observed data, we need a set of covariates C_{MY} that blocks the back-door path between M and Y and contains no descendants of M, with the trio assumptions of consistency, positivity, and unconfoundedness. The DAGs for identifying

$E[Y(a)]$ and $E[Y(m)]$ are illustrated in 4.2:



(a) DAG for identifying $E[Y(a)]$: C_{AY} blocks all back-door paths from A to Y and contains no descendants of A.



(b) DAG for identifying $E[Y(m)]$: C_{MY} blocks all back-door paths from M to Y and contains no descendants of M.

Figure 4.2: Directed acyclic graphs (DAGs) for identifying $E[Y(a)]$ and $E[Y(m)]$. Panel (a) illustrates the treatment–outcome relation with C_{AY}; panel (b) illustrates the mediator–outcome relation with C_{MY}.

In causal mediation analysis, we focus on the trio relationship among A, M and Y and as we illustrated above, we have a sequential order: the occurrence of events is first the treatment A then the mediator M and finally the outcome Y. As the target of causal mediation analysis is always to explain how much the mediator enhances or mitigates the total treatment effect, i.e., the divergence between the potential/interventional outcomes under treated $E[Y(1)]$ and under control $E[Y(0)]$. As the potential outcomes are intervened sequentially by the treatment A and the mediator M, the target outcome can be denoted as $Y(a,m)$ when we set $A = a$ and $M = m$. The assumption of **sequential ignorability** is needed to identify the controlled potential outcome $Y(a,m)$ in observed data (and $Y(a,m)$ in some literature is also called the **controlled response function (CRF)**, see Zhou 2021) be-

cause identification proceeds stage by stage along the temporal order from A to M to Y, and at each stage we need to replace the interventional quantity with an observed conditional. The DAG for the identification can be seen in 4.3:

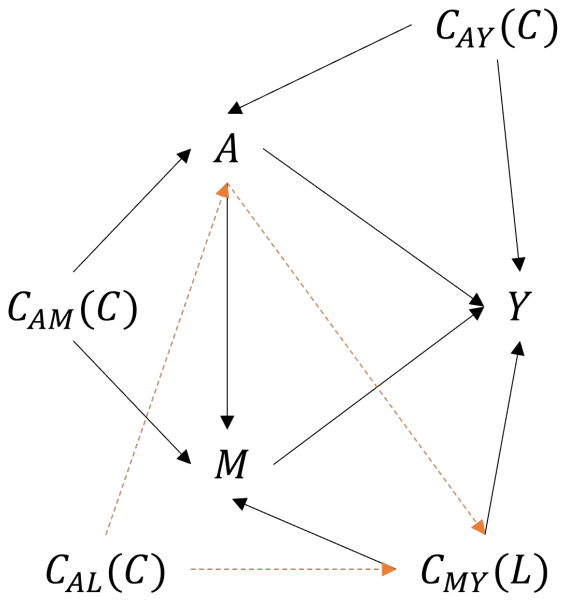


Figure 4.3: DAG for identification of the controlled response function $Y(a,m)$ under sequential ignorability.

Figure 4.3 can be seen as the combination of Figures 4.1 and 4.2, and we add the possible path that A might affect C_{MY} and further affect the outcome Y (we may also have a stronger assumption that A and C_{MY} are independent, which is called the cross-world assumption we will discuss in the following subsection). Like what we did previously, we set a group of covariates C_{AL} to block the back-door path and assume that the covariates C_{AL} contain no descendants of A. Since covariates C_{AM}, C_{AL}, and C_{AY} all do not contain descendants of the treatment, we use C to denote them as the pre-treatment covariates. Covariates C_{MY} contain no descendants of M but are subsequent to A, we use L to denote them later in this thesis and call them post-treatment covariates.

As identification is the process to replace the interventional quantity with the observed conditional, we start with the observed baseline covariates C. Using the $g-$formula, we have:

$$E[Y(a,m)] = \int E[Y(a,m) \mid C = c] f_C(c) dc$$

As the density function for $f_C(c)$ is observable under the pre-treatment positivity conditions, there's no problem with the transformation. However, the conditional expectation on the right side $E[Y(a,m) \mid C = c]$, also based on the $g-$ formula, if we take the post-treatment covariates L into consideration, we further have:

$$E[Y(a,m) \mid C = c] = \int_l E[Y(a,m) \mid L(a) = l, C = c] f(L(a) \mid C = c) dl \tag{4.II.3}$$

Notice this time on the right side of the equation, both $E[Y(a,m) \mid L(a), C]$ and $f(L(a) \mid C = c)$ contain expressions for potential outcomes. Hence, we need the first stage (the treatment stage) ignorability/ unconfoundedness so that with both positivity and consistency assumptions for the treatment, we could have:

$$f(L(a) = l \mid C = c) = f(L(a) = l \mid A = a, C = c) = f(L = l, A = a, C = c)$$

and

$$E[Y(a,m) \mid L(a) = l, C = c] = E[Y(a,m) \mid A = a, L(a) = l, C = c] = E[Y(a,m) \mid A = a, L = l, C = c]$$

And hence the first/treatment-stage ignorability is:

$$\{L(a), M(a), Y(a,m)\} \perp\!\!\!\perp A \mid C$$

Now that the right side of Equation 4.II.3 becomes $\int E[Y(a,m) \mid A = a, L = l, C = c] f(L \mid A = a, C = c) dl$. As $f(L \mid A = a, C = c)$ is now an observed conditional distribution of L among those with $A = a$ which we can estimate from observational data (with the positivity

assumption on L holds: for every L: $f(L = l \mid A = a, C = c) > 0$), the only term that contains the potential outcome expression is the conditional expectation. We hence need the second/mediator stage of sequential ignorability:

$$Y(a,m) \perp\!\!\!\perp M \mid A, L, C$$

so that combining with the positivity and consistency assumption for the mediator:

$$E[Y(a,m) \mid A = a, L = l, C = c] = E[Y(a,m) \mid A = a, M = m, L = l, C = c] = E[Y \mid A = a, M = m, L = l, C = c]$$

Combining the transformation process above, we have:

$$E[Y(a,m)] = \int\int E[Y \mid A = a, M = m, L = l, C = c] f(l \mid A = a, C = c) f(C) \, dl \, dc$$

Which is the extended $g-$ formula for identifying $E[Y(a,m)]$. We now summarize the process of identification and the corresponding consistency, positivity, and ignorability (unconfoundedness) assumptions:

Assumption 4.II.3 (Causal Mediation Assumptions) *Suppose a statistical measurable set $Z = (A, M, Y, L, C)$, in which A denotes the treatment, M denotes the mediator, Y denotes the outcome, C denotes the pre-treatment covariates, and L denotes the post-treatment covariates. To make the statistical conditional expectation $\psi(\mathrm{P}) = E_C E_{L|C} E[Y \mid A = a, M = m, L, C]$ equivalent to the controlled response function $\psi'(\mathrm{P}') = E[Y(a,m)]$ from the hypothetical measurable set $Z' = (A, Y(a,m), M(a), L(a), C$ (where $Y(a,m)$ denotes the potential/interventional outcomes under the treatment $A = a$ and $M = m$, respectively; $M(a)$ and $L(a)$ respectively denote the potential mediator and the potential post-treatment covariates fixing $A = a$. Suppose the treatment and the mediator are both discrete variables; we need the following assumptions:*

1. *Positivity: The probability density functions for the pre-treatment covariates and the post-treatment covariates should be positive (distributions are observable):*

$$f(c) > 0; f(l \mid A = a, C = c) > 0$$

the probability for the treatment defined in $[0,1]$ *should be positive:*

$$P(A = a \mid C = c) > 0;$$

And the probability for the mediator given the treatment and the covariates defined on the mediator domain should be positive[7]*:*

$$P(M = m \mid A = a, L = l, C = c) > 0$$

2. *Consistency: the potential outcome under the treatment and mediator received is the same as the observed outcome:*

$$Y = Y(a,m) \text{ if } A = a, M = m$$

The potential mediator and post-treatment covariates under the treatment received are the same as the observed mediator:

$$M = M(a) \text{ if } A = a; L = L(a) \text{ if } A = a$$

3. *Unconfoundedness: treatment-stage ignorability/ unconfoundedness: conditioned on a set of pre-treatment covariates* C*, the treatment* A *should be independent of the potential values of the outcome* $Y(a,m)$*, the mediator* $M(a)$*, and the post-treatment covariates* $L(a)$

$$\{Y(a,m), M(a), L(a)\} \perp\!\!\!\perp A|C$$

[7] Here we only consider discrete treatment and mediators; indeed, if the treatment and mediator are continuous, we should have the probability density functions $f_{A|C=c}(a) > 0$ so that for some neighborhood N_a we can have positive probability: $P(A \in N_a \mid C = c) > 0$; correspondingly, for the mediator, we have $f_{M|A=a,L=l,C=c}(m) > 0$ so that $P(A = a, M \in N_m \mid L = l, C = c) > 0$.

Mediator-stage ignorability/unconfoundedness: conditioned on the pre-treatment covariates C, the treatment A, and the post-treatment covariates L, the mediator M should be independent to the potential outcome $Y(a,m)$:

$$Y(a,m) \perp\!\!\!\perp M \mid A, L, C.$$

Under Assumption 4.II.3, the four estimands on $Y(a,m)$ (Equations 4.II.4, 4.II.6,4.II.7, and 4.II.8) could be rewritten as:

$$E[Y(a,m)] = \int\int E[Y \mid A=a, M=m, L=l, C=c] f(l \mid A=a, C=c) f(C) dl dc \quad \text{(extended g-formula)} \tag{4.II.4}$$

$$= E_C\left\{E_{L|A=a,C}[E[Y \mid A=a, M=m, C, L]\right\} \quad \text{(pure-imputing)} \tag{4.II.5}$$

$$= E\left[\frac{Y\mathbb{1}(A=a)\mathbb{1}(M=m)}{P(A=a \mid C)P(M=m \mid A, C, L)}\right] \quad \text{(pure-weighting)} \tag{4.II.6}$$

$$= E\left[E[Y \mid A=a, M, L, C]\frac{\mathbb{1}(A=a)}{P(A=a \mid C)}\frac{\mathbb{1}(M=m)}{P(M=m \mid A, C, L)}\right] \quad \text{(imputing-then-weighting)} \tag{4.II.7}$$

$$= E\left[\frac{\mathbb{1}(A=a)}{P(A=a \mid C)}E[Y \mid M=m, A, C, L]\right] \quad \text{(weighting-then-imputing)} \tag{4.II.8}$$

Among the expressions, line 1 is the extended $g-$formula as we derived above (Equation 4.II.4); line 2 (Equation 4.II.5) simply changes the integral into the expression of expectation (which is the definition of expectation); line 3 (Equation 4.II.6) is the inverse probability weighting (IPW) expression when we have a discrete mediator M. This is simply because of the change-of-measure identity: $E[Y\frac{\mathbb{1}(A=a)}{P(A=a|C)}] = E_C[E[Y \mid A=a, C]]$ and we applied the expression twice for A and M, with consistency for the outcome and the mediator holding; line 4 (Equation 4.II.7) is the expansion of the IPW expression, as we first impute the outcome fixing the treatment then use the change-of-measure identity for the mediator; line 5 (Equation 4.II.8) is the reverse of line 4, as we use the change-of-measure identity for the

treatment and then fit the conditional expectation with a fixed mediator.

With observable data to identify the interventional/potential outcome $Y(a,m)$, we can further construct the causal mediation estimators based on the theoretical framing and research interests. For example, in policy-controllable studies, some researchers might be interested in the treatment effect, but with the policy intensity at a specific level. Under these circumstances, researchers may identify the controlled direct effect (CDE), fixing the mediator M at specific levels m and estimating the divergence in outcome between the treatment and the control (Robins and Rotnitzky 1992; Pearl 2009; VanderWeele 2015):

Definition 4.II.2 (Controlled Direct Effects) *The controlled direct effect measures the difference between the outcome of the treatment* $A = 1$) *and the control groups* ($A = 0$), *fixing the mediator at a given level* m*:*

$$CDE(m) = E[Y(A = 1, m) - Y(A = 0, m)] \tag{4.II.9}$$

To make the CDE identifiable, only some of the assumptions in Assumption 4.II.3 will be relaxed: as M is fixed at the manipulated level, for the treatment-stage ignorability, we only need $\{Y(a,m), L(a)\} \perp\!\!\!\perp A \mid C$ and for the consistency assumption we do not have the requirements on the potential mediators. Except for those two, all other assumptions remain the same. With these assumptions, Equations 4.II.5, 4.II.6,4.II.7, and 4.II.8 provide four equivalent identifying functionals for $E[Y(a,m)]$, each of which can be used as the basis for constructing unbiased, regular, and asymptotically linear (RAL) estimators.

In other research settings, researchers might be more interested in how to decompose the total average treatment effect (TATE) $Y(1) - Y(0)$ into the direct effect of the treatment on the outcome and the indirect effect of the treatment on the outcome via the mediator.

However, based on different specifications and assumptions, there are different ways of decomposition. In the following subsection, we will elaborate on the decomposition of the total average treatment effect, based on different "world scenarios."

C. Within-World and Cross-World Scenario

In the subsection above, we discussed the assumptions necessary to identify the controlled response function $Y(a,m)$. As mentioned above, the mediation analysis intrinsically decomposes the total average treatment effect $Y(1)-Y(0)$ into the direct effect of the treatment on the outcome and the indirect effect of the treatment on the outcome via the mediator. Hence, the total average treatment effect, if the mediator is included, is:

$$E[Y(1)-Y(0)] = E[Y(1)] - E[Y(0)] = E[Y(1,M(1))] - E[Y(0,M(0))] \tag{4.II.10}$$

where $M(1)$ and $M(0)$ denote the estimand for the mediator when $A=1$ and $A=0$. When the expression with all subsequent variables after A are from the same treatment value, they are called in the same A world. So the TATE actually compares the mean outcome in the $A=1$ world to that in the $A=0$ world, and it is purely a **within-world** expression: each expectation keeps A and its corresponding mediator from the same world.

Now that we would like to decompose the TATE into the direct effect, which is the direct impact of the treatment on the outcome without going through the mediator, and the indirect effect, which is the part of the treatment's impact on the outcome that goes through the mediator, we may have several theoretical constructions on the estimate: we can compare the outcome under the treatment and control while keeping the mediator at the hypothetical level it would have been under the other exposure (treatment or control); we can also compare the outcome under the treatment and control while randomizing the mediator to follow the distribution it has under a specific exposure (treatment or control). For the first

one, as we take the mediator at the "natural" levels of the treatment, we decompose the total effects into **natural direct and indirect** effects (Robins and Rotnitzky 1992; Pearl 2001); for the second one, as we intervene on the mediator to follow the distribution it has under a specific exposure, we call the decomposed terms **interventional direct and indirect** effects.

For the natural direct and indirect effects, the decomposition of the TATE can be expressed as:

$$TATE = E[Y(1)] - E[Y(0)] = E[Y(1, M(1))] - E[Y(0, M(0))]$$

$$= \underbrace{E[Y(1, M(1))] - E[Y(1, M(0))]}_{NIE}$$
$$+ \underbrace{E[Y(1, M(0))] - E[Y(0, M(0))]}_{NDE}$$

Where NIE stands for the natural indirect effect and NDE stands for the natural direct effect. since we construct the counterfactual outcome under one treatment level and the mediator under another treatment level (for instance, Y(1, M(0))), the subsequent variables M and Y after the treatment A do no longer exist within one world; rather, the potential outcome is in a cross-world in which it contains the treatment from the "treated" world and the mediator from the "control" world. Decomposing the TATE into the natural direct and indirect effects, hence falls in the **cross-world** scenario (Pearl 2012; Richardson and Robins 2013).

For the interventional direct and indirect effects, we intervene on the mediator using a stochastic policy rather than setting it to its unit-specific counterfactual value. Let $G(a)(\cdot \mid \cdot)$ denote a stochastic intervention that draws M from the distribution it would have under $A = a$ (typically conditional on baseline covariates). We use $Y(a, G(a^\star))$ to denote the potential outcome when treatment is set to a and the mediator is generated from the policy $G(a^\star)$.

Under this notation, the interventional indirect effect (IIE) and interventional direct effect (IDE) decompose the contrast

$$E[Y(1,G(1))] - E[Y(0,G(0))],$$

which VanderWeele (2015) refers to as the *overall effect* to distinguish it from the average treatment effect (ATE/TATE). Specifically,

$$\begin{aligned} OE &:= E[Y(1,G(1))] - E[Y(0,G(0))] \\ &= \underbrace{E[Y(1,G(1))] - E[Y(1,G(0))]}_{IIE} + \underbrace{E[Y(1,G(0))] - E[Y(0,G(0))]}_{IDE}. \end{aligned}$$

These interventional effects remain in the **within-world/single-world scenario** because they only involve potential outcomes indexed by a single treatment level at a time (with the mediator generated by a policy), rather than cross-world counterfactuals such as $Y(1,M(0))$ (Andrews and Didelez 2021).

Importantly, in general $Y(a) \neq Y(a,G(a))$ and hence $E[Y(1)] - E[Y(0)]$ (the ATE/TATE) need not equal OE[8]. The ATE/TATE remains defined as $E[Y(1,M(1))] - E[Y(0,M(0))]$, whereas $IIE + IDE$ equals $E[Y(1,G(1))] - E[Y(0,G(0))]$. We re-write Assumption 4.II.3 under the cross-world scenario as follows:

Assumption 4.II.4 (Cross-World Assumptions) *The following assumptions are required to identify the natural direct and indirect effects*

1. *Within-world assumptions:*

[8]The difference between the natural and interventional effects is that identifying natural direct and indirect effects involves cross-world counterfactuals (e.g., $Y(1,M(0))$), which typically requires stronger, cross-world independence assumptions. In settings with post-treatment covariates L that are affected by A and confound the mediator–outcome relationship, natural effects are generally not identified without additional strong assumptions; one sufficient (but restrictive) simplification is to rule out such intermediate confounding (e.g., by assuming A does not affect L).

(a) *Positivity: The probability density functions for the covariates should be positive (distributions are observable):*

$$f(c) > 0; f(l \mid C = c) > 0$$

the probability for the treatment defined in $[0,1]$ *should be positive:*

$$P(A = a \mid C = c) > 0;$$

And the probability for the mediator given the treatment and the covariates defined on the mediator domain should be positive:

$$P(M = m \mid A = a, L = l, C = c) > 0; P(M = m \mid A = a^{\star}.C = c) > 0$$

(b) *Consistency: the potential outcome under the treatment and mediator received is the same as the observed outcome:*

$$Y = Y(a,m) \text{ if } A = a, M = m$$

The potential mediators under the treatment received are the same as the observed mediators:

$$M = M(a) \text{ if } A = a;$$

Covariates L are not affected by the treatment[9]*:*

$$L(a) = L(a^{\star}) = L$$

(c) *Unconfoundedness: treatment-stage ignorability/ unconfoundedness: conditioned on a set of pre-treatment covariates C, the treatment A should be independent of the potential values of the outcome* $Y(a,m)$ *and the mediator* $M(a)$

$$\{Y(a,m), M(a)\} \perp\!\!\!\perp A|C$$

[9]We here do not call L as the post-treatment covariates because L is not affected by A and thus can be treated as the baseline covariates.

Mediator-stage ignorability/unconfoundedness: conditioned on the covariates C and L, the treatment A, the mediator M should be independent to the potential outcome $Y(a,m)$:

$$Y(a,m) \perp\!\!\!\perp M \mid A, L, C.$$

2. *Cross-world assumption: conditioned on the pre-treatment covariates C, the potential mediator in the cross-world $M(a^\star)$ should be independent to the potential outcome $Y(a,m)$:*

$$Y(a,m) \perp\!\!\!\perp M(a^\star) \mid C$$

For the interventional effects, Assumption 4.II.3 satisfies the requirements. We can also show the differences in the DAG for the cross-world and within-world scenarios in Figure 4.4 (VanderWeele 2015). With the assumptions being stated, we obtain the following observa-

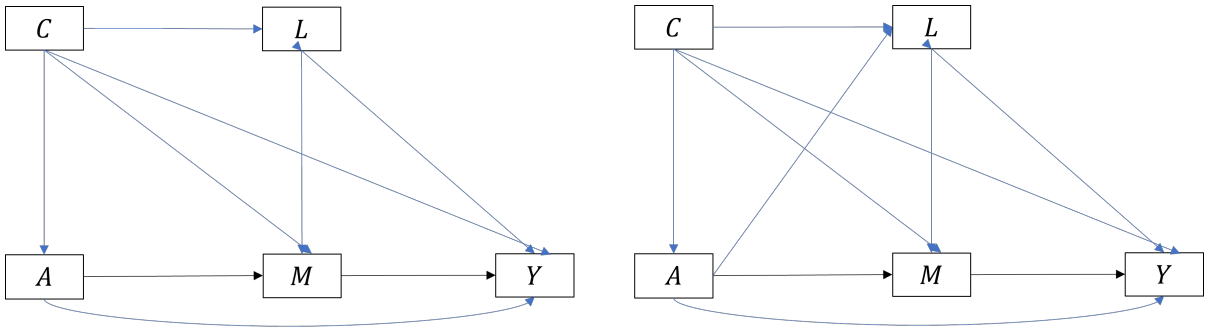


Figure 4.4: Two Scenarios for Unconfoundedness Assumption between the Treatment and the Mediator-Outcome Covariates
Note: Panel a shows the situation in which Assumption 4.II.4 holds (the cross-world scenario); thus, the NDE and the NIE can be yielded. Panel b shows the situation in which Assumption 4.II.3 holds (the within-world scenario), in which we allow the post-treatment covariates L affected by the treatment A, and we can only yield the interventional direct and indirect effects.

tional identifying functions for the NDE and NIE (via the extended $g-$formula or Equation

4.II.5) [10]:

$$\begin{aligned}\text{NDE} &= E_C E_{L|C}[E[Y \mid A=1, M(0), L, C]] - E_C E_{L|C}[E[Y \mid A=0, M(0), L, C]] \\ &= \int\int \sum_m \Big\{E[Y \mid A=1, M=m, C=c, L=l] - E[Y \mid A=0, M=m, C=c, L=l]\Big\} \\ &\qquad \times f_{M|A,C,L}(m \mid 0, c, l)\, f_{L|C}(l \mid c)\, f_C(c)\, dl\, dc, \end{aligned} \tag{4.II.11}$$

$$\text{NIE} = E_C E_{L|C}[E[Y(1, M(1)) \mid L, C]] - E_C E_{L|C}[E[Y(1, M(0)) \mid L, C]] \tag{4.II.12}$$

$$\begin{aligned} &= \int\int \sum_m E[Y \mid A=1, M=m, C=c, L=l] \Big\{f_{M|A,C,L}(m \mid 1, c, l) - f_{M|A,C,L}(m \mid 0, c, l)\Big\} \\ &\qquad \times f_{L|C}(l \mid c)\, f_C(c)\, dl\, dc. \end{aligned} \tag{4.II.13}$$

Similarly, for the interventional effects, we define them as:

$$\begin{aligned}\text{IDE} &= E[Y(1, G(0))] - E[Y(0, G(0))] \\ &= \int\int \sum_m \Big[E[Y \mid A=1, M=m, L=l, C=c]\, f_{M|A,L,C}(m \mid 0, l, c)\, f_{L|A,C}(l \mid 1, c) \\ &\qquad - E[Y \mid A=0, M=m, L=l, C=c]\, f_{M|A,L,C}(m \mid 0, l, c)\, f_{L|A,C}(l \mid 0, c)\Big]\, f_C(c)\, dl\, dc. \end{aligned} \tag{4.II.14}$$

$$\text{IIE} = E[Y(1, G(1))] - E[Y(1, G(0))] \tag{4.II.15}$$

$$\begin{aligned} &= \int\int \sum_m E[Y \mid A=1, M=m, L=l, C=c] \Big[f_{M|A,L,C}(m \mid 1, l, c) - f_{M|A,L,C}(m \mid 0, l, c)\Big] \\ &\qquad f_{L|A,C}(l \mid 1, c)\, f_C(c)\, dl\, dc. \end{aligned} \tag{4.II.16}$$

It is worth noting that the decompositions we illustrated above may not be the only ways to capture the causal mediation effects. For instance, for the NDE and the NIE, we may also consider the cross-world potential as $E[Y(0, M(1))]$ and thus have the NDE as

[10] As M is set to be discrete; if M is continuous, then we just change $\sum_m$ in the following equations into $\int_m$

$E[Y(1,M(1))]-E[Y(0,M(1))]$ as the mediator fixed at the level of $M(1)$ and change the treatment and the NIE as $E[Y(0,M(1))]-E[Y(0,M(0))]$, which is the treatment effect going through the mediator.

Under the within-world scenario, as the post-treatment covariates L are affected by the treatment A, we may also consider a three-way decomposition: the direct treatment effect that does not go through M and L ($A \rightarrow Y$); the indirect effect that goes through M given L ($A \rightarrow M \rightarrow Y$; and the indirect effect that goes through L only $A \rightarrow L \rightarrow Y$. The proposed decomposition framework is quite similar to that which we have multiple mediators (Vansteelandt and Daniel 2017). Let $G_L^a(l \mid c) = f_{L|A,C}(l \mid a,c)$ and $G_M^a(m \mid l,c) := f_{M|A,L,C}(m \mid a,l,c)$, we define:

$$\theta(a;a_L,a_M) = \int\int \sum_m E[Y \mid A=a, M=m, L=l, C=c]\, G_M^{a_M}(m \mid l,c)\, G_L^{a_L}(l \mid c)\, f_C(c)\, dl\, dc.$$

We can have the following three-way decomposition:

$$\text{TATE} = E[Y(1)]-E[Y(0)] = \underbrace{\theta(1;0,0)-\theta(0;0,0)}_{\text{IDE (not via } L \text{ or } M);\psi_{A\rightarrow Y}} + \underbrace{\theta(1;1,0)-\theta(1;0,0)}_{\text{Indirect via } L;\psi_{A\rightarrow L\rightarrow Y}} + \underbrace{\theta(1;1,1)-\theta(1;1,0)}_{\text{Indirect via } M \text{ given } L;\psi_{A\rightarrow M\rightarrow Y|L}}.$$

In which we can have:

$$\begin{aligned}
\psi_{A\rightarrow Y} &= \int\int \sum_m \Big[E[Y \mid A=1, M=m, L=l, C=c]\, f_{M|A,L,C}(m \mid 0,l,c)\, f_{L|A,C}(l \mid 0,c) \\
&\quad - E[Y \mid A=0, M=m, L=l, C=c]\, f_{M|A,L,C}(m \mid 0,l,c)\, f_{L|A,C}(l \mid 0,c)\Big]\, f_C(c)\, dl\, dc, \\
\psi_{A\rightarrow L\rightarrow Y} &= \int\int \sum_m E[Y \mid A=1, M=m, L=l, C=c]\, f_{M|A,L,C}(m \mid 0,l,c) \\
&\quad \times \Big[f_{L|A,C}(l \mid 1,c) - f_{L|A,C}(l \mid 0,c)\Big]\, f_C(c)\, dl\, dc, \\
\psi_{A\rightarrow M\rightarrow Y|L} &= \int\int \sum_m E[Y \mid A=1, M=m, L=l, C=c] \Big[f_{M|A,L,C}(m \mid 1,l,c) - f_{M|A,L,C}(m \mid 0,l,c)\Big] \\
&\quad \times f_{L|A,C}(l \mid 1,c)\, f_C(c)\, dl\, dc.
\end{aligned}$$

According to our derivation above, we can see that in the decompositions, the estimand for the CRF: $Y(a,m)$ and $Y(a,l,m)$ is at the core of all causal mediation decomposition and analysis. Let $\psi'_{am} = Y(a,m)$ and the statistical estimand $\psi_{am} = E[Y \mid A = a, M = m, L, C]$; $\psi'_{alm} = Y(a,l,m)$ and the statistical estimand $\psi_{alm} = E[Y \mid A = a, M = m, L = l, C]$. Our next goal is to derive an efficient/doubly robust estimator with the observational data for ψ_{am} and ψ_{alm}.

III. Review on the Efficient Estimator

Before directly jumping into the estimators for ψ_{am} and ψ_{alm}, we will shortly review how we derive the efficient estimator for $\psi_a = E_C E[Y \mid A = a, C]$. If the readers have walked through the Chapter 1 or are familiar with the techniques, just skip this section and turn to Section IV. In summary, we get $\hat{\psi}_{a,DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_a(C_i) + \frac{\mathbb{1}(A_i=a)}{\hat{\pi}(C_i)}\left(Y_i - \hat{\mu}_a(C_i)\right)\right]$ (where $\mu_a(C_i) = E[Y \mid A_i = a, C_i]$ and $\pi(C_i) = P(A_i = a \mid C_i)$) in two ways: for the saturated model (the observational data without restrictions on the treatment assignment), we use algebraic transformation; for the non-saturated model in which we use the example of the (pseudo-)randomized control trial (RCT), we start with a regular and asymptotically linear (RAL) estimator (but not the efficient one) and project it into sub-tangent spaces to yield the efficient influences on the subspaces and sum them up. We will not give the details but just give a simple outline, as they will later be used for us to find the efficient estimators for ψ_{am} and ψ_{alm}.

A. Saturated Model with Algebraic Transformation

The method for algebraic transformation is quite easy. Because

$$\psi_a = E_C[E[Y \mid A = a, C]] = \int_C E[Y \mid A = a, C]P(C)dc = \int_C \mu_a(c)P(C)dc = \sum_c \mu_a(c)p(c),$$

the efficient influence function $\phi(\psi_a)$ is (using the gradient algebra):

$$\phi(\psi_a) = \sum_c \phi\big(\mu_a(c)\, p(c)\big) = \sum_c \big(\phi(\mu_a(c))\, p(c) + \mu_a(c)\, \phi(p(c))\big). \tag{4.III.17}$$

We know that:

$$\phi\big(\mu_a(c)\big) = \frac{\mathbb{1}(C=c)}{p(c)} \cdot \frac{\mathbb{1}(A=a)}{P(A=a \mid C=c)} \big(Y - \mu_a(c)\big),$$

$$\phi\big(p(c)\big) = \mathbb{1}(C=c) - p(c).$$

Substituting these into the expression for $\phi(\psi_a)$, we obtain:

$$\phi(\psi_a) = \sum_c \left(\mathbb{1}(C=c) \frac{\mathbb{1}(A=a)}{P(A=a \mid C=c)} \big(Y - \mu_a(c)\big) + \mu_a(c) \big(\mathbb{1}(C=c) - p(c)\big) \right).$$

Simplifying the terms and noting that $\sum_c p(c) = 1$, the expression reduces to:

$$\phi(\psi_a) = \frac{\mathbb{1}(A=a)}{P(A=a \mid C)} \big(Y - \mu_a(C)\big) + \mu_a(C) - \psi_a.$$

Therefore, the efficient influence function is:

$$\phi(\psi_a) = \frac{\mathbb{1}(A=a)}{P(A=a \mid C)} \big(Y - \mu_a(C)\big) + \mu_a(C) - \psi_a.$$

And we have the efficient/doubly robust estimator for ψ_a[11] :

$$\hat{\psi}_{a,DR} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathbb{1}(A_i=a)}{\hat{\pi}(C_i)} \big(Y_i - \hat{\mu}_a(C_i)\big) + \hat{\mu}_a(C_i) \right) \tag{4.III.18}$$

B. Non-Saturated Model with Tangent Subspaces Projections

In the saturated model, as there are no restrictions on the data-generating process, the tangent space is the whole L_2 space so that we could calculate any influence function for our

[11] If the readers remember what we discussed in our chapter of the efficient/doubly robust loss function for the truncated-censoring survival data (Chapter 2), the estimator is also called the augmented inverse probability (AIPW) estimator, as it is based on the IPW estimator adding the augmented term of the conditional expectation, weighted by $1 - IPW$.

target parameter ψ and regard it as the efficient influence function. However, in the non-saturated model, since there are restrictions on the data-generating process, for example, under the (pseudo-)RCT, the proportion of cases to be assigned to the treatment group is fixed, and thus, not all the influence functions are the efficient influence function, and not all the RAL estimators are the efficient estimators. Indeed, we could start with an RAL estimator, derive its influence function, and project the influence function onto the tangent space to get the efficient influence function. If the tangent space is complex and the projection is hard to derive, we could use orthogonal decomposition for the tangent space, split it into several orthogonal tangent subspaces, derive the efficient influence function on each of the tangent spaces, and then sum them up.

For $\psi_a = E[Y \mid A = a, C]$, we start with the RAL estimator of the inverse-probability weighting estimator:

$$\hat{\psi}_{a,IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}(A_i = a)}{\hat{\pi}(C_i)} Y_i.$$

Thus, the influence function of the IPW estimator is:

$$\phi_a^{IPW} = \frac{\mathbb{1}(A_i = a)}{\pi(C_i)} Y_i - \psi_a. \tag{4.III.19}$$

The tangent space for the joint distribution of $Y, A = a, C$ can be orthogonally decomposed into :

$$\mathrm{T}_{Y,A=a,C} = \mathrm{T}_{Y|A=a,C} \oplus \mathrm{T}_{A=a} \oplus \mathrm{T}_C.$$

Projecting Equation 4.III.19 on to T_C, we have:

$$\begin{aligned}\phi^{\dagger}_{a\langle \mathrm{T}_C\rangle} &= E[\phi_a^{IPW} \mid C] = E\left[\left(\frac{\mathbb{1}(A_i = a)}{\pi(C_i)} Y_i - \psi_a\right) \,\middle|\, C\right] \\ &= E[Y \mid A = a, C] - \psi_a = \mu_a(C) - \psi_a.\end{aligned}$$

Projecting Equation 4.III.19 onto the tangent space $\mathrm{T}_{A=a}$, the efficient influence function will be zero because the tangent space is orthogonal to ϕ_a^{IPW}:

$$\phi^{\dagger}_{a\langle \mathrm{T}_{A=a}\rangle} = 0,$$

Finally, projecting Equation 4.III.19 onto the tangent space $\mathrm{T}_{Y|A=a,C}$ we get:

$$\begin{aligned}\phi^{\dagger}_{a\langle \mathrm{T}_{Y|A=a,C}\rangle} &= E[\phi_a^{IPW} \mid Y, A=a, C] - E[\phi_a^{IPW} \mid A=a, C] \\ &= \left(\frac{\mathbb{1}(A_i=a)}{\pi(C_i)} Y_i - \psi_a\right) - \left(\frac{\mathbb{1}(A_i=a)}{\pi(C_i)}\mu_a(C_i) - \psi_a\right).\end{aligned}$$

Sum the efficient influence functions in the tangent subspaces up, we have:

$$\begin{aligned}\phi^{\dagger}(\psi_a) &= \phi^{\dagger}_{a\langle \mathrm{T}_C\rangle} + \phi^{\dagger}_{a\langle \mathrm{T}_{A=a}\rangle} + \phi^{\dagger}_{a\langle \mathrm{T}_{Y|A=a,C}\rangle} \\ &= \left[\mu_a(C_i) - \psi_a\right] + 0 + \left[\left(\frac{\mathbb{1}(A_i=a)}{\pi(C_i)} Y_i - \psi_a\right) - \left(\frac{\mathbb{1}(A_i=a)}{\pi(C_i)}\mu_a(C_i) - \psi_a\right)\right] \\ &= \frac{\mathbb{1}(A_i=a)}{P(A_i=a \mid C_i)}\left(Y_i - \mu_a(C_i)\right) + \mu_a(C_i) - \psi_a.\end{aligned}$$

Hence, the efficient estimator for the parameter $\psi_a = E_C E[Y \mid A=a, C]$ is:

$$\hat{\psi}_{a,DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}(A_i=a)}{\hat{\pi}(C_i)}\left(Y_i - \hat{\mu}_a(C_i)\right) + \hat{\mu}_a(C_i)\right).$$

Which is exactly the same as the result we derive from the saturated model.

IV. Doubly Robust Estimator for Static Causal Mediation Analysis

A. Controlled Response Functions

A.1 ψ_{am}

Now, we derive the efficient estimator for the CRF $\psi_{am} = E_C E_{L|A=a,C}[E[Y \mid A=a, M=m, L, C]]$. With the observational data, we can regard what we are doing as deriving the efficient estimator for the saturated model. Thus, we first show the algebraic transformation for the estimator. Notice the pure-imputation expression for ψ_{am}:

$$
\begin{aligned}
\psi_{am} &= E_C\, E_{L|A=a,C}[\,E\{Y \mid A=a, M=m, L, C\}] \\
&= \int \left\{ \int E[Y \mid A=a, M=m, l, c]\; dP(L=l \mid A=a, C=c) \right\} dP(C=c) \\
&= \sum_c P(C=c) \sum_l E[Y \mid A=a, M=m, l, c]\; P(L=l \mid A=a, C=c)\,.
\end{aligned}
$$

The expression above is purely within-world: it uses the observed-law factorization $E_C\, E_{L|A=a,C}\, E[Y \mid A=a, M=m, L, C]$. [12] To make the paper tidier, we define the following notations:

$$
\begin{aligned}
\mu_{am}(L, C) &:= E[Y \mid A=a, M=m, L, C]\,, \\
\pi_a(C) &:= P(A=a \mid C), \\
\pi_m(a, L, C) &:= P(M=m \mid A=a, L, C), \\
\eta_a(C) &:= E_{L|A=a,C}\big[\mu_{am}(L, C)\big]\,.
\end{aligned}
$$

With the nuisance functions, we can consider applying Equation 4.III.18 twice to yield the efficient influence function for $\phi(\psi_{am})$. For the first time, we take the inner-layer to apply Equation 4.III.18, we have:

$$
\underbrace{\frac{\mathbb{1}(A=a)\mathbb{1}(M=m)}{\pi_a(C)\,\pi_m(a, L, C)}\Big\{Y - \mu_{am}(L, C)\Big\}}_{\text{outcome-residual AIPW term}} + \underbrace{\Big(\mu_{am}(L, C) - E_C E_{L|C}[\mu_{am}(L, C)]\Big)}_{\text{remainder to be handled next}}.
$$

The first term is the outcome piece (the outcome-residual augmented inverse probability weighting (AIPW) term, which uses the IPW to correct the outcome regression via the residual. The second term is a remainder that will be transported to the correct law of $(L \mid A, C)$. For the remainder, we take it as a whole entity and substitute it into the Equation 4.III.18 (applying twice):

$$
\underbrace{\frac{\mathbb{1}(A=a)}{\pi_a(C)}\Big\{\mu_{am}(L, C) - \eta_a(C)\Big\}}_{L \mid A, C \text{ transport}} + \underbrace{\eta_a(C) - \psi_{am}}_{\text{centering over } C}.
$$

[12] No cross-world assumption is needed to define ψ_{am} as a statistical estimand; cross-world conditions (e.g., to identify $E\{Y(a, m)\}$) are only required if one wishes to interpret ψ_{am} causally as $E[Y(a, m)]$.

In the above equation, the first part transports the fitted value $\hat{\mu}_{am}(L,C)$ to the distribution of $L \mid A, C$; while the second term centers the expression over C, yielding the mean-zero result. Adding the results together, we can have the efficient influence function for $\phi(\psi_{am})$:

$$\phi(\psi_{am}) = \frac{\mathbb{1}(A=a)\,\mathbb{1}(M=m)}{\pi_a(C)\,\pi_m(a,L,C)}\{Y-\mu_{am}(L,C)\} + \frac{\mathbb{1}(A=a)}{\pi_a(C)}\{\mu_{am}(L,C)-\eta_a(C)\} + \eta_a(C)-\psi_{am}.$$

Using this influence function, we can construct the efficient/doubly robust estimator for ψ_{am}:

$$\hat{\psi}_{am,\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}(A_i=a)\,\mathbb{1}(M_i=m)}{\hat{\pi}_a(C_i)\,\hat{\pi}_m(a,L_i,C_i)}\{Y_i-\hat{\mu}_{am}(L_i,C_i)\} + \frac{\mathbb{1}(A_i=a)}{\hat{\pi}_a(C_i)}\{\hat{\mu}_{am}(L_i,C_i)-\hat{\eta}_a(C_i)\} + \hat{\eta}_a(C_i)\right],$$

In the expression above, the doubly robust characteristic is satisfied as either our specifications of the two weighting nuisance functions ($\hat{\pi}_a(C)$ and $\hat{\pi}_m(A,L,C)$) or the imputation function $\hat{\mu}_{am}(L,C)$ is correctly specified, the estimator will be unbiased.

Now we suppose our causal mediation data are from the (pseudo-)RCT setting. We first derive the RAL estimator for ψ_{am} (the IPW estimation) and derive its influence function, then project the influence function onto the tangent space $\mathrm{T}_{\langle Y,A=a,M=m,L,C\rangle} = \mathrm{T}_{\langle Y|A=a,M=m,L,C\rangle} \oplus \mathrm{T}_{\langle M=m,A=a|L,C\rangle} \oplus \mathrm{T}_{\langle L,C\rangle}$.

The IPW estimator is:

$$\hat{\psi}_{am,\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}(A_i=a)\,\mathbb{1}(M_i=m)}{\hat{\pi}_a(C_i)\,\hat{\pi}_m(a,L_i,C_i)}\,Y_i$$

Thus, the influence function of the IPW estimator for ψ_{am} is:

$$\phi_{am}^{\mathrm{IPW}} = \frac{\mathbb{1}(A=a)\,\mathbb{1}(M=m)}{\pi_a(C)\,\pi_m(a,L,C)}\,Y-\psi_{am}.$$

We start projecting ϕ_{am}^{IPW} onto each component of the tangent space $\mathrm{T}_{Y,A=a,M=m,L,C}$, which is decomposed as:

$$\mathrm{T}_{\langle Y,A,M,L,C\rangle} = \mathrm{T}_{\langle Y|A,M,L,C\rangle} \oplus \mathrm{T}_{\langle M|A,L,C\rangle} \oplus \mathrm{T}_{\langle L|A,C\rangle} \oplus \mathrm{T}_{\langle A|C\rangle} \oplus \mathrm{T}_{\langle C\rangle}.$$

Since projecting onto $\mathrm{T}_{\langle A|C\rangle}$ and $\mathrm{T}_{\langle M|A,L,C\rangle}$ is unnecessary – the EIF is orthogonal to nuisance directions that don't change ψ_{am}, therefore, we first project ϕ_{am}^{IPW} onto $\mathrm{T}_{\langle C\rangle}$:

$$\phi^{\dagger}_{am\langle \mathrm{T}_{\langle C\rangle}\rangle} = E\big[\phi_{am}^{\mathrm{IPW}} \mid C\big] = \eta_a(C) - \psi_{am}.$$

We next project ϕ_{am}^{IPW} onto $\mathrm{T}_{\langle L|A,C\rangle}$:

$$\phi^{\dagger}_{am\langle \mathrm{T}_{\langle L|A,C\rangle}\rangle} = E\big[\phi_{am}^{\mathrm{IPW}} \mid A,L,C\big] - E\big[\phi_{am}^{\mathrm{IPW}} \mid A,C\big] = \frac{\mathbb{1}(A=a)}{\pi_a(C)}\big\{\mu_{am}(L,C) - \eta_a(C)\big\}.$$

Finally, we project ϕ_{am}^{IPW} onto $\mathrm{T}_{\langle Y|A,M,L,C\rangle}$:

$$\phi^{\dagger}_{am\langle \mathrm{T}_{\langle Y|A,M,L,C\rangle}\rangle} = \frac{\mathbb{1}(A=a)\,\mathbb{1}(M=m)}{\pi_a(C)\,\pi_m(a,L,C)}\big(Y - \mu_{am}(L,C)\big).$$

Summing the projected components, we obtain

$$\begin{aligned}\phi^{\dagger}(\psi_{am}) &= \phi^{\dagger}_{am\langle \mathrm{T}_{\langle C\rangle}\rangle} + \phi^{\dagger}_{am\langle \mathrm{T}_{\langle L|A,C\rangle}\rangle} + \phi^{\dagger}_{am\langle \mathrm{T}_{\langle Y|A,M,L,C\rangle}\rangle} \\ &= \big\{\eta_a(C) - \psi_{am}\big\} + \frac{\mathbb{1}(A=a)}{\pi_a(C)}\big\{\mu_{am}(L,C) - \eta_a(C)\big\} + \frac{\mathbb{1}(A=a)\,\mathbb{1}(M=m)}{\pi_a(C)\pi_m(a,L,C)}\big\{Y - \mu_{am}(L,C)\big\}.\end{aligned}$$

Hence, the efficient influence function for ψ_{am} is:

$$\phi^{\dagger}(\psi_{am}) = \frac{\mathbb{1}(A=a)\,\mathbb{1}(M=m)}{\pi_a(C)\,\pi_m(a,L,C)}\big(Y - \mu_{am}(L,C)\big) + \frac{\mathbb{1}(A=a)}{\pi_a(C)}\big\{\mu_{am}(L,C) - \eta_a(C)\big\} + \eta_a(C) - \psi_{am}.$$

And therefore, based on the efficient influence function, the efficient estimator for ψ_{am} is given by:

$$\hat{\psi}_{am,\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}(A_i=a)\,\mathbb{1}(M_i=m)}{\hat{\pi}_a(C_i)\,\hat{\pi}_m(a,L_i,C_i)}\big\{Y_i - \hat{\mu}_{am}(L_i,C_i)\big\} + \frac{\mathbb{1}(A_i=a)}{\hat{\pi}_a(C_i)}\big\{\hat{\mu}_{am}(L_i,C_i) - \hat{\eta}_a(C_i)\big\} + \hat{\eta}_a(C_i)\right]. \tag{4.IV.20}$$

Again, this result is exactly the same as the estimator in the saturated model. The estimator is consistent and asymptotically normal if either (i) both propensity models π_a, π_m are correctly specified, or (ii) the outcome regression μ_{am} and the mechanism used to compute $\eta_a(C) = E_{L|A=a,C}\{\mu_{am}(L,C)\}$ are correctly specified. Hence, this is a doubly robust estimator.

For the special case under the cross-world assumptions, when there is no effect of A on L, we do not have the η terms in our final expression, and the DR estimator should be (Zhou 2021):

$$\hat{\psi}_{am,\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}(A_i = a)\,\mathbb{1}(M_i = m)}{\hat{\pi}_a(C_i)\,\hat{\pi}_m(a, L_i, C_i)}\{Y_i - \hat{\mu}_{am}(L_i, C_i)\} + \hat{\mu}_{am}(L_i, C_i)\right]. \tag{4.IV.21}$$

As either the weighting models π_a, π_m are correctly specified or the outcome regression μ is correctly specified, the estimator is unbiased and efficient [13].

In the remainder of this chapter, we will present only the derivations of the efficient/doubly robust estimators for the saturated model, as in social science, we mainly deal with inference with observational data.

A.2 ψ_{alm}

The efficient/doubly robust estimator derivation for ψ_{alm} is quite similar to that for ψ_{am} under the Assumption 4.II.3, because $\psi_{alm} = E_c[E[Y \mid A = a, M = m, L = l, C]]$, we have the efficient influence function for ψ_{alm}:

$$\phi(\psi_{alm}) = \phi\left(E_c[E[Y \mid A = a, M = m, L = l, C]]\right) = \sum_c \phi(\mu_{alm}(c)\,p(c))$$

[13] In earlier studies like Zhou 2021, Farbmacher et al., 2022, and Tchetgen Tchetgen and Shpitser 2012, they gave rigorous proof on the expression of the CRF ψ_{am} with Neyman orthogonality, and readers who are interested in the proof may refer to their work (but they did not explicitly give out the expressions for the direct and indirect effects as what we do in this chapter).

$$= \sum_c \big(\phi(\mu_{alm}(c))p(c) + \mu_{alm}(c)\phi(p(c))\big)$$

Where $\mu_{alm}(C) = E[Y \mid A = a, M = m, L = l, C]$. By expanding the terms similarly to the ψ_{am} case, we obtain:

$$\phi(\psi_{alm}) = \frac{\mathbb{1}(A = a)\,\mathbb{1}(L = l)\,\mathbb{1}(M = m)}{P(A = a \mid C)\,P(L = l \mid A, C)\,P(M = m \mid A, L, C)}\{Y - \mu_{alm}(C)\} + \mu_{alm}(C) - \psi_{alm}.$$

Using this influence function, we can construct the efficient/doubly robust estimator for ψ_{alm}:

$$\hat{\psi}_{alm,DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}(A_i = a, M_i = m, L_i = l)}{\hat{\pi}_a(C_i)\hat{\pi}_m(A_i, L_i, C_i)\hat{\pi}_l(A_i, C_i)}\big(Y_i - \hat{\mu}_{alm}(C_i)\big) + \hat{\mu}_{alm}(C_i)\right).$$

where $\pi_l(A, C) = P(L = l \mid A, C)$.

Notice here we define the DR estimator for ψ_{alm}, which is **not** the DR estimator for $\theta(a; a_L, a_M)$ because for the latter term we restrict $L = L(a_L)$ (the post-treatment covariates obtain the value under $A = a_L$) and $M = a_m$ (drawing the mediator from its distribution under $A = a_M$ given L. Thus, for $\mu(a; a_L, a_M)$, the target parameter is the nested counterfactual:

$$\psi_{a;a_L,a_M} \equiv E\big[\,Y\big(a,\, M(a_M, L(a_L)),\, L(a_L)\big)\big].$$

Hence, if we set the nuisance functions:

$$r_l(a_L, C) \equiv P(L = l \mid A = a_L, C),$$

$$q_m(a_M, l, C) \equiv P(M = m \mid A = a_M, L = l, C),$$

$$\mu_{aml}(C) \equiv E[Y \mid A = a, M = m, L = l, C].$$

we have:

$$\psi(a; a_L, a_M) = E_C\left[\sum_l \sum_m \mu_{aml}(C)\; r_l(a_L, C)\; q_m(a_M, l, C)\right].$$

We can apply Equation 4.III.18 to the above DR estimator for ψ_{alm} and derive the DR/ efficient estimator for $\psi(a; a_L, a_M)$:

$$\widehat{\psi}^{\text{DR}}_{a;a_L,a_M} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\frac{\mathbb{1}(A_i=a)}{\hat{\pi}_a(C_i)}\Bigg\{\frac{1}{\hat{\pi}_l(a,C_i)}\Bigg(\sum_m \hat{q}_m(a_M,L_i,C_i)\Big[\frac{\mathbb{1}(M_i=m)}{\hat{\pi}_m(a,L_i,C_i)}\{Y_i-\hat{\mu}_{amL_i}(C_i)\}+\hat{\mu}_{amL_i}(C_i)\Big] - \sum_m \hat{q}_m(a_M,L_i,C_i)\,\hat{\mu}_{amL_i}(C_i)\Bigg) + \sum_m \hat{q}_m(a_M,L_i,C_i)\,\hat{\mu}_{amL_i}(C_i) - \sum_l\sum_m \hat{r}_l(a_L,C_i)\,\hat{q}_m(a_M,l,C_i)\,\hat{\mu}_{aml}(C_i)\Bigg\} + \sum_l\sum_m \hat{r}_l(a_L,C_i)\,\hat{q}_m(a_M,l,C_i)\,\hat{\mu}_{aml}(C_i)\Bigg]. \tag{4.IV.22}$$

B. Doubly Robust Estimator for the Direct and Indirect Effects

With the derivation of the doubly robust estimator for the CRF ψ_{am}, we may obtain the efficient/doubly robust estimator for the direct and indirect effects. Namely, as the NDE and the NIE rely on the cross-world assumption (Assumption 4.II.4), we may derive their DR/ efficient estimators based on Equation 4.IV.21; for the CDE, the IDE, and the IIE, as they rely on the within-world assumption (Assumption 4.II.3), we can derive their DR/efficient estimator based on Equation 4.IV.20.

B.1 Doubly Robust Estimator for the Natural Direct and Indirect Effects

As we derived in Equation 4.II.11, the natural direct effect comparing $a=1$ to $a=0$ is:

$$\begin{aligned}\text{NDE} &= E_C E_{L|C}[E[Y \mid A=1, M(0), L, C]] - E_C E_{L|C}[E[Y \mid A=0, M(0), L, C]] \\ &= \int\int \sum_m \Big\{E[Y \mid A=1, M=m, C=c, L=l] - E[Y \mid A=0, M=m, C=c, L=l]\Big\} \\ &\qquad \times f_{M|A,C,L}(m \mid 0, c, l)\; f_{L|C}(l \mid c)\; f_C(c)\; dl\, dc,\end{aligned}$$

which we may rewrite as:

$$\text{NDE} = E_C E_{L|C}\Big[\sum_m \mu_{1m}(L,C)\, q_m(0,L,C)\Big] - E_C E_{L|C}\Big[\sum_m \mu_{0m}(L,C)\, q_m(0,L,C)\Big]$$

Obviously, the main structure that is important to derive the DR/efficient estimator for the NDE is:

$$\theta_{a,t} = E_C E_{L|C}\big[\sum_m \mu_{am}(L,C)\, q_m(t,L,C)\big]$$

and thus we can get the DR estimator for the NDE and NIE: NDE = $\theta_{1,0} - \theta_{0,0}$ and NIE = $\theta_{1,1} - \theta_{1,0}$.. We can understand the expression as using $q(.)$ to reweight the original probability distribution of the mediator for the regression outcome residuals, plus the centering term. According to the Equation 4.IV.21, the original doubly robust estimator for CRF ψ_{am} is:

$$\hat{\psi}_{am,\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}(A_i=a)\,\mathbb{1}(M_i=m)}{\hat{\pi}_a(C_i)\,\hat{\pi}_m(a,L_i,C_i)}\big\{Y_i - \hat{\mu}_{am}(L_i,C_i)\big\} + \hat{\mu}_{am}(L_i,C_i)\right]. \tag{4.IV.23}$$

Thus, when we have the mediator, which is manipulated at the level of mediator when $A = t$, we need to reweight/transport the estimator to the distribution. Hence, the DR estimator for $\theta_{a,t}$ is (note t is the value of the treatment for the manipulated mediator):

$$\begin{aligned}\widehat{\theta}^{\mathrm{DR}}_{a,t} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[& \underbrace{\sum_m \widehat{\mu}_{am}(L_i,C_i)\,\widehat{q}_m(t,L_i,C_i)}_{\text{plug-in}} \\ & + \underbrace{\frac{\mathbb{1}(A_i=a)}{\widehat{\pi}_a(C_i)}\sum_m \frac{\mathbb{1}(M_i=m)}{\widehat{\pi}_m(a,L_i,C_i)}\,\widehat{q}_m(t,L_i,C_i)\Big\{Y_i - \widehat{\mu}_{am}(L_i,C_i)\Big\}}_{\text{(i) outcome-residual correction}} \\ & + \underbrace{\frac{\mathbb{1}(A_i=t)}{\widehat{\pi}_t(C_i)}\left[\sum_m\left\{\frac{\mathbb{1}(M_i=m)}{\widehat{\pi}_m(t,L_i,C_i)} - 1\right\}\widehat{\mu}_{am}(L_i,C_i)\right]}_{\text{(ii) transport correction to the } (A=t) \text{ world}}\Bigg].\end{aligned}$$

The plug-in term here is the basic $g-$ computation; the outcome-residual correction uses only people who actually have $A = a$, and the transport correction term uses only people who actually have $A = t$. Thus,

$$\begin{aligned}\widehat{\mathrm{NDE}}_{\mathrm{DR}} &= \widehat{\theta}^{\mathrm{DR}}_{1,0} - \widehat{\theta}^{\mathrm{DR}}_{0,0} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_m\Big\{\big[\widehat{\mu}_{1m}(L_i,C_i) - \widehat{\mu}_{0m}(L_i,C_i)\big]\,\widehat{q}_m(0,L_i,C_i)\end{aligned}$$

$$
\begin{aligned}
&+\frac{\mathbb{1}(A_i=1)}{\widehat{\pi}_1(C_i)}\frac{\mathbb{1}(M_i=m)}{\widehat{\pi}_m(1,L_i,C_i)}\widehat{q}_m(0,L_i,C_i)\left\{Y_i-\widehat{\mu}_{1m}(L_i,C_i)\right\}\\
&-\frac{\mathbb{1}(A_i=0)}{\widehat{\pi}_0(C_i)}\frac{\mathbb{1}(M_i=m)}{\widehat{\pi}_m(0,L_i,C_i)}\widehat{q}_m(0,L_i,C_i)\left\{Y_i-\widehat{\mu}_{0m}(L_i,C_i)\right\}\\
&+\frac{\mathbb{1}(A_i=0)}{\widehat{\pi}_0(C_i)}\left(\frac{\mathbb{1}(M_i=m)}{\widehat{\pi}_m(0,L_i,C_i)}-1\right)\left[\widehat{\mu}_{1m}(L_i,C_i)-\widehat{\mu}_{0m}(L_i,C_i)\right]\Bigg\}.
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\mathrm{NIE}}_{\mathrm{DR}} &= \widehat{\theta}^{\mathrm{DR}}_{1,1}-\widehat{\theta}^{\mathrm{DR}}_{1,0}\\
&=\frac{1}{n}\sum_{i=1}^{n}\sum_{m}\Bigg\{\widehat{\mu}_{1m}(L_i,C_i)\left[\widehat{q}_m(1,L_i,C_i)-\widehat{q}_m(0,L_i,C_i)\right]\\
&\quad+\frac{\mathbb{1}(A_i=1)}{\widehat{\pi}_1(C_i)}\frac{\mathbb{1}(M_i=m)}{\widehat{\pi}_m(1,L_i,C_i)}\left[\widehat{q}_m(1,L_i,C_i)-\widehat{q}_m(0,L_i,C_i)\right]\left\{Y_i-\widehat{\mu}_{1m}(L_i,C_i)\right\}\\
&\quad+\frac{\mathbb{1}(A_i=1)}{\widehat{\pi}_1(C_i)}\left(\frac{\mathbb{1}(M_i=m)}{\widehat{\pi}_m(1,L_i,C_i)}-1\right)\widehat{\mu}_{1m}(L_i,C_i)-\frac{\mathbb{1}(A_i=0)}{\widehat{\pi}_0(C_i)}\left(\frac{\mathbb{1}(M_i=m)}{\widehat{\pi}_m(0,L_i,C_i)}-1\right)\widehat{\mu}_{1m}(L_i,C_i)\Bigg\}.
\end{aligned}
$$

B.2 Doubly Robust Estimator for the Controlled Direct Effects

For the CDE(m) $= E[Y(1,m)]-E[Y(0,m)]$ and our denotation $\psi_{am}=E[Y(a,m)]$, since in the observable data the post-treatment covariates L may be affected by A, we can simply use Equation 4.IV.20 to derive $\widehat{\psi}_{0m,DR}$ and $\widehat{\psi}_{1m,DR}$:

$$
\begin{aligned}
\widehat{\mathrm{CDE}}_{DR}(m) &= \hat{\psi}_{1m,DR}-\hat{\psi}_{0m,DR}\\
&=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}(A_i=1)\,\mathbb{1}(M_i=m)}{\hat{\pi}_1(C_i)\,\hat{\pi}_m(1,L_i,C_i)}\left\{Y_i-\hat{\mu}_{1m}(L_i,C_i)\right\}+\frac{\mathbb{1}(A_i=1)}{\hat{\pi}_1(C_i)}\left\{\hat{\mu}_{1m}(L_i,C_i)-\hat{\eta}_1(C_i)\right\}+\hat{\eta}_1(C_i)\right]\\
&-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}(A_i=0)\,\mathbb{1}(M_i=m)}{\hat{\pi}_0(C_i)\,\hat{\pi}_m(0,L_i,C_i)}\left\{Y_i-\hat{\mu}_{0m}(L_i,C_i)\right\}+\frac{\mathbb{1}(A_i=0)}{\hat{\pi}_0(C_i)}\left\{\hat{\mu}_{0m}(L_i,C_i)-\hat{\eta}_0(C_i)\right\}+\hat{\eta}_0(C_i)\right]
\end{aligned}
$$

Specifically, if Assumption 4.II.4 holds, simply substituting the expression for $\hat{\psi}_{am,DR}$ using Equation 4.IV.21:

$$
\widehat{\mathrm{CDE}}_{DR}(m)=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}(A_i=1)\mathbb{1}(M_i=m)}{\hat{\pi}_1(C_i)\hat{\pi}_m(A_i,L_i,C_i)}\left(Y_i-\hat{\mu}_{1m}(L_i,C_i)\right)+\hat{\mu}_{1m}(L_i,C_i)\right]
$$

$$-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}(A_i=0)\,\mathbb{1}(M_i=m)}{\hat{\pi}_0(C_i)\hat{\pi}_m(A_i,L_i,C_i)}\big(Y_i-\hat{\mu}_{0m}(L_i,C_i)\big)+\hat{\mu}_{0m}(L_i,C_i)\right]$$

B.3 Doubly Robust Estimator for the Interventional Direct and Indirect Effects

For the interventional direct and indirect effects, there are two distinct ways to yield the DR/efficient estimators: similar to the way we derive the DR/efficient estimator for the CDE, as the interventional effects are also based on the same within-world assumptions, we can directly derive the DR estimator using Equation 4.IV.20: like what we did for the NDE/ NIE, we can understand the expression as the hypothetical distribution of the mediator ($q(.)$) to reweight the original probability distribution of the mediator; besides, we may also regard the IDE and the IIE as operations of $\theta(a;a_L,a_M)$ in our derivation of ψ_{alm}: for the IDE, we can write it as $\theta(1,1,0)-\theta(0,0,0)$ and for the IIE we can express it as $\theta(1,1,1)-\theta(1,1,0)$ so that we can use Equation 4.IV.22 to derive the results.

We start with the derivation based on Equation 4.IV.20. Let Equation 4.IV.20 be the core expression, and we have the core expression to be divided into the following three parts:

$$\hat{\psi}_{am,\mathrm{DR}}=\underbrace{\frac{\mathbb{1}(A_i=a)\,\mathbb{1}(M_i=m)}{\hat{\pi}_a(C_i)\,\hat{\pi}_m(a,L_i,C_i)}\{Y_i-\hat{\mu}_{am}(L_i,C_i)\}}_{(i)}+\underbrace{\frac{\mathbb{1}(A_i=a)}{\hat{\pi}_a(C_i)}\{\hat{\mu}_{am}(L_i,C_i)-\hat{\eta}_a(C_i)\}}_{(ii)}+\underbrace{\hat{\eta}_a(C_i)}_{(iii)},$$

As we discussed when deriving the equation, part (i) is the outcome-residual AIPW term, part (ii) is the term for the distribution $(L\mid A,C)$ transport, and part (iii) is the centering term. Then we reweight the term with $q_m(t,L,C)$:

$$\begin{aligned}\widehat{\theta}_{a,t,\mathrm{DR}}=\frac{1}{n}\sum_{i=1}^{n}\Bigg[&\sum_m\frac{\mathbb{1}(A_i=a)\,\mathbb{1}(M_i=m)}{\widehat{\pi}_a(C_i)\,\widehat{\pi}_m(a,L_i,C_i)}\Big\{Y_i-\widehat{\mu}_{am}(L_i,C_i)\Big\}\,\widehat{q}_m(t,L_i,C_i)\\&+\frac{\mathbb{1}(A_i=a)}{\widehat{\pi}_a(C_i)}\Bigg\{\sum_m\widehat{\mu}_{am}(L_i,C_i)\,\widehat{q}_m(t,L_i,C_i)-\widehat{E}_{L|A=a,C=C_i}\Bigg[\sum_m\widehat{\mu}_{am}(L,C_i)\,\widehat{q}_m(t,L,C_i)\Bigg]\Bigg\}\\&+\widehat{E}_{L|A=a,C=C_i}\Bigg[\sum_m\widehat{\mu}_{am}(L,C_i)\,\widehat{q}_m(t,L,C_i)\Bigg]\Bigg].\end{aligned}\tag{4.IV.24}$$

We can understand the above DR estimator, compared to the DR estimator for ψ_{am} in the natural effects, a term of transport to the distribution of $L \mid A, C$ is added. We can directly derive the DR estimator for the IDE is:

$$
\begin{aligned}
\widehat{\mathrm{IDE}}_{\mathrm{DR}} &= \widehat{\theta}_{1,0,\mathrm{DR}} - \widehat{\theta}_{0,0,\mathrm{DR}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_m \frac{\mathbb{1}(A_i=1)\,\mathbb{1}(M_i=m)}{\widehat{\pi}_1(C_i)\,\widehat{\pi}_m(1,L_i,C_i)}\Big\{Y_i-\widehat{\mu}_{1m}(L_i,C_i)\Big\}\,\widehat{q}_m(0,L_i,C_i) \\
&+ \frac{\mathbb{1}(A_i=1)}{\widehat{\pi}_1(C_i)}\Bigg\{\sum_m \widehat{\mu}_{1m}(L_i,C_i)\,\widehat{q}_m(0,L_i,C_i) - \widehat{E}_{L|A=1,C=C_i}\Big[\sum_m \widehat{\mu}_{1m}(L,C_i)\,\widehat{q}_m(0,L,C_i)\Big]\Bigg\} \\
&+ \widehat{E}_{L|A=1,C=C_i}\Big[\sum_m \widehat{\mu}_{1m}(L,C_i)\,\widehat{q}_m(0,L,C_i)\Big]\Bigg] \\
&- \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_m \frac{\mathbb{1}(A_i=0)\,\mathbb{1}(M_i=m)}{\widehat{\pi}_0(C_i)\,\widehat{\pi}_m(0,L_i,C_i)}\Big\{Y_i-\widehat{\mu}_{0m}(L_i,C_i)\Big\}\,\widehat{q}_m(0,L_i,C_i) \\
&+ \frac{\mathbb{1}(A_i=0)}{\widehat{\pi}_0(C_i)}\Bigg\{\sum_m \widehat{\mu}_{0m}(L_i,C_i)\,\widehat{q}_m(0,L_i,C_i) - \widehat{E}_{L|A=0,C=C_i}\Big[\sum_m \widehat{\mu}_{0m}(L,C_i)\,\widehat{q}_m(0,L,C_i)\Big]\Bigg\} \\
&+ \widehat{E}_{L|A=0,C=C_i}\Big[\sum_m \widehat{\mu}_{0m}(L,C_i)\,\widehat{q}_m(0,L,C_i)\Big]\Bigg]
\end{aligned}
$$

and for the IIE is:

$$
\begin{aligned}
\widehat{\mathrm{IIE}}_{\mathrm{DR}} &= \widehat{\theta}_{1,1,\mathrm{DR}} - \widehat{\theta}_{1,0,\mathrm{DR}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_m \frac{\mathbb{1}(A_i=1)\,\mathbb{1}(M_i=m)}{\widehat{\pi}_1(C_i)\,\widehat{\pi}_m(1,L_i,C_i)}\Big\{Y_i-\widehat{\mu}_{1m}(L_i,C_i)\Big\}\,\widehat{q}_m(1,L_i,C_i) \\
&+ \frac{\mathbb{1}(A_i=1)}{\widehat{\pi}_1(C_i)}\Bigg\{\sum_m \widehat{\mu}_{1m}(L_i,C_i)\,\widehat{q}_m(1,L_i,C_i) - \widehat{E}_{L|A=1,C=C_i}\Big[\sum_m \widehat{\mu}_{1m}(L,C_i)\,\widehat{q}_m(1,L,C_i)\Big]\Bigg\} \\
&+ \widehat{E}_{L|A=1,C=C_i}\Big[\sum_m \widehat{\mu}_{1m}(L,C_i)\,\widehat{q}_m(1,L,C_i)\Big]\Bigg] \\
&- \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_m \frac{\mathbb{1}(A_i=1)\,\mathbb{1}(M_i=m)}{\widehat{\pi}_1(C_i)\,\widehat{\pi}_m(1,L_i,C_i)}\Big\{Y_i-\widehat{\mu}_{1m}(L_i,C_i)\Big\}\,\widehat{q}_m(0,L_i,C_i) \\
&+ \frac{\mathbb{1}(A_i=1)}{\widehat{\pi}_1(C_i)}\Bigg\{\sum_m \widehat{\mu}_{1m}(L_i,C_i)\,\widehat{q}_m(0,L_i,C_i) - \widehat{E}_{L|A=1,C=C_i}\Big[\sum_m \widehat{\mu}_{1m}(L,C_i)\,\widehat{q}_m(0,L,C_i)\Big]\Bigg\} \\
&+ \widehat{E}_{L|A=1,C=C_i}\Big[\sum_m \widehat{\mu}_{1m}(L,C_i)\,\widehat{q}_m(0,L,C_i)\Big]\Bigg]
\end{aligned}
$$

If we derive the DR/efficient estimator from Equation 4.IV.22, as $a_L = a$ and $a_M = t$, Equation 4.IV.22 is intrinsically the same as Equation 4.IV.24. This is because, when $a_L = a$ we average over the same law of L in both formulas:

$$\sum_l \hat{r}_l(a, C)\, g(l, C) \;=\; \widehat{E}\big[g(L, C) \mid A = a, C\big].$$

So both estimators integrate the same quantity over the same L-distribution and are therefore equivalent in expectation. We provide the rigorous derivation in Appendix C for readers who are interested in the mathematical details.

B.4 Summary

In this section, we derive the DR/efficient estimator for the common estimators in causal mediation analysis in static settings. Based on different research scenarios, social scientists may choose different effects for analysis: when the mediator can be directly set to a policy-relevant level, the controlled direct effect is the natural choice; while the goal in research is to decompose the total causal effects into different pathways, researchers may consider using the natural direct/indirect effects (NDE/NIE) or the interventional direct/indirect effects (IDE/IIE).

While all identifications here rely on consistency, positivity, and unconfoundedness for relevant relationships, natural effects additionally require cross-world independence conditions and the assumption that the post-treatment covariates are irrelevant to the treatment. In many research settings for static models, researchers often assume the $A-L$ relation to be negligible. However, for the dynamic models, as we will show in the next section, the correlation between treatment and post-treatment covariates cannot be assumed to be zero due

to the intrinsic treatment-covariates feedback mechanisms; therefore, we can only identify the interventional effects.

V. Dynamic Causal Mediation Models

We extend the causal mediation framework into a time-varying treatment and time-varying mediation setting. In section II, we have discussed the condition under which we can only have the within-world assumptions (Assumption 4.II.3). Under such a scenario, based on VanderWeele et al. 2014, we give two methods to decompose the total average treatment effect (TATE) into the direct and indirect effect of the treatment: in the first method, we count the variability of the mediator-outcome L into the consideration of the CRF: $Y(a,l,m)$ and we could isolate the path specific effect. The effect on the path $A \to Y$ is the direct effect, and the effect through the paths $A \to L \to Y$ and $A \to M \to Y \mid L$ is the indirect effect. A second method is to manipulate the mediator m by drawing the value from a defined distribution spanned by the treatment and the covariates (confounders). Unlike in the natural direct and indirect effects, in which M is defined by the latent generator $M(a)$, we have a randomly selected mediation level from the distribution, and based on the intervention, we can capture the interventional direct and indirect effects.

Now, suppose the mediator-treatment covariates are also included in the covariates C, and L is a mediator which precedes the mediator M. We separately denote the two mediators L, M as M_1, M_2. This becomes the basic time-varying model with multiple mediators. Suppose the models satisfy the conditions of the treatment-outcome unconfoundedness and mediator-outcome confoundedness with appropriate covariate specification. In that case, we can call the models the **marginal structural models (MSM)**(Robins et al.

2000; Robins 2003)[14], for which the effects can be correctly identified via the inverse probability weighting method or conditional expectation method, as we have elaborated above. In time-varying model settings, we use the overline notation to represent a sequence of variables, so here we can denote $\overline{M} = (M_1, M_2)$ as the sequence of mediators (and we use $\overline{m} = (M_1 = m, M_2 = m)$ denoting the same value m for the sequence of mediators).

With the increase in sequence length, using interventional decomposition methods to capture both direct and indirect effects becomes more convenient. This is simply because if we have a sequence of mediators, and if we use the first method, adding all the sequence of mediators in the CRF: $Y(a, \overline{m}) = Y(a, m_1, m_2, \ldots, m_t)$, the terms in the indirect effect will increase vastly, which is hard to calculate (let alone the scenarios where we have sequences of treatments and outcomes). The expressions on the IDE and the IIE will also become complex, but we can derive a general solution to the effects; hence, it will be convenient for us to obtain an efficient/doubly robust/debiased estimator for the effects. Similar to what we did in the static models, this section aims to derive efficient /RAL influence-function representations and corresponding estimators for the IDE and IIE under different time-varying modeling settings. To simplify the derivation, we begin with models that cover two periods.

A. A Two-Period Model

We illustrate the DAG for the two-period model in Figure 4.5. We denote $\overline{A} = (A_1, A_2)$, $\overline{M} = (M_1, M_2)$, and $\overline{Y} = (Y_1, Y_2)$. Further, we set $\overline{A} = \overline{a}$ as treated while $\overline{a^\star}$ denotes the con-

[14]The term of marginal structural model, although covers all the models in our discussions on causal mediation analysis in this and Chapter 5, it appears more commonly in the time-varying causal mediation model settings.

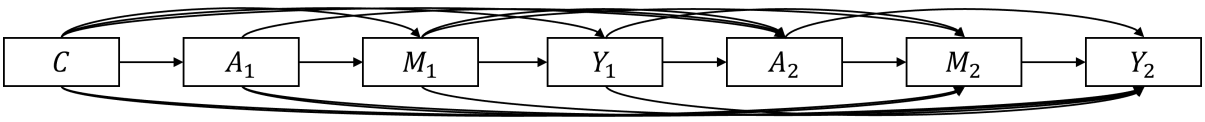


Figure 4.5: Directed Acyclic Graph for Two-period Causal Mediation Analysis

trol [15]. We use $\overline{G}_{\overline{a}}$ to denote the *sequential mediator law* under $\overline{a}$ [16]. It specifies the distribution from which $\overline{M}$ is randomly drawn when the treatment history is fixed at $\overline{a}$. Here, we can analogize that the time-varying history variables M_1, Y_1, and A_2 (the mediator and the outcome in the first wave and the treatment in the second round) play the same role as L in the static model. Now suppose our target is the potential outcome for $Y_2(\overline{a},\overline{m})$, if we still hold the within-world assumption (obviously, the cross-world assumption will never be satisfied in the dynamic models). Recall in Section II, the CRF for $Y(a,m)$ if Assumption 4.II.4 does not hold is given by $E_{L_{\sim L|A=a,C},C}E[Y \mid A=a,M=m,L,C] = \int_C \int_L E[Y \mid A = a,M=m,l,C]dP(L=l \mid A=a,dP(C)$ (in which C denotes both the time-constant and time-varying covariates). Thus, the observational identifying functional for the mean potential outcome $E\{Y_2(\overline{a},\overline{m})\}$ under the intervention $(\overline{A},\overline{M}) = (\overline{a},\overline{m})$ can be written as

$$\begin{aligned}\psi_{\overline{a}\,\overline{m}} &:= E\big\{Y_2(\overline{a},\overline{m})\big\} \qquad (4.V.25)\\ &= \int dP(C=c) \int E\Big[Y_2 \mid \overline{A}=\overline{a},\ \overline{M}=\overline{m},\ Y_1=y_1,\ C=c\Big]\, dP\big(Y_1=y_1 \mid A_1=a_1,\ M_1=m_1,\ C=c\big).\end{aligned}$$

Equation 4.V.25 is an extended g-formula expression for $E\{Y_2(\overline{a},\overline{m})\}$, which is identified under the **sequential causal mediation assumptions**. These assumptions are obtained by extending Assumption 4.II.3 to the time-varying setting, and they are not restricted to the two-period model.We can derive the assumption by extending Assumption 4.II.3 into the time-varying conditions (and the assumptions are not restricted to the two-period model) (VanderWeele et al. 2014):

[15] We choose not to use $A=a=1$ and $A=a^\star=0$ here simply to avoid confusion in the following content.

[16] The Sequential $\overline{G}(.)$ is the multi-period equivalent of $G(.)$ as defined in the within-world scenario as the stochastic policy that draws the mediator from the specific distribution for $\overline{M}$.

Assumption 4.V.1 (Sequential Causal Mediation Assumptions) *Let $t = 1, \ldots, T$. Denote cumulative histories $\overline{A}_t = (A_1, \ldots, A_t)$, $\overline{M}_t = (M_1, \ldots, M_t)$, $\overline{Y}_t = (Y_1, \ldots, Y_t)$, and define $H_t = (\overline{A}_{t-1}, \overline{M}_{t-1}, \overline{Y}_{t-1}, C)$. Let the target be the potential final outcome $Y_T(\overline{a}, \overline{m})$. The marginal structural models will be correctly defined if:*

1. *Sequential Positivity: For any history H_t in the support and all admissible a_t, m_t, the probability of the treatment and the mediator [17] should be positive given the history variables.*

$$P(A_t = a_t \mid H_t) > 0 \quad \textit{and} \quad P\Big(M_t = m_t \mid \overline{A}_t, H_t\Big) > 0,$$

   with probability one on the support of $(\overline{A}_{t-1}, \overline{M}_{t-1}, \overline{Y}_{t-1}, C)$.

2. *Sequential Consistency: the potential outcome under the treatment and mediator received is the same as the observed outcome:*

$$Y_t = Y_t(\overline{a}_t, \overline{m}_t)$$

   and the potential mediator under the treatment and previous mediators is the same as the observed mediator:

$$M_t = M_t(\overline{a}_t, \overline{m}_{t-1}).$$

3. *Sequential Unconfoundedness: we need the following unconfoundedness assumptions to hold, conditioned on the observed time-varying and time-invariant covariates:*

   (a) the treatment-outcome unconfoundedness at wave t: $Y_T(\overline{a}, \overline{m}) \perp\!\!\!\perp A_t \mid H_t$;

   (b) the mediator-outcome unconfoundedness at wave t: $Y_T(\overline{a}, \overline{m}) \perp\!\!\!\perp M_t \mid \overline{A}_t, H_t$;

[17] We here still assume discrete mediators. If the mediators are continuous, its probability density function should be positive, given the history variables and the treatment.

With Assumption 4.V.1, we can identify Equation 4.V.25. Moreover, if we need to construct the stochastic mediator law $\overline{G}_{\overline{a}}$, a further treatment-mediator unconfoundedness assumption is required, so that we can derive the expressions of the IDE and the IIE [18]:

- **Assumption 4.V.2 (Additional condition only when constructing the mediator law $\overline{G}_{\overline{a}}$)**

  The treatment-mediator unconfoundedness at wave t:

  $$M_t(\overline{a}_t) \perp A_t \mid H_t$$

  which identifies the causal mediator distribution under $\overline{a}$ via the observed law $f(M_t \mid A_t, H_t)$.

With the assumptions held, as the IDE and the IIE can be separately expressed as:

$$\text{IDE} = \psi_{\overline{a}\overline{G}_{\overline{a}^\star}} - \psi_{\overline{a}^\star\overline{G}_{\overline{a}^\star}}, \qquad \text{IIE} = \psi_{\overline{a}\overline{G}_{\overline{a}}} - \psi_{\overline{a}\overline{G}_{\overline{a}^\star}}.$$

Therefore, we can have the IDE and the IIE:

$$\begin{aligned}
\text{IDE} = \int \sum_{m_1,m_2} \Bigg[& \int E\Big[Y_2 \mid \overline{A}=\overline{a},\ \overline{M}=(m_1,m_2),\ Y_1=y_1,\ C=c\Big]\, dP\big(Y_1=y_1 \mid A_1=a_1,\ M_1=m_1,\ C=c\big) \\
& - \int E\Big[Y_2 \mid \overline{A}=\overline{a}^\star,\ \overline{M}=(m_1,m_2),\ Y_1=y_1,\ C=c\Big]\, dP\big(Y_1=y_1 \mid A_1=a_1',\ M_1=m_1,\ C=c\big) \Bigg] \\
& \times f_{\overline{G}_{\overline{a}^\star}}(m_1,m_2 \mid C=c)\, dP(C=c).
\end{aligned}$$

$$\begin{aligned}
\text{IIE} &= \psi_{\overline{a}\overline{G}_{\overline{a}}} - \psi_{\overline{a}\overline{G}_{\overline{a}^\star}} \\
&= \int \sum_{m_1,m_2} \Big\{ \int E\Big[Y_2 \mid \overline{A}=\overline{a},\ \overline{M}=(m_1,m_2),\ Y_1=y_1,\ C=c\Big]\, dP\big(Y_1=y_1 \mid A_1=a_1,\ M_1=m_1,\ C=c\big) \Big\} \\
&\qquad \times \Big(f_{\overline{G}_{\overline{a}}}(m_1,m_2 \mid C=c) - f_{\overline{G}_{\overline{a}^\star}}(m_1,m_2 \mid C=c) \Big)\, dP(C=c).
\end{aligned}$$

[18] The additional assumption is not needed to identify Equation 4.V.25 simply because in that equation we fix $\overline{M}=\overline{m}$.

B. Carryover, Feedback, and Full Effects

Now we move to multiple periods. In the two-period model (Figure 4.5), we assume that at each period $t = 1, 2$, A_t, M_t, and Y_t are sequential, and variables that occurred earlier in time had an effect on all subsequent variables (for instance, A_1 affects M_1, Y_1, A_2, M_2, Y_2; M_1 affects Y_1, A_2, M_2, Y_2. In substantive studies, researchers may focus more on the carryover effects or the feedback effects over the other, and ignore one side in their MSM construction.

In the models involving multiple periods, we consider the scenarios of how the variables in the two waves may interact. Figure 4.6 left two panels illustrate the example of the **carryover effects** and the **feedback effects**. The carryover effects mean that researchers assume that earlier treatments and mediators have direct effects on later mediators and outcomes, while the feedback effects stress how the mediator and outcome at the earlier stage will directly affect the later treatments and mediators. Of course, if a model includes both the carryover effects and the feedback effects (like the two-period model denoted in Figure 4.5), then it is a full model. In other words, carryover-only models allow arrows $(A_s, M_s) \rightarrow (M_t, Y_t)$ for $s < t$ but exclude arrows $\bar{Y}_{t-1} \rightarrow (A_t, M_t)$. Feedback-only models allow arrows $(Y_s, M_s) \rightarrow (A_t, M_t)$ for $s < t$ but exclude direct lagged arrows $(A_s, M_s) \rightarrow Y_t$ once A_t, M_t are included. The full model admits both[19].

For the carryover effect and the feedback effect, notably, the compositions of the conditional expectations for the outcome are different, and thus specifications on the estimand for the potential outcome, $(\psi_{\overline{am}})$, the IDE $(\psi_{\overline{a}\overline{G}_{\overline{a}'}} - \psi_{\overline{a}'\overline{G}_{\overline{a}'}})$, and the IIE $(\psi_{\overline{a}\overline{G}_{\overline{a}}} - \psi_{\overline{a}\overline{G}_{\overline{a}'}})$ are different.

[19] In the simplified DAG, we just omit the covariates at each wave t

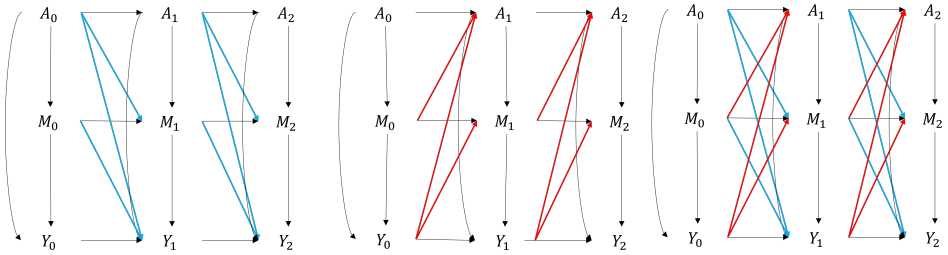


Figure 4.6: carryover and Feedback effects for time-varying causal mediation models
Note: The left panel indicates the model we assume has only carryover effects, the middle panel indicates the model we assume has only feedback effects, and the right panel shows the model for both carryover and feedback effects. We assume the time-varying covariates C affect all waves and have been omitted in the graphs (not in the equations).

For the carryover-only effects, the outcomes are influenced by the previous treatments and mediators, while the mediators are only influenced by the previous treatment but are independent of the previous outcomes. Under Assumption 4.V.1, we have the CRF formula:

$$
\begin{aligned}
\psi_{\bar{a}\,\bar{m}} &= E\{Y_t(\bar{a}_t, \bar{m}_t)\} \\
&= \int p(c)\, dc\, E\big[\, Y_t(\bar{a}_t, \bar{m}_t) \,\big|\, C = c \,\big] \\
&= \int p(c)\, dc \int_{\bar{y}_{t-1}} E\big[\, Y_t(\bar{a}_t, \bar{m}_t) \,\big|\, \bar{Y}_{t-1} = \bar{y}_{t-1}, C = c \,\big]\, p(\bar{y}_{t-1} \mid \bar{a}_{t-1}, \bar{m}_{t-1}, c)\, d\bar{y}_{t-1} \\
&= \int p(c)\, dc \int_{\bar{y}_{t-1}} E\big[\, Y_t \mid \bar{A}_t = \bar{a}_t,\ \bar{M}_t = \bar{m}_t,\ \bar{Y}_{t-1} = \bar{y}_{t-1},\ C = c \,\big]\, p(\bar{y}_{t-1} \mid \bar{a}_{t-1}, \bar{m}_{t-1}, c)\, d\bar{y}_{t-1} \\
&= \int p(c)\, dc \int_{\bar{y}_{t-1}} \mu_t(\bar{a}_t, \bar{m}_t, \bar{y}_{t-1}, c)\, p(\bar{y}_{t-1} \mid \bar{a}_{t-1}, \bar{m}_{t-1}, c)\, d\bar{y}_{t-1} \\
&= \int p(c)\, dc \int_{\bar{y}_{t-1}} \mu_t(\bar{a}_t, \bar{m}_t, \bar{y}_{t-1}, c) \prod_{i=1}^{t-1} p\big(y_i \mid \bar{a}_i,\ \bar{m}_i,\ \bar{y}_{i-1},\ c\big)\, d\bar{y}_{t-1},
\end{aligned}
$$

When we set $\overline{M} = \overline{G}_{\bar{a}^\star}$ with probability mass functions (PMFs) $g_i(m_i \mid \bar{a}_i^\star, \bar{m}_{i-1}, c)$ (no dependence on $\bar{y}_{i-1}$ in carryover-only),

$$
\psi_{\bar{a},\overline{G}_{\bar{a}^\star}} = \int p(c)\, dc \sum_{\bar{m}} \left\{ \int_{\bar{y}_{t-1}} E\big[Y_t \mid \bar{a}_t, \bar{m}_t, \bar{y}_{t-1}, c\big] \prod_{i=1}^{t-1} p\big(y_i \mid \bar{a}_i, \bar{m}_i, \bar{y}_{i-1}, c\big)\, d\bar{y}_{t-1} \right\} \prod_{i=1}^{t} g_i\big(m_i \mid \bar{a}_i^\star, \bar{m}_{i-1}, c\big).
$$

We first define the notations:

$$A_1(\bar{a}, \bar{m}, c) := \int_{\bar{y}_{t-1}} E\big[\, Y_t \mid \bar{a}_t, \bar{m}_t, \bar{y}_{t-1}, c \big] \prod_{i=1}^{t-1} p\big(y_i \mid \bar{a}_i, \bar{m}_i, \bar{y}_{i-1}, c\big)\, d\bar{y}_{t-1},$$

$$B_1(\bar{a}^\star; \bar{m}, c) := \prod_{i=1}^{t} g_i\big(m_i \mid \bar{a}_i^\star, \bar{m}_{i-1}, c\big).$$

Then, the IDE and the IIE are respectively:

$$\text{IDE} = \psi_{\bar{a}, \overline{G}_{\bar{a}^\star}} - \psi_{\bar{a}^\star, \overline{G}_{\bar{a}^\star}} = \int p(c)\, dc \sum_{\bar{m}} \big[A_1(\bar{a}, \bar{m}, c) - A_1(\bar{a}^\star, \bar{m}, c)\big] B_1(\bar{a}^\star; \bar{m}, c),$$

$$\text{IIE} = \psi_{\bar{a}, \overline{G}_{\bar{a}}} - \psi_{\bar{a}, \overline{G}_{\bar{a}^\star}} = \int p(c)\, dc \sum_{\bar{m}} A_1(\bar{a}, \bar{m}, c) \big[B_1(\bar{a}; \bar{m}, c) - B_1(\bar{a}^\star; \bar{m}, c)\big].$$

In the feedback-only effects model, the outcomes are independent of the previous treatments and mediators. Thus, we have the formula for $(\psi_{\overline{am}})$: The CRF becomes

$$\psi_{\bar{a}\,\bar{m}} = \int p(c)\, dc \int_{\bar{y}_{t-1}} E\big[\, Y_t \,\big|\, A_t = a_t,\ M_t = m_t,\ \bar{Y}_{t-1} = \bar{y}_{t-1},\ C = c \,\big] \prod_{i=1}^{t-1} p\big(y_i \mid a_i,\ m_i,\ \bar{y}_{i-1},\ c\big)\, d\bar{y}_{t-1}.$$

Now mediators depend on past outcomes, so for $\overline{G}_{\bar{a}^\star}$,

$$g_i\big(m_i \mid \bar{a}_i^\star,\ \bar{m}_{i-1},\ \bar{y}_{i-1},\ c\big), \qquad d\overline{G}_{\bar{a}^\star}(\bar{m} \mid \bar{y}_{t-1}, c) \rightsquigarrow \sum_{\bar{m}} \prod_{i=1}^{t} g_i\big(m_i \mid \bar{a}_i^\star, \bar{m}_{i-1}, \bar{y}_{i-1}, c\big).$$

Thus,

$$\psi_{\bar{a}, \overline{G}_{\bar{a}^\star}} = \int p(c)\, dc \int_{\bar{y}_{t-1}} \sum_{\bar{m}} E\big[\, Y_t \mid a_t, m_t, \bar{y}_{t-1}, c \big] \prod_{i=1}^{t-1} p\big(y_i \mid a_i, m_i, \bar{y}_{i-1}, c\big)$$
$$\prod_{i=1}^{t} g_i\big(m_i \mid \bar{a}_i^\star, \bar{m}_{i-1}, \bar{y}_{i-1}, c\big)\, d\bar{y}_{t-1}.$$

Similarly, we define the short-hand notations:

$$A_2(a_t, m_t, \bar{a}, \bar{m}, c; \bar{y}_{t-1}) := E\big[\, Y_t \mid a_t, m_t, \bar{y}_{t-1}, c \big] \prod_{i=1}^{t-1} p\big(y_i \mid a_i, m_i, \bar{y}_{i-1}, c\big),$$

$$B_2(\bar{a}^\star; \bar{m}, \bar{y}_{t-1}, c) := \prod_{i=1}^{t} g_i\big(m_i \mid \bar{a}_i^\star, \bar{m}_{i-1}, \bar{y}_{i-1}, c\big).$$

Then we have:

$$\begin{aligned}
\text{IDE} &= \psi_{\bar{a},\overline{G}_{\bar{a}^\star}} - \psi_{\bar{a}^\star,\overline{G}_{\bar{a}^\star}} \\
&= \int p(c)\,dc \int_{\bar{y}_{t-1}} \sum_{\bar{m}} \left[A_2(a_t, m_t, \bar{a}, \bar{m}, c; \bar{y}_{t-1}) - A_2(a'_t, m_t, \bar{a}^\star, \bar{m}, c; \bar{y}_{t-1}) \right] B_2(\bar{a}^\star; \bar{m}, \bar{y}_{t-1}, c)\,d\bar{y}_{t-1}, \\
\text{IIE} &= \psi_{\bar{a},\overline{G}_{\bar{a}}} - \psi_{\bar{a},\overline{G}_{\bar{a}^\star}} \\
&= \int p(c)\,dc \int_{\bar{y}_{t-1}} \sum_{\bar{m}} A_2(a_t, m_t, \bar{a}, \bar{m}, c; \bar{y}_{t-1}) \left[B_2(\bar{a}; \bar{m}, \bar{y}_{t-1}, c) - B_2(\bar{a}^\star; \bar{m}, \bar{y}_{t-1}, c) \right] d\bar{y}_{t-1}.
\end{aligned}$$

Compared to the carryover-only effects, for the feedback-only effects, because $g_i(\cdot)$ depends on $\bar{y}_{i-1}$ (feedback), the mediator measure $d\overline{G}_{\bar{a}^\star}(\bar{m} \mid \bar{y}_{t-1}, c)$ cannot be pulled outside the $\bar{Y}$-integral.Therefore, we cannot simply use the factorization of $\int_{\bar{M}}[A(\cdot)]\,B(\cdot)\,d\bar{m}$ as we did in the carryover-only effects model.

Combining the model settings in the carryover and the feedback effects, we could have the identification for the full model. Again, the CRF is:

$$\psi_{\bar{a}\,\bar{m}} = \int p(c)\,dc \int_{\bar{y}_{t-1}} E\big[\, Y_t \,\big|\, \bar{A}_t = \bar{a}_t,\ \bar{M}_t = \bar{m}_t,\ \bar{Y}_{t-1} = \bar{y}_{t-1},\ C = c \,\big] \prod_{i=1}^{t-1} p\big(y_i \mid \bar{a}_i,\ \bar{m}_i,\ \bar{y}_{i-1},\ c\big)\, d\bar{y}_{t-1}.$$

With feedback in M,

$$\begin{aligned}
\psi_{\bar{a},\overline{G}_{\bar{a}^\star}} = \int p(c)\,dc \int_{\bar{y}_{t-1}} \sum_{\bar{m}} E\big[Y_t \mid \bar{a}_t, \bar{m}_t, \bar{y}_{t-1}, c \big] \prod_{i=1}^{t-1} p\big(y_i \mid \bar{a}_i, \bar{m}_i, \bar{y}_{i-1}, c\big) \\
\prod_{i=1}^{t} g_i\big(m_i \mid \bar{a}_i^\star, \bar{m}_{i-1}, \bar{y}_{i-1}, c\big)\, d\bar{y}_{t-1}.
\end{aligned}$$

For the short-hand denotations:

$$\begin{aligned}
A_3(\bar{a}, \bar{m}, c; \bar{y}_{t-1}) &:= E\big[Y_t \mid \bar{a}_t, \bar{m}_t, \bar{y}_{t-1}, c \big] \prod_{i=1}^{t-1} p\big(y_i \mid \bar{a}_i, \bar{m}_i, \bar{y}_{i-1}, c\big), \\
B_3(\bar{a}^\star; \bar{m}, \bar{y}_{t-1}, c) &:= \prod_{i=1}^{t} g_i\big(m_i \mid \bar{a}_i^\star, \bar{m}_{i-1}, \bar{y}_{i-1}, c\big).
\end{aligned}$$

and the IDE and the IIE are therefore,

$$\text{IDE} = \psi_{\bar{a},\overline{G}_{\bar{a}^\star}} - \psi_{\bar{a}^\star,\overline{G}_{\bar{a}^\star}}$$

$$= \int p(c)\, dc \int_{\bar{y}_{t-1}} \sum_{\bar{m}} \left[A_3(\bar{a}, \bar{m}, c; \bar{y}_{t-1}) - A_3(\bar{a}^\star, \bar{m}, c; \bar{y}_{t-1}) \right] B_3(\bar{a}^\star; \bar{m}, \bar{y}_{t-1}, c)\, d\bar{y}_{t-1}, \tag{4.V.26}$$

$$\begin{aligned} \text{IIE} &= \psi_{\bar{a},\overline{G}_{\bar{a}}} - \psi_{\bar{a},\overline{G}_{\bar{a}^\star}} \\ &= \int p(c)\, dc \int_{\bar{y}_{t-1}} \sum_{\bar{m}} A_3(\bar{a}, \bar{m}, c; \bar{y}_{t-1}) \left[B_3(\bar{a}; \bar{m}, \bar{y}_{t-1}, c) - B_3(\bar{a}^\star; \bar{m}, \bar{y}_{t-1}, c) \right] d\bar{y}_{t-1}. \end{aligned} \tag{4.V.27}$$

We call the Equations 4.V.26 and 4.V.27 the general expressions of the IDE and the IIE, as they are the most precise expressions of the IDE and the IIE. Estimating all the parameters A, M, Y from all the waves in these models is never optimal in terms of time and computational resources, and it is never reasonable to believe that variables in the previous waves have the same weights of probability as the nearer waves. Hence, usually in empirical studies, we assume a **semi-Markov process**, in which we restrict the interactions of the variables in the nearest wave, as Figure 4.6 demonstrates. Specifically, the IDE and the IIE for the two-wave full model are:

$$\begin{aligned} \text{IDE} &= \psi_{\bar{a},\overline{G}_{\bar{a}^\star}} - \psi_{\bar{a}^\star,\overline{G}_{\bar{a}^\star}} \\ &= \int p(c)\, dc \int dy_1 \sum_{m_1} \sum_{m_2} \Big\{ \mu_2(a_1, a_2, m_1, m_2, y_1, c)\, p(y_1 \mid a_1, m_1, c) \\ &- \mu_2(a_1^\star, a_2^\star, m_1, m_2, y_1, c)\, p(y_1 \mid a_1^\star, m_1, c) \Big\} \\ &\times\ g_1(m_1 \mid a_1^\star, c)\, g_2(m_2 \mid a_2^\star, m_1, y_1, c). \end{aligned}$$

$$\begin{aligned} \text{IIE} &= \psi_{\bar{a},\overline{G}_{\bar{a}}} - \psi_{\bar{a},\overline{G}_{\bar{a}^\star}} \\ &= \int p(c)\, dc \int dy_1 \sum_{m_1} \sum_{m_2} \mu_2(a_1, a_2, m_1, m_2, y_1, c)\, p(y_1 \mid a_1, m_1, c) \\ &\quad \times \Big[g_1(m_1 \mid a_1, c)\, g_2(m_2 \mid a_2, m_1, y_1, c) \ -\ g_1(m_1 \mid a_1^\star, c)\, g_2(m_2 \mid a_2^\star, m_1, y_1, c) \Big]. \end{aligned}$$

Based on the general expressions of the IDE and the IIE, we may derive the doubly robust/efficient estimators for them.

VI. Doubly Robust Estimator for the IDE and the IIE in Time-Varying Full Models

For the static models, we have demonstrated the DR/efficient estimator for the IDE/ IIE in Section IV. Now we discuss the DR/efficient estimator for the dynamic models. Due to the complexity of the IDE and the IIE expressions in multi-period models, it is hard to give a precise expression on the efficient estimator for the effects. However, based on our derivation of the IDE and the IIE in the static model, we indeed could have the DR/efficient expression of the IDE/IIE with three parts: the outcome-residual part, the transport part, and the centering part. We first set the nuisance functions:

$$
\begin{aligned}
H_i &= (C, \bar{A}_i, \bar{M}_i, \bar{Y}_{i-1}),\\
H_i^A &= (C, \bar{A}_{i-1}, \bar{M}_{i-1}, \bar{Y}_{i-1}),\\
H_i^M &= (C, \bar{A}_i, \bar{M}_{i-1}, \bar{Y}_{i-1}),\\
\pi_i^A(a_i \mid H_i^A) &= P(A_i = a_i \mid H_i^A),\\
\pi_i^M(m_i \mid H_i^M) &= P(M_i = m_i \mid H_i^M).
\end{aligned}
$$

The mediator intervention is given by known (or modeled) kernels $g_i^{\bar{a}^\star}(m_i \mid a_i^\star, \bar{m}_{i-1}, \bar{y}_{i-1}, c)$. Firstly, we define the outcome regression function:

$$
\mu_i(a_i, m_i, \bar{y}_{i-1}, c) := E\big[\, Y_i \mid A_i = a_i, M_i = m_i, \bar{Y}_{i-1} = \bar{y}_{i-1}, C = c \big].
$$

Hence, we could set the sequential regression as:

$$
\bar{Q}_t^{\bar{a},\bar{a}^\star}(H_t) := \mu_t(a_t, M_t, \bar{Y}_{t-1}, C),
$$

and for $k = t, \ldots, 1$ define recursively

$$
\bar{Q}_{k-1}^{\bar{a},\bar{a}^\star}(H_{k-1}) := \int E\Big[\bar{Q}_k^{\bar{a},\bar{a}^\star}(H_k) \mid A_k = a_k,\ M_k = m_k,\ H_{k-1}\Big]\, g_k^{\bar{a}^\star}\big(m_k \mid a_k^\star, \bar{M}_{k-1}, \bar{Y}_{k-1}, C\big)\, dm_k.
$$

Recall this part should be re-weighted by the counterfactual/interventional distribution of the mediator; thus, we define the IPW weight as:

$$W_k(\bar{a}, \bar{a}^\star) := \prod_{i=1}^{k} \frac{\mathbb{1}(A_i = a_i)}{\pi_i^A(a_i \mid H_i^A)} \cdot \frac{g_i^{\bar{a}^\star}\left(M_i \mid a_i^\star, \bar{M}_{i-1}, \bar{Y}_{i-1}, C\right)}{\pi_i^M\left(M_i \mid H_i^M\right)}.$$

for $k = 1, \ldots, t$. Secondly, we need to identify the transport term (from stage k back to $k-1$). Define the *transport* of a function $f(H_k)$ through the mediator intervention:

$$T_k^{\bar{a},\bar{a}^\star} f(H_{k-1}) := \int E\left[f(H_k) \mid A_k = a_k,\ M_k = m_k,\ H_{k-1}\right] g_k^{\bar{a}^\star}\left(m_k \mid a_k^\star, \bar{M}_{k-1}, \bar{Y}_{k-1}, C\right) dm_k.$$

Then the sequential regressions satisfy

$$\widehat{\bar{Q}}_t^{\bar{a},\bar{a}^\star}(H_t) = \hat{\mu}_t(a_t, M_t, \bar{Y}_{t-1}, C), \qquad \widehat{\bar{Q}}_{k-1}^{\bar{a},\bar{a}^\star} = T_k^{\bar{a},\bar{a}^\star} \widehat{\bar{Q}}_k^{\bar{a},\bar{a}^\star}, \quad k = t, \ldots, 1.$$

Finally, the centering part is the repeated transport of the stage-t regression:

$$\bar{Q}_0^{\bar{a},\bar{a}^\star}(C) = \left(T_1^{\bar{a},\bar{a}^\star} \circ T_2^{\bar{a},\bar{a}^\star} \circ \cdots \circ T_t^{\bar{a},\bar{a}^\star}\right) \mu_t(\cdot).$$

Therefore, we have the DR estimator for $\psi_{\overline{aG}_{\overline{a^*}}}$:

$$\widehat{\psi}^{\mathrm{DR}}_{\bar{a},\bar{G}_{\bar{a}^\star}} = \frac{1}{n} \sum_{r=1}^{n} \left[\sum_{k=1}^{t} \widehat{W}_{k,r} \left\{ \widehat{\bar{Q}}_k(H_{k,r}) - \left(\widehat{T}_k \widehat{\bar{Q}}_k\right)(H_{k-1,r}) \right\} + \left(\widehat{T}_1 \circ \cdots \circ \widehat{T}_t\right) \hat{\mu}_t(C_r) \right],$$

Since $(\widehat{T}_k \widehat{\bar{Q}}_k) = \widehat{\bar{Q}}_{k-1}$ by definition, we could simplify the expression as:

$$\widehat{\psi}^{\mathrm{DR}}_{\bar{a},\bar{G}_{\bar{a}^\star}} = \frac{1}{n} \sum_{r=1}^{n} \left[\sum_{k=1}^{t} \widehat{W}_{k,r}(\bar{a}, \bar{a}^\star) \left\{ \widehat{\bar{Q}}_k^{\bar{a},\bar{a}^\star}(H_{k,r}) - \widehat{\bar{Q}}_{k-1}^{\bar{a},\bar{a}^\star}(H_{k-1,r}) \right\} + \widehat{\bar{Q}}_0^{\bar{a},\bar{a}^\star}(C_r) \right].$$

The double robustness here holds if either the weights or the regression and transport terms are correct. For the IDE and the IIE, therefore,

$$\widehat{\mathrm{IDE}}_{\mathrm{DR}} = \widehat{\psi}^{\mathrm{DR}}_{\bar{a},\bar{G}_{\bar{a}^\star}} - \widehat{\psi}^{\mathrm{DR}}_{\bar{a}^\star,\bar{G}_{\bar{a}^\star}},$$

$$\widehat{\mathrm{IIE}}_{\mathrm{DR}} = \widehat{\psi}^{\mathrm{DR}}_{\bar{a},\bar{G}_{\bar{a}}} - \widehat{\psi}^{\mathrm{DR}}_{\bar{a},\bar{G}_{\bar{a}^\star}}.$$

Specifically, for the carryover-only effects, in both $W(\cdot)$ and $Q(\cdot)$ we rewrite the mediator policy as $g_i^{\bar{a}^\star}\big(M_i \mid a_i^\star, \bar{M}_{i-1}, C\big)$ (since M_i is not affected by Y_{i-1}), and use $\pi_i^M\big(M_i \mid A_i, \bar{M}_{i-1}, C\big)$ together with $\pi_i^A\big(a_i \mid C, \bar{A}_{i-1}\big)$. For the feedback-only effects, we set $g_i^{\bar{a}^\star}\big(M_i \mid a_i^\star, \bar{M}_{i-1}, \bar{Y}_{i-1}, C\big)$, $\pi_i^M\big(M_i \mid A_i, \bar{M}_{i-1}, \bar{Y}_{i-1}, C\big)$, and $\pi_i^A\big(a_i \mid C, \bar{A}_{i-1}, \bar{M}_{i-1}, \bar{Y}_{i-1}\big)$, while Y_i has no direct dependence on (A_{i-1}, M_{i-1}) once (A_i, M_i, Y_{i-1}, C) are included. Since M is discrete in this context, all integrals over m_i in $Q(\cdot)$ and $T(\cdot)$ are replaced by finite (or countable) sums. On the semi-Markov settings, as we only concentrate on the relationships between the two waves, the above derivation still remains the same.

In summary, the algorithms for the doubly robust estimators in causal mediation analysis:

- Identify the target estimand. Based on the substantive or data-driven assumptions, identify the model describing the relationships among the treatment, the mediator, and the outcome (usually with the DAGs); and identify the estimand and its functional form (the CRF, the CDE, the NDE/IDE, or the NIE/IIE).
- Based on the estimand, find the relevant CRF and intermediate terms.
- Derive the nuisance functions, and construct the four nuisance functions from the empirical data *using K-fold cross-fitting*: split the sample into K folds; on each training fold, fit/tune the nuisance models via cross-validation; generate out-of-fold predictions for every observation so that all nuisance estimates are cross-fitted.
- Calculate the relevant efficient influence function for the desired estimand using the cross-fitted nuisance estimates, and based on the efficient influence function, derive the efficient (doubly robust) estimator.

VII. Empirical Studies

This section applies our method to replicate a classic causal mediation analysis in political science: the racial fractionalization theory. In a highly-cited *American Political Science Review (APSR)* paper, Fearon and Laitin (2003) analyzed the post-Second World War civil conflict data. They concluded that racial fractionalization/diversification does not significantly increase the likelihood of civil violence. They argue that conditions that favor insurgencies, like political instability, large populations, and poverty, create the opportunity for rebels to recruit and thus increase civil violence. In other words, the conditions like political instability, large populations, and poverty, from their views, are the mediators impacting the causal effects between racial fractionalization (the treatment) and the onset of civil wars (adding the mediators, the direct causal effects vanished).

Based on the conclusion drawn in Fearon and Laitin's (2003)paper, Acharya et al.(2016) used the traditional linear model method and further examined if political instability, rather than other variables, is the **only** mediator that alleviates the direct causal effect of racial fractionalization on the onset of civil wars. However, they rejected their hypothesis and concluded that putting political instability alone at the position of the mediator could not fully explain the causal effect of racial fractionalization on the onset of civil wars. We first replicate their results under our doubly robust/debiased causal mediation framework in our methodological illustration. Then, we turn this model into a dynamic one, examining the carryover and the feedback effects of political instability mediating the impact of racial fractionalization on civil wars.

A. Static Models

For static models, we aim to test whether political instability is the only mediator of the causal effect of racial fractionalization on the onset of civil wars in post-Second World War countries. Therefore, the outcome variable is a dichotomous measurement of whether a civil war starts in a specific year, and the treatment is the level of racial fractionalization measured by Fearon and Laitin (2003). To simplify our analysis, we transformed the continuous measurement into a dichotomous one, assuming that racial fractionalization is high when the original fractionalization score surpasses 0.5. The mediation variable, political instability, is a dichotomous variable capturing whether the country had a three-or-greater change in the Polity IV Index between $t-2$ and $t-1$, given that the outcome is measured at t. There are 14 missing cases on the mediator (N = 6596), and we drop these cases.

Moreover, we include several pre- and post-treatment confounders. The pre-treatment confounders are the variables that may affect racial fractionalization before the observation year, including the estimated percentage of mountainous terrain in the country, whether the country is a non-contiguous state, whether the country relies on fuel exportation (defined by one-third of the export revenue coming from fuels), and the degree of religious fractionalization of the country. The post-treatment confounders, on the other hand, include the variables that affect the political stability and the onset of civil wars in the country, including whether an ongoing war in the previous year, the country's GDP per capita in the previous year, the logarithm of population, whether the regime is a new one (measured by whether in its first two years of its existence), and the previous year's Polity IV index score. The DAG for the static model can be seen in Figure 4.7:

We begin our modeling with a null linear probability model (LPM), with only racial frac-

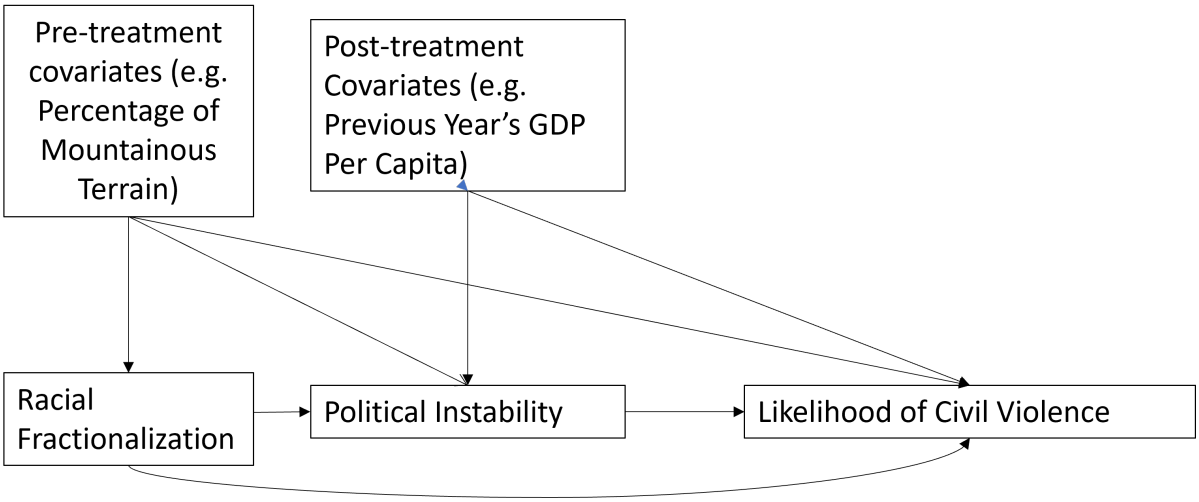


Figure 4.7: Directed Acyclic Graph for the Replication on Fearon and Laitin's Paper

tionalization and political instability as the independent variables. The coefficient of racial fractionalization (shown in the leftmost column in Figure 4.8) suggests the unbalanced treatment effect $E[Y \mid A = 1] - E[Y \mid A = 0]$, and it appears that if we do not consider the confounders between the treatment and the outcome, racial fractionalization does significantly increase the probability of the onset of a civil war, controlling the degree of political instability. We then use the nested model method, adding the confounders into the LPM, as Fearon and Laitin (2003) originally designed, and we call this model the baseline model. The second column in Figure 4.8 shows the results for the baseline model. Although our measurements differ from those in the original paper, the results are the same. After adding all confounders, we couldn't reject the hypothesis that racial fractionalization does not significantly directly affect the probability of the onset of civil wars.

We then test the hypothesis on whether political instability is the only mediator of the direct effect of racial fractionalization on the onset of civil wars. To test this hypothesis, Acharya et al. (2016) adopted the controlled direct effect (CDE) framework: if political in-

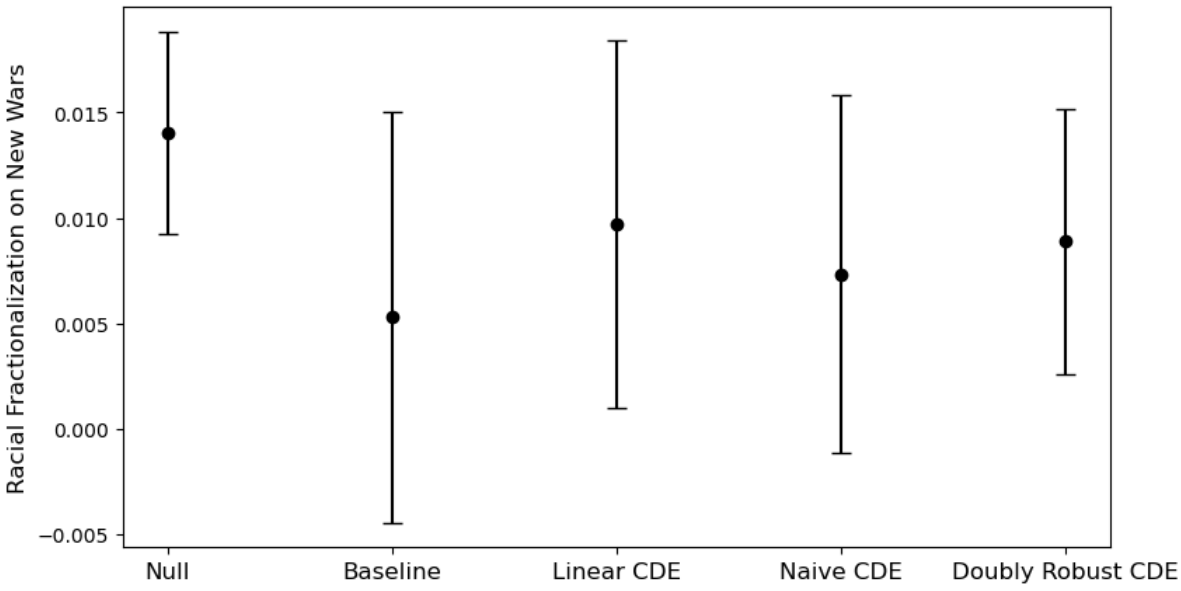


Figure 4.8: Estimation for racial fractionalization on the onset of civil wars (static models)
Note: For the null and baseline models, the standard errors derive directly from the regression models; for the linear CDE, naive plug-in CDE, and doubly robust CDE models, the standard errors are bootstrapped with 100 loops. The basic machine learning model is an XGBoost model (Chen and Guestrin 2016) (with 1000 estimators for each naive estimator).

stability is the only mediator, the direct effect of racial fractionalization on the onset of civil wars should be zero given all countries were politically unstable. In mathematical expressions, our null hypothesis is $CDE(m = 1) = \psi_{1,1} - \psi_{0,1} = 0$. Like their study, we initially attempted to fit the CDE using a linear parametric model, and the results are presented in the third column of Figure 4.8. The parametric model result indicates a significant direct effect, so our null hypothesis is rejected. Therefore, we infer that mechanisms other than political instability also mediate the causal relationship between racial fractionalization and the onset of civil wars.

We further test the hypothesis by establishing a naïve plug-in nonparametric estimator for the CDE. We use a machine learning model to predict the controlled response functions (CRFs) for $\psi_{1,1}$ and $\psi_{0,1}$ and subtract them. The result is shown in the fourth column of Figure 4.8. The estimation from naïve plug-in models is slightly smaller than from the linear model. However, it still illustrates that political instability alone could not mediate all direct effects of racial fractionalization on the onset of civil wars.

The linear parametric model and the naïve plug-in nonparametric models only rely on the pure imputation estimation for $\hat{E}[Y|A = a, M = m, L, C]$, and the result is biased if the imputed expectation is biased. Therefore, we use the aforementioned doubly robust machine learning method to re-estimate. The result is unbiased if either the pure imputation estimator or the pure weighting estimator is unbiased. The result is shown in the rightmost column in Figure 4.8. Similar to the linear estimator and the naïve plug-in estimator results, the doubly robust CDE estimator clearly rejects the null hypothesis.

However, we can infer from Figure 4.8 that the direct treatment effect estimated by the doubly robust model is slightly higher than the effects estimated from the naïve plug-in nonparametric model results, with smaller confidence intervals. This clearly reveals the advantage of the DML estimator in nonparametric marginal structural models over the naive plug-in estimators. Based on the definition of the DML estimator, it is an estimator based on the naive plug-in result, adding a first-order correction term that considers the perturbation of the individuals. Moreover, when the score is the efficient influence function and regularity and nuisance-estimation conditions hold, the DML estimator attains the semiparametric efficiency bound within the class of regular asymptotically linear estimators. The estimation from the linear parametric model, if the model specification is correct, should also be unbiased and efficient in the corresponding parametric model. However, compared to the nonparametric models, the lack of flexibility in the relationship between the outcome and the features (the treatment and the mediator) could be a disadvantage in some studies.

B. Dynamic Models

Next, we extend our discussion towards a dynamic mediation model. The static models could be regarded as the instant mediation effect, and we reject the hypothesis that political instability alone could mediate all direct effects of racial fractionalization on the onset of civil wars. We would further like to test whether the long-lasting political instability alone may explain the dynamic relationship between racial fractionalization and civil wars. To test the assumption, we still adopt the CDE framework.

In the dynamic model, we focus on the cumulative effect. Thus, we change the outcome we measure from the dichotomous measurement on the onset of a civil war to the number of wars in progress in the country at the observation year. Besides the outcome, the mea-

surements of the other variables remain the same in the dynamic models. We take their values in the observation year and the previous years into the models for the treatment, the mediator, and the outcome. To simplify our analysis, we only consider a two-period dynamic model.

We apply two MSM settings: the carryover and feedback effect models. In the dynamic models, besides the identifications in the static models among the treatment, the mediator, and the outcome, we also assume that the previous racial fractionalization, political instability, and the number of wars affect the corresponding variables in the observation year. Besides, for the carryover effect model, we also assume that racial fractionalization and political instability in the previous year affect the political instability and the number of wars in the observation year. For the feedback effect model, on the other hand, we assume that the previous mediator and outcome affect the treatment and the mediator in the next round.

We only elaborate on the results from the doubly robust estimator for the carryover and feedback effects models. Figure 4.9 shows the results. As can be seen, the estimations from both models reject the hypothesis and suggest that political instability alone could not eliminate all the direct effects of racial fractionalization and the number of civil wars. However, after controlling the mediation effect of political instability, the direct effect of racial fractionalization on civil conflict is larger in the feedback models, suggesting that controlling the carryover effects of the previous racial fractionalization and political instability on the war status in the next round explains more of the indirect effect than controlling the feedback effects of the previous mediator and treatment on the treatment and mediator in the next round. However, it is also worth noting that the feedback model's estimates are much less precisely estimated, with wider confidence intervals and correspondingly weaker statistical significance.

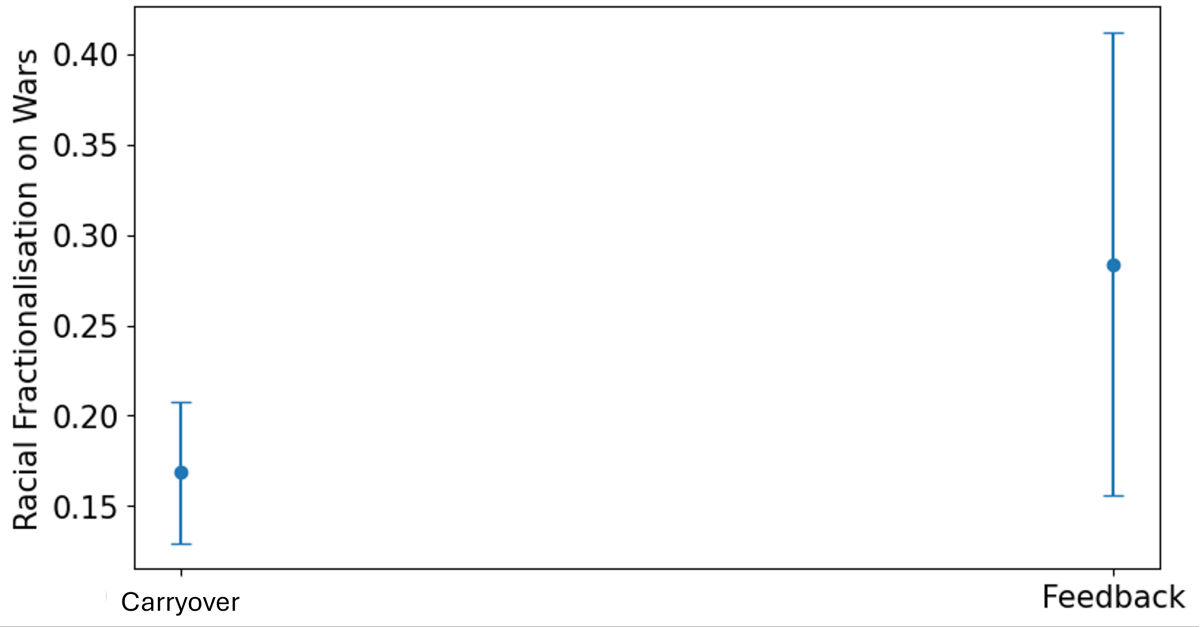


Figure 4.9: Estimation for racial fractionalization on civil war processes (Dynamic models)
Note: Standard errors are bootstrapped with 100 loops. The basic machine learning model is an XGBoost model (with 1000 estimators for each naive estimator).

VIII. Conclusion

In this chapter, we discuss the efficient/doubly robust estimations for the causal mediation analysis. We first discussed the "do-calculus" (DoC) framework for causal inference, which is different but has a tight connection with the Neyman-Rubin causal framework we discussed in the previous chapters. Based on the background knowledge of efficient/doubly robust estimators from the introductory chapter, we derive the efficient/doubly robust estimators for the controlled response function (CRF); and use the CRF as the starting point to derive the efficient/doubly robust estimators for the controlled direct effect, the natural direct and indirect effect, and the interventional direct and indirect effects.

The core question researchers aim to tackle with the causal mediation analysis is how

much the mediator plays a role in mediating the causal relationship between the treatment and the outcome. Specifically, the questions include: 1) how much of the total causal effect is accounted for by the causal effect through the mediating variable (by the decomposition of the direct and indirect effects); 2) how the causal effects, measured by the potential outcome under treatment minus under control, would change under different mediation levels; 3) what the direct and the indirect effect would be if we intervene on the distribution of the mediators. Different statistical indicators are required to answer different questions, and the indicators are based on different sets of model specifications and, most importantly, causal identification assumptions. Among all the causal assumptions, the assumption of unconfoundedness is the most crucial. For the first question, the natural direct and indirect effects are required, and they have the most strict unconfoundedness assumption settings: they not only require the unconfoundedness between the treatment and the outcome, the treatment and the mediator, the mediator and the outcome, but also require no association between the treatment and the post-treatment confounders (the mediator-outcome controlled covariates). For the second question, the controlled direct effect (CDE) will be most commonly applied, but the assumption requirements are the weakest: as we fix the level of the mediators, we only need no confoundedness between the fixed mediator and the outcome, and between the treatment and the outcome. For the third question, we typically use the interventional direct and indirect effects. The assumptions for interventional effects require the unconfoundedness between the treatment and the outcome, between the treatment and the mediator, and between the mediator and the outcome.

In many cases, researchers seek to apply the causal mediation framework to the time-varying treatments and time-varying mediators settings. We derive the algorithm for the doubly robust CRF for time-varying models from the $g-$formula in parametric MSMs. We

differentiate the carryover and feedback effects based on different confounding assumptions for the time-varying treatment, mediator, and outcome: in the carryover effect, the previous treatment and mediator restrict the mediator and outcome in the next period, while in the feedback effect, the previous mediator and outcome affect the treatment and mediator in the next period. The CRF is based on cumulative treatment and mediator in the time-varying settings. Moreover, because the previous outcomes are affected by the treatment, and they influence the outcomes in later periods, the unconfounding assumption between the treatment and the post-treatment covariates is violated. We could only examine the interventional direct and indirect effects in the time-varying settings.

Indeed, decomposing the total treatment effect into the direct and indirect effects is only a straightforward way to isolate the effects of the mediator on the treatment effects, as we also have shown the effect of ψ_{alm} when there are confounding effects between the post-treatment covariates and the treatment and we derived the efficient/doubly robust estimator for the CRF. Moreover, we may also adopt a three-way decomposition method to decompose the total effect into the direct effect, the indirect effect, and the interactional effect, allowing the intersection between the treatment and the mediator:

$$\begin{aligned} ATE &= \psi_{1,m(1)} - \psi_{0,m(0)} \\ &= \underbrace{(\psi_{1,0} - \psi_{0,0})}_{(:a)} + \underbrace{(\psi_{1,m(1)} - \psi_{1,m(0)})}_{(:b)} + \underbrace{P(m(0))\Big[(\psi_{1,1} - \psi_{0,1}) - (\psi_{1,0} - \psi_{0,0})\Big]}_{(:c)} \\ &+ \mathrm{Cov}\big(P(m(0)), (\psi_{1,1} - \psi_{0,1}) - (\psi_{1,0} - \psi_{0,0})\big) \end{aligned} \tag{4.VIII.28}$$

The framework for causal mediation analysis has recently been applied to calculating the reduced disparities or gap-closing effects of the outcome between two subgroups (Jackson and VanderWeele 2018; Lundberg 2024). In Pearl's (2009) causal analysis framework, we

have to be able to manipulate the levels of the treatment and the mediators to derive the potential outcomes based on the manipulation. In social science, we are sometimes interested in how an action (for instance, the execution of a policy) intervenes in the existing gaps between different social groups. The causal mediation framework could also be applied to these studies. For example, our previous study (Zhou and Pan 2023) discusses how college graduation shrinks the earnings gap for Black and White college attendees. Therefore, we set A as the race, M as the college graduation, and the outcome Y represents their earnings. We applied and transformed the decomposition framework illustrated in Equation 4.VIII.28 and decomposed the total earnings gap into the earnings gap between attendees, the proportional gap in completion ($P(M(1))$, and the earnings gap between completion ($CDE(1) - CDE(0)$), and the covariance between the completion proportion and earnings gap (we have no part (b) in Equation 4.VIII.28 simply because our analytical subjects are all college attendees).

The causal mediation analysis framework applies to the conditions where researchers are more interested in the causal effects on the group level (the "global" effects) rather than on the individual level (the "local" effects, for instance, measuring the change in a person's life course). In Chapter 5, we will apply the model we developed here in an empirical study to examine the marriage and parenthood penalties and premiums in earnings and discuss how the earnings gaps between groups are mediated by the time devoted to the labor market for both genders.

Chapter 5

From Static Models to Dynamic Models: Reconsidering Carryover and Feedback Effects in Marriage and Parenthood Penalties and Premiums

I. Introduction

Marriage and parenthood have divergent causal effects on men's and women's wages (England 2005; Waldfogel 1998; Gough and Noonan 2013; Angelov et al. 2016). Previous literature suggested that becoming a wife and becoming a mother negatively affect heterosexual women's wages (Glauber 2007; Waldfogel 1997; Budig and England 2001; Crittenden 2001; England et al. 2016; Correll et al. 2007; Cheng 2016) while becoming husbands or fathers gave wage premiums to heterosexual men (Korenman and Neumark 1991; Dougherty 2006; Killewald 2013; Killewald and Gough 2013a; Killewald and Lundberg 2017; Ludwig and Brüderl 2018; Loughran and Zissimopoulos 2009; Lundberg and Rose 2000; Glauber 2018: 2008; Lundberg and Rose 2002; Hodges and Budig 2010). According to the literature, women's marital and motherhood penalty and men's marital and fatherhood premiums are the primary causes of the gender wage gap in the labor market.

Being married and becoming parents will doubtlessly impact the time arrangements for domestic work and labor market participation, especially for women. Compared with men, women tend to have less time in the labor market and switch their attention to domestic work after becoming mothers and getting married due to the interruption of childbirth (Becker 1981; Anderson et al. 2002: 2003; Waldfogel 1997; Grimshaw and Rubery 2015), the specialization arrangement within the couples(Chun and Lee 2001; Killewald and Gough 2013a), and institutional stereotypes and discrimination in the labor market (Correll et al. 2007; Glauber 2018; Yu and Kuo 2017; Yu and Hara 2021; Luhr 2020). For men, however, marriage and childbirth do not impede them from labor market work; nevertheless, due to specialization arrangements and the need to make up wives' labor market loss, they will devote more time to their professional work (Glauber 2008; Lundberg and Rose 2000: 2002; Killewald 2013; Hersch and Stratton 2000). In this regard, we can preliminarily understand that labor market participation is a crucial mediator in the causal relationship between marriage, parenthood, and wage returns.

Most previous work has noticed the causal effects of marriage and parenthood on wages for men and women and has pointed out the mediation effect of labor market participation: they rely on linear fixed-effect (LN-FE) models to compare the wage and labor market time losses/gains before and after the event of marriage and childbirth to estimate penalties and premiums Budig and England (2001); Killewald and Gough (2013a); Cheng (2016); Ludwig and Brüderl (2018). However, the LN-FE models measure the average treatment effects for the treated (ATT), which is the effect of marriage and childbirth on the wage for individuals who experienced marriage and childbearing, and people without marriage or childbearing experiences (the never-takers) were excluded from identifying the treatment effects (Gough and Noonan 2013; Blau and Kahn 2017; Ludwig and Brüderl 2018; Vagni and Breen 2021).

For distributional and inequality questions, the estimand of interest is the average treatment effects (ATE) over all labor market participants, not only the switchers.

A second limitation is dynamic: standard fixed-effects specifications ignore heterogeneity in the time and persistence of treatment, mediator, and outcome, and assume time-invariant unobserved factors absorb all confounding (Kim and Imai 2019; Zhou and Wodtke 2020). With the model assumptions, researchers do not allow (i)carryover of past treatment into current outcomes or mediators, nor (ii) treatment-confounder feedback in which prior outcomes or mediators influence subsequent treatment. However, previous research on the timing of first births and marriages indicates that individuals' labor market experiences and positions could influence the age at which they have children and get married. In addition, several recent studies on marital and parenthood premiums and penalties suggest that marriage and childbearing continuously influence an individual's trajectory in the labor market several years after the event. Hence, designs that accommodate dynamic treatment and mediators that allow carryover and feedback effects are required to capture the mechanisms of how prior marriage/parenthood affects current labor market participation and wages, and how prior participation and wages in turn shape future family decisions.

In this paper, we take a causal-inference approach using nonparametric marginal structural models (NPMSM, Robins et al. 2000) to estimate the causal effects of marriage and parenthood on wages for American men and women, with labor market participation treated as a mediator. To situate these effects in a labor market perspective and inequality framework, we adopt the gap-closing estimand (Jackson and VanderWeele 2018; Lundberg 2024) quantifying how wage disparities would change under hypothetical interventions on marital status (married vs. unmarried) and parental status (with children vs. childless). For mediation,

we follow VanderWeele's (2015) and estimate controlled direct effects (CDE) by fixing labor market participation at specified levels, revealing how the marriage/parenthood effect on wages varies when participation is held constant. We also decompose the total (average) treatment effect (TATE) into direct and indirect components attributable to labor market participation.

We analyze both static and dynamic specifications. Firstly, we establish the static models to compare with the results from the LN-FE models, treating effects as time-invariant. We then turn to the dynamic models that accommodate (i) carryover effects of prior marital/parental status on current participation and wages, and (ii) feedback effects of prior participation and wages on subsequent marriage and fertility decisions. This dynamic formulation directly addresses treatment-mediator-outcome dependencies that static LN-FE models rule out.

Our contributions are twofold. Substantively, we provide dynamic estimates that incorporate carryover and feedback effects between family status, labor-market participation, and wages, extending a literature that has primarily relied on static LN-FE designs. Methodologically, we pair NPMSMs with debiased/double machine learning (DML; Chernozhukov et al., 2018a) to flexibly learn high-dimensional nuisance functions and deliver orthogonalized, robust estimation of causal and mediation parameters, complementing the developments in Chapter 4.

II. Literature Review

A. Marriage and Parenthood, Premium and Penalty

Parenthood and marriage are central drivers of gender inequality in wages (Waldfogel 1998; Kleven et al. 2019; Gough and Noonan 2013; Vagni and Breen 2021; Grimshaw and Rubery

2015; Glauber 2007). The literature documents "motherhood penalties" and women's marital wage penalties, alongside evidence of "fatherhood" and male marital premiums. In U.S. data, linear fixed-effects (LN-FE) studies consistently find motherhood penalties—on the order of 6% per additional child in many settings (Budig and England 2001; Glauber 2007; Anderson et al. 2002: 2003; Budig and Hodges 2010; Crittenden 2001; Gangl and Ziefle 2009)—while estimates for women's marriage effects are mixed, ranging from null or negative (Anderson et al. 2003; Loughran and Zissimopoulos 2009; Van der Klaauw 1996) to positive once fertility is controlled (Budig and England 2001; Glauber 2007; Taniguchi 1999; Waldfogel 1997; Killewald and Gough 2013a). For men, several studies report non-negative fatherhood effects and frequently a premium (Glauber 2008; Hersch and Stratton 2000; Lundberg and Rose 2000: 2002; Killewald 2013); LN-FE estimates often attribute 5–7% higher wages to marriage (Cornwell and Rupert 1997; Dougherty 2006; Killewald 2013; Killewald and Gough 2013a), with life-course analyses pointing to larger late-career gains (Dougherty 2006; Cheng 2016). Yet more recent FE specifications (e.g., linear splines or FE individual slopes) attenuate or eliminate men's marital premia (Killewald and Lundberg 2017; Ludwig and Brüderl 2018).

Crucially, the LN-FE designs identify the **average treatment effects on the treated (ATT)** among switchers who change marital or parental status within the panel. Individuals who never marry or never become parents do not identify the treatment coefficient (although they can inform time effects), which raises interpretive concerns if ATT is interpreted as a population-wide average treatment effect (ATE) (Gough and Noonan 2013; Blau and Kahn 2017; Vagni and Breen 2021). Recently, extensions like FE individual slopes (FEIS) (Ludwig and Brüderl 2018) or synthetic control studies (Vagni and Breen 2021) mitigate some selection and trend-heterogeneity, but the estimand remains the effect for individuals who

transition (an ATT among switchers), rather than a population ATE. [1].

By contrast, some studies also adopt experimental or audit designs and compare outcomes between groups (married vs. unmarried; parents vs. non-parents) under randomized or quasi-randomized signals and thus estimate **average treatment effects (ATE)** at the population level (Correll et al. 2007; Pedulla 2016; Duguet et al. 2005; Petit 2007; Neumark 2018). For example, Correll et al. (2007) randomized parental-status cues on matched résumés and found sizable motherhood penalties (and some fatherhood premia), directly addressing group disparities.

Studies report that the ATT and the ATE have distinct sociological implications. Studies that report the ATT estimated from within-person changes before and after marriage or parenthood among those who actually transition, indeed speak to life-course and family-process questions: the opportunity cost or gain an individual experiences upon entering these roles. Studies with the ATE as the estimand, on the other hand, address population-level labor market differentiation and discrimination by comparing outcomes between parents and non-parents and between married and unmarried under hypothetical interventions. Because we work with observational longitudinal data, we employ nonparametric marginal structural models (NPMSMs; Robins et al. 2000), a class of g-methods that reweight observed treatment/mediator histories to address time-varying confounding, accommodate carryover and feedback between family status, labor-market participation, and wages, and thus align the estimand with the population question. In tandem with a gap-closing es-

[1]For the FEIS models, they add a person-specific time trend so switchers are compared to their own pre-trend. This reduces bias from differential trends, but the treatment coefficient is still identified only by people who change status; never-takers don't identify it. For the synthetic control studies, they build a synthetic counterfactual for each treated unit from donors matched on pre-trends. The effect is computed for treated units and then averaged across them, which is still an ATT for those who receive treatment, not an ATE.

timand (Jackson and VanderWeele 2018; Lundberg 2024), this framework asks how between-group wage gaps would change under hypothetical interventions on marital or parental status. However, we have a clear trade-off: relative to LN-FE's strength in life-trajectory questions, NPMSMs have stricter requirements for unconfoundedness and positivity.

B. Gap Closing Estimand for Marriage and Parenthood Effects

Since our focus is on group disparities (the ATE), which refers to how population-level wage gaps would change under hypothetical interventions, we use the gap-closing estimand in conjunction with nonparametric marginal structural models (NPMSMs) (Jackson and VanderWeele 2018; Lundberg 2024; Robins et al. 2000). The gap-closing perspective asks a concrete counterfactual: how much would the observed wage gap between two groups shrink or enlarge if we intervened on a manipulable factor? Indeed, this framing connects directly to labor market inequality and potential labor market discrimination.

A crucial step in the gap-closing framework is to specify which gap and which intervention. For prior studies that estimate ATT among those who transition, they usually put within-person variables in the model as controls and implicitly hold person-fixed characteristics constant. In our setting, the gap-closing estimand requires us to define the between-group comparison (parents vs. non-parents; married vs. unmarried) and the manipulable factor we will intervene on. We define the groups as the attribute whose disparity we seek to explain, and the manipulable factor as the policy lever we intervene on; we then compare between-group wage gaps under fixed intervention regimes, yielding controlled disparities for parenthood and for marriage. In this sense, we actually use the controlled direct effect (CDE) (VanderWeele 2015) for the gap-closing estimand.

Due to our research question on the causal impact of parenthood and marriage on wages, we analyze each gender separately and report two complementary gap-closing contrasts. For the parenthood effect, we compare the wage gaps among parents with different numbers of children, with marital status fixed. The estimator isolates the parenthood effect when marriage is held constant, aiming to answer how different numbers of children affect individual wages for the married and unmarried subgroups. Secondly, for the marriage effect, we fix the number of children and compare the wage gap between the married and unmarried. This isolates the marriage effect net of parity, aligning questions about the labor market's perception of marital status.

We use controlled disparities to echo the intuition of LN-FE while targeting a population-level ATE. In LN-FE estimates of the parenthood effect, marital status is typically included as a time-varying covariate (and, conversely, parity is included when estimating a marriage effect). These specifications identify an ATT among individuals who change status and condition on the other family status at its observed values. To provide an analogous contrast at the population level, our gap-closing analysis defines explicit intervention regimes: we fix marital status and compare parents with non-parents within the same gender (a controlled parenthood contrast), and we fix parity and compare married individuals with unmarried individuals within the same gender (a controlled marriage contrast). This retains the "maximally comparable" intuition while shifting from within-person conditioning to interventional, ATE-level contrasts. Whereas LN-FE controls can induce post-treatment bias when marriage and parenthood interact, our approach addresses this issue by design, aligning the estimand with our inequality question.

As the gap-closing estimand is explicitly interventional, we use nonparametric marginal

structural models (NPMSMs) (Robins et al. 2000; Pearl 2009) to address how the between-group wage disparities would change under a specified fix. NPMSMs operationalize counterfactual policies using observational panels by reweighting observed histories; hence, this framework enables the implementation of controlled disparities central to our research design and delivers population-level ATE contrasts that directly address inequality. We will go over the details of the gap-closing estimand in the Analytical Strategy section.

C. Mediation Effects of Labor Market Participation

Previously, LN-FE analyses have argued that differences in labor market participation mediate much of the observed wage "premiums" and "penalties" associated with parenthood and marriage for men and women (Gough and Noonan 2013; Grimshaw and Rubery 2015; Killewald and Gough 2013a). Relying on human capital (Mincer and Polachek 1974; Waldfogel 1997), household specialization (Killewald 2013; Hodges and Budig 2010), and labor market discrimination frameworks (Musick et al. 2020; Correll et al. 2007; Rivera 2017), this literature links the reallocation of time between domestic work and market work following marriage and childbirth to subsequent wage changes.

Human capital theory (Becker 1981) tries to explain the divergent effects of **parenthood** on wages for men and women. According to the human capital theory, women experience the motherhood penalty after giving birth to a child because childbearing interrupts their work time, thereby reducing their human capital in the labor market (Anderson et al. 2002: 2003; Mincer and Polachek 1974; Waldfogel 1997; Davies and Pierre 2005; Meurs et al. 2010). Angelov et al. (2016) report a sharp decline in women's wages soon after childbearing. Meanwhile, several studies have noted that after returning to the labor market, women's wages often fail to return to their pre-pregnancy levels (Angelov et al. 2016; Musick et al.

2020). The long-term motherhood penalty suggests that the effect of the interruption of labor market participation is not limited to the time of childbearing; instead, women increasingly withdraw from the labor market over time after having a child (Hakim 2002; Munasinghe et al. 2008). Women would rather switch to less productive, more time-flexible, and more "mother-friendly" jobs to satisfy their childcare needs (Petersen and Morgan 1995; Budig and England 2001; Korenman and Neumark 1992; Staff and Mortimer 2012). For men, since childbearing does not interrupt their time in the labor market, becoming a father does not penalize their wages. On the contrary, due to the sacrifices made by mothers, in married couple families, fathers can spend more time in the labor market, leading to a fatherhood premium (Glauber 2007; Hersch and Stratton 2000; Lundberg and Rose 2000: 2002; Killewald 2013).

Specialization theory discusses why **marriage** may affect the wages of men and women differently. For a married couple, specialization in labor division is considered a "rational choice" to maximize the total income for the family. The traditional gender role model stereotypes women with better housework abilities and men with better performance in the labor market; hence, married women invest and specialize in home production, while married men put their effort into labor market activities (Gupta 1999; Hersch and Stratton 2000). Since these labor division arrangements could only be feasible for married couples while unmarried men and women could not specialize, married men and women may have different labor market outcomes than their unmarried counterparts (Killewald and Gough 2013a). Due to their specialization in the labor market, married men will have a wage premium from marriage over unmarried men (Glauber 2008; Chun and Lee 2001; Hodges and Budig 2010). On the contrary, due to their specialization in housework and less time in the labor market, married women have a marital wage penalty compared to unmarried women

(Waldfogel 1997; Noonan 2001).

Finally, a complementary line of evidence suggests that hiring, promotion, and pay are unequally treated based on marital and parental status (Correll et al. 2007; Rivera 2017; Duguet et al. 2005; Petit 2007; Pedulla 2016). Discrimination is usually against women and sets the invisible “glass ceilings" for their career choices. Women always suffer wage penalties for motherhood and marriage across different firms (Yu and Hara 2021), sectors (Luhr 2020), and occupational characteristics (Yu and Kuo 2017). In more market-oriented sectors, women face greater penalties, while men tend to receive more premiums. Moreover, women are discriminated against not only by their current marriage and parenthood statuses but also by their potential time cost of childcare, even for those who are unmarried or without children (Duguet et al. 2005; Petit 2007).

Due to the impact of marriage and parenthood on labor market participation and wages, individuals may deliberately postpone marriage (Glick and Landau 1950) and childbearing Goisis and Sigle-Rushton (2014); Testa and Toulemon (2006); Loughran and Zissimopoulos (2009) to balance their career and family better. Indeed, individuals may make their marriage and childbearing decisions with consideration of the consequences for future labor market participation and wages, and the information from their previous labor market experience and wages. Thus, the impact of marriage and childbearing on wages is not only about whether the decision is made, but also when/at what age the decision is made. Motivated by this evidence, we extend mediation analyses of the wage effects of marriage/parenthood from a static to a dynamic perspective, allowing for carryover and feedback between family status, labor-market participation, and wages.

D. From Static Models to Dynamic Models

In linear fixed-effects (LN-FE) models, researchers identify the causal relationship between marriage/parenthood and wage and the mediation effect of labor market participation without differentiating the heterogeneity in the treatment effects among marital/parenthood ages. Furthermore, the LN-FE models assume that the unobserved covariates on the individual level will not bias the estimation (Arkhangelsky and Imbens 2018; Wooldridge 2005). However, as mentioned above, we presume that the causal identification assumption needs to be revised because the relationships among the variables are hardly time-independent. First, marriage and parenthood decisions at a specific time (the treatment variables) are affected by the previous labor market participation status (the mediator) and previous wage status (the outcome variable). In addition, the current marital and parenthood status may affect future labor market participation and wages. If the treatment and the mediator are affected by the mediator and the outcome from the previous rounds, we call the effects in the dynamic models the “feedback effects"; oppositely, if the earlier treatment and mediators affect the mediator and outcomes in the later rounds, we call the dynamic model contains the “carryover effects" (Kim and Imai 2019; Zhou and Wodtke 2020).

From the theoretical perspective, we believe that the dynamic models could more appropriately describe the formation of premiums and penalties separately from marriage and parenthood for men and women (Oppenheimer 1997; Killewald and Gough 2013b; Juhn and McCue 2017). Consider the specialization theory in a dynamic process (see Figure 5.1). Suppose a married couple with children makes a specialization arrangement (the wife tends to devote more time to domestic work while the husband puts more time into the labor market). As a result of the differing time commitments to the labor market, the wife tends

to reduce her time on the labor market compared with an unmarried, childless woman, while the husband tends to participate more in the labor market as he needs to make up for the wife's loss in the labor market. Due to the time difference devoted to the labor market, the husband is more likely to earn a higher position in the labor market (compared to an unmarried, childless man), while the wife is less likely to do so. Consequently, the husband receives the wage premium (compared to the unmarried man), while the wife experiences the penalty. The wage changes further stimulate the wife to devote more time to domestic work and push the husband to invest more in the labor market, reinforcing a specialization cycle and its effects.

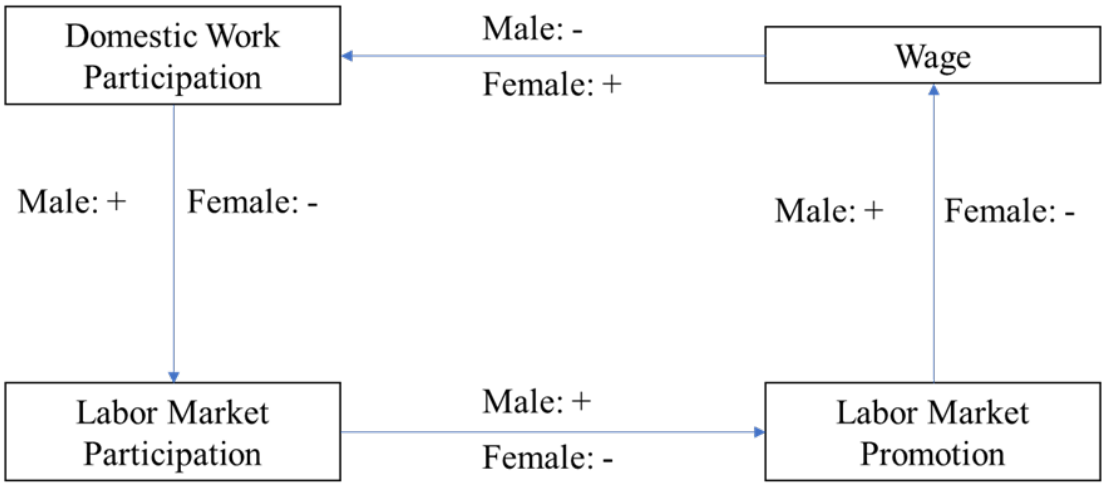


Figure 5.1: Dynamic Process of Specialization in Domestic and Labor Market Work
Note: + and – are compared with individuals with no children or unmarried.

Consistent with our gap-closing analysis, we continue to use NPMSMs to estimate both static and dynamic contrasts with observational panel data. NPMSMs explicitly accommodate the carryover and feedback effects. Identification rests on sequential unconfoundedness and positivity given rich observed histories; in practice, we therefore construct high-

dimensional, time-indexed covariate sets, monitor overlap and weight stability, and use debiased/double machine learning (Chernozhukov et al. 2018a; Semenova and Chernozhukov 2020) for nuisance estimation to mitigate regularization bias and support valid inference. This strategy aligns the estimand with our population-level inequality question and addresses the dynamic confounding that static LN-FE designs leave unresolved.

III. Analytical Strategy

In the previous sections, we have made it clear that we will apply the NPMSMs to analyze the population-level ATE for parenthood and marriage effects, and the mediation effects of labor market participation on wages. In this section, we focus on the detailed analytical strategies with these NPMSMs settings. Specifically, we will demonstrate the multivariate causal mediation relationships with the directed acyclic graphs (DAG) to help the readers understand our presumed relationships among the treatment, mediation, and outcome variables (Pearl 1995: 2009; VanderWeele 2015; Kim and Imai 2019).

A. Static Models

A.1 Gap-Closing Models for the Treatment Effects

We use a gap-closing estimand to ask how between-group wage gaps would change under explicit interventions on a manipulable factor (Jackson and VanderWeele 2018; Lundberg 2024). Figure 5.2a denotes the DAG for this model. In Figure 5.2a, G represents the groups, A represents the variables to be fixed (the manipulable factor), and Y denotes the outcome variable (the divergence). For legibility, measured confounders C are omitted in the diagram, but identification requires consistency, positivity, and no unmeasured confounding

for the relevant treatment–outcome and group–outcome relations.

We estimate two complementary contrasts within each gender in this paper. First, we analyze the gap-closing effects of marriage. As previously noted, we set parity A to a level a for everyone and compare the expected wage of married vs. unmarried. This gives the disparity between marriage groups if everyone had the same parity a. Figure 5.2b illustrates the process. Second, we analyze the parenthood effect. In this study, we set marital status A to a for everyone and compare the expected wage across parity levels $g = (0, 1, 2, 3+)$. The process is illustrated in Figure 5.2c.

We use nonparametric marginal structural models with doubly robust, cross-fitted learners, implemented within gender strata and using out-of-fold predictions. For any policy "set $A = a$," the target causal mean is

$$\mu(a, g) := E\big[\, Y(a) \mid G = g \,\big],$$

with estimator $\hat{\mu}(a, g)$. For binary G (the marital gap), the post-intervention gap is

$$\Delta(a) := \mu(a, g{=}1) - \mu(a, g{=}0), \quad \widehat{\Delta}(a) := \hat{\mu}(a, 1) - \hat{\mu}(a, 0).$$

When G has multiple categories (the parity gap), we report $\hat{\mu}(a, g)$ for each g and prespecified contrasts $\widehat{\Delta}(a; g, g') = \hat{\mu}(a, g) - \hat{\mu}(a, g')$: specifically, the contrasts between zero and one child, between one and two children, and between two and three or more children.

To obtain $\widehat{\mu}(.)$ from observational data, we first fit the nuisance functions:

$$\mu_a^g(c) := E\big[\, Y \mid A = a, C = c, G = g \,\big], \qquad \pi_a^g(c) := P\big(A = a \mid C = c, G = g\big).$$

We use cross-fitting techniques to fit $\widehat{\mu}_a^g(\cdot)$ and $\widehat{\pi}_a^g(\cdot)$. Specifically, we partition the data into K folds, stratified by G and A. For each fold k, fit $\mu_a^g(\cdot)$ and $\pi_a^g(\cdot)$ on the $K-1$ comple-

mentary folds and predict $\widehat{\mu}_a^g(C_i)$ and $\widehat{\pi}_a^g(C_i)$ on held-out units i in fold k. After cycling, every observation has out-of-fold predictions. We then have the doubly robust estimator for the causal mean $\mu(a,g)$ as:

$$\widehat{\mu}(a,g) = \frac{1}{n_g} \sum_{i:G_i=g} \left\{ \widehat{\mu}_a^g(C_i) + \frac{\mathbb{1}(A_i = a)}{\widehat{\pi}_a^g(C_i)} \big(Y_i - \widehat{\mu}_a^g(C_i)\big) \right\}.$$

where $n_g = \sum_{i=1}^n \mathbb{1}(G_i = g)$. The estimator $\widehat{\mu}(a,g)$ is consistent if either $\mu_a^g(\cdot)$ is correctly specified or $\pi_a^g(\cdot)$ is correctly specified and hence it is a doubly robust one. The process illustrated here is exactly the semiparametric modeling we described in Chapter 4.

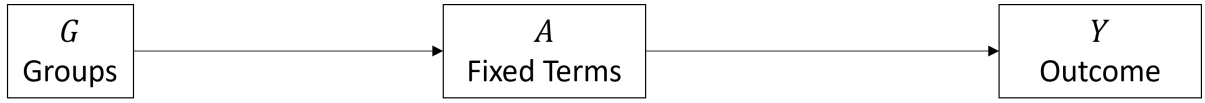
A.2 Mediation Model Among Treatment, Mediation, and Outcome Variables

We then turn to the causal mediation models in which labor market participation is seen as the mediator. Instead of asking how much a pre-intervention gap would change, we examine how the intervention's effect is transmitted. We decompose the total (treatment) effect into a direct component not mediated by labor-market participation and an indirect component that is mediated by it.

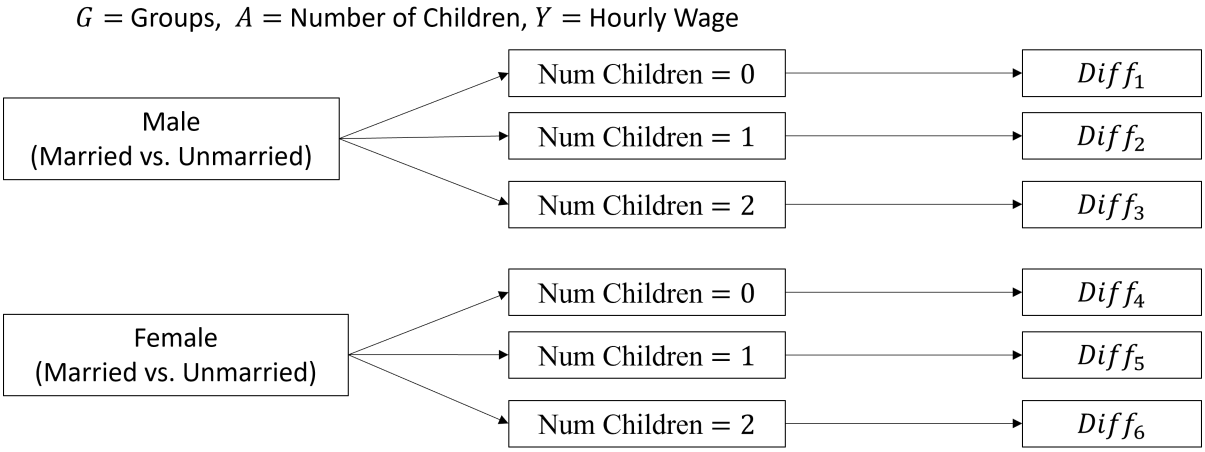
In the static models, we assume a cross-world mediation setup for both the marriage and parenthood analyses, with labor market participation M as the mediator and A as the intervention (treatment). The DAG for the static models is illustrated in Figure 5.3. When we set up the cross-world scenario, we can identify the natural direct and indirect effects (NDE and NIE), with the following assumptions (VanderWeele 2015):

Assumption 5.III.1 (Assumptions to Identify the Natural Direct and Indirect Effects) •

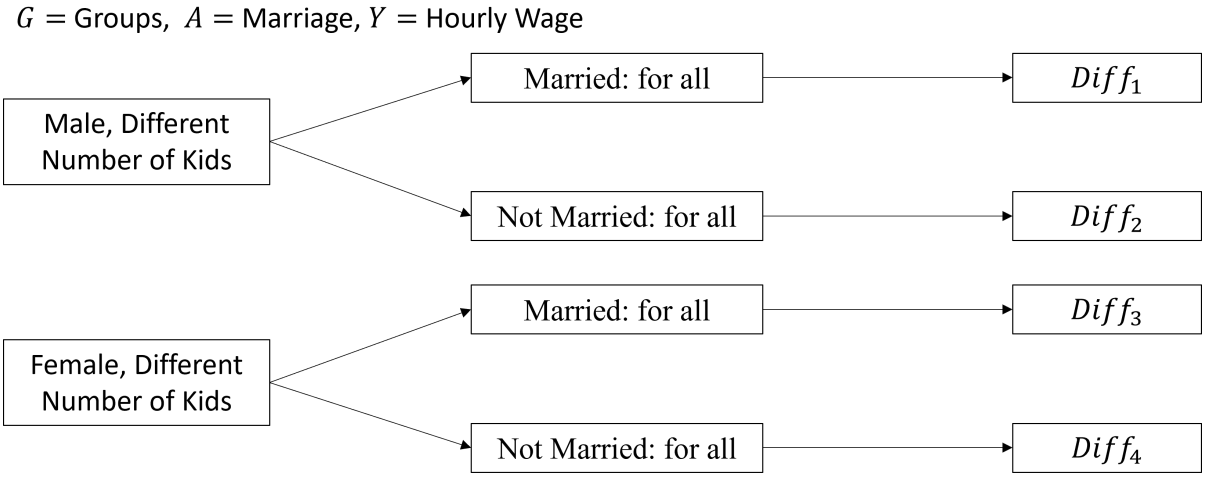
Consistency for the mediator and outcome: if $A = a$ and $M = m$, then $Y = Y(a,m)$ and $M = M(a)$.



(a) General Framework



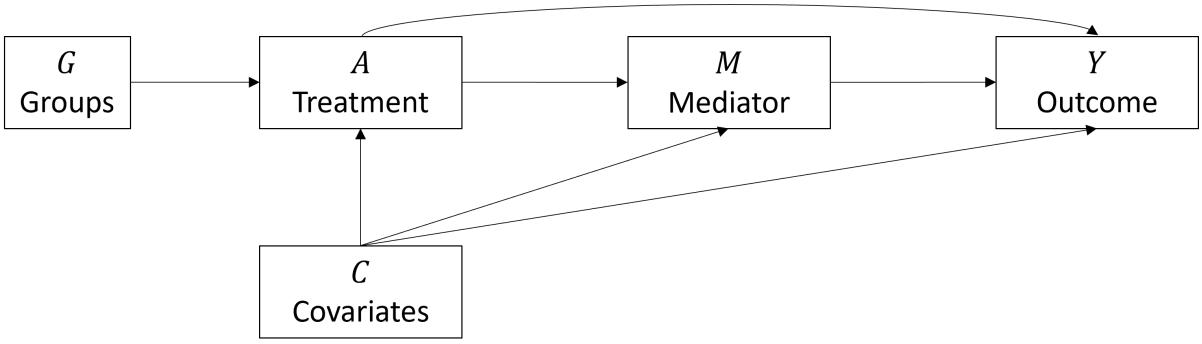
(b) Model for Marital Differences (Fixing Parenthood)



(c) Model for Parenthood Differences (Fixing Marital Status)

Figure 5.2: Directed Acyclic Graph for Gap-Closing Static Model between Treatment and Outcome (Confounders Omitted)

- *Positivity for the treatment and mediator:* $0 < P(A = a \mid C) < 1$ *and* $0 < P(M = m \mid A = a, C) < 1$ *for all relevant* (a, m, C).



(a) General Framework

Figure 5.3: Directed Acyclic Graph for Static Mediation Model

- *No unmeasured confounding for the treatment and mediator:*

$$\{Y(a,m), M(a)\} \perp\!\!\!\perp A \mid C \quad \textit{and} \quad Y(a,m) \perp\!\!\!\perp M \mid A = a, C.$$

- *Cross-world implication: under consistency, positivity, and unconfoundedness, the cross-world counterfactual $Y\big(a, M(a')\big)$ is well-defined and identified from the observed data.*

When we decompose the mediation of labor-market participation on marriage effect, as the treatment (marital status) is binary (married vs. unmarried), the total average treatment effect (TATE) from the treatment to the control decomposes into:

$$\begin{aligned}
\text{TATE}(a,a') &= E\{Y(a)\} - E\{Y(a')\} \\
&= \underbrace{\Big(E\{Y(a,M(a))\} - E\{Y(a,M(a'))\}\Big)}_{\text{NIE}(a,a')\text{: change in } Y \text{ due to } A\text{'s effect on } M} \\
&\quad + \underbrace{\Big(E\{Y(a,M(a'))\} - E\{Y(a',M(a'))\}\Big)}_{\text{NDE}(a,a')\text{: change in } Y \text{ not through } M}.
\end{aligned}$$

The NIE is the part of the marriage effect that works through participation: we hold marital status conceptually fixed and then change only participation from the level it would take if a person were unmarried to the level it would take if the same person were married. Hence,

the change in expected wage is the proportion transmitted along the path "marriage → participation → wages." The NDE is the part that does not operate through participation: we hold participation fixed at the level it would take if the person were unmarried, and then switch the marital status from unmarried to married; the change in expected wages captures pathways other than participation. The total effect of marriage on wages equals the sum of these two components.

For parity effects, since the treatment has multiple levels, we can calculate the pairwise natural effects using the same cross-world formulas. Namely, $A = \{0, 1, 2, 3+\}$, and for any two levels $a, a' \in A$, we have:

$$\begin{aligned}
\text{NDE}(a, a') &= E\{Y(a, M(a'))\} - E\{Y(a', M(a'))\}, \\
\text{NIE}(a, a') &= E\{Y(a, M(a))\} - E\{Y(a, M(a'))\}, \\
\text{TATE}(a, a') &= E\{Y(a)\} - E\{Y(a')\} \;=\; \text{NDE}(a, a') + \text{NIE}(a, a').
\end{aligned}$$

In our analysis, we report the mediation decomposition results for the adjacent/marginal contrasts: we report NDE$(m{+}1, m)$, NIE$(m{+}1, m)$, and TATE$(m{+}1, m)$ for $m \in \{0, 1, 2\}$.

Given the expressions of the NDE and the NIE, we could yield their DR estimator. As we elaborated in Chapter 4, let

$$\theta_{a,t} := E_C\left[\int \mu_a(m, C)\, f_t(m \mid C)\, dm\right],$$

where

$$\mu_a(m, C) := E[Y \mid A = a,\, M = m,\, C],\;\; f_t(m \mid C) := f_{M|A=t,C}(m \mid C).$$

Suppose we have the nuisance functions:

$$\mu_a(m, C) := E[Y \mid A = a, M = m, C], \qquad \pi_a(C) := P(A = a \mid C), \qquad f_a(m \mid C) := f_{M|A=a,C}(m \mid C).$$

Define the conditional density ratio:

$$r_{t,a}(m,C) := \frac{f_t(m \mid C)}{f_a(m \mid C)}, \qquad \eta_{a,t}(C) := \int \mu_a(m,C)\, f_t(m \mid C)\, dm.$$

Similar to what we derived in Chapter 4, with cross-fitted $\widehat{\mu}_a$, $\widehat{\pi}_a$, $\widehat{f}_a$ (and $\widehat{r}_{t,a}$), and $\widehat{\eta}_{a,t}$, the DR estimator for $\theta_{a,t}$ is (notice both a and t are the possible values for the treatment):

$$\widehat{\theta}^{\mathrm{DR}}_{a,t} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\eta}_{a,t}(C_i) + \frac{\mathbb{1}(A_i = a)}{\widehat{\pi}_a(C_i)}\widehat{r}_{t,a}(M_i,C_i)\left\{Y_i - \widehat{\mu}_a(M_i,C_i)\right\} + \frac{\mathbb{1}(A_i = t)}{\widehat{\pi}_t(C_i)}\left\{\widehat{\mu}_a(M_i,C_i) - \widehat{\eta}_{a,t}(C_i)\right\}\right].$$

The DR estimator is very similar to what we derived in Chapter 4, with the only difference here being that we have a continuous mediator and a possible multinomial treatment (in the case of parenthood effect; in the case of marital effect, the treatment is still dichotomous). Like what we introduced in Chapter 4,the first part of the estimator is the plug-in estimator, the second part is the outcome-residual correction, and the third part is the transport correction in the $(A = t)$ world. The DR estimators for the NDE and the NIE, are:

$$\widehat{\mathrm{NDE}}_{\mathrm{DR}} = \widehat{\theta}^{\mathrm{DR}}_{a,a'} - \widehat{\theta}^{\mathrm{DR}}_{a',a'},$$
$$\widehat{\mathrm{NIE}}_{\mathrm{DR}} = \widehat{\theta}^{\mathrm{DR}}_{a,a} - \widehat{\theta}^{\mathrm{DR}}_{a,a'}.$$

This estimator is doubly robust: if either (i) μ_a and $\eta_{a,t}$ are correctly specified (e.g., through correct μ_a and f_t), or (ii) π_a, π_t and the density ratio $r_{t,a}$ (equivalently f_a and f_t) are correctly specified.

For the multi-valued treatment case of the parenthood effect, we set $\pi_a(C)$ as a multinomial distribution, and let a, t be the two adjacent parity contrasts.

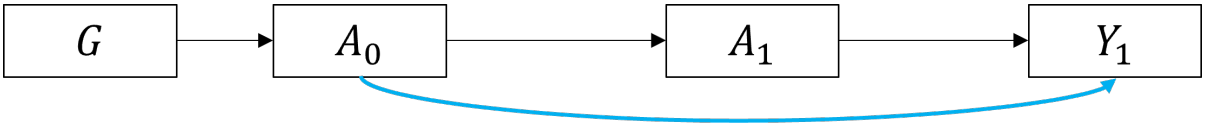
B. Dynamic Models

We then turn to the dynamic models to analyze the causal relations first. Our dynamic models assume that the causal and mediation effects are heterogeneous at different marital and

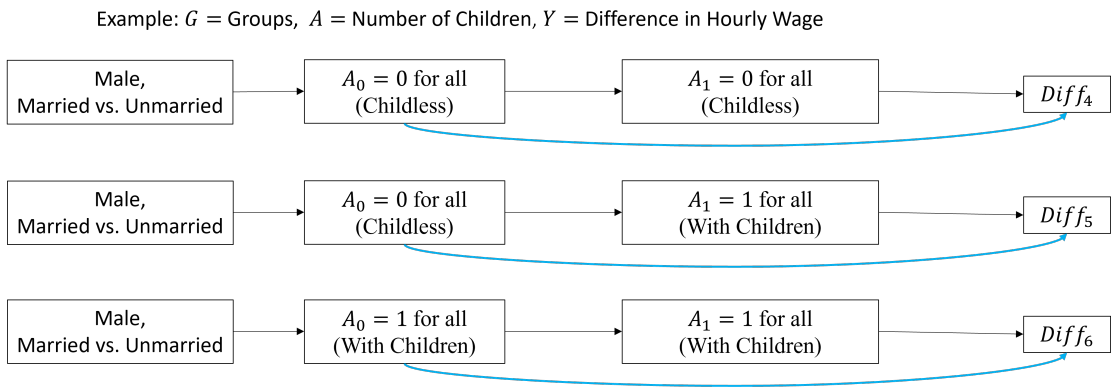
childbearing ages. Therefore, we set the cutoff age at 21, 25, 29, and 33 and established the models separately for $22-25, 26-29$, and $30-33$ groups (So for the 22-25 age group model, their $t-1$ statuses refer to their statuses during age 18-21). Let $\overline{A_t} = (A_1, A_2, \ldots, A_t)$ and $\overline{M_t} = (M_1, M_2, \ldots, M_t)$ denote the history of treatment and mediator up through time t. To make our analysis more convenient, we make a semi-Markov assumption that only the latest statuses impact the statuses at the next time point. As mentioned above, due to the advantages of the NPMSM, we adopt this method to construct our dynamic models. Due to the difficulty in interpreting coefficients, if we still set up multinomial adjacent contrasts for parenthood effects, in the dynamic models, we regard both marital effects and parenthood effects as binary.

B.1 Gap-Closing Models

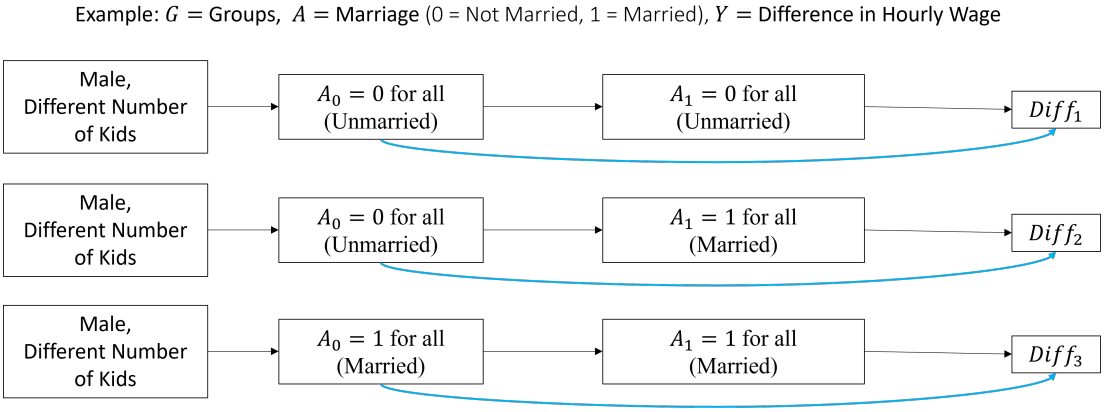
We start with the gap-closing models. Similar to the gap-closing models for the static model, we fix the effect of parenthood, explore the disparities between different marital groups, and fix the effect of marriage to show the divergence in parenthood. However, in the dynamic models, we allow the variables to be fixed across two rounds. The general idea of the gap-closing estimand is presented in Figure 5.4a, in which G denotes the groups, A is the variable to be fixed, and Y is the between-group disparity (the outcome). As can be seen from Figure 5.4a, we use the blue line to denote the carryover effect, allowing the value of the fixed term in the previous round t_0 to have an effect on our outcome in t_1 (the time-invariant covariates and time-varying covariates in both t_0 and t_1 were omitted from the DAG as a visual illustration). Likewise, we use Figure 5.4b and Figure 5.4c to illustrate the dynamic gap-closing estimand in our specific scenarios for men (for women, the analytical strategy is exactly the same). As shown in Figure 5.4b, we first group men in the specific age range



(a) General Framework



(b) Dynamic Model for Marital Differences (Fixing Parenthood)



(c) Both Carryover and Feedback Effects (Full Model)

Figure 5.4: Directed Acyclic Graph for Gap-Closing Dynamic Model between Treatment and Outcome (Confounders Omitted)

(18-21, 22-25, 26-29, and 30-33) altogether, fitting a double machine learning (DML) model for their wages with the time-invariant variables, time-varying variables at both t_0 and t_1, and most importantly, the parenthood status at t_1 and t_0. With the fitted model, we then manipulate the parenthood status for the three scenarios: in the first line of Figure 5.4b, we set at t_0 and t_1 both as childless for men, which gives the difference in wages between

married and unmarried men under the circumstance of no children. In the second line, we manipulate the parenthood status, setting t_0 as childless and t_1 as having children, estimating the wage gap between married and unmarried men shortly after becoming a father (the instant effect). The third line shows the gap where men are parents at both t_0 and t_1, and hence, compared to the settings in the first line, this scenario shows the wage gap between married and unmarried when the individual has been a parent for at least four years.

Similarly, in Figure 5.4c, we aim to capture the gap between individuals who are childless and those who have children, fixing the marital status. In Figure 5.4c, the first line indicates we fix the marriage status as not married (neither married four years ago nor married at the observation time), and the second line denotes the newly married (just married in the past four years), and finally, in the third line, we have the long-time married (married at least for four years).

Since only two waves are taken into our consideration, we can define the group-specific post-intervention mean at t_1 as:

$$\psi_g(\bar{a}) := E\big[\, Y_1\big(A_0 = a_0, A_1 = a_1\big) \,\big|\, G = g \,\big],$$

Hence, the dynamic gap-closing estimand (target mean gap) is the difference

$$\Delta(\bar{a}) := \psi_{g_1}(\bar{a}) - \psi_{g_0}(\bar{a}),$$

for any two comparison groups g_1, g_0 (e.g., married vs. unmarried, or childless vs. with children). As we denoted in Figures 5.4b and 5.4c, we have three typical value sets for $\overline{a}$: (0,0): fixed “no" (not married/no children) at both waves; (0,1): switch on between t_0 and t_1; and (1,1): fixed “yes" (married; with children) at both waves. Suppose:

$$H_{0i} := (C_{0i}, G_i), \qquad H_{1i} := (C_{0i}, G_i, A_{0i}, C_{1i}),$$

where C_0 are baseline confounders and C_1 are time-varying confounders measured before A_1 and Y_1. Under the following identification assumptions:

Assumption 5.III.2 (Identification Assumptions for Dynamic Gap-Closing Models) *The following assumptions are required to identify the gap-closing estimand in the dynamic models:*

1. *Consistency: if* $A_{0i} = a_0$ *and* $A_{1i} = a_1$, $Y_{1i} = Y_{1i}(A_0 = a_0, A_1 = a_1)$*; if* $A_{0i} = a_0$, $C_{1i} = C_{1i}(A_0 = a_0)$.

2. *Positivity:* $\pi_0(a_0 \mid H_0) > 0$ *and* $\pi_1(a_1 \mid H_1) > 0$ *on the relevant support.*

3. *Sequential ignorability (unconfoundedness):*

$$Y_1\big(A_0 = a_0, A_1 = a_1\big) \perp\!\!\!\perp A_1 \,\big|\, H_1, \qquad Y_1\big(A_0 = a_0, A_1 = a_1\big) \perp\!\!\!\perp A_0 \,\big|\, H_0,$$

We can set

$$\psi_g(\bar{a}) \;=\; E\big[\, \mu_0^{\bar{a}}(H_0) \,\big|\, G = g\big],$$

where

$$\mu_0^{\bar{a}}(H_0) \;:=\; E\big[\, \mu_1^{a_1}(H_1) \big|\, A_0 = a_0,\; H_0\big], \quad \mu_1^{a_1}(H_1) \;:=\; E\big[\, Y_1 \,\big|\, A_1 = a_1,\; H_1\big].$$

We can have the nuisance functions for the treatment models

$$\pi_0(a_0 \mid H_0) \;=\; P(A_0 = a_0 \mid H_0), \qquad \pi_1(a_1 \mid H_1) \;=\; P(A_1 = a_1 \mid H_1),$$

and the outcome regressions

$$\mu_1^{a_1}(H_1) = E[Y_1 \mid A_1 = a_1, H_1] \qquad \mu_0^{\bar{a}}(H_0) = E[\mu_1^{a_1}(H_1) \mid A_0 = a_0, H_0]$$

for $t \in \{0, 1\}$. With cross-fitted estimates $\widehat{\pi}_0, \widehat{\pi}_1, \widehat{\mu}_1^{a_1}, \widehat{\mu}_0^{\bar{a}}$, we have:

$$\widehat{\psi}_g^{\mathrm{DR}}(\bar{a}) = \frac{1}{n_g} \sum_{i=1}^{n} \mathbb{1}(G_i = g) \left[\widehat{\mu}_0^{\bar{a}}(H_{0i}) + \frac{\mathbb{1}(A_{0i} = a_0)}{\widehat{\pi}_0(a_0 \mid H_{0i})} \left\{ \widehat{\mu}_1^{a_1}(H_{1i}) - \widehat{\mu}_0^{\bar{a}}(H_{0i}) + \frac{\mathbb{1}(A_{1i} = a_1)}{\widehat{\pi}_1(a_1 \mid H_{1i})} \big(Y_{1i} - \widehat{\mu}_1^{a_1}(H_{1i})\big) \right\} \right].$$

The estimator is consistent if either (i) both outcome regressions $\mu_1^{a_1}, \mu_0^{\bar{a}}$ are correctly specified, or (ii) both treatment models π_0, π_1 are correctly specified. Hence, the dynamic gap-closing estimand is:

$$\widehat{\Delta}^{\mathrm{DR}}(\bar{a}) = \widehat{\psi}_{g_1}^{\mathrm{DR}}(\bar{a}) - \widehat{\psi}_{g_0}^{\mathrm{DR}}(\bar{a}).$$

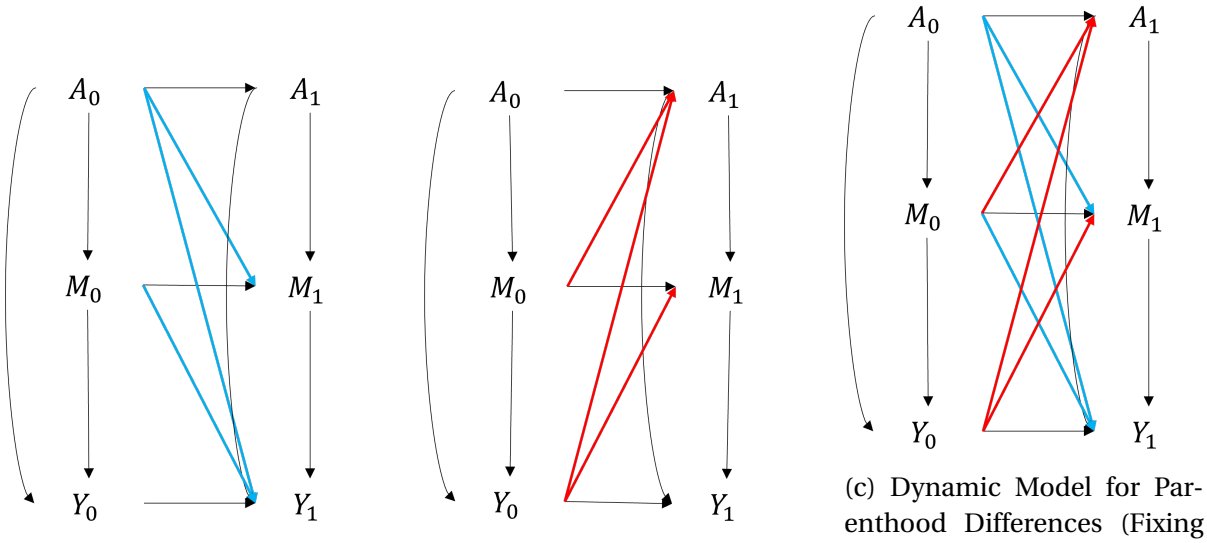
B.2 Dynamic Causal Mediation Models, Carryover, Feedback, and Full Effects

At last, we discuss the dynamic models for the mediation effect of labor market participation on the causal relationship between marriage/parenthood and wage. In this regard, we still decompose the marriage/parenthood effect on wages into the indirect effect, which goes through labor market participation, and the direct effect, which does not go through the mediator. However, in dynamic models, we cannot assume the natural effect, as the post-treatment variables (i.e., outcomes in the same wave, treatments, and mediators in subsequent waves) will be affected by the first-wave treatment. But we can still make assumptions based on the “one-world" scenario under the sequential consistency, positivity, and unconfoundedness assumptions and identify the interventional direct and indirect effects. The interventional direct effects (IDE) capture how marriage/parenthood changes wages, holding the entire time-path distribution of participation fixed to that under a reference path, and it reflects mechanisms not operating through participation. The interventional indirect effects (IIE), on the other hand, attribute wage changes specifically to the pre-specified (hypothetical) intervention on the participation trajectories while keeping treatment exposure fixed.

We set up different dynamic models so that the variables in the two waves will have different paths/mechanisms affecting each other. Namely, we aim to compare which one of the following three scenarios best describes the mediation of labor market participation

on the relationship between marriage/parenthood and wage: 1) carryover-only effects, 2) feedback-only effects, and 3)both carryover and feedback effects. Figure 5.5 presents the Directed Acyclic Graphs (DAGs) for the three scenarios with different assumptions.

In Figure 5.5a, we only consider the carryover-only effects: the treatment at $t-1$ affects the mediator and the outcome at t, and the mediator at $t-1$ affects the outcome at t. In Figure 5.5b, we only consider the feedback-only effects of the previous mediator and the outcome of the current treatment. Finally, Figure 5.5c denotes the most complex circumstance in which both carryover and feedback effects are included (the full model). As we



(a) Only Carryover Effects

(b) Only Feedback Effects

(c) Dynamic Model for Parenthood Differences (Fixing Marital Status)

Figure 5.5: Directed Acyclic Graph for Dynamic Mediation Model

take a semi-Markovian approach, we only consider the interaction between the adjacent two waves in the dynamic model. Hence, in the two-wave models, we consider the baseline variables (A_0, M_0, Y_0, C) and end-of-horizon outcome Y_1. Let $\bar{a} = (a_0, a_1)$ denote the treatment path and $\Gamma_{a_1^\star}$ the mediator policy (intervention) at wave 1 [2]. The mediator variables

[2] $\Gamma(.)$ is the $G(.)$ function denoted in Chapter 4. However, in this chapter, as we denote groups as G, we use

M_0, M_1 are continuous, and therefore, all mediator terms below are densities. Like in the dynamic gap-closing estimands, we denote histories:

$$H_0 := (C, A_0, M_0, Y_0), \qquad H_1 := (H_0, A_1, M_1).$$

For wave-1 treatment propensity, based on the DAGs, we have:

$$\pi_1^{A,\text{co}}(a_1 \mid C, A_0), \qquad \pi_1^{A,\text{fb}}(a_1 \mid C, A_0, M_0, Y_0), \qquad \pi_1^{A,\text{full}}(a_1 \mid C, A_0, M_0, Y_0).$$

separately for the carryover-only, feedback-only, and full models.

When modeling the carryover-only effects, the target parameter is $\psi^{\text{co}}_{\bar{a},\Gamma_{a_1^\star}} := E\left[Y_1^{\bar{a},\Gamma_{a_1^\star}} \right]$, with $\gamma_1^{\text{co}}(m_1 \mid a_1^\star, M_0, C)$ as the wave-1 intervention kernels to define the counterfactual. The identification assumptions include:

Assumption 5.III.3 (Assumptions for Carryover-only Dynamic Causal Mediation Models) *We need the following assumptions to identify* $\psi^{\text{co}}_{\bar{a},\Gamma_{a_1^\star}}$*:*

(i) Consistency/ SUTVA: $Y_1 = Y_1(\bar{A}, \bar{M}), \quad M_1 = M_1(A_1, M_0)$

(ii) Positivity: $\pi_1^{A,\text{co}}(a_1 \mid C, A_0) > 0$, $\pi_1^{M,\text{co}}(m_1 \mid a_1, M_0, C) > 0$ *whenever* $\gamma_1^{\text{co}}(m_1 \mid a_1^\star, M_0, C) > 0$.

(iii) Sequential unconfoundedness/ignorability for A_1*:* $A_1 \perp\!\!\!\perp Y_1^{\bar{a},\Gamma_{a_1^\star}} \mid (C, A_0)$.

(iv) No unmeasured confounding for $M_1 \to Y_1$*:* $M_1 \perp\!\!\!\perp Y_1^{\bar{a},\Gamma_{a_1^\star}} \mid (A_1, M_0, A_0, Y_0, C)$*, and* γ_1^{co} *excludes* Y_0.

We set up a set of nuisance functions:

$$\mu_1^{\text{co}}(a_1, m_1; A_0, M_0, Y_0, C) := E[Y_1 \mid A_1 = a_1,\, M_1 = m_1,\, A_0, M_0, Y_0, C],$$

$\Gamma(.)$ for the intervention function on the mediator and the kernel is denoted as $\gamma(.)$.

$$\pi_1^{A,\mathrm{co}}(a_1 \mid C, A_0) := P(A_1 = a_1 \mid C, A_0),$$

$$\pi_1^{M,\mathrm{co}}(m_1 \mid a_1, M_0, C) := f_{M_1 \mid A_1, M_0, C}(m_1 \mid a_1, M_0, C),$$

$$\gamma_1^{\mathrm{co}}(m_1 \mid a_1^\star, M_0, C),$$

$$Q_0^{\mathrm{co}}(H_0; a_1, a_1^\star) := \int \mu_1^{\mathrm{co}}(a_1, m_1; A_0, M_0, Y_0, C)\, \gamma_1^{\mathrm{co}}(m_1 \mid a_1^\star, M_0, C)\, dm_1.$$

Based on our derivation in Chapter 4, the DR estimator for $\psi^{\mathrm{co}}_{\bar{a},\Gamma_{a_1^\star}}$ is:

$$\widehat{\psi}^{\mathrm{DR,co}}_{\bar{a},\Gamma_{a_1^\star}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\underbrace{\frac{\mathbb{1}(A_{1i} = a_1)}{\widehat{\pi}_1^{A,\mathrm{co}}(a_1 \mid C_i, A_{0i})} \cdot \frac{\widehat{\gamma}_1^{\mathrm{co}}(M_{1i} \mid a_1^\star, M_{0i}, C_i)}{\widehat{\pi}_1^{M,\mathrm{co}}(M_{1i} \mid a_1, M_{0i}, C_i)}}_{\widehat{W}_{1i}^{\mathrm{co}}}$$
$$\Big\{\widehat{\mu}_1^{\mathrm{co}}(a_1, M_{1i}; A_{0i}, M_{0i}, Y_{0i}, C_i) - \widehat{Q}_0^{\mathrm{co}}(H_{0i}; a_1, a_1^\star)\Big\} + \widehat{Q}_0^{\mathrm{co}}(H_{0i}; a_1, a_1^\star)\Bigg].$$

and the IDE, the IIE are:

$$\widehat{\mathrm{IDE}}^{\mathrm{co}}_{\mathrm{DR}} = \widehat{\psi}^{\mathrm{DR,co}}_{(a_0,a_1),\Gamma_{a_1'}} - \widehat{\psi}^{\mathrm{DR,co}}_{(a_0',a_1'),\Gamma_{a_1'}}, \qquad \widehat{\mathrm{IIE}}^{\mathrm{co}}_{\mathrm{DR}} = \widehat{\psi}^{\mathrm{DR,co}}_{(a_0,a_1),\Gamma_{a_1}} - \widehat{\psi}^{\mathrm{DR,co}}_{(a_0,a_1),\Gamma_{a_1'}}.$$

For the target in feedback-only effect models $\psi^{\mathrm{fb}}_{\bar{a},\Gamma_{a_1^\star}} := E\left[Y_1^{\bar{a},\Gamma_{a_1^\star}}\right]$, the interventional mediator kernel is $\gamma_1^{\mathrm{fb}}(m_1 \mid a_1^\star, M_0, Y_0, C)$ (which includes Y_0). Identification assumptions are:

Assumption 5.III.4 (Assumptions for Feedback-only Dynamic Causal Mediation Models) *The following assumptions are needed to identify* $\psi^{\mathrm{fb}}_{\bar{a},\Gamma_{a_1^\star}}$*:*

(i) Consistency/SUTVA: $Y_1 = Y_1(\bar{A}, \bar{M}), \quad M_1 = M_1(A_1, Y_0)$

(ii) Positivity: $\pi_1^{A,\mathrm{fb}}(a_1 \mid C, A_0, M_0, Y_0) > 0$, $\pi_1^{M,\mathrm{fb}}(m_1 \mid a_1, M_0, Y_0, C) > 0$ *whenever* $\gamma_1^{\mathrm{fb}}(m_1 \mid a_1^\star, Y_0, C) > 0$.

(iii) Sequential unconfoundedness/ignorability for A_1*:* $A_1 \perp\!\!\!\perp Y_1^{\bar{a},\Gamma_{a_1^\star}} \mid (C, A_0, M_0, Y_0)$.

(iv) No unmeasured confounding for $M_1 \rightarrow Y_1$*:* $M_1 \perp\!\!\!\perp Y_1^{\bar{a},\Gamma_{a_1^\star}} \mid (A_1, M_0, Y_0, C)$*, and* γ_1^{fb} *excludes* M_0*.*

The nuisance functions are:

$$\mu_1^{\text{fb}}(a_1, m_1; Y_0, C) := E[Y_1 \mid A_1 = a_1,\, M_1 = m_1,\, Y_0, C],$$
$$\pi_1^{A,\text{fb}}(a_1 \mid C, A_0, M_0, Y_0) := P(A_1 = a_1 \mid C, A_0, M_0, Y_0),$$
$$\pi_1^{M,\text{fb}}(m_1 \mid a_1, M_0, Y_0, C) := f_{M_1 \mid A_1, M_0, Y_0, C}(m_1 \mid \cdot),$$
$$\gamma_1^{\text{fb}}(m_1 \mid a_1^\star, M_0, Y_0, C),$$
$$Q_0^{\text{fb}}(H_0; a_1, a_1^\star) := \int \mu_1^{\text{fb}}(a_1, m_1; Y_0, C)\, \gamma_1^{\text{fb}}(m_1 \mid a_1^\star, M_0, Y_0, C)\, dm_1.$$

and the DR estimator is:

$$\widehat{\psi}^{\text{DR,fb}}_{\bar{a},\Gamma_{a_1^\star}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\underbrace{\frac{\mathbb{1}(A_{1i} = a_1)}{\widehat{\pi}_1^{A,\text{fb}}(a_1 \mid C_i, A_{0i}, M_{0i}, Y_{0i})} \cdot \frac{\widehat{\gamma}_1^{\text{fb}}(M_{1i} \mid a_1^\star, M_{0i}, Y_{0i}, C_i)}{\widehat{\pi}_1^{M,\text{fb}}(M_{1i} \mid a_1, M_{0i}, Y_{0i}, C_i)}}_{\widehat{W}_{1i}^{\text{fb}}}$$
$$\Big\{\widehat{\mu}_1^{\text{fb}}(a_1, M_{1i}; Y_{0i}, C_i) - \widehat{Q}_0^{\text{fb}}(H_{0i}; a_1, a_1^\star)\Big\} + \widehat{Q}_0^{\text{fb}}(H_{0i}; a_1, a_1^\star)\Bigg].$$

and the IDE and the IIE are:

$$\widehat{\text{IDE}}^{\text{fb}}_{\text{DR}} = \widehat{\psi}^{\text{DR,fb}}_{(a_0,a_1),\Gamma_{a_1'}} - \widehat{\psi}^{\text{DR,fb}}_{(a_0',a_1'),\Gamma_{a_1'}}, \qquad \widehat{\text{IIE}}^{\text{fb}}_{\text{DR}} = \widehat{\psi}^{\text{DR,fb}}_{(a_0,a_1),\Gamma_{a_1}} - \widehat{\psi}^{\text{DR,fb}}_{(a_0,a_1),\Gamma_{a_1'}}.$$

Finally, we have the full model, which is the combination of the set-ups for the carryover-only effect models and the feedback-only effect models. γ_1^{full} should take the same format as in the feedback-only effect model, and the identification assumptions are the combination of the carryover-only and the feedback-only assumptions:

Assumption 5.III.5 (Assumptions for Full Dynamic Causal Mediation Models) *The following assumptions are needed to identify* $\psi^{\text{full}}_{\bar{a},\Gamma_{a_1^\star}}$*:*

(i) Consistency/SUTVA: $Y_1 = Y_1(\bar{A}, \bar{M}), \quad M_1 = M_1(A_1, M_0, Y_0)$

(ii) Positivity: $\pi_1^{A,\text{full}}(a_1 \mid C, A_0, M_0, Y_0) > 0$, $\pi_1^{M,\text{full}}(m_1 \mid a_1, A_0, M_0, Y_0, C) > 0$ *whenever* $\gamma_1^{\text{full}}(m_1 \mid a_1^\star, A_0, M_0, Y_0, C) > 0$.

(iii) Sequential unconfoundedness/ignorability for A_1*:* $A_1 \perp\!\!\!\perp Y_1^{\bar{a}, \Gamma_{a_1^\star}} \mid (C, A_0, M_0, Y_0)$.

(iv) No unmeasured confounding for $M_1 \to Y_1$*:* $M_1 \perp\!\!\!\perp Y_1^{\bar{a}, \Gamma_{a_1^\star}} \mid (A_1, A_0, M_0, Y_0, C)$.

Nuisance functions are:

$$\mu_1^{\text{full}}(a_1, m_1; A_0, M_0, Y_0, C) := E[Y_1 \mid A_1 = a_1,\, M_1 = m_1,\, A_0, M_0, Y_0, C],$$

$$\pi_1^{A,\text{full}}(a_1 \mid C, A_0, M_0, Y_0) := P(A_1 = a_1 \mid C, A_0, M_0, Y_0),$$

$$\pi_1^{M,\text{full}}(m_1 \mid a_1, A_0, M_0, Y_0, C) := f_{M_1 \mid A_1, A_0, M_0, Y_0, C}(m_1 \mid \cdot),$$

$$\gamma_1^{\text{full}}(m_1 \mid a_1^\star, M_0, Y_0, C),$$

$$Q_0^{\text{full}}(H_0; a_1, a_1^\star) := \int \mu_1^{\text{full}}(a_1, m_1; A_0, M_0, Y_0, C)\, \gamma_1^{\text{full}}(m_1 \mid a_1^\star, M_0, Y_0, C)\, dm_1.$$

And the DR estimator is:

$$\widehat{\psi}^{\text{DR,full}}_{\bar{a}, \Gamma_{a_1^\star}} = \frac{1}{n} \sum_{i=1}^{n} \Bigg[\underbrace{\frac{\mathbb{1}(A_{1i} = a_1)}{\widehat{\pi}_1^{A,\text{full}}(a_1 \mid C_i, A_{0i}, M_{0i}, Y_{0i})} \cdot \frac{\widehat{\gamma}_1^{\text{full}}(M_{1i} \mid a_1^\star, M_{0i}, Y_{0i}, C_i)}{\widehat{\pi}_1^{M,\text{full}}(M_{1i} \mid a_1, A_{0i}, M_{0i}, Y_{0i}, C_i)}}_{\widehat{W}_{1i}^{\text{full}}}$$

$$\Big\{ \widehat{\mu}_1^{\text{full}}(a_1, M_{1i}; A_{0i}, M_{0i}, Y_{0i}, C_i) - \widehat{Q}_0^{\text{full}}(H_{0i}; a_1, a_1^\star) \Big\} + \widehat{Q}_0^{\text{full}}(H_{0i}; a_1, a_1^\star) \Bigg].$$

We have the expressions for the IDE and the IIE:

$$\widehat{\text{IDE}}^{\text{full}}_{\text{DR}} = \widehat{\psi}^{\text{DR,full}}_{(a_0, a_1), \Gamma_{a_1'}} - \widehat{\psi}^{\text{DR,full}}_{(a_0', a_1'), \Gamma_{a_1'}}, \qquad \widehat{\text{IIE}}^{\text{full}}_{\text{DR}} = \widehat{\psi}^{\text{DR,full}}_{(a_0, a_1), \Gamma_{a_1}} - \widehat{\psi}^{\text{DR,full}}_{(a_0, a_1), \Gamma_{a_1'}}.$$

IV. Data and Variables

A. Data

The data source for this study comes from the National Longitudinal Survey of Youth, 1979 (NLSY79). The NLSY79 is a nationally representative longitudinal survey that follows the American youth born between 1957 and 1964. In the base wave of 1979, the study interviewed 12,686 individuals (6,403 men and 6,283 women) when the respondents were 15 to 22 years old. Between 1980 and 1993, the study returned to interview the respondents annually, and since 1994, researchers have conducted the survey biennially. Due to our research interests, we extracted the respondents' information from the ages of 16 to 35. For our analysis, except for the dynamic model, we reshaped the data to long-form (one record represents one individual in one year). We then drop missing values in any of the variables in our analysis. We have 4,538 male respondents with 57,201 records and 4,858 female respondents with 59,078 records, with the data restriction process. As many previous papers also adopt the NLSY79 data (Budig and England 2001; Killewald and Gough 2013b; Killewald and Lundberg 2017; Ludwig and Brüderl 2018; Cheng 2016), after data cleaning, we compare the descriptive statistics with the ones reported in the previous research and find there are few differences between ours and the previously reported results.

The variables in our analysis include the outcome variable, the treatment variables, the mediation variables, the time-constant covariates, and the time-varying covariates. The time-constant variables have the same value for a person throughout the period, while the time-varying variables are different at different time points. However, we do not differentiate the two types of covariates in the static models.

Outcome Variable – The outcome variable in this research is the rank of the respondent's hourly wage from his/her primary job. Previous research often uses the logarithm of the absolute wage. The main difference between the absolute measurement (log wage) and the relative measurement (wage rank) is the way to deal with the zero values – for log wage, researchers always add a small amount (for instance, add one) so that the individuals with zero wage would not be excluded during the log transformation. However, the operation might be quite different for different researchers— for instance, different researchers might choose different "small amounts" to add on, and some might first calculate the CPI inflation and then add the small amount and take the log transformation, while others might first log-transform the data and then take the CPI inflation into account. Therefore, it might not be a huge problem if there are not many zero values in the dataset. However, as our research has to include many individuals who only take part-time or freelance jobs after marriage and childbearing and even devote all their time to domestic work, we could not simply drop the zero values, and such insufficiency might lead to a biased estimation of the causal effect. If we take wage rank as the outcome, the causal effect we estimate tells us how much an individual's marriage/parenthood affects his/her position in the wage distribution of the population [3], and we would not misspecify the zero values.

Treatment variables — we have two treatment variables: marital status and parenthood status. We only consider whether the specific individual is married or not at the specific year for marital status. The NLSY79 provides the synthesis variable of one's marital status at the specific year, and we also mark the individual as married if one of his/her marriages

[3]Using wage rank as the dependent variable is compatible with SUTVA under the measurement convention used here: ranks are calculated relative to the observed empirical wage distribution in each year, and the hypothetical intervention changes an individual's potential wage rank relative to that fixed reference distribution. If ranks were instead recomputed after a population-wide intervention that changed the whole wage distribution, this would require an additional equilibrium or interference assumption. The same convention applies to the working-hours-rank mediator. Similar discussions could be seen in Lundberg (2024).

started before that year and ended after it. For parenthood status, we measure how many coresident children an individual has in a specific year. To clean the data, we turn to the household records and calculate the number of children. Following the operation in the previous research, in the static models, we treat the parenthood status as an ordered variable with four levels: zero children, one child, two children, and three or more children; while in the dynamic models, as there will be too many results to present, we only consider the comparison between the with children and childless groups [4].

Mediation Variable – Our mediator variable is the rank of total hours worked for the individual. The number of total hours devoted to the job could symbolize whether the respondent takes a full-time or part-time job and how he/she balance the time devoted to domestic work and work in the market. The ranks of the working hours are more skewed in distribution compared with the original data, and thus we measure the time in the labor market on a 1 to 100 scale.

Covariates— Covariates include the time-varying covariates and time-constant covariates. The time-varying covariates include additional information concerning labor market status and one's educational status at a particular year. We have the total tenure weeks in the primary job, potential years of work experience, and the primary job sector (private vs. public) as the additional variables to measure labor market status. Moreover, we include a dummy of whether the individual was a full-time student and dummies (not graduate from high school, high school graduation, college non-graduates, college graduates and post-college graduates) for the highest level of education ever achieved at the specific year as

[4]The NLSY79 also provides the childbirth year information for us to calculate the number of children: we could compare the childbirth year with the current year and calculate the number. The distributions with these two measurements are pretty analogous. However, considering coresidence might be the main effect of domestic labor devotion, we still adopt the measurement from the household record.

the indicators for educational status. On the other side, demographic variables, parental status, cognitive and self-control abilities scales, and several expectation variables of the individuals are set as the time-constant covariates in our analysis. The demographic variables include race (dummies for black and Hispanic), number of siblings, and region (a dummy for south vs. north and a dummy for urban vs. rural). Variables on the respondents' parents include whether the father was absent and whether the mother was absent in childhood, parental earnings rank when the respondent was 18-20 years old, and the highest education grade completed by parents (mainly the mother's education completion, if the mother's information is missing, we impute with the father's). The variable measuring cognitive ability is the AFQT score measured in the 1981 wave and the Rotter score (the ability to control one's own life) measured in the 1979 wave. Finally, the self-expectation on the highest grade to complete measured in the wave 1979 and the respondent's self-expectation on the number of children in the future measured in the 1979 wave are the expectation variables included in the analysis.

Table 5.1 summarizes the descriptive statistics.

V. Results

A. Results for Static Models

A.1 Results for Static Gap-Closing Estimand

We first present the results from the static models. Table 5.2 shows the results for the gap-closing estimand on marriage by comparing the difference between the married and not-married stratified by the number of children (illustrated in Figure 5.2b). Thus, for each gender, we compare the difference between the upper and lower rows. As can be inferred from the table, for men with zero, one, two, and three and more children, marriage provides them

Table 5.1: Descriptive Table for the Variables (NLSY79)

Variable	Men		Women	
	Mean	Standard Deviation	Mean	Standard Deviation
Person				
Ever Married	0.79	0.15	0.85	0.20
Age at first marriage	25.40	5.34	23.70	5.29
Ever Parent	0.72	0.14	0.78	0.15
Age at entry to parenthood	27.12	6.05	25.22	5.18
Person-Year				
Dependent Variable				
Rank of hourly pay	49.41	22.92	39.72	23.12
Mediation Variable				
Rank of hours worked	61.72	25.79	49.30	27.18
Exposure Variable				
Married	0.46	0.50	0.40	0.49
Number of children	1.58	0.89	1.74	0.95
Number of children: zero	0.64	0.48	0.55	0.50
Number of children: one	0.18	0.38	0.22	0.42
Number of children: two	0.13	0.34	0.16	0.37
Number of children: three and more	0.05	0.22	0.06	0.24
Time-Varying Covariates				
Tenure weeks	147.28	156.71	136.46	147.47
potential years of experience	6.51	4.11	6.11	4.03
Private sectors	0.84	0.36	0.82	0.39
Highest education grade completed	13.09	2.44	13.87	2.41
Highest education level: primary	0.17	0.37	0.08	0.27
Highest education level: junior high	0.48	0.50	0.46	0.50
Highest education level: senior high	0.20	0.40	0.26	0.44
Highest education level: college	0.16	0.37	0.20	0.40
Dummy for full-time student	0.02	0.13	0.02	0.13
Time-Constant Covariates				
Race: black	0.21	0.41	0.22	0.42
Race: Hispanic	0.17	0.38	0.18	0.38
Number of siblings	3.59	2.50	3.62	2.48
Region: south	0.34	0.47	0.36	0.48
Region: Urban	0.75	0.43	0.77	0.42
Age at first interview	17.25	2.12	17.14	2.06
Mother absent	0.10	0.31	0.11	0.31
Father absent	0.21	0.41	0.22	0.42
Highest education grade for parents	10.92	3.15	10.89	3.09
Parental income rank (18-20)	45.33	28.99	44.67	28.85
Rotter score	8.61	2.35	8.74	2.33
AFQT score	0.45	0.29	0.47	0.27
Expected highest education grade	13.71	2.37	14.19	2.14
Expected number of children	1.66	0.93	1.49	0.74

with a wage rank premium of 3.80, 4.41, 12.05, and 3.56 points, respectively. Therefore, marriage always brings men a wage premium, and the premium is at the highest value when they have two children. For women, however, marriage does not always give them a penalty in wages. For women with no child or one child, marriage decreases their wage rank by 1.98 and 2.17 points; however, if they raise more children, marriage still provides them with a premium: for women with two children, the wage premium in rank is 3.24, and for women with three or more children, the value for the premium is 6.10.

Table 5.2: Gap-Closing Estimand Controlling Number of Children (NLSY79)

Marital Status/	**Number of Children**			
Gender	**Zero**	**One**	**Two**	**Three and More**
Men				
Married	52.96	53.33	54.55	50.77
	(52.83, 53.08)	(53.23, 53.42)	(54.38, 54.71)	(50.31, 51.44)
Not Married	49.16	48.92	42.50	47.21
	(49.15, 49.17)	(48.44, 49.46)	(41.81, 43.25)	(45.97, 48.45)
Difference	3.80	4.41	12.05	3.56
Women				
Married	42.16	39.77	36.98	36.75
	(41.83, 42.41)	(39.64, 39.91)	(36.62, 37.32)	(35.83, 37.72)
Not Married	44.14	41.94	33.74	30.65
	(44.11, 44.17)	(41.66, 42.22)	(33.36, 34.10)	(30.23, 30.96)
Difference	-1.98	-2.17	3.24	6.10

Note: The numbers in parentheses indicate the 2.5% and 97.5% values of the estimation with 500-times bootstrapping.

We then turn to discuss the effect of parenthood when fixing marital status. Table 5.3 presents the empirical results for the model illustrated in Figure 5.2c. To measure the impact of parenthood, we should compare the difference between the two adjacent columns in the table. As inferred from the table, there's no parenthood penalty for them if they have one child. Married men will only suffer a disadvantage in wage rank for fatherhood after

they have three or more children, and unmarried men start to suffer the parenthood penalty of 6 points per child, starting at two children. The case is totally different for women. As the lower panel of Table 5.3 shows, married and unmarried women suffer motherhood penalties from their first child. For married women, the first, second, and third child bring their motherhood penalty to 6.4, 5.2, and 4.3 points in their wage ranks; unmarried women have 7.0, 3.1, and 4.1 points. The results showed almost no fatherhood premium for men. Still, they verified the motherhood penalty for women and suggested that women suffer the highest motherhood penalty at the birth of their first child.

Table 5.3: Gap-Closing Estimands Fixing Marital Status (NLSY79)

Marital Status/	**Number of Children**			
Gender	**Zero**	**One**	**Two**	**Three and More**
Men				
Married	53.41	53.35	54.00	49.91
	(53.26, 53.61)	(53.33, 53.38)	(53.98, 54.01)	(49.63, 50.16)
Not Married	49.88	50.60	44.86	38.92
	(49.79, 49.95)	(50.07, 51.06)	(44.11, 45.81)	(37.50, 40.92)
Women				
Married	44.95	37.35	32.14	27.76
	(44.83, 45.07)	(37.28, 37.41)	(32.02, 32.38)	(27.56, 27.91)
Not Married	46.59	39.63	36.48	32.35
	(46.43, 46.80)	(39.47, 39.81)	(36.27, 36.66)	(32.14, 32.67)

Note: The numbers in parentheses indicate the 2.5% and 97.5% values of the estimation with 500-times bootstrapping.

A.2 Results for Static Causal Mediation Analysis

We then examine how labor market working hours mediate the marital and parenthood effects. Table 5.4 presents the results for men. In summary, if we do not fix the marginal distribution for marriage and parenthood when analyzing the other, both marriage (shown in the first line) and parenting (the 2nd to 4th lines) give men a wage premium. For the analysis of marriage, the NDE makes up 76% of the ATE/total effects, indicating that a quarter

of the wage difference between married and unmarried men can be attributed to a change in working hours (if we change the working hours from the distribution of the unmarried to the married). For the mediation impact on the fatherhood premium, the NDE makes up around 90% for every increased child in the family, suggesting that men's labor market time change due to the increased number of children could only explain 10% of the total effect of the parenthood premium. Therefore, we may conclude that having more children does bring a fatherhood premium if marital status is not controlled. Still, the premium is subtle due to the changes in the amount of time fathers participate in the labor market.

Table 5.4: Natural Direct Effect and Natural Indirect Effect of Working Hours on Wage Returns for Men

Treatment	**Control**	$E[\widehat{\theta}_{a,a}]$	$E[\widehat{\theta}_{a,a'}]$	$E[\widehat{\theta}_{a',a'}]$	**NDE**	**NIE**	**ATE**	**NDE Percentage**	**NIE Percentage**
Married	Not Married	49.75	49.63	49.24	0.39	0.12	0.51	76.99%	23.01%
		(48.95, 50.56)	(48.82, 50.53)	(48.46, 50.42)	(-0.57, 1.25)	(-0.36, 0.55)	(-0.51, 1.38)		
One Kid	Zero Kid	49.53	49.49	49.15	0.34	0.04	0.38	90.10%	9.90%
		(48.00, 50.83)	(47.99, 50.71)	(48.57, 49.71)	(-1.02, 1.16)	(-0.31, 0.55)	(-1.08, 1.39)		
Two Kids	One Kid	49.80	49.76	49.42	0.34	0.04	0.38	89.11%	10.89%
		(48.44, 51.83)	(48.78, 51.74)	(47.99, 50.71)	(-0.49, 2.74)	(-0.76, 0.48)	(-0.80, 2.54)		
Three Kids and More	Two Kids	50.40	50.36	49.96	0.40	0.04	0.44	91.02%	8.98%
		(48.17, 52.53)	(48.37, 52.65)	(48.93, 51.44)	(-1.95, 2.40)	(-0.85, 0.50)	(-2.21, 2.27)		

Note: The numbers in parentheses indicate the 2.5% and 97.5% values of the estimation with 100-times bootstrapping.

Finally, we discuss how labor market time mediates the wage penalties faced by women in marriage and motherhood. Table 5.5 elaborates on the results. Still, if we do not restrict the marginal distributions for marriage when analyzing motherhood effects, nor fix motherhood when studying marriage effects, we could see women suffer wage penalties both due to having more children, while the effects of marriage are not statistically significant (and the total effects are smaller than in the previous gap-closing models). For the analysis of marriage, as shown in the first line of Table 5.5, the NDE accounts for 86% of the ATE, indicating that around 14% of the wage penalty can be attributed to the change in women's labor market time after marriage. For motherhood penalties, as could be seen from the

third and the fourth line, the NDEs for the second child compared to the first and the third compared to the first make up 90% of the total effects, meaning that from the first to the second and from the second to the third, change in labor market time also attributes to a subtle effect of the corresponding motherhood penalties. However, it is worth noting that for the birth of the first child, the NDE accounts for only 68%, meaning that almost a third of the motherhood penalty in wages for the first child can be attributed to the difference in labor market time for mothers.

Table 5.5: Natural Direct Effect and Natural Indirect Effect of Working Hours on Wage Returns for Women

Treatment	**Control**	$E[\widehat{\theta}_{a,a}]$	$E[\widehat{\theta}_{a,a'}]$	$E[\widehat{\theta}_{a',a'}]$	**NDE**	**NIE**	**ATE**	**NDE Percentage**	**NIE Percentage**
Married	Not Married	36.51	36.68	37.73	-1.05	-0.17	-1.22	86.07%	13.93%
		(35.80, 38.37)	(35.79, 38.53)	(36.58, 38.73)	(-1.64, 0.63)	(-0.47, 0.22)	(-1.70, 0.66)		
One Kid	Zero Kid	37.35	37.53	37.93	-0.40	-0.18	-0.58	68.79%	31.21%
		(36.06, 38.56)	(36.27, 39.10)	(36.56, 38.94)	(-1.09, 0.86)	(-1.15, 0.17)	(-1.98, 0.72)		
Two Kids	One Kid	36.56	36.67	37.78	-1.11	-0.11	-1.22	90.99%	9.01%
		(35.50, 37.53)	(35.62, 37.73)	(36.27, 39.10)	(-2.56, 0.61)	(-0.64, 0.33)	(-2.71, 0.45)		
Three Kids and More	Two Kids	35.59	35.75	37.11	-1.36	-0.16	-1.52	89.56%	10.44%
		(33.29, 37.86)	(33.64, 37.92)	(36.17, 38.31)	(-3.35, 0.73)	(-0.94, 0.43)	(-4.15, 0.63)		

Note: The numbers in parentheses indicate the 2.5% and 97.5% values of the estimation with 100-times bootstrapping.

B. Results for Dynamic Models

B.1 Results for Dynamic Gap-Closing Estimand

Now, we turn to the results for the dynamic models. We first elaborate on the results for the gap-closing estimand, fixing the parenthood status. Hence, we may examine how, in different gender, age, and parenthood status groups, marriage affects the wage rank. A first glance at the table shows the expected result that with the increase in age between 21 and 33, the wage rank for both men and women is increasing, and obviously, there is a gender gap in wages in the labor market.

Table 5.6: Gap-Closing Estimands Fixing Parenthood Status (NLSY79)

Parenthood Status	**Male, Mean Wage Rank 21-25**		**Female, Mean Wage Rank 21-25**	
	Not Married	**Married**	**Not Married**	**Married**
Childless 21, Childless 25	48	50.1	41	40.3
	(46.6,49.2)	(48.4,52.0)	(39.6,42.3)	(38.3,41.8)
Childless 21, With Children 25	49	51.2	38.9	38.3
	(47.0,50.0)	(49.1,52.3)	(37.2,40.3)	(36.5,40.1)
With Children 25, With Children 29	49.5	51.7	37.9	37.4
	(47.9,51.3)	(50.1,53.9)	(36.1,39.5)	(35.6,39.3)
Parenthood Status	**Male, Mean Wage Rank 26-29**		**Female, Mean Wage Rank 26-29**	
Childless 25, Childless 29	59.8	59.8	55.2	52.2
	(58.3,60.9)	(57.9,61.7)	(53.2,57.1)	(50.0,54.3)
Childless 25, With Children 29	61	60.8	53.6	50.9
	(59.8,62.3)	(59.0,62.3)	(51.4,55.5)	(48.8,53.0)
With Children 25, With Children 29	60.9	60.7	50.2	47.6
	(59.3,62.6)	(58.5,61.6)	(47.7,53.1)	(45.4,49.7)
Parenthood Status	**Male, Mean Wage Rank 30-33**		**Female, Mean Wage Rank 30-33**	
Childless 29, Childless 33	64	63.2	60.6	56.4
	(62.5,65.2)	(61.5,64.8)	(58.7,62.6)	(54.5,58.5)
Childless 29, With Children 33	64.9	64.1	61	56.6
	(63.6,65.9)	(62.2,65.7)	(51.4,55.5)	(48.8,53.0)
With Children 29, With Children 33	65.2	64.4	56	52
	(53.8,66.3)	(62.8,65.8)	(54.2,57.8)	(49.6,53.9)

We then delve into the marriage effects for both men and women. For men, as indicated by the left two columns of Table 5.6, after accounting for crossover parenthood effects, only those who marry before age 25 experience an insignificant wage rank premium of about 2.2 percentiles. However, for those marrying at ages 29 and 33, marriage does not result in any wage premium. These findings are consistent across childless individuals, those who had their first child within four years, and those who have raised at least one child for over four years. For women who marry before age 25, marriage does not impose a penalty on their earnings. However, the penalty increases with age: if a woman remains unmarried until ages 29 or 33, she enjoys a 3 and 4-percentile advantage in wage rank, respectively, compared to those who married at those ages. Therefore, for men, early marriage offers a slight wage rank benefit, while for women, delaying marriage and motherhood reduces la-

bor market penalties.

Table 5.7: Gap-Closing Estimands Fixing Marital Status (NLSY79)

Marital Status	**Male, Mean Wage Rank 21-25**		**Female, Mean Wage Rank 21-25**	
	No Children	**With Children**	**No Children**	**With Children**
Unmarried 21, Unmarried 25	48.2	50.3	44	35
	(47.0, 49.3)	(48.6, 52.1)	(42.8, 45.4)	(33.5, 36.5)
Unmarried 21, Married 25	48.5	50.7	44.1	35.2
	(47.0, 50.0)	(49.1, 52.3)	(42.8, 45.6)	(33.7, 37.0)
Married 21, Married 25	48.9	51	44	35.2
	(47.3, 50.8)	(49.2, 52.6)	(42.0, 45.4)	(33.8, 37.0)
Marital Status	**Male, Mean Wage Rank 26-29**		**Female, Mean Wage Rank 26-29**	
Unmarried 25, Unmarried 29	60.7	60	59.7	45.8
	(59.2, 61.9)	(58.4, 61.2)	(58.1, 61.7)	(44.2, 47.5)
Unmarried 25, Married 29	60.7	60.1	60.6	46.7
	(59.2, 62.2)	(58.8, 61.7)	(58.5, 62.3)	(45.2, 48.2)
Married 25, Married 29	60.7	60	60.6	46.9
	(59.3, 62.6)	(58.5, 61.6)	(58.6, 62.4)	(45.4, 48.3)
Marital Status	**Male, Mean Wage Rank 30-33**		**Female, Mean Wage Rank 30-33**	
Unmarried 29, Unmarried 33	64.2	64.6	64.8	52.5
	(62.6, 65.9)	(63.0, 65.7)	(62.7, 67.0)	(51.1, 54.0)
Unmarried 29, Married 33	64.3	64.6	64.9	52.6
	(62.7, 66.1)	(63.0, 65.8)	(62.7, 66.9)	(51.1, 54.2)
Married 29, Married 33	64.2	64.6	65	52.6
	(62.8, 65.8)	(63.2, 65.7)	(63.0, 67.0)	(51.2, 54.2)

Note: The numbers in parentheses indicate the 2.5% and 97.5% values of the estimation with 100-times bootstrapping.

Next, we control for marital status and examine the differences between men and women who have children before a specific age versus those who do not. The results are presented in Table 5.7. The results reveal that for relatively young (under 25) men, for all marital categories, having children will bring a 2.1-2.2 percentile premium on wages for them if they become fathers. For men at age 29 or 33, for all fixed marital groups (instant marriage and accumulative marriage), there's no significant divergence in wages if they have or do not have a child. Furthermore, the wage rank will be slightly higher for earlier married men, while there's no distinct difference in the age 29 and age 33 groups.

For women, we find substantial motherhood penalties for all age groups with all marital statuses. Women with children suffer 8.8-9 percentile in wage penalties at 25, 13.7-13.9 percentile at 29, and 12.3-12.4 percentile at 33, indicating that women between 25 and 29 suffer the most penalties from motherhood for the NLSY 79 cohort (in the 1980s). The penalties were reduced slightly when they were observed at age 33.

Hence, we could summarize that while motherhood penalties ubiquitously exist for women ages 25 to 33, a fatherhood premium for men only exists for men younger than 25.

B.2 Results for Dynamic Causal Mediation Analysis

We then present the results for dynamic causal mediation models. As we discussed in Section III, we present the results assuming the carryover effects (assuming the treatment and the mediator in the previous round will affect the mediator and the outcome in the next round), the feedback effects (assuming the mediator and the outcome in the previous round will affect the values of the treatment and the mediator in the second round), and the full

model settings. Again, the total effects in the tables are just the sum of the IDE and the IIE; it is not the unbiased estimation of the causal effect of parenthood/marriage on wage rank (the total average treatment effects, TATE). If the effect of the IDE and the IIE go in the same direction (which means they separately make a partial contribution to the total effect), we will show their representative percentages in the total effect. However, if the IDE and the IIE have the opposite effect (meaning the path $A \rightarrow Y$ and $A \rightarrow M \rightarrow Y$ offset the effect of each other), we will omit the percentage and dig further into the meaning of the positive or negative effects.

Table 5.8 shows the results for the causal mediation results for working hours mediating the causal effect of marriage on wages (treatment group married vs. the control group unmarried). Although the total effects are not an accurate estimation of the TATE, we can still see that the marital premium for men can only be observed at 25 (from the results in the total effect column in Table 5.8), and the premium shrinks (to statistically insignificant) as men age. In both the 22-25 and 26-29 age ranges, the direct and indirect effects are positive, suggesting that the direct effect from marriage and the indirect effect from marriage through labor market participation both increase men's wages, whereas the percentage of the indirect effect increases at age 29 compared to its proportion in age 25 under the full model. Comparing the carryover, the feedback, and the full model in the three age stages, it is clear that the feedback effects count more in the total effects. In contrast, the carryover effects (without specifying the previous round's mediator and the outcome's effect on the treatment and mediators) overestimate the direct effect. Thus, for modeling the causal effect of men's marital status on wages via labor market participation, we need to take the feedback effects into consideration. Finally, in the model for the age 33 group, we can see that the direct effect is negative, although the indirect effect is still positive, although they

are not statistically significant, indicating that the direct effect of marriage will not bring men premiums in their wages.

Table 5.8: Dynamic Interventional Direct and Indirect Effects of Working Hours on the Marital Causal Effects for Men

Group	Model Type	$E[\widehat{\theta}_{a,a}]$	$E[\widehat{\theta}_{a,a'}]$	$E[\widehat{\theta}_{a',a'}]$	Total Effect	Direct Effect	Indirect Effect	IDE Percentage	IIE Percentage
Age 25	carryover	51.9	51.82	50.21	1.68	1.61	0.08	95.53%	4.47%
		(50.25, 53.51)	(50.06, 53.47)	(49.17, 51.49)	(0.30, 3.16)	(0.22, 3.06)	(-0.24, 0.43)		
	feedback	51.79	51.61	50.5	1.29	1.11	0.17	86.54%	13.46%
		(50.34, 53.42)	(50.27, 53.07)	(49.58, 51.78)	(0.13, 2.70)	(0.14, 2.35)	(-0.53, 0.86)		
	full	52.07	51.88	50.47	1.6	1.41	0.19	88.32%	11.68%
		(50.49, 54.00)	(50.33, 53.57)	(49.41, 51.69)	(0.12, 3.30)	(0.07, 2.87)	(-0.47, 0.86)		
Age 29	carryover	62.74	62.71	61.41	1.33	1.30	0.03	97.82%	2.18%
		(61.52, 63.99)	(61.46, 64.00)	(60.29, 62.47)	(-0.22, 2.91)	(-0.20, 2.87)	(-0.26, 0.33)		
	feedback	63.13	62.36	61.57	1.56	0.79	0.77	50.69%	49.31%
		(62.01, 64.10)	(61.12, 63.38)	(60.49, 62.52)	(0.42, 2.66)	(-0.03, 1.82)	(0.02, 1.51)		
	full	63.18	62.35	61.69	1.49	0.65	0.83	43.90%	56.10%
		(61.83, 64.33)	(61.28, 63.55)	(60.54, 62.72)	(0.30, 2.60)	(-0.46, 1.70)	(0.15, 1.68)		
Age 33	carryover	66.25	66.31	65.74	0.51	0.57	-0.06	n.a.	n.a.
		(64.92, 67.48)	(65.05, 67.52)	(64.47, 67.07)	(-0.68, 1.55)	(-0.46, 1.69)	(-0.33, 0.29)		
	feedback	66.46	65.74	65.89	0.57	-0.15	0.72	n.a.	n.a.
		(65.14, 67.62)	(64.51, 67.02)	(64.65, 66.93)	(-0.53, 1.62)	(-0.97, 0.36)	(-0.16, 1.75)		
	full	66.40	65.71	65.84	0.56	-0.14	0.70	n.a.	n.a.
		(65.12, 67.60)	(64.44, 67.14)	(64.64, 67.01)	(-0.54, 1.61)	(-0.89, 0.48)	(-0.25, 1.57)		

Note: The numbers in parentheses indicate the 2.5% and 97.5% values of the estimation with 100-times bootstrapping. n.a.: Percentage decomposition is omitted because the IDE and IIE have opposite signs; one percentage would be negative while the other would exceed 100%.

We then discuss if there's a fatherhood premium for men and if the working hours in the labor market mediate the effect. The results are presented in Table 5.9. Similarly, if we compare the three models, we could easily conclude that omitting the feedback effects will overestimate the direct effects of fatherhood on earnings. So, we make our inferences based on the feedback and full models. The results suggest that for men becoming fathers before age 25, there's a direct fatherhood premium on their wages (as the confidence interval of the direct effect is higher than 0). In contrast, for the age group 29 to 33, the direct effects are reduced to insignificance, while the indirect effect via labor market participation is prominent (makes around 83% of the total effect in age 29 to 33). This is consistent with the

speculations from the specialization theory, at least for men at 29 and 33, that the wage premiums they enjoy from fatherhood are because they devote more time to the labor market even after becoming fathers. Then, we turn to discuss the causal mediation effect of labor

Table 5.9: Dynamic Interventional Direct and Indirect Effects of Working Hours on the Parenthood Causal Effects for Men

Group	Model Type	$E[\widehat{\theta}_{a,a}]$	$E[\widehat{\theta}_{a,a'}]$	$E[\widehat{\theta}_{a',a'}]$	Total Effect	Direct Effect	Indirect Effect	IDE Percentage	IIE Percentage
Age 25	carryover	53.04	52.76	49.7	3.34	3.06	0.28	91.61%	8.39%
		(51.21, 55.02)	(50.94, 54.73)	(48.55, 50.98)	(1.34, 5.86)	(1.17, 5.44)	(-0.22, 0.78)		
	feedback	52.7	51.49	50.23	2.48	1.27	1.21	51.12%	48.88%
		(51.44, 54.61)	(50.41, 53.02)	(48.92, 51.32)	(0.94, 4.30)	(0.28, 2.57)	(0.25, 2.22)		
	full	53.29	51.98	50.2	3.09	1.77	1.32	57.35%	42.65%
		(51.64, 55.41)	(50.38, 53.92)	(49.02, 51.45)	(1.02, 5.30)	(0.23, 3.41)	(0.28, 2.59)		
Age 29	carryover	63.06	62.88	60.5	2.56	2.38	0.18	92.97%	7.03%
		(61.95, 64.28)	(61.81, 64.02)	(59.30, 61.47)	(1.10, 3.88)	(1.01, 3.64)	(-0.20, 0.62)		
	feedback	63.55	61.91	61.18	2.37	0.74	1.63	31.05%	68.95%
		(62.44, 64.66)	(60.95, 62.95)	(59.96, 62.21)	(1.37, 3.56)	(-0.06, 1.65)	(0.95, 2.59)		
	full	63.29	61.67	61.35	1.94	0.32	1.63	16.28%	83.72%
		(62.01, 64.52)	(60.63, 62.79)	(60.06, 62.33)	(0.66, 3.52)	(-0.88, 1.50)	(0.85, 2.58)		
Age 33	carryover	66.64	66.52	64.49	2.16	2.04	0.12	94.42%	5.58%
		(65.70, 67.98)	(65.32, 68.02)	(62.93, 66.13)	(1.03, 3.88)	(0.82, 3.52)	(-0.27, 0.54)		
	feedback	66.89	65.42	65.12	1.78	0.30	1.48	16.69%	83.31%
		(65.86, 68.18)	(63.98, 66.84)	(63.72, 66.55)	(0.58, 2.81)	(-0.34, 1.02)	(0.64, 2.52)		
	full	66.83	65.39	65.08	1.75	0.31	1.45	17.43%	82.57%
		(65.76, 67.99)	(63.89, 66.83)	(63.52, 66.48)	(0.48, 2.87)	(-0.49, 1.08)	(0.67, 2.40)		

Note: The numbers in parentheses indicate the 2.5% and 97.5% values of the estimation with 100-times bootstrapping.

market participation for women. Table 5.10 presents the results. As can be seen from the table, in all age ranges for all (carryover, feedback, full) models, the effects of both the direct effect of marriage on wages and the indirect impact of marriage via labor market participation on wages are negligible (only in the full model in the age 33 group, the direct effect shows some negative causation of marriage which is close to the significant level). The results are consistent with our previous findings in both the static causal mediation model and the dynamic gap-closing estimands, which state that marriage is, at least, not the most crucial factor contributing to women's disadvantaged earnings in the labor market.

Finally, we present the results of the mediation effects on the motherhood penalty for women in the dynamic model in Table 5.11. Like the findings in the previous models, we clearly find evidence of the motherhood penalty for women in all age groups. With the

Table 5.10: Dynamic Interventional Direct and Indirect Effects of Working Hours on the Marital Causal Effects for Women

Group	Model Type	$E[\widehat{\theta}_{a,a}]$	$E[\widehat{\theta}_{a,a'}]$	$E[\widehat{\theta}_{a',a'}]$	Total Effect	Direct Effect	Indirect Effect	IDE Percentage	IIE Percentage
Age 25	carryover	39.98	40.02	40.21	-0.24	-0.19	-0.05	79.84%	20.16%
		(38.64, 41.51)	(38.75, 41.57)	(39.19, 41.22)	(-1.49, 1.01)	(-1.33, 1.01)	(-0.44, 0.29)		
	feedback	40.20	40.45	40.18	0.02	0.28	-0.25	n.a.	n.a.
		(38.96, 41.52)	(39.08, 41.53)	(39.10, 41.09)	(-1.09, 1.24)	(-0.58, 1.38)	(-0.93, 0.25)		
	full	40.04	40.24	40.19	-0.15	0.05	-0.20	n.a.	n.a.
		(38.58, 41.45)	(38.92, 41.73)	(39.11, 41.32)	(-1.55, 1.30)	(-1.04, 1.30)	(-0.87, 0.41)		
Age 29	carryover	53.61	53.53	52.43	1.18	1.10	0.09	92.73%	7.27%
		(51.80, 54.94)	(51.59, 54.80)	(50.92, 53.89)	(-0.49, 2.72)	(-0.50, 2.55)	(-0.26, 0.43)		
	feedback	52.91	52.82	52.44	0.47	0.38	0.09	80.34%	19.66%
		(51.39, 54.26)	(51.27, 54.21)	(50.90, 53.79)	(-0.56, 1.66)	(-0.75, 1.52)	(-0.59, 0.77)		
	full	53.10	52.94	52.77	0.32	0.17	0.16	51.56%	48.44%
		(51.49, 54.56)	(51.27, 54.33)	(51.18, 54.19)	(-1.09, 1.48)	(-1.06, 1.16)	(-0.57, 0.85)		
Age 33	carryover	57.76	57.62	57.81	-0.05	-0.20	0.15	n.a.	n.a.
		(56.28, 59.20)	(56.01, 58.98)	(56.43, 59.18)	(-1.19, 0.98)	(-1.26, 0.87)	(-0.22, 0.60)		
	feedback	57.76	57.09	57.75	0.01	-0.66	0.67	n.a.	n.a.
		(56.19, 59.43)	(55.66, 58.53)	(56.15, 59.24)	(-1.33, 1.39)	(-1.59, 0.17)	(-0.32, 1.61)		
	full	57.70	56.99	57.91	-0.21	-0.92	0.71	n.a.	n.a.
		(56.38, 59.33)	(55.69, 58.68)	(56.26, 59.41)	(-1.78, 1.09)	(-1.92, 0.01)	(-0.37, 1.60)		

Note: The numbers in parentheses indicate the 2.5% and 97.5% values of the estimation with 100-times bootstrapping. n.a.: Percentage decomposition is omitted because the IDE and IIE have opposite signs; one percentage would be negative while the other would exceed 100%.

results of decomposition from the full models, we find that the proportion of the indirect effect (the motherhood penalty via women's reduced labor market participation) increases as the observational age increases (from 23.5% at the age of 25 to around 40% at the age of 29 and to around 60% at the age of 33). The results are once again consistent with the parenthood specialization theory: as the time of motherhood increases, women withdraw their energy from the labor market and devote it to domestic jobs, leading to their wage penalties in the labor market, and then the process is reinforced, women devote less time to the labor market but more time to domestic work, and their lack of labor market participation gradually becomes the main component in their motherhood penalty. On the other hand, as we explained in Table 5.9, after childbearing, men increase their time in the labor market, making the indirect effect of labor market participation the main component in their fatherhood premium (the proportions for the IIE in the TE for men at 25, 29, and 33 are

respectively 42%, 84%, and 83%).

Table 5.11: Dynamic Interventional Direct and Indirect Effects of Working Hours on the Parenthood Causal Effects for Women

Group	Model Type	$E[\widehat{\theta}_{a,a}]$	$E[\widehat{\theta}_{a,a'}]$	$E[\widehat{\theta}_{a',a'}]$	Total Effect	Direct Effect	Indirect Effect	IDE Percentage	IIE Percentage
Age 25	carryover	37.7	37.99	41.49	-3.8	-3.5	-0.3	92.21%	7.79%
		(36.12, 39.36)	(36.49, 39.56)	(40.32, 42.46)	(-5.80, -2.12)	(-5.27, -1.93)	(-0.67, 0.03)		
	feedback	37.97	38.85	41.55	-3.57	-2.7	-0.87	75.56%	24.44%
		(36.58, 39.55)	(37.46, 40.35)	(40.31, 42.77)	(-5.21, -1.88)	(-4.25, -1.15)	(-1.58, -0.05)		
	full	37.62	38.54	41.54	-3.92	-3	-0.92	76.50%	23.50%
		(35.96, 39.16)	(37.14, 40.35)	(40.47, 42.56)	(-5.72, -2.28)	(-4.51, -1.49)	(-1.87, -0.09)		
Age 29	carryover	49.67	50.32	54.84	-5.16	-4.52	-0.64	87.52%	12.48%
		(47.45, 51.53)	(48.16, 52.25)	(53.12, 56.28)	(-7.66, -3.24)	(-7.06, -2.39)	(-1.42, -0.21)		
	feedback	50.32	52.3	54.28	-3.96	-1.98	-1.98	50.00%	50.00%
		(48.20, 51.81)	(50.63, 53.69)	(52.54, 55.48)	(-5.81, -2.52)	(-3.36, -0.83)	(-3.24, -0.98)		
	full	49.71	51.67	54.71	-4.99	-3.04	-1.96	60.78%	39.22%
		(47.72, 51.55)	(49.77, 53.42)	(52.79, 56.15)	(-6.92, -3.26)	(-4.89, -1.85)	(-3.23, -0.91)		
Age 33	carryover	55.48	55.75	60.10	-4.62	-4.35	-0.27	94.14%	5.86%
		(53.82, 57.27)	(53.97, 57.43)	(58.47, 61.83)	(-6.77, -2.71)	(-6.37, -2.40)	(-0.85, 0.19)		
	feedback	55.93	58.99	59.93	-4.01	-0.94	-3.07	23.44%	76.56%
		(54.38, 57.52)	(57.41, 60.60)	(58.20, 61.57)	(-5.73, -2.26)	(-2.30, -0.09)	(-4.33, -1.63)		
	full	55.42	58.45	60.44	-5.02	-1.99	-3.02	39.74%	60.26%
		(53.85, 56.87)	(56.79, 60.20)	(58.78, 62.05)	(-6.91, -2.96)	(-3.40, -0.75)	(-4.46, -1.81)		

Note: The numbers in parentheses indicate the 2.5% and 97.5% values of the estimation with 100-times bootstrapping.

VI. Conclusions and Further Discussions

In this study, we present a novel research design investigating the classic topic in family sociology: the causal effects of parenthood and marital status on wages for both men and women. Using the same dataset as most previous studies did for this topic, we introduce an innovative approach; whereas previous analyses have primarily focused on the ATT for the switchers (which compares the effect of marriage and parenthood for individuals), our approach utilizes the ATE to offer a more comprehensive understanding of these dynamics between different subpopulation groups. Our main findings provide evidence consistent with the motherhood penalty, showing that women often experience a wage decrease after having children. The effects of marriage on earnings are smaller and less persistent than the static estimates alone might suggest, especially after accounting for dynamic carryover and

feedback processes. However, we observe a fatherhood premium exclusively among young men under 25, with this premium being the highest within this group. Conversely, the wage penalty is lowest for women who are unmarried and without children after the age of 33, suggesting age and parental status are critical factors influencing women's earnings.

Substantively, our research contributes to the existing literature by highlighting the global causal effects of parenthood on wages and the significant role of labor market participation. Our study is based on the labor market perspective (instead of the individual's life trajectory, as previous studies capture the ATT), which avoids the selection problem in previous studies. From our perspective, this approach is more suitable if the researchers aim to understand the effects of a specific demographic event on the supply and demand in the labor market. Meanwhile, we introduce a dynamic analytical framework to our research, and our findings are consistent with the specialization theory in parenthood: men tend to allocate more time to the labor market and less to domestic work after childbearing, resulting in a wage premium. Women, in contrast, often reduce their labor market participation to accommodate increased domestic responsibilities, leading to a wage penalty. This divergence emphasizes how traditional gender roles continue to influence economic outcomes in the context of parenthood.

Methodologically, we advance the field by employing Non-Parametric Marginal Structural Models (NPMSM) and Double Machine Learning (DML) techniques for gap-closing estimands and time-varying causal mediation analysis. This is an example of applying the framework of the efficient estimator in the causal mediation analysis, as we discussed in the previous chapter. The idea of controlled disparity was first introduced in epidemiological research (Jackson and VanderWeele 2018) and further adopted by sociologists (Lundberg

2024) as the gap-closing estimand which aims to tackle the causal effects for variables that could not be treated as “treatments" (in Pearl’s [2009] term, in which he views all treatment as a “do"). In this paper, we further extend the idea of “fixing the group" in a time-varying setting and discuss the application in a dynamic model. The time-varying causal mediation framework is also a mature area (VanderWeele and Tchetgen Tchetgen 2017), and we here combine it with the efficient/doubly robust/Neyman-orthogonal estimator and apply it to the empirical research. We believe our approach, compared to the methods in the previous literature, better addresses the dynamic process in the production of inequality.

Conclusion

In this thesis, we discussed the debiased machine learning/doubly robust/efficient estimator used for causal inference and showed how it could be applied in social science. We first reviewed the literature concerning the foundations in the areas of causal inference and debiased machine learning/efficient estimation theory, giving a comprehensive and relatively easy-to-understand derivation for the estimation. We focus on applying the estimation to the observational (survey) data and pseudo-randomized controlled trial settings—areas of significant interest to social scientists. Next, we focus on two specific areas: survival analysis and mediation analysis. In particular, we developed debiased machine learning (doubly robust/efficient) estimators for causal inference in survival data and causal mediation analysis.

For the DML with the left-truncated-right-censored (LTRC) survival data, we developed the twice doubly robust estimator, which directly constructs the counterfactual survival curves and estimates the average and heterogeneous treatment effect on the mean survival time for observational/survey data. In this estimator, one doubly robust estimation is used for the causal effect identification, and another doubly robust estimator is used to tackle the survival function estimation. We compared the advantages of estimation over the methods adopted in social science, especially the method of marginal hazard ratio (MHR), and showed its robustness in estimating the heterogeneous treatment effects for the sur-

vival outcomes simulations.

We elaborate on the DML estimator for survival data in an empirical study concerning the average and heterogeneous causal effect of widowhood on mortality for US elders. Applying the DML method for causal identification, we find evidence consistent with a causal effect of widowhood in reducing the surviving spouse's life expectancy under the stated identification assumptions. Meanwhile, we show that such a causal effect has heterogeneity: 1) individuals who are less prepared for their spouse's death experience a more pronounced impact on their life expectancy; 2) men and women under different educational and family wealth statuses suffer differently from losing their spouse.

In the second part, we turn to derive the DML estimator for causal mediation analysis. We reviewed the differences between mediators, colliders, and confounders and the classic decomposition methods to measure the direct causal effects on the outcome and the indirect causal effects from the treatment via the mediator on the outcome. Based on the prerequisite knowledge of the causal mediation analysis, we derive the DML/efficient estimators for the direct and indirect effects in both the static and the dynamic models and further apply our method to a classic social science research issue analyzing how political instability mediates the causal effect of racial fractionalization on the onset of wars.

The last chapter in this thesis applies the DML framework in causal mediation analysis to reanalyze the causal effects of parenthood and marriage on wages for men and women and how labor market participation mediates the effects. Unlike the previous research focusing on the local causal effects of marriage and childbearing on individuals, we apply the nonparametric marginal structural models (NPMSM) aiming to tackle the global treatment

effects– the divergences between different groups in the labor market. In this study, we use the "gap-closing estimand" and the traditional causal mediation analysis framework and find substantial evidence on the motherhood penalty, while the other marital and parenthood effects on wages for men and women are less clear. Furthermore, to verify the specialization assumption on gender labor division, we designed the time-varying model to reveal the dynamic process of childbearing, marriage, and labor market participation, leading to gender wage inequality in the labor market.

Intrinsically, we deploy the DML/efficient estimator method to tackle the missing data problem in social science. In causal inference, the missing part is the counterfactual outcome, and in survival analysis, the missing part is the survival function for the truncated and censored individuals. To predict or impute the missing data, we always need to keep in mind that it is infeasible if we do not have assumptions about the underlying data-generating process (without assumptions, no inferences). The assumptions define the scope of validity/generalizability of our conclusions. Even though all methods discussed in this thesis are based on the principle of unbiasedness, the model will still produce incorrect conclusions if the assumptions do not align with reality.

For many quantitative social researchers, Occam's razor is a principle that should always be remembered: *"entia non sunt multiplicanda sine necessitate"* (entities must not be multiplied beyond necessity). A more complex model usually requires better data quality, further assumptions on the relationships among variables, and the feasibility of the models' algorithms. Indeed, an unbiased, consistent, and efficient estimator is a better measure than a biased, non-convergent, and more error-prone one. However, adding unnecessary conditions and complicating the original problem to achieve such metrics can be counter-

productive.

Finally, we want to point out that although this thesis focuses on developing the methodological side of sociology, methodology should remain subordinate to sociological theory and substantive explanation in social science studies. This does not mean that we can forgo a rigorous and careful examination when applying statistical methods; rather, as Otis Duncan once said, "the application of statistical models and methods should be strictly subordinate to the central scientific task" (Duncan and Stenbeck 1988). Specifically, the validity of the estimator can never be ensured without theory-driven identification. The causal effects we pursue in the empirical studies in this thesis have practical significance and academic value only when they can be interpreted through, and contribute back to, debates in social demography, political sociology, family sociology, and social stratification and mobility. In this sense, as Duncan and Stenbeck said, the rhetorical strategy for methodological studies is to engender skepticism about statistical models, as they still fall short of providing a comprehensive scientific explanation on their own; future research should therefore combine stronger identification, better data, and richer theory.

Appendix A

Appendix to Chapter 1

A.1. Appendix I: Proof on Regularity and Score Functions

Score Function Transformations Suppose we have two ($\sigma-$finite) measures $P(Z)$ and $\tilde{P}(Z)$ where $P(Z)$ dominates $\tilde{P}(Z)$ (or $\tilde{P}(Z)$ is absolutely continuous with respect to $P(Z)$). Now imagine we have a **differentiable path** starting from $P(Z)$ and ending at $\tilde{P}(Z)$. The most convenient method to define it is linear interpolation: let $\tilde{P}_\epsilon(Z) = \epsilon\tilde{P}(Z)+(1-\epsilon)P(Z)(\epsilon \in [0,1])$. Therefore, suppose the probability density function for $P(Z)$ and $\tilde{P}(Z)$ are respectively $p(z)$ and $\tilde{p}(z)$, then the density function for $\tilde{p}_\epsilon(z) = \epsilon\tilde{p}(z) + (1-\epsilon)p(z)$.

PROOF A.A.1.1 *We have the definition of a probability density function $p(z)$ for a measure P such that for any measurable set Z:*

$$P(Z) = \int_Z p(z)dz$$

Thus,

$$\tilde{P}_\epsilon(Z) = \epsilon\int_Z \tilde{p}(z)\,dz + (1-\epsilon)\int_Z p(z)\,dz = \int_Z \big(\epsilon\tilde{p}(z) + (1-\epsilon)p(z)\big)\,dz$$

And therefore,

$$\tilde{p}_\epsilon(z) = \epsilon\tilde{p}(z) + (1-\epsilon)p(z).$$

With the differentiable path, we define the score function corresponding to the path from $P(Z)$ towards $\tilde{P}(Z)$ as the rate change of the log-likelihood at the starting point $P(Z)$ (the gradient of the log-likelihood function):

$$s_{\epsilon_0}(z) = \frac{\partial log\, \tilde{p}_\epsilon(z)}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} \tag{A.A.1.1}$$

Usually, in our discussion, we set $\epsilon_0 = 0$. According to Equation 1.III.12[1], we may rewrite $\tilde{P}_\epsilon(Z)$ and $\tilde{p}_\epsilon(z)$ as: $\tilde{P}_\epsilon(Z) = \int_Z (1+\epsilon s(z))\, dP(Z)$ and $\tilde{p}_\epsilon(z) = (1+\epsilon s(z))\, p$.

PROOF A.A.1.2 *We first prove that the definition of* $s(z) = \frac{\partial log\, \tilde{p}_\epsilon(z)}{\partial \epsilon}\Big|_{\epsilon=0}$ *can be rewritten as:*

$$s(z) = \frac{\tilde{p}(z)}{p(z)} - 1 \tag{A.A.1.2}$$

Using the chain rule to differentiate:

$$\begin{aligned}\frac{\partial log\, \tilde{p}_\epsilon(z)}{\partial \epsilon}\Big|_{\epsilon=0} &= \frac{\partial log\, \tilde{p}_\epsilon(z)}{\partial \tilde{p}_\epsilon(z)} \cdot \frac{\partial \tilde{p}_\epsilon(z)}{\partial \epsilon}\Big|_{\epsilon=0} = \frac{1}{\tilde{p}_\epsilon(z)} \cdot \frac{\partial}{\partial \epsilon}(\epsilon \tilde{p}(z) + (1-\epsilon)\, p(z))\Big|_{\epsilon=0} \\ &= \frac{1}{p(z)}(\tilde{p}(z) - p(z)) = \frac{\tilde{p}(z)}{p(z)} - 1\end{aligned}$$

Therefore, $\tilde{p}(z) = (s(z)+1)\, p(z)$.

We could rewrite $\tilde{p}_\epsilon(z)$ as $\tilde{p}_\epsilon(z) = (1-\epsilon)\, p(z) + \epsilon(s(z)+1)\, p(z)$. Simplifying the expression:

$$\tilde{p}_\epsilon(z) = (1 - \epsilon + \epsilon s(z) + \epsilon)\, p(z) = (1 + \epsilon s(z))\, p(z) \tag{A.A.1.3}$$

Further, to prove that $\tilde{P}_\epsilon(Z) = \int_Z (1+\epsilon s(z))\, dP(Z)$, we need the **Radon-Nikodym theorem**:

Theorem A.A.1.1 (Radon-Nikodym Theorem) *Let* (Ω, F, P) *be a probability space, and let* $\tilde{P}_\epsilon$ *be another probability measure on* (Ω, F) *such that measure* $\tilde{P}_\epsilon$ *is absolutely continuous with*

[1] Generally, if $\epsilon_0 = 0$, we simplify the score function $s_\epsilon(z)$ as $s(z)$, or further, s.

respect to measure P. Then there exists a P-integrable function $f : \Omega \to [0, \infty)$ such that for every $A \in F$,

$$\tilde{P}_\epsilon(A) = \int_A f \, dP.$$

The function f is called the Radon-Nikodym derivative and is often denoted by $\frac{d\tilde{P}_\epsilon}{dP}$.

The theorem gives us a toolbox to "rescale" the measure $\tilde{P}_\epsilon$ with the measure P and the **Radon-Nikodym derivative** function f. From Equation A.A.1.3, the scale is indeed $(1+\epsilon s(z)$, and thus,

$$\tilde{P}_\epsilon(Z) = \int_Z (1 + \epsilon s(z)) \, p(z) dz = \int_Z (1 + \epsilon s(z)) \, dP(Z). \tag{A.A.1.4}$$

The definition and Equation A.A.1.3 indeed reveal that the essence of the score function is a "direction pointer" (or compass). We could use the score function to specifically point the direction of the difference between the empirical estimator and the true statistical estimand. Suppose we have the estimand as $\psi(P)$, the pathway derivative, or the gradient of the estimand in the direction of the score function[2], can be defined as:

$$\lim_{\epsilon \to 0} \frac{\psi(\tilde{P}_\epsilon) - \psi(P)}{\epsilon} = \nabla_s \psi(P) \tag{A.A.1.5}$$

Where we could regard $\psi(\tilde{P}_\epsilon)$ as a perturbed version of the true estimand $\psi(P)$, adjusted by a small amount of ϵ in the direction of some perturbation.

Factorization Suppose we have a bivariate joint distribution $P(Y, X)$, clearly, we have the Bayesian rule $P(Y, X) = P(Y \mid X)P(X)$. We define the score function for $P(Y \mid X)$ and $P(X)$ separately as $s_{Y|X}(x, y)$ and $s_X(x)$. The score functions satisfy $E[s_{Y|X}(x, y) \mid X] = 0$ and

[2]The gradient here is defined on the direction of the score function for the estimator. Indeed, the score function is also a gradient, but the gradient on the measurable set Z: $s(z) = \frac{\partial \log \tilde{p}_\epsilon(z)}{\partial \epsilon}\Big|_{\epsilon=0} = \nabla_z \log p(z)$.

$E[s_X(x)] = 0$. Therefore, $s_{Y|X}(x,y) \perp s_X(x)$.

PROOF A.A.1.3 *The target is to prove the expectation* $E[s_{Y|X}(x,y)s_X(x)] = 0$. *Rewrite the expectation with the law of conditional expectation:*

$$\begin{aligned} E[s_{Y|X}(x,y)s_X(x)] &= E[E[s_{Y|X}(x,y)s_X(x) \mid X]] \\ &= E[s_X(x)E[s_{Y|X}(x,y) \mid X]] = E[s_X(x) \cdot 0] = 0. \end{aligned}$$

With the orthogonal relationship between $s_{Y|X}(x,y)$ and $s_X(x)$, we could have $s_{X,Y}(x,y) = s_{Y|X}(x,y) + s_X(x)$ and $\nabla_{s_{X,Y}}\psi = \nabla_{s_{Y|X}}\psi + \nabla_{s_X}\psi$. Therefore, we could calculate the score function for the marginal distribution $P(X)$ and the conditional distribution $P(Y|X)$ and sum them separately to get the score function for the joint distribution if the score function for the joint distribution is hard to capture.

PROOF A.A.1.4 *The joint distribution* $P(Y,X)$ *can be expressed using the chain rule of probability:*

$$p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x) \iff \log p_{X,Y}(x,y) = \log p_{Y|X}(y|x) + \log p_X(x)$$

Applying the definition of the score function, we differentiate both sides with respect to ϵ *and evaluate at* $\epsilon = 0$*:*

$$s_{X,Y}(x,y) = \frac{\partial \log \tilde{p}_{X,Y,\epsilon}(x,y)}{\partial \epsilon}\Big|_{\epsilon=0} = \frac{\partial \log \tilde{p}_{Y|X,\epsilon}(y|x)}{\partial \epsilon}\Big|_{\epsilon=0} + \frac{\partial \log \tilde{p}_{X,\epsilon}(x)}{\partial \epsilon}\Big|_{\epsilon=0}$$

Thus:

$$s_{X,Y}(x,y) = s_{Y|X}(x,y) + s_X(x) \tag{A.A.1.6}$$

By the definition of the gradient in the direction of the score function, we have:

$$\nabla_{s_{X,Y}}\psi = \lim_{\epsilon\to 0}\frac{\psi(\tilde{P}_\epsilon(X,Y)) - \psi(P(X,Y))}{\epsilon}$$

Since $p_{X,Y}(x,y) = p_{Y|X}(x,y)p_X(x)$, we have:

$$\nabla_{s_{X,Y}}\psi = \lim_{\epsilon\to 0}\frac{\psi(\tilde{P}_\epsilon(X,Y)\tilde{P}_\epsilon(X)) - \psi(P(Y|X)P(X))}{\epsilon}$$

Given that ψ is influenced by the score functions $s_{Y|X}$ and s_X independently, we have:

$$\nabla_{s_{X,Y}}\psi = \lim_{\epsilon\to 0}\frac{\psi(\tilde{P}_\epsilon(Y|X)) - \psi(P(Y|X)}{\epsilon} + \lim_{\epsilon\to 0}\frac{\psi(\tilde{P}_\epsilon(X)) - \psi(P(X))}{\epsilon}$$

Thus, we get:

$$\nabla_{s_{X,Y}}\psi = \nabla_{s_{Y|X}}\psi + \nabla_{s_X}\psi \tag{A.A.1.7}$$

A.2. Appendix II: Proof on Asymptotic Linearity and Influence Functions

Central Identity for Influence Functions Recall our definition of the gradient in the direction of the score function on the estimator, we have:

$$\nabla_s\psi(P) = \lim_{\epsilon\to 0}\frac{\psi(\tilde{P}_\epsilon) - \psi(P)}{\epsilon} = \frac{\partial}{\partial\epsilon}\psi(\tilde{P}_\epsilon)\Big|_{\epsilon=0}$$

Similarly, using distributional Taylor expansion, we have:

$$\psi(\tilde{P}_\epsilon)\Big|_{\epsilon=0} \approx \psi(P) + \epsilon\,\frac{\partial\psi}{\partial\epsilon}\Big|_{\epsilon=0} + O(\epsilon)$$

And use the chain rule to analyze the first-order term :

$$\frac{\partial\psi}{\partial\epsilon}\Big|_{\epsilon=0} = \int\frac{\delta\psi}{\delta p(z)}\,\frac{\partial\tilde{p}_\epsilon(z)}{\partial\epsilon}\Big|_{\epsilon=0}dz.$$

Based on Equation A.A.1.3 describing the relationship between $p(z)$ and $\tilde{p}_\epsilon(z)$:$\tilde{p}_\epsilon(z) = (1+\epsilon s(z))\, p(z)$, we have:

$$\frac{\partial \tilde{p}_\epsilon(z)}{\partial \epsilon}\bigg|_{\epsilon=0} = \frac{\partial}{\partial \epsilon}\left(1+\epsilon s(z)\right) p(z)\bigg|_{\epsilon=0} = s(z)p(z).$$

Therefore,

$$\frac{d\psi}{d\epsilon}\bigg|_{\epsilon=0} = \int \frac{\delta \psi}{\delta p(z)} s(z)p(z)\, dz.$$

Note the right side of the expression can be expressed as the form of expectation under the original distribution $p(z)$:

$$\int \frac{\delta \psi}{\delta p(z)} s(z)p(z)\, dz = E_P\left[\frac{\delta \psi}{\delta p(z)} s(z)\right].$$

As we mentioned, the form $\frac{\delta \psi(P)}{\delta p(z)}$ can be regarded as the continuous form of the influence function $\phi(\psi; P; z)$[3]. Therefore, we finally obtain:

$$\nabla_s \psi(P) = E_P\left[\phi(\psi; P; z) s(z)\right]. \tag{A.A.2.8}$$

A.3. Appendix III: Proof during Efficient Influence Function Derivation

Proof Sketch on the Cramer-Rao Bound

PROOF A.A.3.1 (Proof sketch on the Cramer-Rao Bound) *Consider a regular one-dimensional parametric model* $\{P_\psi\}$ *with score* $s_\psi(Z) = \partial \log p_\psi(Z)/\partial \psi$ *and Fisher information* $I(\psi) = E_\psi[s_\psi(Z)^2]$. *If* $\hat{\psi}$ *is unbiased for* ψ, *then* $E_\psi[\hat{\psi}] = \psi$. *Differentiating this identity with respect to* ψ *and using the regularity conditions that justify interchanging differentiation and integration gives*

$$1 = \frac{\partial}{\partial \psi} E_\psi[\hat{\psi}] = E_\psi\left[(\hat{\psi}-\psi) s_\psi(Z)\right].$$

[3] In some literature, the definition of the influence function with the expectation expression is the score-based definition of the influence function, and the score function can also be regarded as $s_0(z) = \frac{\partial}{\partial \epsilon} \log[p(z) + \epsilon(\tilde{p}_\epsilon(z) - p(z))]\big|_{\epsilon=0} = \frac{\tilde{p}_\epsilon(z)-p(z)}{p(z)}$, which is the same as the definition in Equation 1.III.13.

By the Cauchy-Schwarz inequality,

$$1^2 \leq E_\psi[(\hat{\psi}-\psi)^2]E_\psi[s_\psi(Z)^2] = \mathrm{Var}_\psi(\hat{\psi})I(\psi).$$

Therefore,

$$\mathrm{Var}_\psi(\hat{\psi}) \geq \frac{1}{I(\psi)}.$$

Proof on EIF on Tangent Space A crucial property for the EIF $\phi^\dagger$ is that it lies in the tangent space. This is because only the influence function in the tangent space obtains the lowest variance: $\phi^\dagger \perp h^\perp$, where $h^\perp$ stands for the lines orthogonal to any element in the tangent space $h \in T$.

PROOF A.A.3.2 *We suppose we could decompose the influence function into two orthogonal parts: one is the projection on the tangent space h, and one is orthogonal to the tangent space, $\phi - h$. For any score function s on the tangent space, according to the definition, we have: $E[(\phi-h)s] = 0$* [4]*. Thus, for the variance of the influence function:*

$$\begin{aligned} Var(\phi) &= Var(h+(\phi-h)) = Var(h) + Var(\phi-h) \text{ (orthogonality, no covariance)} \\ &= E[h^2] - (E[h])^2 + E[(\phi-h)^2] + (E[\phi-h])^2 \\ &= E[h^2] + E[(\phi-h)^2] \end{aligned}$$

Since h is constant (as it is the projection of the influence functions onto the tangent space), the target function $\phi^\dagger = \arg\min_\phi(\phi-h)^2 \Rightarrow \phi^\dagger = h$. Therefore, the influence function should lie on the tangent space to be efficient.

Therefore, EIF can also be called as the "**canonical gradient** ".

[4]The influence function and the score function are both Hilbert space (since they are both L_2^0 space as we mentioned before, and we could also use the inner product of Hilbert space to represent it, $\langle \phi - h, s\rangle = 0, \forall s \in T(P)$. The orthogonal decomposition is justified by the Riesz Representation Theorem, which states that for every continuous linear functional F on the Hilbert space H, there is always an element $h \in H$ such that $F(f,h) = \langle f, h\rangle \ \forall f \in H$.

EIF for Unconditional Mean We first derive the EIF for $\psi(P) = E_P[X] = \int x dP(x)$. We start with the Gateaux derivative definition of the influence function. The perturbed distribution is set as:

$$\tilde{P}_\epsilon = (1-\epsilon)P + \epsilon\delta_x$$

Therefore, we construct $\psi(\tilde{P}_\epsilon)$:

$$\begin{aligned}\psi(\tilde{P}_\epsilon) &= \int x d\tilde{P}_\epsilon = \int x d\big((1-\epsilon)P + \epsilon\delta_x\big) \\ &= (1-\epsilon)\int x dP + \epsilon \int x d\delta_x \\ &= (1-\epsilon)\psi(P) + \epsilon x \\ &= \psi(P) + \epsilon(x - \psi(P))\end{aligned}$$

Since $\int d\delta_x = 1$. Thus, we have the influence function $\phi(\psi = E_P[X]; P, x)$:

$$\phi(E_P[X]) = \frac{\partial}{\partial\epsilon}\psi(\tilde{P}_\epsilon)\big|_{\epsilon=0} = x - \psi(P) = x - E_P[X] \tag{A.A.3.9}$$

EIF for Conditional Expectation We then derive the EIF for $\psi(P) = E_P[Y|X=x] = \int_y y dP(y|x)$. Still, $\tilde{P}_\epsilon = (1-\epsilon)P + \epsilon\delta_{y|x}$. The hard part we need to derive here is $P(y|x)$. We may recall the Bayesian rule $P(y|x) = \frac{P(y,x)}{P(x)}$. Therefore,

$$\tilde{P}_\epsilon(y|x) = \frac{\tilde{P}_\epsilon(y,x)}{\tilde{P}_\epsilon(x)} = \frac{(1-\epsilon)P(y,x) + \epsilon\delta_{y,x}}{(1-\epsilon)P(x) + \epsilon\delta_x}$$

given by the equation above. Since the influence function defined as $\frac{\partial\psi(\tilde{P}_\epsilon)}{\partial\epsilon}\big|_{\epsilon=0}$ is a gradient, we first deal with the part to be integrated ($d\tilde{P}_\epsilon(y|x)$), which we have transformed with the above equation. Recall the gradient chain rule for division: for two functions u and v, $\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}$. Therefore,

$$\tilde{P}_\epsilon(y|x)\big|_{\epsilon=0} = \frac{[(1-\epsilon)P(y,x) + \epsilon\delta_{y,x}]'[(1-\epsilon)P(x) + \epsilon\delta_x] - [(1-\epsilon)P(y,x) + \epsilon\delta_{y,x}][(1-\epsilon)P(x) + \epsilon\delta_x]'}{[(1-\epsilon)P(x) + \epsilon\delta_x]^2}$$

$$= \frac{[\delta_{y,x} - P(y,x)][(1-\epsilon)P(x) + \delta_x] - [\delta_x - P(x)][(1-\epsilon)P(y,x) + \delta_{y,x}]}{[(1-\epsilon)P(x) + \epsilon\delta_x]^2}$$
$$= \frac{P(y,x)\delta_x - \delta_{y,x}P(x)}{P(x)^2} \text{ (since } \epsilon = 0)$$

Therefore, The EIF is:

$$\begin{aligned}
\frac{\partial \tilde{P}_\epsilon}{\partial \epsilon}\Big|_{\epsilon=0} &= \int y d\Big[\frac{P(y,x)\delta_x - \delta_{y,x}P(x)}{P(x)^2}\Big] \\
&= \frac{\delta_x}{p(x)}\Big[\int y d\frac{\delta_{y,x}}{\delta_x} - \int y d\frac{P(y,x)}{P(x)}\Big] \\
&= \frac{\mathbb{1}_x}{p(x)}\Big[\int y d\delta_{y|x} - \int y dP(y|x)\Big] \text{ (Since } \frac{P(y,x)}{P(x)} = P(y|x); \int d\delta = \mathbb{1}) \\
&= \frac{\mathbb{1}_x}{p(x)}[y - E_P[Y|X=x]]
\end{aligned}$$

We get:

$$\phi(\psi = E_P[Y|A=a, X=x], P_X(y,a,x)) = \frac{\mathbb{1}_{(a,x)}}{p[A=a,x]}\Big[y - E_P[Y|A=a,X]\Big] \tag{A.A.3.10}$$

Equation 1.III.19 is the EIF for the conditional expectation.

Validation on the EIF for the ATE with Central Identity We could use the central identity for the influence function (Equation 1.III.14) to verify if our EIF is correct.

PROOF A.A.3.3 *Suppose we are verifying the EIF for ψ_0. Using the factorizing of the tangent space technique, we could decompose the gradient of the estimator in any direction into:*

$$\nabla_s \psi_0 = \nabla_{s_{Y|A,X}} \psi_0 + \nabla_{s_{A|X}} \psi_0 + \nabla_{s_X} \psi_0$$

Similarly, for the expectation of the influence and the score function, we could also decompose it through the algebra of expectations:

$$E[\phi s] = E[\phi s_{Y|A,X}] + E[\phi s_{A|X}] + E[\phi s_X]$$

Therefore, we need to prove that $\nabla_{s_{Y|A,X}}\psi_0 = E[\phi s_{Y|A,X}]$, $\nabla_{s_{A|X}}\psi_0 = E[\phi s_{A|X}]$, *and* $\nabla_{s_X}\psi_0 = E[\phi s_X]$, *correspondingly. We begin with* $\nabla_{s_X}\psi_0 = E[\phi s_X]$.

$$\begin{aligned}\nabla_{s_X}\psi_0 &= \frac{\partial}{\partial\epsilon}\int_x\int_y \tilde{p}_\epsilon(y|0,x)\,dy\,\tilde{p}_\epsilon(x)dx = \int_x\int_y p(y|0,x)\,dy\frac{\partial}{\partial\epsilon}(1+\epsilon s)\,p(x)dx \\ &= \int_x\int_y p(y|0,x)\,dy\,s_X\,p(x)dx = \int_x \mu_0(x)\,s_X\,p(x)dx\end{aligned}$$

And,

$$\begin{aligned}E[\phi s_X] &= E\left[\left(\frac{\mathbb{1}(A=0)}{\pi_0(x)}[y-\mu_0(x)]+\mu_0(x)\right)s_X\right] \\ &= \int_x\int_y\left[\left(\frac{\mathbb{1}(A=0)}{\pi_0(x)}[y-\mu_0(x)]+\mu_0(x)\right)s_X\right]p(y|x,a)\,dy\,p(x)\,dx \\ &= \int_x s_X\left[\int_y \frac{\mathbb{1}(A=0)}{\pi_0(x)}[y-\mu_0(x)]\,p(y|x,a)\,dy + \mu_0(x)\int_y p(y|x,a)\,dy\right]p(x)\,dx\end{aligned}$$

Notice that $\int_y p(y|x,a)\,dy = 1$ *and* $\int_y [y-\mu_0(x)]\,p(y|x,a)\,dy = E[y-\mu_0(x)|x,0] = 0$, *thus,*

$$E[\phi s_X] = \int_x s_X\mu_0(x)\,p(x)dx = \nabla_{s_X}\psi_0$$

Similarly,

$$\begin{aligned}\nabla_{s_{A|X}}\psi_0 &= \int_x\int_y p(y\mid 0,x)\,dy\frac{\partial}{\partial\epsilon}(1+\epsilon s_{A|X})\,p(x)d(x) \\ &= \int_x \mu_0 s_{A|X}\,p(x)dx = 0\end{aligned}$$

$$\begin{aligned}E[\phi s_{A|X}] &= \int_x\int_a \phi s_{A|X}(a|x)\,p(a|x)\,p(x)\,da\,dx \\ &= \int_x \phi\underbrace{\left(\int_a s_{A|X}(a|x)\,p(a|x)\,da\right)}_{=0}p(x)dx \\ &= 0 = \nabla_{s_{A|X}}\psi_0\end{aligned}$$

At last,

$$\nabla_{s_{Y|A,X}}\psi_0 = \int_x\int_y p(y|x,0)\,s_{Y|A,X}\,dy\,p(x)dx$$

$$= \int_x p(x)\Big(\int_y p_0(y|x) s_{Y|A,X} dy\Big) dx$$
$$= \int_x (y|A=0) s_{Y|A,X} dx$$

$$E[\phi s_{Y|A,X}] = \int_x \int_y \left[\left(\frac{\mathbb{1}(A=0)}{\pi_0(x)}[y-\mu_0(x)] + \mu_0(x)\right) s_{Y|A,X}\right] p(y|x,a)\, dy\, p(x)\, dx$$
$$= \int_x \left[\frac{\mathbb{1}(A=0)}{\pi_0(x)} y s_{Y|A,X}\right] p(x) dx$$
$$= \int_x (y|A=0) s_{Y|A,X} p(x) dx = \nabla_{s_{Y|A,X}} \psi_0.$$

Therefore, $E[\phi s] = \nabla_s \psi$ and our derivation on the EIF is correct.

project the influence function on T_X First, we try to project the influence function of the IPW estimator on the tangent space T_X. The projection function is defined to find the score function on the tangent space for which its mean square error with the influence function from the IPW is minimal[5]. For any influence function,

$$Proj_{\mathrm{T}_X}(\phi) = \arg\min_{h(X)\in \mathrm{T}_X} E[(\phi - h(X))^2].$$

Let $h^\dagger(X) = \arg\min_{h(X)\in \mathrm{T}_X} E[(\phi - h(x)^2]$, therefore, the residual $\phi - h^\dagger(X)$ should be orthogonal to functions on T_X:

$$E[(\phi - h^\dagger(X))h(X)] = 0, \forall h(X) \in \mathrm{T}_X.$$

Since $h^\dagger(X)$ satisfies the orthogonality condition, it should be the conditional expectation of ϕ given X:

$$h^\dagger(X) = E[\phi|X].$$

[5]We have a sketch Figure A.1 illustrating the projection process for the readers' reference for understanding the algebraic process here.

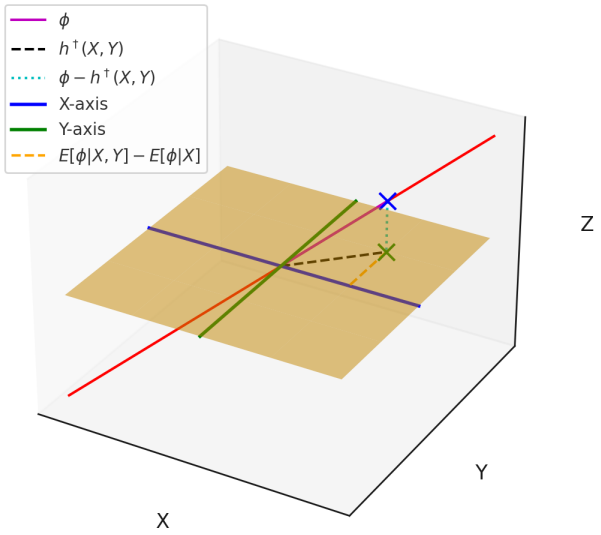


Figure A.1: Illustrations on projections of influence functions

Note: The yellow plane denotes the tangent space. The figure shows projecting the (inefficient) influence function (not on the tangent space) towards the joint distribution of (X, Y) and the marginal distribution of $(Y|X = 0)$.

Therefore, we have the projection of the influence function for the IPW estimator on the tangent space T_X as its conditional expectation on the X axis:

$$\begin{aligned}
\phi^{\dagger}_{0\langle \mathrm{T}_X \rangle} &= E[\phi_0^{IPW} \mid X] = E\left[\left(\frac{\mathbb{1}(A_i = 0)}{\pi_0(X_i)} Y_i - \psi_0\right) \mid X\right] \\
&= E\left[E\left[\frac{\mathbb{1}(A_i = 0)}{\pi_0(X_i)} Y_i \mid A_i, X\right] \mid X\right] - \psi_0 \\
&= E\left[E\left[\frac{Y_i}{\pi_0(X_i)} \mid A_i = 0, X\right] \mathbb{1}(A_i = 0) \mid X\right] - \psi_0 \\
&= E\left[Y_i | A_i = 0, X\right] - \psi_0 = \mu_0(X_i) - \psi_0
\end{aligned}$$

A.4. Appendix IV: Proof on Second and higher Order Term Convergence

Proof on Multidimensional CLT We first have the lemma on the multidimensional central limit theorem (CLT): proving the multidimensional CLT requires prior knowledge of characteristic functions, Levy's Continuity Theorem, and the Cramer-Wold device. The characteristic function of the real-valued random variable defines the probability distribution– in other words, two distinct distributions with the same characteristic function are identically the same distribution. Levy's continuity theorem states that if the characteristic functions of a sequence of random variables pointwise converge towards the characteristic function of a limiting random variable, then the sequence of random variables converges in distribution towards the limiting random variable. Cramer-Wold device states that a sequence of random variables converging to a limiting random variable is equivalent to the scalar of the sequence of random variables converging to the scalar of the limiting random variable. With these backgrounds, we have the proof:

We first define the normalized sample mean vector: given that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where X_i are i.i.d. random variables with mean $\mu = E[X_i]$ and covariance matrix $\Sigma = E[(X_i - \mu)(X_i - \mu)^T]$, let $Z_n = \sqrt{n}(\bar{X}_n - \mu)$. Then we find the characteristic function of Z_n:

$$\varphi_{Z_n}(t) = E\left[e^{it^T Z_n}\right] = E\left[e^{it^T\sqrt{n}(\bar{X}_n-\mu)}\right] = E\left[e^{it^T\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i-\mu\right)}\right]$$

Using the Taylor expansion of the exponential function $e^{ix} \approx 1 + ix - \frac{x^2}{2}$, we get:

$$E\left[e^{it^T\frac{X_i-\mu}{\sqrt{n}}}\right] \approx 1 + it^T\frac{E[Y_i]}{\sqrt{n}} - \frac{1}{2}t^T\frac{E[(X_i-\mu)(X_i-\mu)^T]}{n}t$$

Since $E[X_i - \mu] = 0$ and $E[(X_i - \mu)(X_i - \mu)^T] = \Sigma$, this simplifies to:

$$E\left[e^{it^T\frac{X_i-\mu}{\sqrt{n}}}\right] \approx 1 - \frac{1}{2}t^T\frac{\Sigma}{n}t$$

Therefore,

$$\varphi_{Z_n}(t) \approx \left(1 - \frac{1}{2} t^T \frac{\Sigma}{n} t\right)^n \to e^{-\frac{1}{2} t^T \Sigma t} \text{ as } n \to \infty$$

By Levy's Continuity Theorem, this implies that Z_n converges in distribution to a multivariate normal distribution:

$$Z_n \xrightarrow{d} N(0, \Sigma)$$

Therefore, we have proved the multidimensional Central Limit Theorem:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma)$$

For the specific details, see van der Vaart 1998: Ch.2, pp.12-16.

Empirical Process Convergence We can prove the effect of sample splitting on controlling the convergence rate of the empirical influence function towards the influence function from the estimand with the multidimensional CLT. To make the case simple, consider splitting the dataset $(Z_1, \ldots, Z_n)$ only into two subsets: $S_1 = \{Z_1, Z_2, \ldots, Z_{n_1}\}$ be the first subset, and $S_2 = \{Z_{n_1+1}, Z_{n_1+2}, \ldots, Z_n\}$ be the second subset.

We first use S_1 to construct an initial estimator $\hat{\psi}_0$ of the parameter ψ, which, presumably, is the consistent estimator of the estimand $\hat{\psi}$. We may write the one-step estimator with sample splitting as:

$$\hat{\psi}_{\text{1-step}} = \hat{\psi}_0 + \frac{1}{n_2} \sum_{i \in S_2} \phi(\hat{\psi}_0, \mathbb{P}_{n_2}, Z_i)$$

Notice that for the naive estimator, we use the result from S_1, while for the first-order bias correction term, we use estimators from S_2 (as n_2 is the size of the split S_2, and $\mathbb{P}_{n_2} = \frac{1}{n_2}\sum_{i\in S_2}\delta_{Z_i}$ is the empirical measure based on S_2). Obviously, the one-step estimator is consistent, as the law of large numbers indicates $\hat{\psi}_0 \xrightarrow{prob.} \psi$ and the correction term $\frac{1}{n_2}\sum_{i\in S_2}\phi(\hat{\psi}_0, \mathbb{P}_{n_2}, Z_i)$ will converge in probability to zero if $\phi(\psi, \mathbb{P}_n, Z_i)$ is well-defined. Meanwhile, since the naive estimator and the first-order bias correction term come from independent datasets, the construction of $\hat{\psi}_0$ is irrelevant to the data used to make the correction.

Now we turn to the asymptotic normality. Due to the CLT, we have:

$$\frac{1}{\sqrt{n_2}}\sum_{i\in S_2}\phi(\psi, \mathbb{P}_{n_2}, Z_i) \xrightarrow{d} \mathrm{N}(0, \sigma^2)$$

Since $\hat{\psi}_0 \xrightarrow{prob.} \psi$ and $\phi(\hat{\psi}_0, \mathbb{P}_{n_2}, Z_i) \approx \phi(\psi, \mathbb{P}_{n_2}, Z_i)$, we have:

$$\frac{1}{\sqrt{n_2}}\sum_{i\in S_2}\phi(\hat{\psi}_0, \mathbb{P}_{n_2}, Z_i) \xrightarrow{d} \mathrm{N}(0, \sigma^2) \quad \text{(Slutsky's Theorem)}$$

Therefore, the one-step estimator with the sample splitting method is asymptotically normal with mean ψ and variance σ^2/n_2, and the control on the convergence rate at $1/\sqrt{n}$ is accomplished.

In summary, when the second (and higher) order remainders and the empirical process become insignificant under the large sample size scenario, we can use the one-step estimator: the empirical naive estimator plus the first-order bias correction term to robustly and efficiently infer the true estimand.

Proof on Convergence for DR Estimator on the ATE From the perspective of the convergence rate, we will find that the estimator also satisfies double robustness with its higher-order Distributional Taylor Expansion. According to Equation 1.III.16, We can derive the second-order remainder of the estimator ψ_a :

$$
\begin{aligned}
R_2 &= \psi(\tilde{P}_\epsilon) - \psi(P) + \frac{\partial}{\partial \epsilon}\psi(\tilde{P}_\epsilon)\Big|_{\epsilon=1} \\
&= \psi(\tilde{P}_\epsilon) - \psi(P) + \frac{1}{n}\sum_{i=1}^{n}\phi(\psi(\tilde{P}_\epsilon)) \\
&= \hat{\psi}_a - E[\mu_a(X_i)] + E\left[\frac{\mathbb{1}(A_i = a)}{\hat{\pi}_a(X_i)}(Y_i - \hat{\mu}_a(X_i)) + \hat{\mu}_a(X_i)\right] - E\left[\hat{\psi}(a)\right] \\
&= E\left[\frac{\pi_a(X_i)}{\hat{\pi}_a(X_i)}\mu_a(X_i) - \mu_a(X_i)\right] + E\left[\hat{\mu}_a(X_i) - \frac{\pi_a(X_i)}{\hat{\pi}_a(X_i)}\hat{\mu}_a(X_i)\right] \\
&= E\left[\frac{1}{\hat{\pi}_a(X_i)}\left(\pi_a(X_i) - \hat{\pi}_a(X_i)\right)\left(\mu_a(X_i) - \hat{\mu}_a(X)\right)\right] \\
&\leq ||\pi_a(X_i) - \hat{\pi}_a(X_i)||\,||\mu_a(X_i) - \hat{\mu}_a(X_i)|| \text{ (Cauchy-Schwarz Inequality)}
\end{aligned}
$$

As R_2 denotes the second-order remainder. We could see from this formation that when either $\pi_a(X_i) = \hat{\pi}_a(X_i)$ or $\mu_a(X_i) - \hat{\mu}_a(X_i)$, the second-order remainder will be zero. Therefore, when either $\hat{\pi}_a(X_i)$ or $\hat{\mu}_a(X_i)$ is correctly specified, our estimator is unbiased.

Further, this inequality also suggests the convergence rate of the doubly robust estimator. Since we require the second-order remainder to converge at the rate of $o_p(n^{-1/2})$, we could require both $||\pi_a(X_i) - \hat{\pi}_a(X_i)||$ and $||\mu_a(X_i) - \hat{\mu}_a(X_i)||$ to converge at the rate of $o_p(n^{-1/4})$. Regularized machine learning methods and cross-validation need to be used for both the propensity score model $\hat{\pi}_a(X_i)$ and the outcome model $\hat{\mu}_a(X_i)$. For instance, we could choose gradient boosting machines or random forests to predict the propensity and generalized additive models (GAM) or random forests to predict the outcome[6], and then

[6] A pretty useful programming package for model choice is called super-learner(Polley et al. 2023), which could be convenient for social scientists choose the appropriate machine learning models for model fitting.

use regularized techniques like lasso, ridge, or elastic nets to control the complexity of the models.

Appendix B

Appendix to Chapter 3

B.1. Appendix V: Results from Marginal Hazard Ratio Models

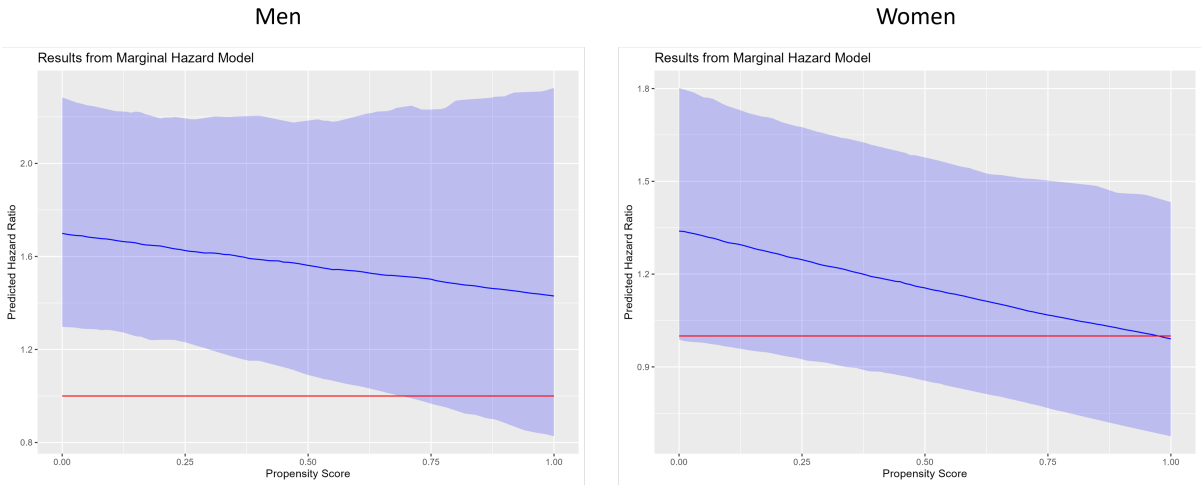


Figure B.1: Heterogeneous Treatment Effect Using Marginal Hazard Models
Note: The red line indicates the hazard ratio equals one, while the blue ribbon indicates the confidence interval for the heterogeneous treatment effect of widowhood.

According to Figure B.1, in the HRS data we use, the average hazard ratio of mortality for widowed against non-widowed for men is around 1.6, and for women is around 1.2, indicating that although both widowed men and widowed women suffer widowhood penalty in mortality risks, men on average suffer more than women, which is in agreement with results from the previous literature and our findings in the main paper. We also find that for

women with the highest preparedness scores, the marginal hazard ratio between widowed and non-widowed approaches one, indicating that the widowhood penalty at this end is eliminated. The result also verifies our findings in the main paper.

B.2. Appendix VI: Technical Details for Survival Function Estimation

For Cox Proportional Hazard (Cox-PH) models, we first suppose the covariates do not change with time; in general, at time t, the functional parametric model is written as:

$$\frac{h(t \mid X)}{h_0(t)} = exp(X\beta) \tag{B.B.2.1}$$

On the left side of Equation B.B.2.1, $h(t \mid X)$ indicates the conditional hazard function at time t conditioned on covariates X. According to the definition, the hazard function represents the instantaneous rate of failure at a given time, while the survival function represents the probability of surviving beyond a given time $S(t) = P(T > t)$. Thus, we have the relationship between $h(t)$ and $S(t)$:

$$h(t) = \frac{f(t)}{S(t)} = \frac{-\frac{dS(t)}{dt}}{S(t)} = -\frac{dS(t)}{S(t)}\frac{1}{dt} = -\frac{d\log S(t)}{dt}; \tag{B.B.2.2}$$

On the right side of Equation B.B.2.1, β represents the coefficients for the covariates.

Therefore, to obtain $\hat{S}(t \mid X)$, we need three steps:

The first step is to estimate $\exp(X\hat{\beta})$. As X is known, the target here is $\hat{\beta}$.To do so, consider the maximum likelihood estimation to yield $\hat{\beta}$ as:

$$\hat{\beta} = \arg\max_{\beta} L(\beta) = \arg\max_{\beta} \prod_{i \in D} \frac{\exp(X\beta)}{\sum_{j \in R_i} \exp(X_j\beta)}$$

In the expression above, D denotes the set of individuals who experienced the event, i, j refers to two different individuals, R_i is the risk set at the time of the event for individual i. By maximizing the likelihood function, we find the estimates of parameters β (simply, we could try with Newton-Raphson algorithm).

Then, we consider the estimation on $\hat{h}_0(t)$, which is the baseline hazard at time t. A simple way to do so is with the Breslow estimator (Breslow 1975). Suppose the cumulative baseline function $H_0(t) = \int_0^t h_0(u)du$. The Breslow estimator assumes that $\hat{H}_0(t)$ is given by:

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

In the equation above, d_i denotes the number of events at time t_i while n_i represents the number at risk just before time t_i. Then we could have $\hat{h}_0(t) = \frac{d\hat{H}_0(t)}{dt}$.

The final step is to transfer $\hat{h}(t)$ to $\hat{S}(t)$. According to Equation B.B.2.2, we have:

$$\hat{S}(t) = \exp\left(-\int_0^t \hat{h}(u)du\right).$$

Now, we consider the covariates in the model to be time-varying. Therefore, the parametric model for the hazard function should be:

$$\frac{h(t \mid X(t))}{h_0(t)} = exp(X(t)\beta(t))$$

And the likelihood function for $\beta(t)$ is:

$$\hat{\beta(t)} = \arg\max_{\beta} L(\beta(t)) = \arg\max_{\beta} \prod_{i \in D} \frac{\exp(X(t)_i \beta(t))}{\sum_{j \in R_i} \exp(X(t)_j \beta(t))}.$$

In our analysis, we take the fixed-β assumption, assuming that the coefficients β do not change over time. The simplification can still handle some time dependence, but it still requires that the effect of the covariates on the hazard is constant. After estimating $\hat{\beta}(t)$, the second and third steps are the same as in the time-constant models (see Appendix B.5 for the tests).

B.3. Appendix VII: Results from White Subsample

Since in the HRS sample, most respondents (86.73% of men and 86.75% of women, see the descriptive table in Appendix B.4) are white, in order to examine the robustness of the results in the main paper, we present the results from the white subsample. As mentioned in the main paper, except for white women in the lowest strata of assets, patterns in other subgroups are consistent with the results presented in the main paper. We suspect the abnormality of women in the lowest strata is caused by the limited sample size for white women in the lowest strata.

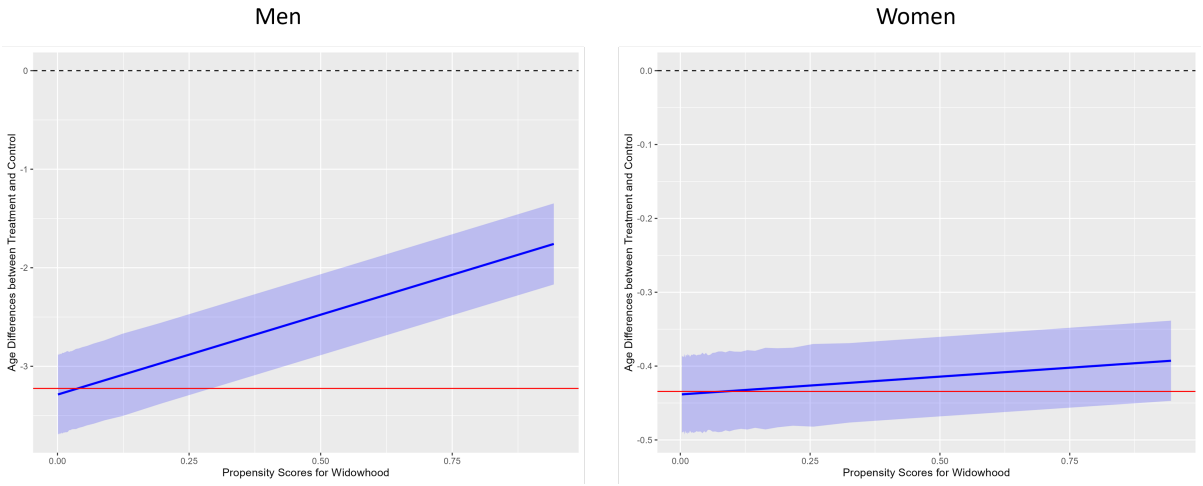


Figure B.2: Average Treatment Effect and Heterogeneous Treatment Effect of Widowhood Effects

Note: Red and blue lines separately indicate the linear approximation of the ATE and the HTE, while the black dash indicates the horizontal line of difference is 0. The blue ribbon denotes the 95% confidence interval for the HTE estimation from 100 bootstraps.

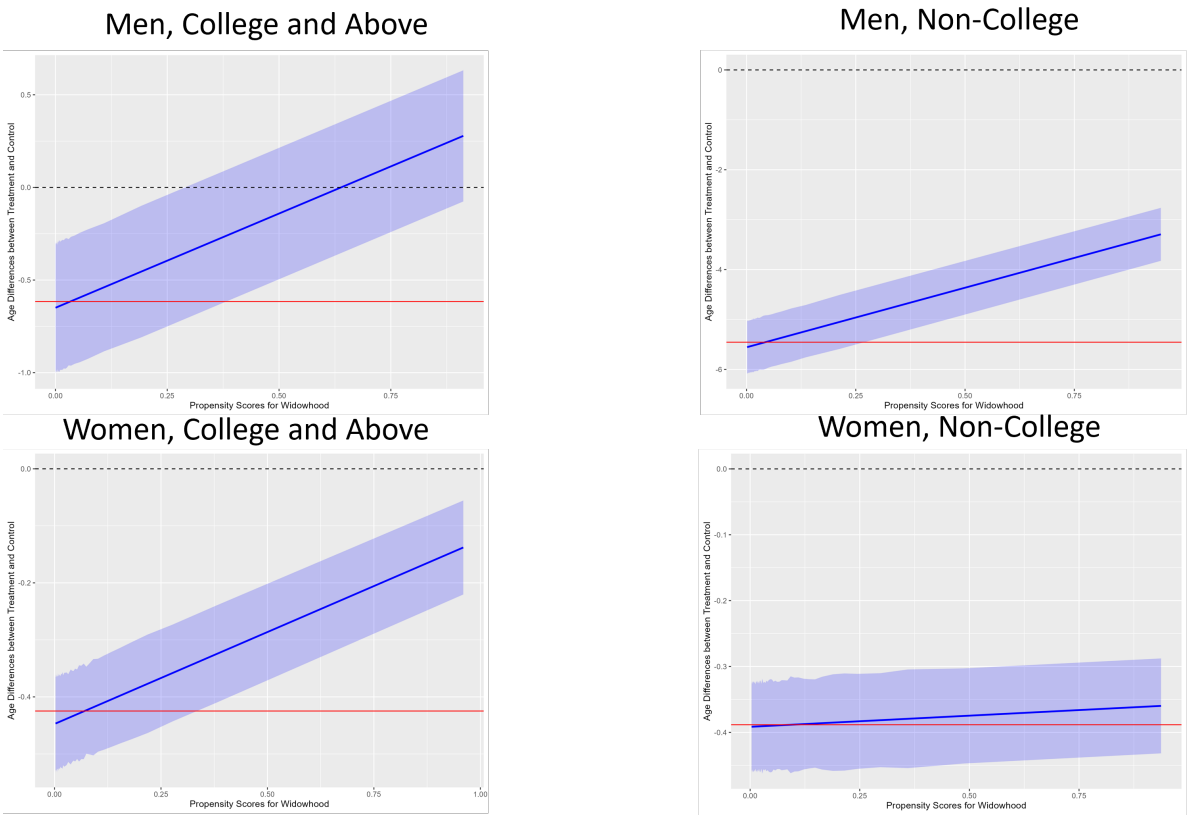


Figure B.3: College Education and Widowhood Effects

Note: Red and blue lines separately indicate the linear approximation of the ATE and the HTE, while the black dash indicates the horizontal line of difference is 0. The blue ribbon denotes the 95% confidence interval for the HTE estimation from 100 bootstraps.

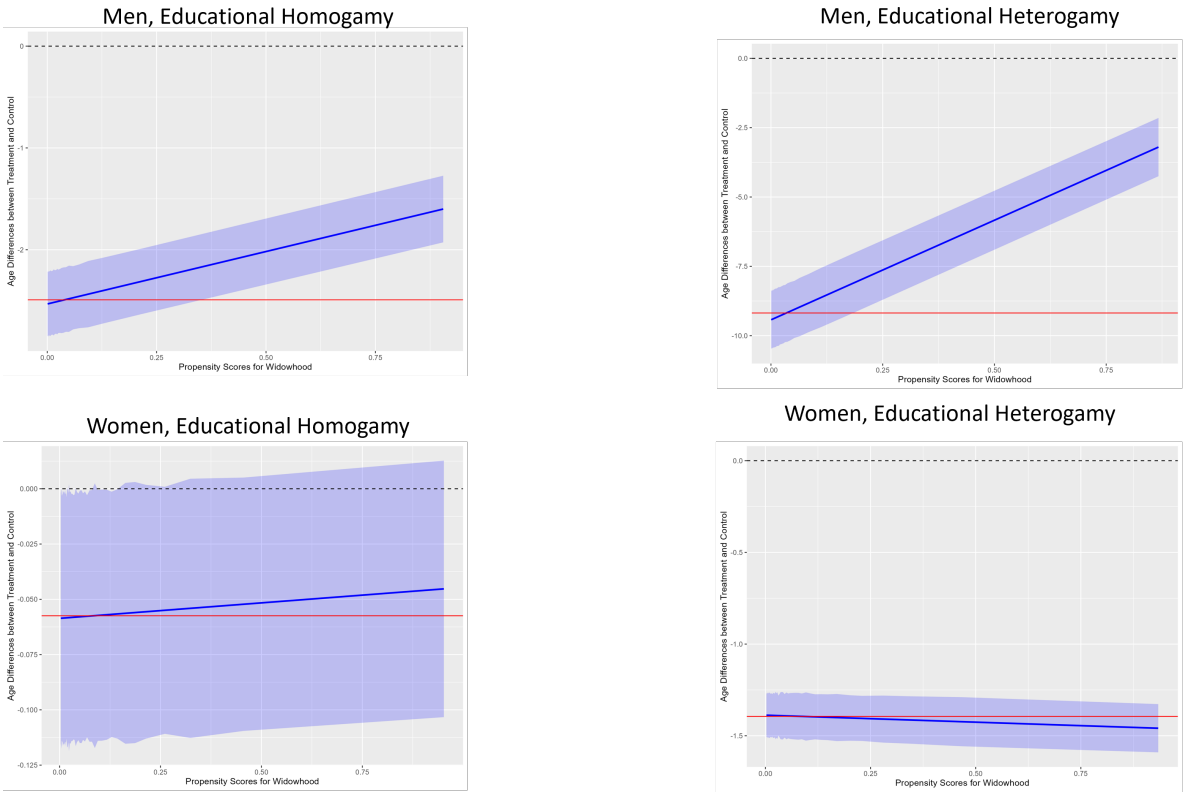


Figure B.4: Educational Homogamy and Widowhood Effects

Note: Red and blue lines separately indicate the linear approximation of the ATE and the HTE, while the black dash indicates the horizontal line of difference is 0. The blue ribbon denotes the 95% confidence interval for the HTE estimation from 100 bootstraps.

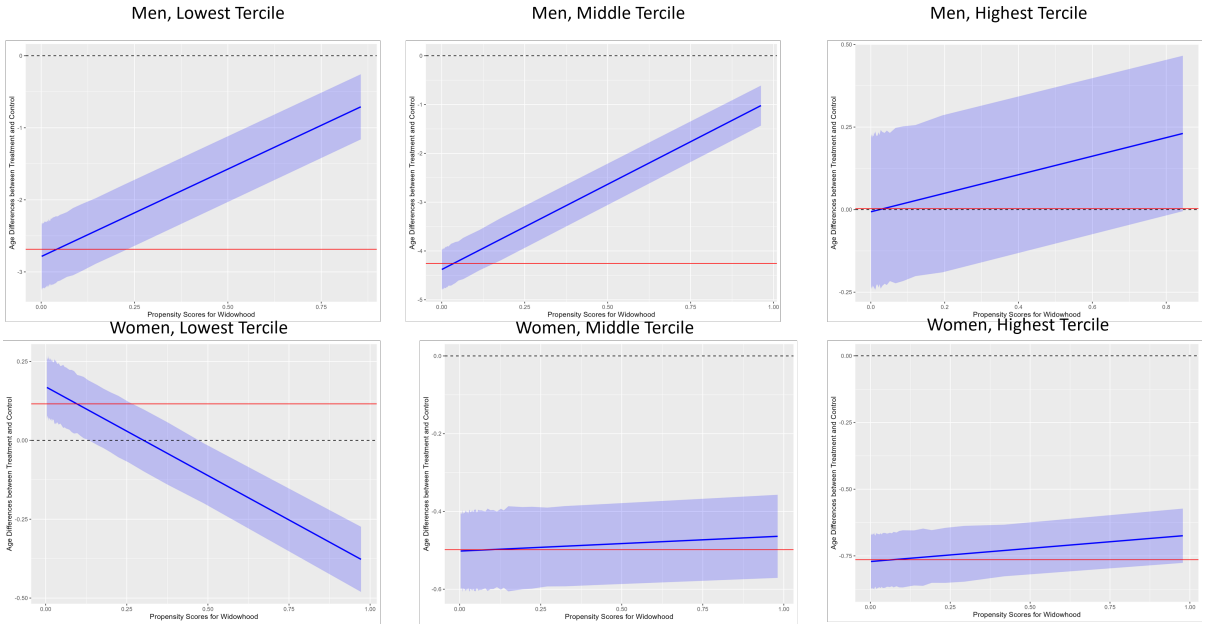


Figure B.5: Wealth and Widowhood Effects

Note: Red and blue lines separately indicate the linear approximation of the ATE and the HTE, while the black dash indicates the horizontal line of difference is 0. The blue ribbon denotes the 95% confidence interval for the HTE estimation from 100 bootstraps.

B.4. Appendix VIII: Heterogeneous Results for Race and Racial Homogamy

In this section, we present the results for race and racial homogamy. Since in our analytical sample of the HRS data, non-white samples only make up 14%, the results may not be stable, so we just put them in the Appendix, and we hope the results could encourage more researchers to dig into the patterns of heterogamous widowhood effects on race.

Like the divergent results on education and the widowhood effects, results for the widowhood effect and race are still inconsistent. If we only compare the widowhood effect between the black and the white, different papers with different datasets and various model settings yield opposite conclusions. Elwert and Christakis (2006) used several longitudinal datasets from the 1990s with continuous-time Cox Proportional Hazard (Cox-PH) models. Men reported that white men suffer significantly higher widowhood effects on mortality, while the hazard ratios between widowed and non-widowed black men are insignificant. Women reported that only white women who lose their white partners suffer significantly higher mortality risks, while the widowhood effects on black women are negligible. On the contrary, Liu et al. (2020) used the Health and Retirement Study (HRS) 1992 – 2016 data and applied the discrete-time (logistic) hazard models. For men, they witnessed significant increases in hazard ratios of death for both black and white widowers, although the two effects are not differentiable. For women, they also reported a significant increase in risks for both black and white widows, but they concluded that black women suffer more from the widowhood effect than white women.

In summary, while Liu et al. (2020) suggested that minorities [1] suffer the widowhood

[1] In Liu et al. (2020) paper, they also reported the results from the Hispanic group. Indeed, they captured

effects more than white people, Elwert and Christakis (2006) implied that white men and women are more vulnerable to the spouse's death while black men and women are unaffected. Interestingly, because of their divergent results, they have disparate explanations of the racial difference in the widowhood effects: Liu et al. (2020) applied a Karlson-Holm-Breen (KHB) decomposition after the comparison and suggested that more considerable financial constraints after widowhood for minorities expose them to higher risks of mortality. Meanwhile, Elwert and Christakis (2006) attributed the resilience to the widowhood effects for the black to marital cultural and marital contextual differences between the black and white subgroups.

Table B.1 presents the descriptive statistics for race and racial homogamy in our sample.

Results in Figure B.6 summarise the average and heterogeneous treatment effect of widowhood on mortality for different gender and racial groups. The figure intuitively elaborates that for both white men and women, better preparedness has a positive impact on shrinking the widowhood penalty, although for white women, the effect is not significant; to our surprise, the more non-white men and women prepared for the spouse's loss, the widowed has lower life expectancy compared to the non-widowed.

We first take a closer look at the average treatment effect. For white men, life expectancy is 3.1 years lower for widowed men than non-widowed men. However, we observed that for non-white men, those who experienced widowhood had an average of 2.7 years longer in life expectancy than those who were not widowed. For women, both white and non-white widows have shorter life expectancies than non-widows: for whites, the average widowhood

the highest and most significant (different from white) widowhood effects for Hispanic men and women.

Table B.1: Sample Size By Gender and Race (HRS 1998 - 2018)

Characteristics	Total	%	Not Widowed Dead	Not Widowed Alive	Widowed Dead	Widowed Alive
Men (N = 6,667)						
White	5782	86.73%	2687	2238	465	392
Non-White	885	13.27%	408	338	70	69
Husband White, Wife White	5637	84.42%	2632	2167	459	379
Husband White, Wife Non-White	145	2.17%	55	71	6	13
Husband Non-White, Wife White	111	1.66%	45	54	6	6
Husband Non-White, Wife Non-White	774	11.59%	363	284	64	63
Women (N = 6,145)						
White	5331	86.75%	1256	1923	880	1272
Non-White	814	13.25%	200	269	124	221
Wife White, Husband White	5205	84.70%	1235	1857	869	1244
Wife White, Husband Non-White	126	2.05%	21	66	11	28
Wife Non-White, Husband White	85	1.38%	18	30	9	28
Wife Non-White, Husband Non-White	729	11.86%	182	239	115	193

effect is 0.44 years, while for non-whites, it is 0.18 years.

We then turn to the heterogeneous treatment effect. For white men, while the worst-prepared widowers have 3.2 years shorter life expectancy than the non-widowers, the best-prepared widowers have around 1.8 years shorter life expectancy, suggesting that preparedness improves widowers' life expectancy by 1.4 years. For white women, better preparedness only mitigates the widowhood effect from -0.45 years to -0.38 years, and the effect is

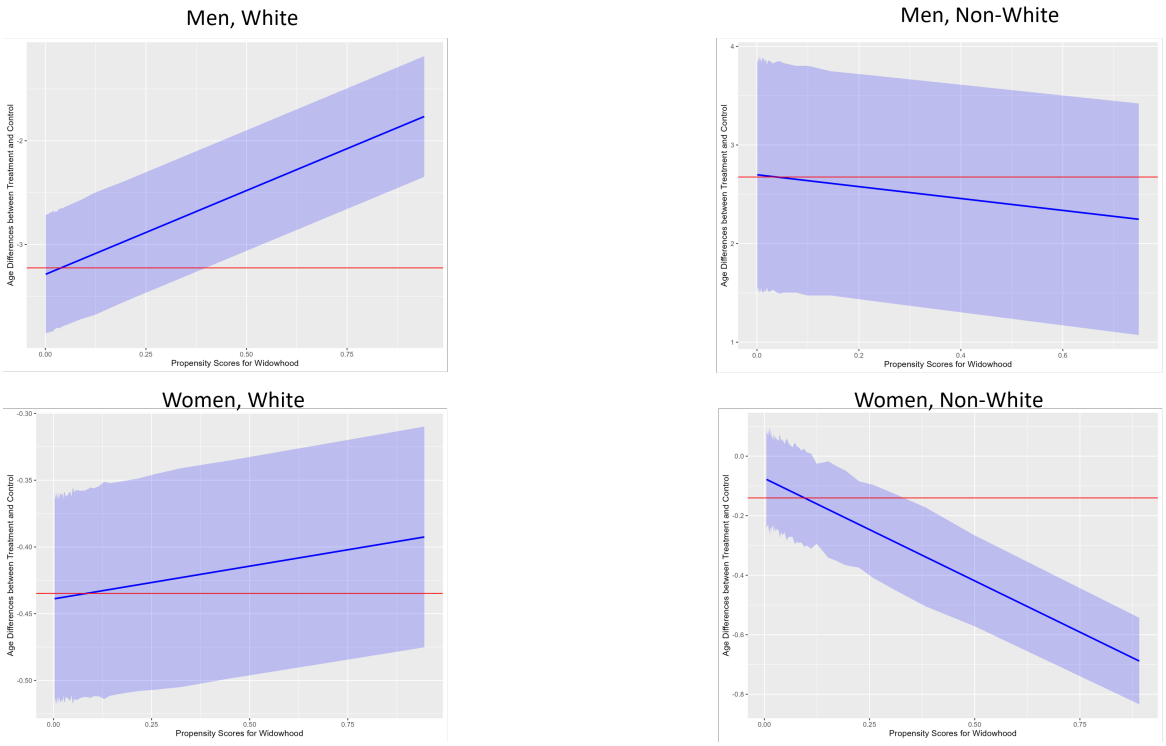


Figure B.6: Race and Widowhood Effects

Note: Red and blue lines separately indicate the linear approximation of the ATE and the HTE. The blue ribbon denotes the 95% confidence interval for the HTE estimation from 100 bootstraps.

not statistically different from the ATE. For non-white men, the worst-prepared widowers surpass the non-widowers in life expectancy by an average of 3.7 years, and the gap is reduced to 2.2 years. However, the shrinkage in the gap is still not significantly different from the ATE. Finally, for non-white women, widowed women have around 0.1 years shorter life expectancy at the worst-prepared end than non-widowed women, but the gap enlarges to 0.7 years on the best-prepared end.

Next, we discuss the heterogeneity in racial homogamy. Since the cases of racial heterogamy and the cases of other racial homogamy (i.e., black-black, Hispanic-Hispanic, Asian-Asian) in our sample are insufficient, we only compare the white-white homogamy with

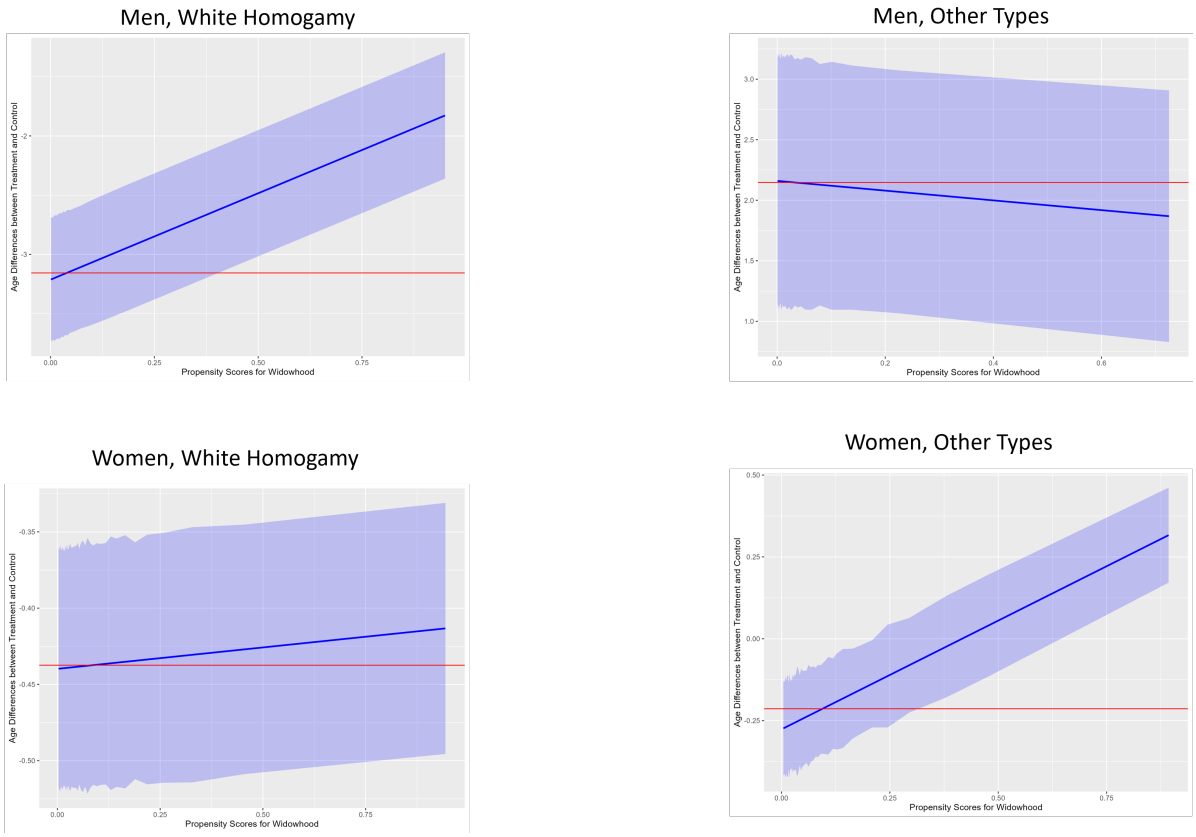


Figure B.7: Racial Homogamy and Widowhood Effects

Note: Red and blue lines separately indicate the linear approximation of the ATE and the HTE. The blue ribbon denotes the 95% confidence interval for the HTE estimation from 100 bootstraps.

other groups. Figure B.7 shows the results. As can be seen from the graph at first glance, except for the non-white-homogamy men, the patterns in all other groups meet our expectation that better preparedness scores indicate more minor widowhood penalties. We describe the patterns in detail as follows.

First, we present the results for the average treatment effect. On the men's side, the white-white homogamous widowers suffer an average penalty of 3.1 years of life expectancy. However, widowed men have an average of 2.2 years longer in life expectancy for other racial marriage groups than non-widowed men, based on our counterfactual estimations.

For women, the widowhood penalty for the white homogamy group is 0.44 years of life expectancy. It is 0.24 years for other groups, suggesting that white-white homogamy women suffer a larger widowhood penalty than the other groups.

Next, we analyze the effect of heterogeneous treatment on the racial homogamous and heterogamous groups. For homogamous men, the increase in preparedness scores for widowhood mitigates the widowhood effect in life expectancy difference from 3.1 years shorter than the non-widowers to 1.8 years shorter. For the non-white-homogamous men, the widowhood effect ranges from 2.3 years longer life expectancy than the non-widowers to 1.8 years longer, although it is not statistically different from the ATE. For white homogamous women, better preparedness scores reduce the widowhood penalty from 0.44 years of life expectancy to 0.41 years, and it is not significantly different from the ATE. For the other women, the worst-prepared widows have widowhood penalties of 0.25 years in life expectancy compared to non-widows, while at the best-preparedness end, widows have a longer life expectancy of 0.3 years than non-widows.

Our comparative analysis of the widowhood effect across diverse racial groups reveals nuanced patterns. For white men and women, better preparedness appears to mitigate the negative impact of widowhood on life expectancy. However, the opposite trend emerges among non-white individuals. We hypothesize that these differences may arise from varying cultural interpretations of widowhood across racial communities. Furthermore, our heterogeneity analysis focusing on racial homogamy shows that improved preparedness significantly extends life expectancy for widowed white men in homogamous relationships and non-white women in similar circumstances. These findings challenge previous research by indicating that the effects of widowhood on life expectancy are not uniform but

vary based on a combination of gender and race factors.

B.5. Appendix IX: Schoenfeld Tests for Proportional Hazard Assumptions

The proportional hazard assumption is crucial for using the survival function with the Cox-PH model. Thus, we need to test the assumption to ensure the validity of the model. We mainly check if the preparedness variable violates the assumption.

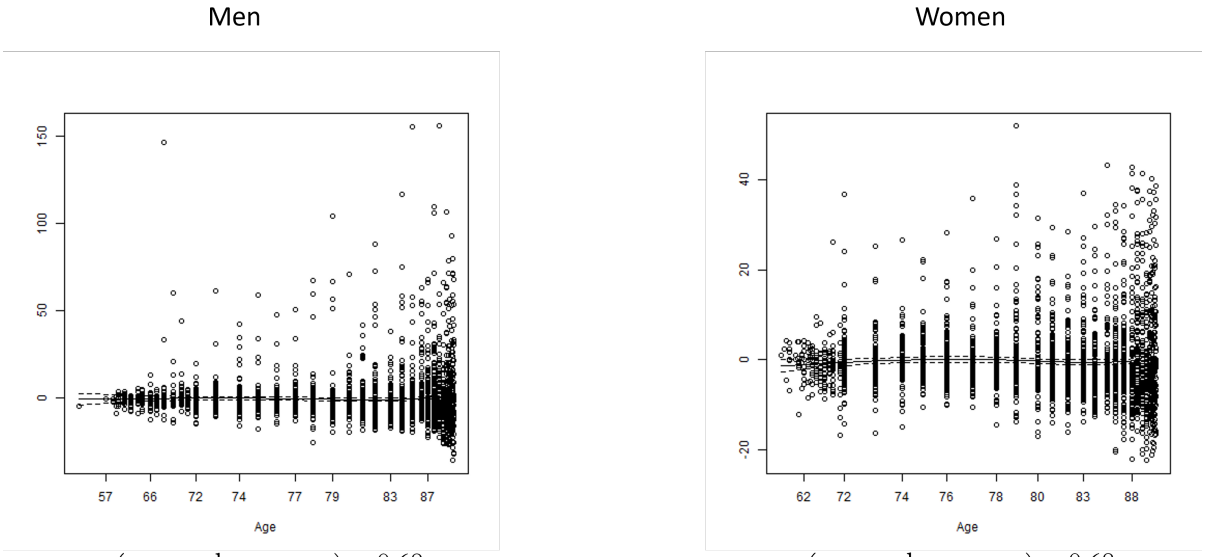


Figure B.8: Schoenfeld Test Results for Cox Proportional Hazard Model

As can be seen from the test results, we find no strong evidence against the proportional hazard assumption in these models.

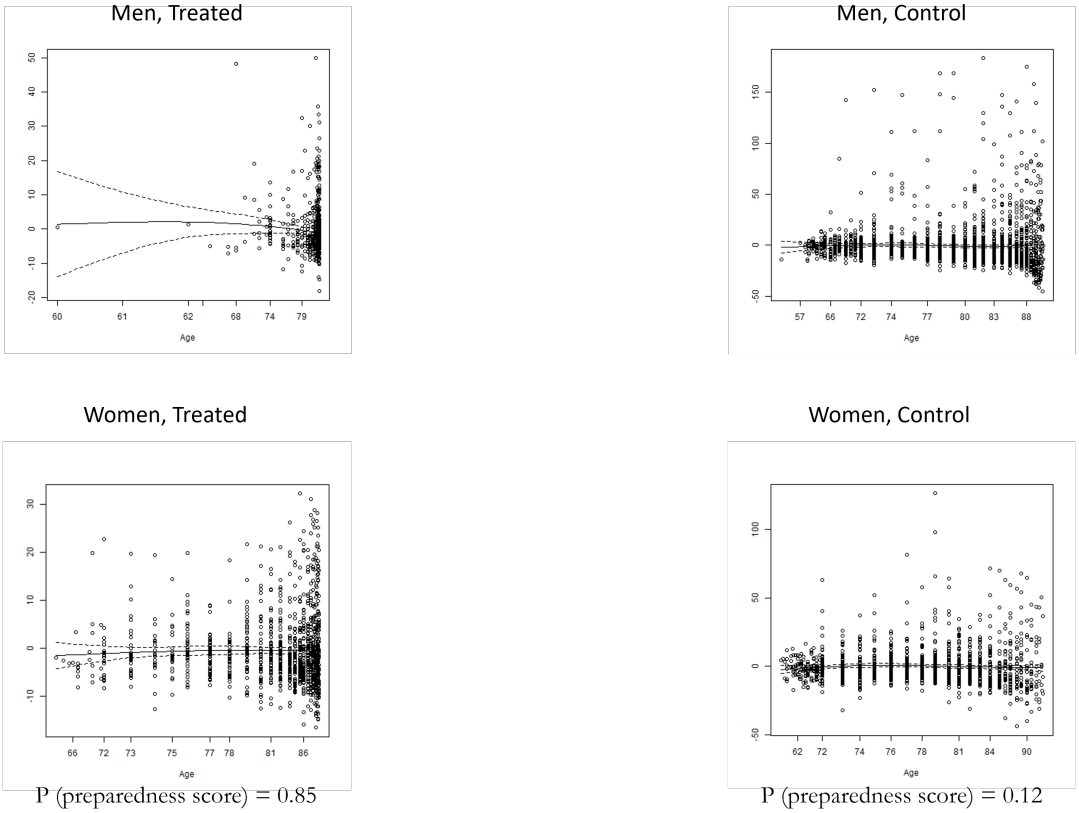


Figure B.9: Schoenfeld Test for Proportional Hazard Assumption, by Treatment

Appendix C

Appendix to Chapter 4

C.1. Appendix X: Equivalence in Expectation of Equations 4.IV.24 and 4.IV.22 when $A_L = a$

We start with Equation 4.IV.22:

$$\widehat{\psi}^{\mathrm{DR}}_{a;a,a_M} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\frac{\mathbb{1}(A_i=a)}{\hat{\pi}_a(C_i)}\Bigg\{\frac{1}{\hat{\pi}_{L_i}(a,C_i)}\Big(\sum_m \hat{q}_m(t,L_i,C_i)\big[\frac{\mathbb{1}(M_i=m)}{\hat{\pi}_m(a,L_i,C_i)}\{Y_i-\hat{\mu}_{am}(L_i,C_i)\}+\hat{\mu}_{am}(L_i,C_i)\big]$$
$$-\sum_m \hat{q}_m(t,L_i,C_i)\hat{\mu}_{am}(L_i,C_i)\Big)+\sum_m \hat{q}_m(t,L_i,C_i)\hat{\mu}_{am}(L_i,C_i)-\sum_l\sum_m \hat{\pi}_l(a,C_i)\hat{q}_m(t,l,C_i)\hat{\mu}_{aml}(C_i)\Bigg\}$$
$$+\sum_l\sum_m \hat{\pi}_l(a,C_i)\hat{q}_m(t,l,C_i)\hat{\mu}_{aml}(C_i)\Bigg].$$

$$\frac{1}{\hat{\pi}_{L_i}(a,C_i)}\Big(\sum_m \hat{q}_m(t,L_i,C_i)\hat{\mu}_{am}(L_i,C_i)-\sum_m \hat{q}_m(t,L_i,C_i)\hat{\mu}_{am}(L_i,C_i)\Big)=0.$$

Hence the inner parentheses reduce to

$$\frac{1}{\hat{\pi}_{L_i}(a,C_i)}\underbrace{\sum_m \hat{q}_m(t,L_i,C_i)\frac{\mathbb{1}(M_i=m)}{\hat{\pi}_m(a,L_i,C_i)}\{Y_i-\hat{\mu}_{am}(L_i,C_i)\}}_{=:\widehat{\zeta}_i}.$$

Thus,

$$\widehat{\psi}^{\mathrm{DR}}_{a;a,t} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\frac{\mathbb{1}(A_i=a)}{\hat{\pi}_a(C_i)}\Big\{\frac{1}{\hat{\pi}_{L_i}(a,C_i)}\widehat{\zeta}_i+\hat{\phi}(L_i,C_i)-\widehat{E}_{L|A=a,C=C_i}\big[\hat{\phi}(L,C_i)\big]\Big\}+\widehat{E}_{L|A=a,C=C_i}\big[\hat{\phi}(L,C_i)\big]\Bigg],$$

where

$$\hat{\phi}(L,C) := \sum_m \hat{\mu}_{am}(L,C)\,\hat{q}_m(t,L,C), \qquad \widehat{\zeta}_i := \sum_m \hat{q}_m(t,L_i,C_i)\,\frac{\mathbb{1}(M_i=m)}{\hat{\pi}_m(a,L_i,C_i)}\{Y_i - \hat{\mu}_{am}(L_i,C_i)\}.$$

If we compare with Equation 4.IV.24:

$$\widehat{\theta}_{a,t,\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_m \frac{\mathbb{1}(A_i=a)\,\mathbb{1}(M_i=m)}{\hat{\pi}_a(C_i)\,\hat{\pi}_m(a,L_i,C_i)}\big\{Y_i - \hat{\mu}_{am}(L_i,C_i)\big\}\,\hat{q}_m(t,L_i,C_i) + \frac{\mathbb{1}(A_i=a)}{\hat{\pi}_a(C_i)}\Big\{\hat{\phi}(L_i,C_i) - \widehat{E}_{L|A=a,C=C_i}[\hat{\phi}(L,C_i)]\Big\} + \widehat{E}_{L|A=a,C=C_i}[\hat{\phi}(L,C_i)]\Bigg].$$

Define the difference of the two influence terms:

$$\Delta_i := \frac{\mathbb{1}(A_i=a)}{\hat{\pi}_a(C_i)}\left[\frac{1}{\hat{\pi}_{L_i}(a,C_i)}\widehat{\zeta}_i - \sum_m \frac{\mathbb{1}(M_i=m)}{\hat{\pi}_m(a,L_i,C_i)}\{Y_i - \hat{\mu}_{am}(L_i,C_i)\}\,\hat{q}_m(t,L_i,C_i)\right].$$

By the law of iterated expectations and the usual IPW residual identity,

$$E\left[\frac{\mathbb{1}(M=m)}{\pi_m(a,L,C)}\{Y-\mu_{am}(L,C)\}\,\middle|\,A=a,L,C\right]=0,$$

thus also

$$E\big[\widehat{\zeta}_i \,\big|\, A_i=a,L_i,C_i\big]=0.$$

Therefore,

$$E[\Delta_i \mid A_i=a,L_i,C_i]=0 \quad\Longrightarrow\quad E[\Delta_i]=0.$$

Consequently,

$$\widehat{\psi}^{\mathrm{DR}}_{a;a,t} = \widehat{\theta}_{a,t,\mathrm{DR}} + \frac{1}{n}\sum_{i=1}^{n}\Delta_i, \qquad E[\Delta_i]=0,$$

This is to say, the two estimators are equivalent in expectation up to a conditionally mean-zero augmentation.

References

Aalen, O. O., R. J. Cook, and K. Røysland (2015). Does cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime data analysis 21*, 579–593.

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies 72*(1), 1–19.

Acharya, A., M. Blackwell, and M. Sen (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review 110*(3), 512–529.

Adler, N. E. and K. Newman (2002). Socioeconomic disparities in health: pathways and policies. *Health Affairs (Millwood) 21*(2), 60–76.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology 13*, 61–98.

Allison, P. D. (2014). *Event History and Survival Analysis: Regression for Longitudinal Event Data* (2 ed.). SAGE Publications.

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. Springer.

Anderson, D. J., M. Binder, and K. Krause (2002). The motherhood wage penalty: Which mothers pay it and why? *American Economic Review 92*(2), 354–358.

Anderson, D. J., M. Binder, and K. Krause (2003). The motherhood wage penalty revisited: Experience heterogeneity, work effort, and work-schedule flexibility. *ILR Review 56*(2), 273–294.

Andrews, R. M. and V. Didelez (2021). Insights into the cross-world independence assumption of causal mediation analysis. *Epidemiology 32*(2), 209–219.

Angelov, N., P. Johansson, and E. Lindahl (2016). Parenthood and the gender gap in pay. *Journal of Labor Economics 34*(3), 545–579.

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review 80*(3), 313–336.

Angrist, J. D. and A. B. Krueger (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives 15*(4), 69–85.

Angrist, J. D. and V. Lavy (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics 114*(2), 533–575.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Arkhangelsky, D. and G. W. Imbens (2018). The role of the propensity score in fixed effect models. Working Paper w25162, National Bureau of Economic Research.

Autor, D. H., D. Dorn, and G. H. Hanson (2013). The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review 103*(6), 2121–2168.

Baker, S. G., B. S. Kramer, and K. S. Lindeman (2016). Latent class instrumental variables: a clinical and biostatistical perspective. *Statistics in medicine 35*(1), 147–160.

Barry, L. C., S. V. Kasl, and H. G. Prigerson (2002). Psychiatric disorders among bereaved persons: The role of perceived circumstances of death and preparedness for death. *The American Journal of Geriatric Psychiatry 10*(4), 447–457.

Becker, G. S. (1981). *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Berkman, N. D., S. L. Sheridan, K. E. Donahue, D. J. Halpern, and K. Crotty (2011). Low health literacy and health outcomes: an updated systematic review. *Annals of Internal Medicine 155*(2), 97–107.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York: Wiley-Interscience.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association 90*(430), 443–450.

Bowling, A. (1987). Mortality after bereavement: A review of the literature on survival periods and factors affecting survival. *Social science & medicine 24*(2), 117–124.

Bowling, A. (1989). Who dies after widow(er)hood? a discriminant analysis. *OMEGA-Journal of Death and Dying 19*(2), 135–153.

Box-Steffensmeier, J. M. and B. S. Jones (2004). *Event History Modeling: A Guide for Social Scientists*. Cambridge University Press.

Boyle, P. J., Z. Feng, and G. M. Raab (2011). Does widowhood increase mortality risk? testing for selection effects by comparing causes of spousal death. *Epidemiology 22*(1), 1–5.

Braveman, P., S. Egerter, and D. R. Williams (2011). The social determinants of health: coming of age. *Annual Review of Public Health 32*, 381–398.

Breslow, N. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review 43*(1), 45–58.

Brown, D. C., R. A. Hummer, and M. D. Hayward (2014). The importance of spousal education for the self-rated health of married adults in the united states. *Population Research and Policy Review 33*, 127–151.

Budig, M. J. and P. England (2001). The wage penalty for motherhood. *American Sociological Review 66*(2), 204–225.

Budig, M. J. and M. J. Hodges (2010). Differences in disadvantage: Variation in the motherhood penalty across white women's earnings distribution. *American Sociological Review 75*(5), 705–728.

Capiński, M. and P. E. Kopp (2004). *Measure, Integral and Probability* (2nd ed.). London: Springer.

Card, D. (1999). The causal effect of education on earnings. In *Handbook of Labor Economics*, Volume 3, pp. 1801–1863. Elsevier.

Card, D. and A. B. Krueger (1994). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. *The American Economic Review 84*(4), 772–793.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM.

Cheng, S. (2016). The accumulation of (dis)advantage: The intersection of gender and race in the long-term wage effect of marriage. *American Sociological Review 81*(1), 29–56.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018b). Double/debiased/neyman machine learning of treatment effects. *American Economic Review 108*(5), 899–927.

Christenson, B. A. and N. E. Johnson (1995). Educational inequality in adult mortality: An assessment with death certificate data from michigan. *Demography 32*, 215–229.

Chun, H. and I. Lee (2001). Why do married men earn more: Productivity or marriage selection? *Economic Inquiry 39*(2), 307–319.

Cornwell, C. and P. Rupert (1997). Unobservable individual effects, marriage, and the earnings of young men. *Economic Inquiry 35*(2), 285–294.

Correll, S. J., S. Benard, and I. Paik (2007). Getting a job: Is there a motherhood penalty? *American Journal of Sociology 112*(5), 1297–1339.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Cox, D. R. (1997). Some remarks on the analysis of survival data. In D. Y. Lin and T. R. Fleming (Eds.), *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, pp. 1–9. New York: Springer.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman and Hall.

Crittenden, A. (2001). *The Price of Motherhood: Why the Most Important Job in the World Is Still the Least Valued*. New York, NY: Macmillan.

Davies, R. and G. Pierre (2005). The family gap in pay in europe: A cross-country study. *Labour Economics 12*(4), 469–486.

Dell, M., B. F. Jones, and B. A. Olken (2009). Temperature and income: Reconciling new cross-sectional and panel estimates. *American Economic Review 99*(2), 198–204.

Diamond, A. and J. S. Sekhon (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics 95*(3), 932–945.

Dougherty, C. (2006). The marriage earnings premium as a distributed fixed effect. *Journal of Human Resources 41*(2), 433–443.

Duguet, E., P. Petit, and P. Petit (2005). Hiring discrimination in the french financial sector: An econometric analysis on field experiment data. *Annales d'Economie et de Statistique 78*, 79–102.

Duncan, O. D. and M. Stenbeck (1988). Panels and cohorts: Design and model in the study of voting turnout. *Sociological Methodology 18*, 1–35.

Durkheim, E. (1951). *Suicide: A study in sociology [1897]*. The Free Press, Glencoe, Illinois.

Efron, B. (1988). Logistic regression, survival analysis, and the kaplan–meier curve. *Journal of the American Statistical Association 83*(402), 414–425.

Elo, I. T. (2009). Social class differentials in health and mortality: Patterns and explanations in comparative perspective. *Annual Review of Sociology 35*, 553–572.

Elwert, F. and N. A. Christakis (2006). Widowhood and race. *American Sociological Review 71*(1), 16–41.

Elwert, F. and N. A. Christakis (2008a). The effect of widowhood on mortality by the causes of death of both spouses. *American journal of public health 98*(11), 2092–2098.

Elwert, F. and N. A. Christakis (2008b). Wives and ex-wives: A new test for homogamy bias in the widowhood effect. *Demography 45*, 851–873.

England, P. (2005). Gender inequality in labor markets: The role of motherhood and segregation. *Social Politics: International Studies in Gender, State & Society 12*(2), 264–288.

England, P., J. Bearak, M. Budig, and M. J. Hodges (2016). Do highly paid, highly skilled women experience the largest motherhood penalty? *American Sociological Review 81*(6), 1161–1189.

Espinosa, J. and W. N. Evans (2008). Heightened mortality after the death of a spouse: Marriage protection or marriage selection? *Journal of health economics 27*(5), 1326–1342.

Fan, W. and Y. Qian (2019). Rising educational gradients in mortality among us whites: What are the roles of marital status and educational homogamy? *Social science & medicine 235*, 112365.

Faraggi, D. and R. Simon (1995). A neural network model for survival data. *Statistics in Medicine 14*(1), 73–82.

Farbmacher, H., M. Huber, L. Lafférs, H. Langen, and M. Spindler (2022). Causal mediation analysis with double machine learning. *The Econometrics Journal 25*(2), 277–300.

Farr, W. (1858). The influence of marriage on the mortality of the french people. In *Transactions of the National Association for the Promotion of Social Science*, pp. 504–513.

Fearon, J. D. and D. D. Laitin (2003). Ethnicity, insurgency, and civil war. *American Political Science Review 97*(1), 75–90.

Fisher, A. and E. H. Kennedy (2021). Visually communicating and teaching intuition for influence functions. *The American Statistician 75*(2), 162–172.

Fletcher, J. M. and B. L. Wolfe (2009). The effects of teenage childbearing on the short-and long-term health behaviors of mothers. *Journal of Population Economics 22*(3), 575–597.

Gangl, M. and A. Ziefle (2009). Motherhood, labor force behavior, and women's careers: An empirical assessment of the wage penalty for motherhood in britain, germany, and the united states. *Demography 46*(2), 341–369.

Gelman, A. and J. Hill (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Glauber, R. (2007). Marriage and the motherhood wage penalty among african americans, hispanics, and whites. *Journal of Marriage and Family 69*(4), 951–961.

Glauber, R. (2008). Race and gender in families and at work: The fatherhood wage premium. *Gender & Society 22*(1), 8–30.

Glauber, R. (2018). Trends in the motherhood wage penalty and fatherhood wage premium for low, middle, and high earners. *Demography 55*(5), 1663–1680.

Glick, P. C. and E. Landau (1950). Age as a factor in marriage. *American Sociological Review 15*(4), 517–529.

Goisis, A. and W. Sigle-Rushton (2014). Family size and children's educational outcomes in the uk: Cross-cohort evidence from the 1970 british cohort study and the millennium cohort study. *Demography 51*(4), 1529–1554.

Goldman, N. and Y. Hu (1993). Excess mortality among the unmarried: A case study of japan. *Social science & medicine 36*(4), 533–546.

Goldman, N., S. Korenman, and R. Weinstein (1995). Marital status and health among the elderly. *Social science & medicine 40*(12), 1717–1730.

Gough, M. and M. Noonan (2013). A review of the motherhood wage penalty in the united states. *Sociology Compass 7*(4), 328–342.

Greene, W. H. (2012). *Econometric Analysis* (7th ed.). Upper Saddle River, NJ: Pearson Education.

Grimshaw, D. and J. Rubery (2015). The motherhood pay gap. Technical report, International Labour Organization.

Gruber, J. (1994). The incidence of mandated maternity benefits. *The American Economic Review 84*(3), 622–641.

Grundy, E. and O. Kravdal (2008). Reproductive history and mortality in late middle age among norwegian men and women. *American journal of epidemiology 167*(3), 271–279.

Gupta, S. (1999). The effects of transitions in marital status on men's performance of housework. *Journal of Marriage and the Family 61*(3), 700–711.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica 12*(Supplement), iii–115.

Hahn, J., P. Todd, and W. V. der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

Hakim, C. (2002). Lifestyle preferences as determinants of women's differentiated labor market careers. *Work and Occupations 29*(4), 428–459.

Hart, C. L., D. J. Hole, D. A. Lawlor, G. D. Smith, and T. F. Lever (2007). Effect of conjugal bereavement on mortality of the bereaved spouse in participants of the renfrew/paisley study. *Journal of Epidemiology & Community Health 61*(5), 455–460.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Hauksdóttir, A., G. Steineck, C. J. Fürst, and U. Valdimarsdóttir (2010). Long-term harm of low preparedness for a wife's death from cancer—a population-based study of widowers 4–5 years after the loss. *American Journal of Epidemiology 172*(4), 389–396.

Hebert, R. S., H. G. Prigerson, R. Schulz, and R. M. Arnold (2006). Preparing caregivers for the death of a loved one: A theoretical framework and suggestions for future research. *Journal of Palliative Medicine 9*(5), 1164–1171.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998). Characterizing selection bias using experimental data. *Econometrica 66*(5), 1017–1098.

Heckman, J. and R. Pinto (2024). Econometric causality: The central role of thought experiments. *Journal of Econometrics 243*(1–2), 105719.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.) 21*(1), 13.

Hernán, M. A., B. Brumback, and J. M. Robins (2002). Estimating the causal effect of zidovudine on cd4 count with marginal structural models. *Statistics in Medicine 21*(12), 1689–1709.

Hernán, M. A. and J. M. Robins (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

Hersch, J. and L. S. Stratton (2000). Housework, wages, and the division of housework time for employed spouses. *American Economic Review 90*(2), 337–341.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Hodges, M. J. and M. J. Budig (2010). Who gets the daddy bonus? organizational hegemonic masculinity and the impact of fatherhood on earnings. *Gender & Society 24*(6), 717–745.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

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

House, J. S., K. R. Landis, and D. Umberson (1988). Social relationships and health. *science 241*(4865), 540–545.

Hu, L., J. Ji, and F. Li (2021). Estimating heterogeneous survival treatment effect in observational data using machine learning. *Statistics in Medicine 40*(14), 3428–3444.

Hu, Y. and N. Goldman (1990). Mortality differentials by marital status: An international comparison. *Demography 27*, 233–250.

Ichimura, H. and W. K. Newey (2022, 1). The influence function of semiparametric estimators. *Quantitative Economics 13*(1), 29–61.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics 142*(2), 615–635.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge: Cambridge University Press.

Iwasaki, M., T. Otani, R. Sunaga, H. Miyazaki, L. Xiao, N. Wang, S. Yosiaki, and S. Suzuki (2002). Social networks and mortality based on the komo-ise cohort study in japan. *International Journal of Epidemiology 31*(6), 1208–1218.

Jackson, J. W. and T. J. VanderWeele (2018). Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology 29*(6), 825–835.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning: with Applications in R*. New York: Springer.

Jenkins, S. P. (1995). Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics 57*(1), 129–138.

Jin, L. and N. A. Christakis (2009). Investigating the mechanism of marital mortality reduction: The transition to widowhood and quality of health care. *Demography 46*, 605–625.

Jones, D. R. and P. O. Goldblatt (1987). Cause of death in widow(er)s and spouses. *Journal of Biosocial Science 19*(1), 107–121.

Joung, I. M., J. J. Glerum, F. W. van Poppel, J. W. Kardaun, and J. P. Mackenbach (1996). The contribution of specific causes of death to mortality differences by marital status in the netherlands. *The European Journal of Public Health 6*(2), 142–149.

Juhn, C. and K. McCue (2017). Specialization then and now: Marriage, children, and the gender earnings gap across cohorts. *Journal of Economic Perspectives 31*(1), 183–204.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data* (2 ed.). John Wiley & Sons.

Kalmijn, M. (1998). Intermarriage and homogamy: Causes, patterns, trends. *Annual Review of Sociology 24*(1), 395–421.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice Hall.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In H. He, P. Wu, and D.-G. Chen (Eds.), *Statistical Causal Inferences and Their Applications in Public Health Research*, pp. 141–167. Cham: Springer.

Kennedy, E. H. (2022a). Semiparametric doubly robust targeted double machine learning: A review. arXiv preprint arXiv:2203.06469; last revised on January 26, 2023.

Kennedy, E. H. (2022b). Semiparametric doubly robust targeted double machine learning: A review. arXiv preprint arXiv:2203.06469; last revised on January 26, 2023.

Kennedy, E. H. (2023). Semiparametric doubly robust targeted double machine learning: a review.

Kiecolt-Glaser, J. K. and S. J. Wilson (2017). Lovesick: How couples' relationships influence health. *Annual Review of Clinical Psychology 13*, 421–443.

Killewald, A. (2013). A reconsideration of the fatherhood premium: Marriage, coresidence, biology, and fathers' wages. *American Sociological Review 78*(1), 96–116.

Killewald, A. and M. Gough (2013a). Does specialization explain marriage penalties and premiums? *American Sociological Review 78*(3), 477–502.

Killewald, A. and M. Gough (2013b). Does specialization explain marriage penalties and premiums? *American Sociological Review 78*(3), 477–502.

Killewald, A. and I. Lundberg (2017). New evidence against a causal marriage wage premium. *Demography 54*(3), 1007–1028.

Kim, I. S. and K. Imai (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science 63*(2), 467–490.

King, G. and R. Nielsen (2019). Why propensity scores should not be used for matching. *Political Analysis 27*(4), 435–454.

Kitagawa, E. M. and P. M. Hauser (1973). *Differential mortality in the United States: A study in socioeconomic epidemiology*. Harvard University Press.

Klein, J. P. and M. L. Moeschberger (1992). Bounds on net survival probabilities for left-truncated and right-censored data. *Biometrics 48*(4), 1143–1150.

Klein, J. P. and M. L. Moeschberger (2003). *Survival Analysis: Techniques for Censored and Truncated Data* (2nd ed.). New York: Springer.

Kleven, H., C. Landais, J. Posch, A. Steinhauer, and J. Zweimüller (2019). Child penalties across countries: Evidence and explanations. In *AEA Papers and Proceedings*, Volume 109, pp. 122–126. American Economic Association.

Kolip, P. (2005). The association between gender, family status and mortality. *Journal of Public Health 13*, 309–312.

Korenman, S. and D. Neumark (1991). Does marriage really make men more productive? *Journal of Human Resources 26*, 282–307.

Korenman, S. and D. Neumark (1992). Marriage, motherhood, and wages. *Journal of Human Resources 27*(2), 233–255.

Kravdal, O. (2003). Children, family and cancer survival in norway. *International Journal of Cancer 105*(2), 261–266.

Kristensen, P., L. Weisæth, and T. Heir (2012). Bereavement and mental health after sudden and violent losses: A review. *Psychiatry: Interpersonal & Biological Processes 75*(1), 76–97.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data* (2nd ed.). Hoboken, NJ: Wiley.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature 48*(2), 281–355.

Leggett, A. N., A. J. Sonnega, and M. C. Lohman (2020). Till death do us part: Intersecting health and spousal dementia caregiving on caregiver mortality. *Journal of aging and health 32*(7-8), 871–879.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2nd ed.). New York: Springer.

Lillard, L. A. and C. W. Panis (1996). Marital status and mortality: The role of health. *Demography 33*, 313–327.

Lillard, L. A. and L. J. Waite (1995). 'til death do us part: Marital disruption and mortality. *American Journal of Sociology 100*(5), 1131–1156.

Liu, H., D. Umberson, and M. Xu (2020). Widowhood and mortality: Gender, race/ethnicity, and the role of economic resources. *Annals of epidemiology 45*, 69–75.e1.

Loughran, D. S. and J. M. Zissimopoulos (2009). Why wait?: The effect of marriage and childbearing on the wages of men and women. *Journal of Human Resources 44*(2), 326–349.

Ludwig, V. and J. Brüderl (2018). Is there a male marital wage premium? new evidence from the united states. *American Sociological Review 83*(4), 744–770.

Luhr, S. (2020). Signaling parenthood: Managing the motherhood penalty and fatherhood premium in the us service sector. *Gender & Society 34*(2), 259–283.

Lundberg, I. (2024). The gap-closing estimand: A causal approach to study interventions that close disparities across social categories. *Sociological Methods & Research 53*(2), 507–570.

Lundberg, S. and E. Rose (2000). Parenthood and the earnings of married men and women. *Labour Economics 7*(6), 689–710.

Lundberg, S. and E. Rose (2002). The effects of sons and daughters on men's labor supply and wages. *Review of Economics and Statistics 84*(2), 251–268.

Luo, S. and E. C. Klohnen (2005). Assortative mating and marital quality in newlyweds: A couple-centered approach. *Journal of Personality and Social Psychology 88*(2), 304–326.

Lusyne, P., H. Page, and J. Lievens (2001). Mortality following conjugal bereavement, belgium 1991-96: The unexpected effect of education. *Population Studies 55*(3), 281–289.

Makela, P., T. Valkonen, and T. Martelin (1997). Contribution of deaths related to alcohol use to socioeconomic variation in mortality: register based follow up study. *BMJ 315*(7102), 211–216.

Malyutina, S., M. Bobak, G. Simonova, V. Gafarov, Y. Nikitin, and M. Marmot (2004). Education, marital status, and total and cardiovascular mortality in novosibirsk, russia: A prospective cohort study. *Annals of epidemiology 14*(4), 244–249.

Manor, O. and Z. Eisenbach (2003). Mortality after spousal loss: Are there socio-demographic differences? *Social science & medicine 56*(2), 405–413.

Manor, O., Z. Eisenbach, E. Peritz, and Y. Friedlander (1999). Mortality differentials among israeli men. *American journal of public health 89*(12), 1807–1813.

Mao, H., L. Li, W. Yang, and Y. Shen (2018). On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Statistics in Medicine 37*(26), 3745–3763.

Martikainen, P. and T. Valkonen (1998). Do education and income buffer the effects of death of spouse on mortality? *Epidemiology 9*(5), 530–534.

Meeker, W. Q. (1998). *Statistical Methods for Reliability Data*. New York: Wiley.

Meurs, D., A. Pailhé, and S. Ponthieux (2010). Child-related career interruptions and the gender wage gap in france. *Annals of Economics and Statistics/Annales d'Économie et de Statistique 99/100*, 15–46.

Meyer, B. D., W. K. Viscusi, and D. L. Durbin (1995). Workers' compensation and injury duration: Evidence from a natural experiment. *The American Economic Review 85*(3), 322–340.

Mincer, J. and S. Polachek (1974). Family investments in human capital: Earnings of women. *Journal of Political Economy 82*(2, Part 2), S76–S108.

Monden, C. W., F. Van Lenthe, N. D. De Graaf, and G. Kraaykamp (2003). Partner's and own education: Does who you live with matter for self-assessed health, smoking and excessive alcohol consumption? *Social science & medicine 57*(10), 1901–1912.

Montez, J. K., R. A. Hummer, M. D. Hayward, H. Woo, and R. G. Rogers (2011). Trends in the educational gradient of us adult mortality from 1986 through 2006 by race, gender, and age group. *Research on Aging 33*(2), 145–171.

Morgan, S. L. and C. Winship (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). New York: Cambridge University Press.

Munasinghe, L., T. Reif, and A. Henriques (2008). Gender gap in wage returns to job tenure and experience. *Labour Economics 15*(6), 1296–1316.

Musick, K., M. Doherty Bea, and P. Gonalons-Pons (2020). His and her earnings following parenthood in the united states, germany, and the united kingdom. *American Sociological Review 85*(4), 639–674.

Naimi, A. I., S. R. Cole, and E. H. Kennedy (2017). An introduction to g methods. *International Journal of Epidemiology 46*(2), 756–762.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature 56*(3), 799–866.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics 5*(2), 99–135.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–1382.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472. Translated from the 1923 Polish original by Dorota M. Dabrowska and Terence P. Speed.

Noonan, M. C. (2001). The impact of domestic work on men's and women's wages. *Journal of Marriage and Family 63*(5), 1134–1145.

Oppenheimer, V. K. (1997). Women's employment and the gain to marriage: The specialization and trading model. *Annual Review of Sociology 23*(1), 431–453.

Oreopoulos, P. (2006). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review 96*(1), 152–175.

Ostergren, O., S. Fors, and J. Rehnberg (2022). Excess mortality by individual and spousal education for recent and long-term widowed. *The Journals of Gerontology: Series B 77*(5), 946–955.

Parkes, C. M., B. Benjamin, and R. G. Fitzgerald (1969). Broken heart: A statistical study of increased mortality among widowers. *Br med J 1*(5646), 740–743.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–688.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann Publishers Inc.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J. (2012). The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention Science 13*(4), 426–436.

Pearl, J. (2015). Trygve haavelmo and the emergence of causal calculus. *Econometric Theory 31*(1), 152–179.

Pedulla, D. S. (2016). Penalized or protected? gender and the consequences of nonstandard and mismatched employment histories. *American Sociological Review 81*(2), 262–289.

Petersen, T. and L. A. Morgan (1995). Separate and unequal: Occupation-establishment sex segregation and the gender wage gap. *American Journal of Sociology 101*(2), 329–365.

Petit, P. (2007). The effects of age and family constraints on gender hiring discrimination: A field experiment in the french financial sector. *Labour Economics 14*(3), 371–391.

Polley, E., A. E. Hubbard, and M. J. van der Laan (2023). *SuperLearner: Super Learner Prediction*. CRAN. R package version 2.0-31.

Rahman, O., A. Foster, and J. Menken (1992). Older widow mortality in rural bangladesh. *Social science & medicine 34*(1), 89–96.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (2nd ed.). New York: Wiley.

Richardson, T. S. and J. M. Robins (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. Technical Report 128, Center for Statistics and the Social Sciences, University of Washington.

Rivera, L. A. (2017). When two bodies are (not) a problem: Gender and relationship status discrimination in academic hiring. *American Sociological Review 82*(6), 1111–1138.

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

Robins, J. M. (1994). Correcting for noncompliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods 23*(8), 2379–2412.

Robins, J. M. (2003). Semantics of causal dag models and the identification of direct and indirect effects. *Causal Inference in Epidemiology: Past, Present, and Future 112*, 139–145.

Robins, J. M., M. A. Hernán, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*(5), 550–560.

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

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association 84*(408), 1024–1032.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rosenbaum, P. R. and D. B. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79*(387), 516–524.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Rubin, D. B. (1980). Bias reduction using mahalanobis-metric matching. *Biometrics 36*(2), 293–298.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science 5*(4), 472–480.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning representations by back-propagating errors. *Nature 323*(6088), 533–536.

Samuelsson, G. and O. Dehlin (1993). Family network and mortality: Survival chances through the lifespan of an entire age cohort. *The International Journal of Aging and Human Development 37*(4), 277–295.

Sasson, I. (2016). Trends in life expectancy and lifespan variation by educational attainment: United states, 1990–2010. *Demography 53*(2), 269–293.

Schaefer, C., C. P. Quesenberry Jr, and S. Wi (1995). Mortality following conjugal bereavement and the effects of a shared environment. *American journal of epidemiology 141*(12), 1142–1152.

Schuler, A. and M. J. van der Laan (2024). Introduction to modern causal inference. Accessed: 2024-09-28.

Schwartz, C. R. and R. D. Mare (2005). Trends in educational assortative marriage from 1940 to 2003. *Demography 42*, 621–646.

Semenova, V. and V. Chernozhukov (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal 23*(3), C1–C26.

Shah, S. M., I. M. Carey, T. Harris, S. DeWilde, C. R. Victor, and D. G. Cook (2013). The effect of unexpected bereavement on mortality in older couples. *American journal of public health 103*(6), 1140–1145.

Shkolnikov, V. M., D. Jasilionis, E. M. Andreev, D. A. Jdanov, V. Stankuniene, and D. Ambrozaitiene (2007). Linked versus unlinked estimates of mortality and length of life by education and marital status: Evidence from the first record linkage study in lithuania. *Social science & medicine 64*(7), 1392–1406.

Shor, E., D. J. Roelfs, M. Curreli, L. Clemow, M. M. Burg, and J. E. Schwartz (2012). Widowhood and mortality: A meta-analysis and meta-regression. *Demography 49*(2), 575–606.

Shurtleff, D. (1955). Mortality and marital status. *Public Health Reports 70*(3), 248.

Shurtleff, D. (1956). Mortality among the married. *Journal of the American Geriatrics Society 4*(7), 654–666.

Singer, J. D. and J. B. Willett (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics 18*(2), 155–195.

Singer, J. D. and J. B. Willett (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press.

Snowden, J. M., S. Rose, and K. M. Mortimer (2011). Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology 173*(7), 731–738.

Staff, J. and J. T. Mortimer (2012). Explaining the motherhood wage penalty during the early occupational career. *Demography 49*(1), 1–21.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*(3), 557–586.

Steingrimsson, J. A., L. Diao, A. M. Molinaro, and R. L. Strawderman (2016). Doubly robust survival trees. *Statistics in Medicine 35*(20), 3595–3612.

Steingrimsson, J. A. and S. Morrison (2020). Deep learning for survival outcomes. *Statistics in Medicine 39*(17), 2339–2349.

Stock, J. H. and F. Trebbi (2003). Who invented instrumental variable regression? *Journal of Economic Perspectives 17*(3), 177–194.

Stock, J. H., J. H. Wright, and M. Yogo (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics 20*(4), 518–529.

Stock, J. H. and M. Yogo (2005). Testing for weak instruments in linear iv regression. In D. W. K. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pp. 80–108. New York: Cambridge University Press.

Strawderman, R. L. (2000). Estimating the mean of an increasing stochastic process at a censored stopping time. *Journal of the American Statistical Association 95*(452), 1192–1208.

Stringhini, S., S. Sabia, M. Shipley, E. Brunner, H. Nabi, M. Kivimäki, and A. Singh-Manoux (2010). Association of socioeconomic position with health behaviors and mortality. *JAMA 303*(12), 1159–1166.

Stroebe, M., H. Schut, and W. Stroebe (2007). Health outcomes of bereavement. *The Lancet 370*(9603), 1960–1973.

Subramanian, S., F. Elwert, and N. Christakis (2008). Widowhood and mortality among the elderly: The modifying role of neighborhood concentration of widowed individuals. *Social science & medicine 66*(4), 873–884.

Sullivan, A. R. and A. Fenelon (2014). Patterns of widowhood mortality. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences 69*(1), 53–62.

Suresh, K., C. Severn, and D. Ghosh (2022). Survival prediction models: An introduction to discrete-time modeling. *BMC medical research methodology 22*(1), 207.

Tang, S. T., W.-C. Chang, W.-C. Chou, C.-H. Hsieh, J.-S. Chen, and F.-H. Wen (2021). Family caregivers' emotional preparedness for death is distinct from their cognitive prognostic awareness for cancer patients. *Journal of Palliative Medicine 24*(3), 405–412.

Taniguchi, H. (1999). The timing of childbearing and women's wages. *Journal of Marriage and the Family 61*(4), 1008–1019.

Taubman, S. L., J. M. Robins, M. A. Mittleman, and M. A. Hernán (2009). Intervening on risk factors for coronary heart disease: An application of the parametric g-formula. *International Journal of Epidemiology 38*(6), 1599–1611.

Tchetgen Tchetgen, E. J. and I. Shpitser (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *The Annals of Statistics 40*(3), 1816–1845.

Testa, M. R. and L. Toulemon (2006). Family formation in france: Individual preferences and subsequent outcomes. *Vienna Yearbook of Population Research 2006*, 41–75.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.

Thierry, X. (2000). Risks of mortality and excess mortality during the first ten years of widowhood. *Population: An English Selection 12*(1), 81–109.

Thistlethwaite, D. L. and D. T. Campbell (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology 51*(6), 309–317.

Thomas, L. and E. M. Reyes (2014). Tutorial: Survival estimation for cox regression models with time-varying coefficients using sas and r. *Journal of Statistical Software, Code Snippets 61*(1), 1–23.

Treml, J., V. Schmidt, M. Nagl, and A. Kersting (2021). Pre-loss grief and preparedness for death among caregivers of terminally ill cancer patients: A systematic review. *Social Science & Medicine 284*, 114240.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Vable, A. M., S. Subramanian, P. M. Rist, and M. M. Glymour (2015). Does the "widowhood effect" precede spousal bereavement? results from a nationally representative sample of older adults. *The American Journal of Geriatric Psychiatry 23*(3), 283–292.

Vagni, G. and R. Breen (2021). Earnings and income penalties for motherhood: Estimates for british women using the individual synthetic control method. *European Sociological Review 37*(5), 834–848.

Van der Klaauw, W. (1996). Female labour supply and marital status decisions: A life-cycle model. *The Review of Economic Studies 63*(2), 199–235.

van der Laan, M. J. and J. M. Robins (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

VanderWeele, T. J. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. J. and E. J. Tchetgen Tchetgen (2017). Mediation analysis with time-varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(3), 917–938.

VanderWeele, T. J., S. Vansteelandt, and J. M. Robins (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology 25*(2), 300–306.

Vansteelandt, S. and R. M. Daniel (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology 28*(2), 258–265.

Waldfogel, J. (1997). The effect of children on women's wages. *American Sociological Review 62*(2), 209–217.

Waldfogel, J. (1998). The family gap for young women in the united states and britain: Can maternity leave make a difference? *Journal of Labor Economics 16*(3), 505–545.

Wen, F.-H., W.-C. Chou, C.-H. Hsieh, J.-S. Chen, W.-C. Chang, and S. T. Tang (2021). Distinct death-preparedness states by combining cognitive and emotional preparedness for death and their evolution for family caregivers of terminally ill cancer patients over their last six months of life. *Journal of Pain and Symptom Management 62*(3), 503–511.

Willett, J. B. and J. D. Singer (1995). It's déjà vu all over again: Using multiple-spell discrete-time survival analysis. *Journal of Educational and Behavioral Statistics 20*(1), 41–67.

Wolfers, J. (2006). Did unilateral divorce laws raise divorce rates? a reconciliation and new results. *The American Economic Review 96*(5), 1802–1820.

Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Review of Economics and Statistics 87*(2), 385–390.

Wortman, C. B., R. C. Silver, and R. C. Kessler (1993). The meaning of loss and adjustment to bereavement. In M. S. Stroebe, W. Stroebe, and R. O. Hansson (Eds.), *Handbook of bereavement: Theory, research, and intervention*, pp. 349–366. Cambridge University Press.

Wu, L. L. and F. Wen (2022). Hazard Versus Linear Probability Difference-in-Differences Estimators for Demographic Processes. *Demography 59*(5), 1911–1928.

Yu, W.-h. and Y. Hara (2021). Motherhood penalties and fatherhood premiums: Effects of parenthood on earnings growth within and across firms. *Demography 58*(1), 247–272.

Yu, W.-h. and J. C.-L. Kuo (2017). The motherhood wage penalty by work conditions: How do occupational characteristics hinder or empower mothers? *American Sociological Review 82*(4), 744–769.

Zhang, Z., J. Reinikainen, K. A. Adeleke, M. E. Pieterse, and C. G. Groothuis-Oudshoorn (2018). Time-varying covariates and coefficients in cox regression models. *Annals of translational medicine 6*(7), 121–121.

Zhou, X. (2021). *Some Doubly and Multiply Robust Estimators of Controlled Direct Effects*. Ph. D. thesis, Harvard University.

Zhou, X. and G. Pan (2023). Higher education and the black-white earnings gap. *American Sociological Review 88*(1), 154–188.

Zhou, X. and G. T. Wodtke (2020). Residual balancing: A method of constructing weights for marginal structural models. *Political Analysis 28*(4), 487–506.