UNIVERSITY OF CALIFORNIA,
IRVINE

Enhancing the Utility of Instrumental Variables in Observational Research

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Roy S. Zawadzki

Dissertation Committee:
Professor Daniel Gillen, Chair
Assistant Professor Tianchen Qian
Professor David Richardson

2024

© 2024 Roy S. Zawadzki

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Daniel L. Gillen, not only for his unwavering guidance and support throughout the development of this dissertation but, additionally, for his constant willingness to pass on a plethora of invaluable insights and wisdom gained from years of teaching and practicing statistics in clinical settings. I look to Dan as an exemplary role model for what it means to be a world-class scientific investigator and communicator.

Furthermore, I would like to graciously thank my reading committee of Joshua D. Grill, David B. Richardson, and Tianchen Qian as well as Scott Bartell and Hengrui Cai, who served on my advancement committee. Their insightful feedback, without a doubt, has significantly enhanced the quality of this work.

Throughout my time at UCI, I was supported by the following grants: National Institutes of Health (NIH) National Institute on Aging (NIA) P30AG066519 and 1RF1AG075107. Additionally, I thank the support and mentorship from the NIH/NIA T32 Training in Translational ADRD Neuroscience (TITAN) grant AG073088-03.

Part of the content in Chapter 1 and Chapter 2 has been published in *BMC Medical Research Methodology.*

Data used in the preparation of this dissertation were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The ADNI dataset is avaliable to the public at http://adni.loni.usc.edu via the completion of an online application form and acceptance of data use agreement. All participants in ADNI provided informed consent. For details regarding the ethics statement of the ADNI study population please see: https://adni.loni.usc.edu.

Data collection and sharing for this dissertation was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The

Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

VITA

Roy S. Zawadzki

EDUCATION

Doctor of Philosophy in Statistics **November 2024**
University of California, Irvine *Irvine, CA*

Master of Science in Statistics **June 2022**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Statistics **June 2020**
California Polytechnic State University, San Luis Obispo *San Luis Obispo, CA*

RESEARCH EXPERIENCE

Graduate Research Assistant **2021–2023**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **Winter 2018**
University of California, Irvine *Irvine, California*

Teaching Assistant **Winter 2017**
University of California, Irvine *Irvine, California*

Teaching Assistant **Fall 2016**
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Long-Term Outcomes After Lung Transplantation in Children With Intellectual Disabilities **2024**

Pediatric Transplantation

Frameworks for Estimating Causal Effects in Observational Settings: Comparing Confounder Adjustment and Instrumental Variables **2023**

BMC Medical Research Methodology

Disparities in hepatocellular carcinoma survival by Medicaid-status: a national population-based risk analysis **2023**

European Journal of Surgical Oncology

Development and evaluation of a dietary bisphenol A (BPA) exposure risk tool **2023**

BMC Nutrition

Evaluating Predicted Heart Mass in Adolescent Heart Transplantation **2022**

Journal of Heart and Lung Transplantation

Weight Matching in Infant Heart Transplantation: Analysis of A National Registry Analysis **2022**

Annals of Thoracic Surgery

Unexplained Mortality During the U.S. COVID-19 Pandemic: Retrospective Analysis of Death Certificate Data and Critical Assessment of Excess Death Calculations **2021**

BMJ Open

Where do we go from here?: A Framework for using SIR Models for Policymaking in Emerging Infectious Diseases **2021**

Value in Health

A US population health survey on the impact of COVID-19 using the EQ-5D-5L **2021**

Journal of General Internal Medicine

Adults with Mild to Moderate Congenital Heart Disease Demonstrate Measurable Neurocognitive Deficits **2020**

Journal of the American Heart Association

REFEREED CONFERENCE PUBLICATIONS

Interpretable Clustering for Patient Phenotyping using Advanced Machine Learning Models **October 2024**

Computing in Cardiology 2024 (IEEE)

Synthetic Data Generation in Small Datasets to Improve Classification Performance for Chronic Heart Failure Prediction **October 2023**

Computing in Cardiology 2023 (IEEE)

Patient Phenotyping using Interpretable Clustering to Study Clinical Outcome in TAVR Patients **October 2022**

Computing in Cardiology 2022 (IEEE)

ABSTRACT OF THE DISSERTATION

Enhancing the Utility of Instrumental Variables in Observational Research

By

Roy S. Zawadzki

Doctor of Philosophy in Statistics

University of California, Irvine, 2024

Professor Daniel Gillen, Chair

A central goal in health research is to estimate the causal effect of a treatment or exposure on an outcome of interest. When randomization cannot be achieved due to ethical, feasibility, or monetary constraints, we must turn to observational studies to isolate causal effects. One core challenge in this setting is controlling for confounding, or extraneous factors that cause both the exposure and outcome. Observational studies are prone to bias due to unmeasured confounding, which renders methods like confounder adjustment and propensity scores ineffective. This has motivated the instrumental variable (IV) approach where we use a variable that influences the exposure but is otherwise not associated with the outcome, to quasi-randomize the exposure, hence producing unbiased causal effects. This dissertation makes three important contributions to enhance the utility of IVs in their application to observational data. First, in the linear setting, we analytically quantify the relative trade-offs between the confounder and the IV approach under the violation of key causal identification assumptions including unmeasured confounding, the exclusion restriction, independence of the IV, and unmeasured treatment effect heterogeneity. We further provide guidelines for practice and develop a sensitivity analysis procedure to quantify these relative trade-offs. In the next contribution, we move to the topic of nonparametric identification of the local average treatment effect (LATE), the estimand targeted by IVs, by developing an influence function (IF) based estimator to incorporate unknown sampling weights to replicate causal estimates across populations – an important facet of enhancing confidence in observational study findings. Via the use of cross-fitting, our method is able to use machine learning (ML) to flexibly model nuisance functions, including the sampling weights. Furthermore, we extend this framework to provide weighted bounds on the ATE. Our final contribution extends the nonparametric, IF-based framework for identifying the LATE to the time-to-event setting. With time-to-event outcomes, causal inference with IVs is often limited by the proportional hazards assumption and the non-collapsibility of the hazard ratio (HR). Therefore, rather than targeting the HR, we extend the accelerated failure time (AFT) model and the Buckley-James (BJ) imputation procedure to nonparametrically identify the percentage difference in the median survival time among compliers for a binary exposure. With this approach, we are able to circumvent several issues involving the application of IVs to estimate the causal HR and, furthermore, protect against misspecification via the incorporation of ML with cross-fitting and double-robustness properties.

Chapter 1

Introduction

A common goal in healthcare is to estimate the causal effect of an exposure, treatment, or intervention on a particular medical outcome. The gold standard for this task remains a well-controlled randomized control trial (RCT). In this setting, randomization allows us to establish cause and effect estimates because, on average, observed differences between treatment arms will be due to either the assigned treatment or chance.

RCTs may not always be feasible due to ethical, logistical, or monetary constraints. When this is the case, we may turn to observational studies in an attempt to isolate causal effects of interest. Observational studies can provide hypothesis-generating evidence to help inform future investigations of treatments such as extensions to different populations or use cases. In this case, observational studies give researchers real-world evidence surrounding the effectiveness and safety of a treatment to augment RCT findings (e.g. Phase 4 trials). Another common use of observational data, particularly in the epidemiological setting, is to quantify the causal impact of biomarkers such as blood lipids on disease incidence and progression. The recent increase in the volume of electronic health records (EHRs) and insurance claims data coupled with the heightening demand on healthcare systems have created many opportunities for potentially valuable observational studies.

A fundamental challenge to causal inference in observational settings, and one that is central to this dissertation, is the fact that the exposure, treatment, or invention (herein referred to as "the treatment") is not randomized. A common source of bias due to this is confounding by indication, or treatment selection bias, where factors affect both the assignment of treatment and the targeted medical condition. These factors, called confounders, range from patient characteristics to other concurrent treatments.

Two overarching approaches have been developed in the literature for these purposes: adjusting for observed confounders and pseudo-randomization through instrumental variables (IVs). Briefly, confounder approaches aim to "adjust" for all factors that both explain treatment assignment and the outcome. In contrast, IVs determine only the assignment of the treatment but, otherwise, are not associated with the outcome. IVs are used to define a subset of the population whose treatment assignment is free from confounding. More formally, an IV is defined by three main conditions: (i) it influences the treatment assignment (relevance), (ii) it is not a cause of the outcome after conditioning on treatment assignment (exclusion restriction), and (iii) it is not associated with unobserved confounders (independence). The use of IV approach can be thought to theoretically relax the assumption that all confounders are adequately measured and correctly modeled. Even so, both approaches are characterized by untestable assumptions in order to isolate causal effects of interest.

Traditionally, among statisticians and epidemiologists, the confounder approach is heavily utilized; however, IVs are rising in interest. In this dissertation, we focus on optimizing the use of IVs in health research. More specifically, we aim to develop methodology that assists analysts to effectively estimate causal effects that are consistent for the true causal effect under several important and realistic scenarios.

In Chapter 3, we study the behavior of existing confounder and IV approaches under potential assumption violations. Because, as mentioned, both approaches are subject to untestable assumptions, it may be unclear which assumption violation scenarios one method is superior in terms of mitigating inconsistency for the a target estimand such as the average causal effect (ACE). Although general guidelines exist, direct theoretical comparisons of the trade-offs between CAC and the IVAC assumptions are limited. Using ordinary least squares (OLS) for CAC and two-stage least squares (2SLS) for IVAC, two popular methods, we analytically compare the relative inconsistency for the ACE of each approach under a variety of assumption violation scenarios and discuss rules of thumb for practice. Additionally, a sensitivity framework is proposed to guide analysts in determining which approach may result in less inconsistency for estimating the ACE with a given dataset. We demonstrate our findings both through simulation and by revisiting Card's analysis of the effect of educational attainment on earnings, which has been the subject of previous discussion on instrument validity. The implications of our findings on causal inference practice are discussed, providing guidance for analysts to judge whether CAC or IVAC may be more appropriate for a given situation.

In Chapter 4, we shift to the development of novel methodology that utilizes IVs to optimize causal inference in the setting of replicability and generalizability of causal estimates. Replicating causal estimates across different cohorts is crucial for increasing the integrity of epidemiological studies. However, strong assumptions regarding unmeasured confounding and effect modification often hinder this goal. By employing an IV approach and targeting the local average treatment effect (LATE), these assumptions can be relaxed to some degree; however, little work has addressed the replicability of IV estimates. In this chapter, we propose a novel survey weighted LATE (SWLATE) estimator that incorporates unknown sampling weights and leverages machine learning for flexible modeling of nuisance functions, including the weights. Our approach, based on influence function theory and cross-fitting, provides a doubly-robust and efficient framework for valid inference, aligned with the growing "double machine learning" literature. We further extend our method to provide bounds on a target population ATE. The effectiveness of our approach, particularly in non-linear settings, is demonstrated through simulations and applied to a Mendelian randomization analysis of the relationship between triglycerides and cognitive decline.

In Chapter 5, we focus on another common area in health research: examining how certain treatments or exposures affect time-to-event outcomes. Besides handling right-censoring, in the observational setting we must mitigate the effect of unmeasured confounding. One approach to this is through the use of instrumental variable (IV) methods. As is common in the time-to-event setting, we may seek to estimate the causal hazard ratio. Yet, existing IV methods are are limited due to non-collapsibility of the Cox proportional hazards model and violations in the proportional hazard assumption. The accelerated failure time (AFT) model offers an alternative approach to the setting of proportional hazards, expressing the effects as multiplicative changes in survival time. Traditional IV approaches within AFT models are limited to linear settings and inefficiently handle censored data. To overcome these limitations, we propose a novel, nonparametric estimator for the local average treatment effect (LATE) based on the non-linear AFT in conjunction with the Buckley-James imputation procedure to effectively include censored observations. Specifically, our approach estimates the percent difference in median survival time for compliers in studies featuring binary exposures and IVs. By leveraging influence functions and sample-splitting, our estimator can accommodate machine learning techniques to estimate nuisance functions and is doubly-robust for key nuisance functions, protecting against bias due to model misspecification. We demonstrate the performance of our estimator via simulation in both linear and non-linear AFT settings. Furthermore, we apply our proposed methods to a Mendelian randomization analysis examining the relationship between between high-density lipoprotein and progression of cognitive decline among those with mild cognitive impairment.

In this work, by addressing three key issues in causal inference, assumption violations, replicability, and survival analysis, we have meaningfully increased the usefulness of IV methods in health research. Importantly, through the use of IVs, the impact of unmeasured confounding is mitigated, which promises higher-quality causal estimates. First, analysts will be able to use our findings to examine whether an IV approach is justified, both theoretically and empirically via sensitivity analysts. Then, they may enhance the replicability, and, thus, external validity, of IV estimates via nonparametric survey-weighted estimates. And, lastly, extend nonparametric IV methodology into time-to-event settings, which comprise a large portion of health research.

Chapter 2

Background

In this Chapter, we review the fundamentals of causal inference in observational settings as it pertains to the content of this dissertation. We begin by providing a theoretical overview of the confounder and IV causal identification approaches as well as the key assumptions behind each approach. We then rigorously define the causal estimand targeted by the IV approach, the local average treatment effect and discuss common applications of IVs. Following this, we discuss traditional methodology to estimate causal effect including regression adjustment and propensity scores for confounder methods and two-stage least squares for IV methods. These methods will be at the center of Chapter 3. Moving to more modern casual inference approaches, we discuss the use of influence functions to nonparametrically identify causal estimates, which includes the local average treatment effect. This framework will be central to Chapters 4 and 5 as they allow us to flexibly model key quantities with black-box machine learning methods with straightforward inference via sample-splitting. Finally, we review causal inference in time-to-event settings, which is the focus of Chapter 5, for both the confounder and instrumental variable settings.

2.1 Two Identification Approaches: Confounders and Instruments

We begin by considering the simple scenario depicted in Figure 3.1. Such figures are called directed acyclic graphs (DAGs) where the nodes are variables, the edges represent a directed causal effect, and the greek letters represent the magnitude of the causal effect of the respective edge. Contextually, D is an indicator for a binary treatment, Y represents the outcome of interest, U represents a confounding variable, and Z is an IV.

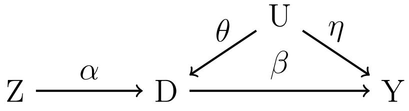


Figure 2.1: A Directed Acyclic Graph with One Confounder and One IV

We may define the causal estimand of interests in two ways: using the potential outcomes framework [101] and Pearl's notation.[86] For the potential outcomes, let $Y_i(1)$ be the outcome if subject i had taken the treatment and $Y_i(0)$ be the outcome if subject i had taken the control. Therefore, we are interested in isolating $\beta = E[Y_i(1) - Y_i(0)]$ in Figure 3.1, or the average treatment effect (ATE). In Pearl's notation, the equivalent quantity is $\beta = \frac{\partial}{\partial D}E[Y|do(D)]$ where $do(D)$ means manipulating the $D - Y$ edge whilst holding the other edges constant. Throughout this background section, we will use the potential outcomes notation.

We cannot observe both potential outcomes for any individual and thus we must use observed values from those who were prescribed either treatment option. In order to identify β, we must make a series of assumptions. First, is stable unit treatment values assumption (SUTVA) of consistency and no interference. Consistency means that everyone receives the same versions of the treatment or control while no interference provides that the potential outcome of one individual does not affect the potential outcome of another person. The second assumption is treatment assignment ignorability or that the treatment assignment is as good as randomized: $Y(0), Y(1) \perp\!\!\!\perp D$. Thirdly, positivity of treatment assignment or $0 < P(D = d) < 1$ with

$d = 0, 1$. Under these assumptions $\beta = E[Y(1) - Y(0)] = E[Y|D = 1| - E[Y|D = 0]$

Our main focus throughout this dissertation pertains to issues surrounding when the assumption of ignorability fails to hold. Clearly, we have a violation because U is a confounder when $\theta \neq 0$ and $\eta \neq 0$ resulting in $Y(0), Y(1) \not\perp\!\!\!\perp D$. Simply estimating the difference in the group means as the above will not recover β due to $E[Y(1) - Y(0)] \neq E[Y|D = 1| - E[Y|D = 0]$. In other words, indirect paths from D to Y (via U) pose major issues in obtaining causal estimates.

There are two ways we will describe how the confounder approach can isolate β. First, we need to find a set of variables $\mathbf{X}$ such that conditional ignorability, $D \perp\!\!\!\perp Y(0), Y(1)|\mathbf{X}$ is achieved. Alternatively, we need to condition upon variables $\mathbf{X}$ such that all alternative or "backdoor" paths from D to Y are blocked.[85] Visually, if we think of a DAG as a set of pipes, the flow of water or "information" will only be through the pipe flowing from D to Y. These two notions of confounding lead us to conclude that in Figure 3.1 it is sufficient and necessary to condition on U and use the finite estimator of $E[Y|D = 1, U] - E[Y|D = 0, U]$ to obtain β. Note that there are many other descriptions of confounding and we will focus on the conditional ignorability and graphical definitions. [116][43]

Alternatively, we may identify the treatment effect through an IV denoted as Z. For simplicity, we assume for now that only one IV is sufficient. Briefly, the definition of an IV is that Z must influence D (relevance), or $\alpha \neq 0$, does not cause Y conditioning on X (exclusion restriction), and is not associated with any unobserved confounders (independence). In Figure 3.1, Z is an IV because $\alpha \neq 0$ and there are no other arrows in or out of Z that go to Y. Figure 2.2 demonstrates how this latter notion can be violated if either δ, ϵ, or ϕ are non-zero. In this case, Z is, in fact, a confounder but if $\delta = 0$ and we are able to condition upon U, then Z is re-classified as an IV. Compared with the confounder approach, if we have access to an IV, we may be able to obtain a causal estimate without having to account for all possible confounders, which is a major potential advantage of using IVs over confounder adjustment methods.

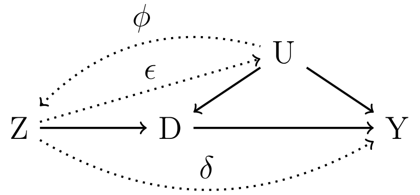


Figure 2.2: Dotted lines disqualify Z from being an IV

The building of a valid causal network requires both knowledge of all variables in the network and the arrows between them. Unfortunately, we cannot be sure that a posited DAG is correct using data. For example, to suggest unobserved confounding requires knowledge that goes beyond the dataset at hand. Nevertheless, we can still use DAGs to capture assumption violations as clearly as possible and move towards the best option in a given scenario. By first thinking outside the scope of the current dataset, one captures a fuller picture of the study.

The untestability of causal assumptions suggests that users of the confounder and IV approaches must think relatively and not in absolutes. For example, rather than arguing a set of confounders is sufficient for conditional ignorability, one should instead find confounders to condition upon that potentially bring the estimate closer to the true causal estimand. For IVs, rather than justifying whether have a true IV or not, we can think about how strongly the treatment is identified relative to potential violations in the assumptions of Z and the hypothesized overall magnitude of confounding.

2.1.1 Identifying and Using Confounders

Without directly modeling the response, in order to avoid spurious results from multiple testing, a reasonable strategy to identify confounders is to first postulate variables that affect the response and then distinguish which of these variables may influence treatment assignment.[20] In the latter step, caution must be taken in the direction of causality: mistakenly adjusting for mediating variables on the pathway from D to Y may produce unintended consequences such as attenuation of the estimated treatment effect. To see this, consider Figure 2.3 – a scenario where W is a mediating variable.

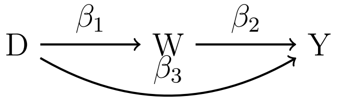


Figure 2.3: DAG of a Mediator, W

For a more formal attenuation scenario, denote the edge weights in Figure 2.3 as $\{\beta_i : i = 1, 2, 3\}$ (the decomposed β in Figure 1). If we assume linear relationships between variables then $\beta = \beta_1\beta_2 + \beta_3$. If $|\beta| > 0$ and $sign(\beta_1\beta_2)) = sign(\beta_3)$, then adjusting for W will yield β_3 but there will be attenuation as $|\beta| > |\beta_3|$.

The first and strongest reference for determining causal links for confounders should be the underlying scientific mechanism. Such information can be based on prior basic science, epidemiological findings, and historical trials. These sources often help one to identify a vast majority of relevant confounding factors. Another source but of potentially lesser quality is past empirical studies done on predictors of the response. One should assess the quality of these studies in terms of replicability, precision, and study design before choosing to use the associated information.

Given a set of potential confounders, it may not always be advantageous to select all of them in the data analysis. In reality, much of the confounding may be captured by a few variables such as basic demographics (e.g. age and sex), commonly collected lifestyle factors (e.g. smoking and alcohol use), and comorbidities (e.g. chronic disease and corresponding medication use). With each confounder included in the analysis, we must weigh moving towards conditional ignorability against overfitting (i.e. increased imprecision and Type II error), interpretability, and reproducibility. Consider that under non-linearity, adjusting for confounders may change the interpretation of the estimate of the treatment effect.[43] In the linear setting, a similar scenario can occur under treatment effect heterogeneity, which is when the effect of the treatment differs across the values of one or more factors.[7]

2.1.2 Identifying and Using Instrumental Variables

Unlike confounders, the use of IVs is not as straightforward and often requires more technical knowledge to employ effectively. The crux of the IV approach is that we use variation independent from confounding to identify treatment assignment. Because IVs, by definition, cannot be determined by other variables in the causal paradigm (there are exceptions: for example, see Figure 2.4), we can assume that the IV values are effectively randomized. As a result, the values of the treatment assignment generated by IVs, denoted D_{IV}, are also randomized. It follows that the estimate using D_{IV}, $\hat{\beta}_{IV}$, is theoretically free of unobserved confounding.

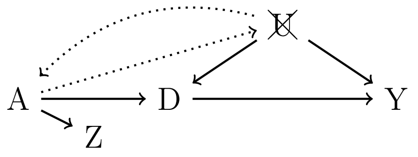


Figure 2.4: Z is a valid IV because we have blocked the path via U and it is a proxy of an effectively randomized variable A. By varying Z, we allow variation in A to influence D.

Complexity arises in using IVs mainly because D and D_{IV} are not technically the same variable. This means that β_{IV} is a different estimand than β and the IVs cannot be used to directly calculate the ATE. However, under some conditions we will soon detail, we may be able to identify "Local Average Treatment Effect" (LATE) where we are "local" to variation in the IVs.[61] Fortunately, if there is no treatment effect heterogeneity and the assumptions for an IV are met, $\hat{\beta}_{IV}$ is consistent for β or, in other words, the LATE will equal the ATE.

As a simple demonstration of the above notions, the form of the LATE with a binary IV and binary treatment can be given by the Wald estimand in Eq. 2.1:

$$\beta_{IV} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}. \tag{2.1}$$

Besides adding more intuition behind IV-derived treatment effects, this equation introduces the importance of IV strength or "predictive power" captured by $E[D|Z=1] - E[D|Z=0]$ or α in Figure 3.1. Heuristically, when α is small then the IV is "weak" and if α is sufficiently large then the IV is "strong." In the linear setting, the finite sample bias of $\hat{\beta}_{IV}$ is partially a function of α where we incur large bias for β with small values of α.[19]

The impacts of weak IVs are not just limited to finite samples. Recall that we cannot confirm we have a true IV using observed data and so we must assume our IV estimate is inconsistent for β. In this case, as an IV becomes weaker, the sensitivity of the corresponding estimate $\hat{\beta}_{IV}$ to IV independence assumption violations increases.[131] To elucidate this, suppose we had two IV candidates Z_1 and Z_2 with corresponding strengths α_1 and α_2, where $|\alpha_1| > |\alpha_2|$. For the same degree of violation in the independence assumptions (e.g. in Figure 2.2 $\delta = c > 0$ where c is some constant) the inconsistency of an estimate derived from Z_2 would be greater than from using Z_1.

Table 2.1: Potential Treatment Assignment

Sub-population	$D(0)$	$D(1)$
Always Takers	1	1
Compliers	0	1
Defiers	1	0
Never Takers	0	0

When there is treatment effect heterogeneity, even when we have a valid IV, $\hat{\beta}_{IV}$ can be inconsistent for β because β_{IV} is an estimand for a subset of the original population. Table 2.1 summarizes four distinct sub-populations related to the IVs: always-takers, compliers, defiers, and never-takers. We can use potential outcomes once again but for treatment assignment: let Z be a binary instrument and $D(0)$ be the treatment assignment had the value of the IV been 0 and $D(1)$ had the values of the IV been 1.

In Table 2.1, it is clear that changing values of the IVs results in changing values of the treatment assignment only for compliers and defiers. Therefore, we cannot directly identify the treatment effects for always-takers and never-takers using IVs. In addition, to achieve the LATE we must impose a further assumption that the defier population does not exist for a given IV. This notion is called "monotonicity" where $D(1) \geq D(0)$ or vice-versa. Thus, we conclude that the subpopulation identified by the IVs are the compliers. To explain why we require monotonicity, we can rewrite Eq. 2.1 as[5]

$$\begin{aligned} E[Y|Z=1] - E[Y|Z=0] &= E[\big(Y(1)-Y(0)\big)\big(D(1)-D(0)\big)] \\ &= \Big(E[Y(1)-Y(0)|D(1)>D(0)]P(D(1)>D(0)) \\ &\quad - E[Y(1)-Y(0)|D(1)<D(0)]P(D(1)<D(0))\Big). \end{aligned}$$

Because of treatment effect heterogeneity, the ATE is differential depending on the subpopulation and there is the potential for a non-zero treatment effect in each group to cancel out. Of course, this does not occur if $P(D(1) < D(0)) = 0$ (no defiers) or $E[Y(1) - Y(0)|D(1) > D(0)] = E[Y(1) - Y(0)|D(1) < D(0)]$ (no treatment effect heterogeneity). If monotonicity is violated, then our resulting estimand is no longer the LATE and generally ambiguous.

Through a similar derivation in the denominator of Eq. 2.1 as above, we arrive at a clearer definition of the LATE: $\beta_{IV} = E[Y(1) - Y(0)|D(1) > D(0)]$ or the ATE for compliers.[61] This LATE, changes with chosen IV. If there are multiple IVs, then the LATE is a weighted average of LATEs characterized by each IV. When we have covariates included to establish the validity of Z or decrease error in predicting D, then the LATE is an estimand defined on a population conditional on these covariates. Furthermore, unless the model is saturated, always and never-takers are included.[2, 18] As most models in practice include covariates, the interpretability of IV models can be nebulous.

With the potential outcomes notation, we may formalize the core IV assumptions in the more general case where must must control for confounders X to satisfy the independence assumption. One common way to control for measured confounders is via the PS.[100] In the context of IVs, we utilize the instrument propensity score (IPS) $e(X) = P(Z = 1|X)$ to control for IV-related confounding. The conditions to identify the LATE via the IPS are similar to the that of ATE: we require positivity of the IPS and for the set of confounders to achieve strong ignorability, or "independence," of the potential outcomes. Additionally, the use of confounders additionally modifies the standard IV assumptions of relevance and monotonicity to be strata-specific. Formally, we may define the assumptions to identify the LATE conditional on a set of confounders X as follows:

Assumption 1 (Positivity of IPS): $0 < e(X) < 1$ a.s. for all $x \in \mathcal{S}_X$

Assumption 2 (Independence): $Z \perp\!\!\!\perp \{Y(0,0), Y(0,1), Y(1,0), Y(1,1), D(1), D(0)\} \mid X$

Assumption 3 (Exclusion Restriction): $Y(d,0) = Y(d,1)$ for $d \in \{0,1\}$

Assumption 4 (Relevance): $P[D(1) = 1|X] > P[D(0) = 1|X]$ a.s. for all $x \in \mathcal{S}_X$

Assumption 5 (Monotonicity): $P[D(1) \geq D(0)|X] = 1$ a.s. for all $x \in \mathcal{S}_X$

where $\mathcal{S}_X$ refers to the support of X and $Y(d, z)$ refers to the potential outcome under treatment assignment d and IV assignment z.

In practice, there usually exists treatment effect heterogeneity so before using IVs we must determine if it is reasonable to target the LATE, or whatever the IV estimand may be, as a proxy for the ATE estimand. Consider the following scenarios. First, when treatment effect heterogeneity is unrelated to the choice of treatment then the estimates for compliers will not be systematically different from the other subpopulations. Next, when the first-stage is strong, the population characterized by the IVs will be relatively close to the overall population of the study. For example, if we have found a strong genetic determinant of some condition that was a valid IV, it is plausible that, for the vast majority of the population, the occurrence of the condition would vary with the assignment of the gene. It follows that if the IV is weak, the LATE will only capture a small subset of the original population, introducing significant inconsistency in estimating the ATE. Lastly, there is a developing set of literature that relaxes the assumptions for the Wald estimand to equal the ATE such as requiring heterogeneity in the outcome caused by the treatment assignment to be independent from heterogeneity in treatment assignment caused by the IV as well as the IV itself.[6, 122, 47] Note that in these cases, we are not necessarily concerned with monotonicity if we are targeting the ATE not the LATE. In Chapter 3 of this dissertation, we will focus on identifying the ATE using the LATE whereas in Chapters 4 and 5, we will directly estimate the LATE without concern of whether it is equal to the ATE, though it may be.

Identifying potential IVs is significantly less straightforward than identifying confounders, which is a main limitation of the approach. While there is usually abundant literature on predictors of a medical condition, the factors that determine the assignment of a treatment are difficult to study and are not usually studied. One reason for writing this paper is to generate exposure to biomedical researchers to IVs such that potential IVs can be shared in the literature similar to how predictors are. In a similar vein to the confounder adjustment approach, we may begin by determining factors that predict treatment assignment and then prune those that affect the outcome. Determining confounders first is helpful as variables that were once invalid IVs may become valid after holding certain confounders constant.

One popular source of IVs is variation in medical practice as it is well known that practice differs across physicians and regions across a wide variety of medical conditions.[128, 127, 35] If appropriate, we could use factors such as regional variation, facility prescribing patterns, attitudes to certain contraindications, physician preference, and calendar time as IVs. [27, 21] For example, with access to the relevant data, physician preference can be quantified by tabulating the proportion of patients under each physician who were prescribed the treatment of interest. Following this, we can use these proportions to predict which treatment a new patient who sees any of these physicians will receive.

Even still, the validity of prescriber preference as an IV can be questioned. It could be that certain types of patients tend to select a physician that they know is more likely to give them the treatment (graphically, Figure 2.2 edges from Z to U). Furthermore, geographic variation in general population health could necessitate higher utilization of treatments in some regions compared to others. Herein lies the value of identifying confounders in IV analyses: perhaps controlling for patient characteristics will block these pathways and greatly reduce assumption violations (e.g. Figure 2.4). One takeaway, however, is that IV analysis can easily suffer from issues related to unobserved confounding.

Given a set of IVs, we should characterize each subpopulation. For medical practice patterns, most likely, some patients would not comply with a doctor's opinions; some patients could insist to get the treatment (alway-takers) and others would refuse under all circumstances (never-takers). One that does the opposite of what the doctor says (defiers) is possible and we will have to assume that they do not exist, which is practically untestable but can be reasoned as unlikely. Under this assumption, the LATE would roughly be those who follow the doctors' orders. All of this considered, the analyst should determine whether the complier treatment effect is of scientific value or the sufficient conditions have been attained vis-a-vis treatment effect heterogeneity to directly target the ATE.

Another common class of IV, particularly in the epidemiological setting, is Mendelian Randomization (MR) wherein known genetic associations with a key exposure of interest are utilized as IVs.[102] Common examples include the examining the effect of BMI and cardiovascular disease or the role of cholesterol in the risk of dementia. where genome wide association studies (GWAS) have identified one or more single nucleotide polymorphisms (SNPs) predicted the level of the exposure (i.e. the relevance assumption). These SNPs can be combined to form an IV if they meet the causal assumptions described above. The complier subpopulation is thus those whose level of the exposure varies monotonically with the level oof gene expression. As with any application area, there are several challenges in particular to MR including linkage disequilibrium, canalization, winner's curse in GWAS, reverse causation, and gene-environment interactions.[117] In our MR analyses throughout this work, we will assume these challenges have been mitigated as methodologically addressing them is outside of the scope of this dissertation.

2.1.3 Interactions of the Confounder and IV Approaches

The confounder and IVs approaches are deeply related. Therefore, even if an analyst decided to pursue one approach over another, awareness of the principles of the other approach is important. One pervasive issue in this vein is adjusting for an IV as if it was a confounder. Widely-cited guidelines such as Hirano and Imbens (2001) state that variables that are predictive of treatment assignment should be selected for confounder methods like propensity scores,[52] which risks adjusting for IVs and mediators. In the best case, treating IVs as confounders decreases precision because it does not explain variation in the response. Even worse, when there is unobserved confounding, existing inconsistency is amplified, also named "bias amplification."[17, 132, 37, 87] By adjusting for IVs, we reduce variation in the treatment that is uncorrelated with the unobserved confounding. Thus, variation in the treatment produced by unobserved confounding proportionally increases, which causes more bias in the treatment effect.

The impact of adjusting for IVs and mediators demonstrates why one should avoid a purely "kitchen sink," data-driven approach to variable selection for causal inference. Simply because the estimate of the treatment effect changes when a variable is introduced does not necessarily mean it should be adjusted for. This is one reason why we advocate that confounders largely be sourced *a priori* by first hypothesizing predictors of the outcome. If one is reasonably certain that a variable is predictive of the outcome but is unlikely to be associated with the predictor of interest, one has a "precision variable," which still may be of use. Specifically in the linear model setting, adjusting for such a variable will decrease standard errors in the treatment effect estimate with no cost to bias.[22, 43]

2.2 Traditional Methodology for Estimating Causal Effects

In this section, we discuss three popular approaches for estimating causal effects: regression adjustment using ordinary least squares (OLS) or generalized linear models, propensity score weighting with inverse probability of treatment weighting (IPTW), and utilizing two-stage least squares (2SLS) with IVs. Regression and IPTW are used in the confounder approach while 2SLS serves as a methodology for the IV approach.

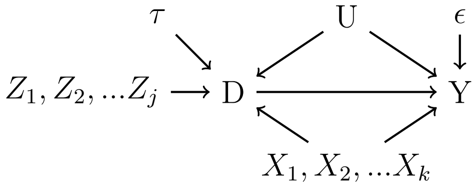


Figure 2.5: A DAG with more IVs (Z's) and observed confounders (X's) as well as relevant stochastic errors τ and ϵ for D and Y, respectfully

A more sophisticated version of Figure 1 is presented in Figure 5. We have added a vector of IVs of length j $(Z_1, Z_2, ...Z_j)$ and a vector of observed confounders of length k $(X_1, X_2, ...X_k)$. In addition, we have there are now stochastic errors for D and Y. For simplicity, assume the effect of each IV is the same magnitude and similarly for each confounder and that the outgoing arrows capture from the joint effect. Furthermore, U captures all unobserved confounding though, in reality, there are likely many variables. Assuming the relationships between variables are linear, we can write the following system of relevant structural equations:

$$d_i = \gamma_0 + x_i^T \gamma_X + z_i^T \gamma_Z + u_i \gamma_U + \tau_i, \tag{2.2}$$

$$y_i = \beta_0 + \beta_D d_i + x_i^T \beta_X + u_i \beta_U + \epsilon_i. \tag{2.3}$$

Eq. 2.2 depicts the treatment assignment or "first stage" while Eq. 2.3 depicts the outcome or "second stage." The estimand of interest is β_D. Because the treatments are usually binary variables, the functional form of the treatment assignment is commonly characterized using a logit or probit model. For ease of exposition, however, we will assume a linear probability model (LPM) as in Eq. 2.3.

Using the above two equations, we can provide a high-level overview of the three methods in practice. Regression methods fit Eq. 2.3 with the treatment and the observed confounders to estimate β_D. U is not observed so the estimate is inconsistent due to misspecification. Meanwhile, IPTW first fits Eq. 2.2 with only the confounders to predict the propensity score for all subjects. The propensity scores will be used to compute a weighted sum that will allow us to estimate β_D. Because U is missing, the predicted propensity scores will not be correct nor adequate to achieve ignorability conditional on the propensity score.

2SLS will fit Eq. 2.2 as stated (except for U) and use the predictions to construct $\hat{D}$. We then effectively substitute D in Eq. 2.3 and fit an OLS model to estimate the coefficient in front of $\hat{D}$. Importantly, the omission of U does not affect the consistency of this estimate under the conditions of Figure 2.5 and if there is no treatment effect heterogeneity then we have a consistent estimate of β_D. If there is then, at the least, the estimate is not affected by U.

2.2.1 Regression Adjustment and Propensity Score Methods

While OLS and IPTW are different mathematically, they are conceptually similar: both seek to isolate variation in the outcome caused by the treatment by eliminating variation caused by confounding factors. Regression adjustment can be thought of as blocking paths in Figure 2.5, which is another way of stating that we are holding X constant in order to isolate the effect of D on Y and obtain the direct causal pathway with the following estimand:

$$\beta_D^{OLS} = E[Y|D=1, X=x] - E[Y|D=0, X=x].$$

On the other hand, IPTW weights outcomes based on the probability of receiving $(p(X) = P(D=1|X)$, creating a pseudo-population that balances confounders across the treatment groups in a similar rationale to randomization. IPTW has the following estimand:

$$\beta_D^{IPTW} = E\left[\frac{DY}{p(X)}\right] - E\left[\frac{(1-D)Y}{1-p(X)}\right]. \tag{2.4}$$

For a more concrete example of how a pseudo-population is constructed, suppose an individual in the treatment group had a propensity score of 0.1. In other words, this individual is very likely to receive the control and has many similar subjects in the control group. The weighted outcome of this individual can represent the counterfactual outcomes of the comparable control group subjects and so we would weight that individual's contribution to the treatment effect ten times. As a consequence of such a procedure, in Figure 2.5 the pseudo-population will theoretically no longer contain edges for the X's because the treatment assignment cannot be explained by covariate imbalance. A balance of observed confounders, however, does not imply a balance of unobserved confounders: the act of balancing observed confounders may increase the imbalance of these unobserved confounders.[23]

Propensity scores are often used to match individuals across treatment groups. Once one or more suitable matches are found for each subject in the treatment groups, we can compute the differences in outcomes, and average them to obtain the average treatment effect on the treated (ATT), $E[Y(1) - Y(0)|D = 1]$. We focus on IPTW and not propensity score matching because, under unobserved confounding, the ATT will not equal the ATE:

$$\begin{aligned}\beta^{ATT} &= E[Y(1) - Y(0)|D = 1] = E[Y(1)|D = 1] - E[Y(0)|D = 1] \\ &\neq E[Y(1)|D = 1] - E[Y(0)|D = 0] \ (Y(0) \not\perp\!\!\!\perp D) \\ &= \beta^{ATE}.\end{aligned}$$

While our discussion of methodology in this dissertation mainly centers around the ignorability assumption, it is important to briefly touch upon the implications of positivity assumption violations. In propensity score methods this means each subject has a positive probability of receiving the treatment given each level of the covariates or $0 < P(D = d|\boldsymbol{X} = \boldsymbol{x}) < 1$ for all $d \in D$ and $\boldsymbol{x} \in \boldsymbol{X}$. Another way to conceptualize the positivity assumption is the notion of "common support" where there must be full overlap in each group's distribution of propensity scores or, by extension, their observed covariates. Even when there is no unobserved confounding, positivity violations can arise when we fail to observe certain variables that are needed to create overlap. Therefore, for the subpopulations lacking overlap, extrapolating counterfactual claims can lead to erroneous conclusions.

Conceptually, a violation of the positivity assumption means that we are dividing by 0 in Eq. 2.4. In practice, this results in extreme weights, which both increases variability in the parameter estimates but also impacts finite sample bias because the estimate is weighted towards the few extreme observations.[90]. In addition, near-positivity violations, or individuals who are extremely unlikely to receive the treatment or placebo, pose similar issues for estimation. When using OLS, positivity violations pose a similar bias of the estimand as IPTW. Nevertheless, OLS does not face the same degree of finite sample estimation issues as IPTW when there are positivity or near-positivity violations.

While IPTW and OLS target the same estimand, the ATE, and both are vulnerable to inconsistency via unobserved confounding, there are differences to consider in practice. If there are no extreme weights, by encapsulating many covariates in the propensity score, IPTW will generally be more efficient than OLS because of the degrees of freedom saved. If there are extreme weights, however, the instability of the variance of estimators will be larger than that of adjustment-based regression methods.

One common solution to extreme weights in propensity score methods is to trim extreme weights. This procedure, however, risks estimating the treatment effect for a population different than the original target population. In other words, for a decrease in variance, there is a potential increase in bias. Furthermore, the direction of this bias is difficult to determine because one must define a new population resulting from truncation. Though one could argue that the bias due to positivity violations could be advantageously traded-off with the bias due to truncation[129]. Another common solution to extreme weights could be to use stabilized weights as opposed to conventional inverse probability weights.[97]

OLS and IPTW also have potential differences in ease of reproducibility and interpretability. Because propensity scores are the result of fitting a model for treatment assignment in order to generate propensity scores, methods ranging from logistic regression to random forests may be used. An issue, however, arises when different models produce different sets of propensity scores resulting in different pseudo-populations. This poses challenges for reproducibility across different studies of the same population. The simplicity of OLS arguably reduces the risk of this since basic adjustment can be easily communicated. On the other hand, when unaccounted for treatment effect heterogeneity exists, OLS will generate a marginal treatment effect estimate that is implicitly weighted by the covariance structure in the observed data sample as opposed to explicit weighting in IPTW.[7, 5]

One advantage of the propensity score is that data-driven selection of the confounders to model treatment assignment is done separately from the fitting of Eq. 2.3. In contrast, adding and removing confounders in OLS in a data-driven fashion will also affect the estimand, estimate, and corresponding inference for the treatment effect. Therefore, IPTW is able to control inflation in Type I error from repeated testing of the treatment effect coefficient as a result of fitting several models.

Though it may be tempting to cast estimating propensity scores as a prediction problem, this may lead to unintended consequences. The original philosophy of propensity scores from Rosenbaum and Rubin is not to fit the first-stage as well as possible; rather, it is to find a balancing score sufficient to achieve ignorability.[100, 9, 119, 57] Furthermore, measures of model performance like the C-statistic do not provide useful information to suggest unobserved confounding is mitigated more in one model than another.[126]. IVs are predictors of treatment assignment and, yet, they are adverse to causal estimation if included in the model.[17, 9] In addition, including variables that are predictive of the outcome but not the treatment can help improve the efficiency of treatment effect point estimates.[22]. Therefore, we suggest that variable selection for propensity score modeling is not conducted purely by optimizing out-of-sample model prediction error.

In contrast, the simplicity of using one equation in regression methodology assists in avoiding confusion between prediction and inference goals because there is no intermediate step in obtaining an estimate for β_D. Indeed, an optimized model MSE may reduce coefficient standard errors and yet could lead to issues in internal validity. For example, if we fit Eq. 2.3 with LASSO to select confounders that optimized out-of-sample prediction error, the elimination of confounders via shrinkage to zero leads to potential omitted variable bias because we have unblocked a path in the DAG.[29]

A combination of the IPTW and regression methodology is the augmented IPTW (AIPTW) with the so-called "doubly robust" property: a consistent estimate is obtained if either the propensity score (Eq. 2.2) or the outcome equation (Eq. 2.3) are correctly specified. Certainly, AIPTW offers more robustness than IPTW but its practical advantage is unclear. Firstly, in the case where one suspects the propensity score equation is misspecified but the outcome equation is not, OLS also will result in a consistent estimate and could theoretically be more efficient. Secondly, an unobserved confounder would cause misspecification in both equations, rendering any estimate inconsistent. As such, it is unclear how the inconsistency due to unobserved confounding in the AIPTW compares to that of IPTW or OLS.

2.2.2 Two-stage Least Squares

2SLS is the most commonly used and well-studied IV method. The motivations of 2SLS largely stem from the inadequacy of OLS to provide a consistent estimate of β_D. Under the reduced form in Eq. 2.5, omitting U leads to inconsistency because $\alpha_D = \beta_D + \frac{Cov(D,U)}{Var(D)}$.

$$y_i = \alpha_0 + \alpha_D d_i + x_i^T \alpha_X + \phi_i \tag{2.5}$$

Another name for this scenario is that there is "endogeneity" or correlation of D with the error term. This is because $\phi_i = \beta_U + \epsilon_i$, which is correlated with D and $E[\phi_i|D, X] \neq 0$. IV methods will identify a re-characterized treatment, D_{IV}, using exogenous variation (uncorrelated with the error term) such that $Cov(D_{IV}, \phi_i) = 0$.

In the first stage of 2SLS, we regress the X's and Z's (design matrix $\boldsymbol{F}$) on D and use the projection matrix $P_Z = \boldsymbol{F}(\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}$ to obtain predicted values $\hat{D}_{IV}$. In the second stage, we use the P_Z to obtain the coefficients $\boldsymbol{\beta} = (\boldsymbol{S}^T P_Z \boldsymbol{S})^{-1}\boldsymbol{S}^T P_Z Y$ with design matrix $\boldsymbol{S}$ consisting of the X's and D. By including the X's in both the first and second stages, we can improve the prediction of D and covariates serve as their own IVs. If all assumptions are met, and there is no treatment effect heterogeneity, then β_D^{IV} is consistent for β_D but not necessarily unbiased. In fact, in the case where we only have one IV for one endogenous variable, the first moment does not exist.[67]

There are considerable trade-offs in the IV analysis: consistency comes at the cost of increased standard errors compared to OLS since we only use the exogenous variation in the treatment and, thus, have less "information" to calculate the treatment effect.[131] In the case where the IVs are weak, the finite bias will move towards OLS as weakness increases (i.e. first-stage coefficients go towards 0) and inflate estimator standard errors.[4, 19] This is because our treatment effect is determined only by compliers, a subset of the overall population. If the IVs are correlated with the second stage error term (i.e. U) not only will the estimates be inconsistent but the magnitude of inconsistency is greatly affected by the IV strength.[131] This can be observed by deriving the form of the 2SLS estimand under this condition.

$$\beta_D^{IV} = \beta_D + \frac{Cov(Z, \epsilon)}{Cov(Z, D)}. \tag{2.6}$$

$Cov(Z, \epsilon)$ captures the degree of the violation in the independence of the IV, and $Cov(Z, D)$

captures the first stage strength. Rewriting covariances as correlations in Eq. 2.6 and in the OLS estimate α_D, we obtain:

$$\beta_D^{IV} = \beta_D + \frac{\sigma_\phi}{\sigma_D} \frac{Corr(Z, \phi)}{Corr(Z, D))},$$

$$\alpha_D = \beta_D + \frac{\sigma_\phi}{\sigma_D} Corr(D, \phi).$$

It is clear that when the IVs are weak, 2SLS inconsistency can be greater than even that of OLS if $|Corr(D, \phi)||Corr(D, Z)| < |Corr(Z, \phi)|$. Considering that we can never confirm the IV independence assumption, the use of weak IVs may be perilous.

Resuming our assumption of valid IVs, another perhaps helpful perspective is that the first-stage is chiefly a prediction task. In this interpretation, weak IVs lead to inaccurate first-stage predictions, which leads to finite sample bias because we are unable to adequately capture the treatment assignment of the original population of interest. Simply adding more weak IVs to the first-stage rarely improves the issue; indeed, packing the first-stage with too many instruments will lead to overfitting and, hence, finite sample bias.[99] These issues are partially mitigated by large sample sizes.

Unlike many of the assumptions discussed, the degree of instrument relevance is somewhat determinable using the data. Because 2SLS utilizes OLS for the first stage, the F-statistic is commonly used to measure the joint strength of the IVs with 10 being a "rule of thumb" for sufficient strength. In the case of heteroskedasticity, a robust F-statistic can be used but variance estimates may be noisy.[134] Nevertheless, using a sample statistic to infer upon assumptions could be problematic. For instance, Young 2017 points out that there is a relatively high chance of spuriously obtaining a high F-statistic and unreliability of "guaranteed bounds" of size and bias of weak IVs tests. In addition, weak IV tests assume that the IVs are valid in the first place or else coverage will be incorrect.[44]

Data-driven fitting of the first-stage may prompt concerns about external validity. In theory, we could select variables in the data that optimized a cross-validated F-statistic. In this process, however, we would fail to address the impacts on the interpretability of the LATE. This scenario introduces difficult situations where one set of IVs could have a worse F-statistic but is more interpretable for the desired use case. This is why we advocate for first thinking through the conceptual soundness of the selection of IVs given a set of theoretically strong predictors of the treatment.

Although the independence of the IVs is untestable, there are a few interesting falsification tests that one could employ. One straightforward test is to compare the values of potential confounders across values of the IVs similar to how one examines values of potential confounders across levels of the treatment.[10] Imbalance of IVs across an observed confounder is problematic because the observed confounders could be related to an unobserved confounder. Another test that assumes observed confounders may be related to unobserved confounders involves negative control outcomes, or populations constructed to falsify IV independence.[36] Of course, these tests cannot ensure IV assumptions are met and are subject to data availability.

In the multiple IV case, given there is one valid IV, one may test if additional IVs influence the outcome through Sargan's J-statistic. [103] By running OLS on the residuals of 2SLS on the IVs, if one or more coefficients are not 0, we have some sort of violation. This test may have deceivingly high p-values and poor power when one IV is weak but valid and the others are strong but invalid.[68] The scenario of a mix of strength and validity of IVs poses an interesting question about whether one would choose a strong but slightly invalid IV over a weak but valid IV.

2.2.3 Complexities that Arise with Non-Linearity

So far, our discussion has been focused on the case where the outcome is continuous and, thus, we can reasonably assume linear structural equations. When the outcome is not continuous, some of the statements we have previously made must be modified. In particular, we will revisit the consequences of adjusting for confounders, IVs, and precision variables, marginal versus conditional estimands, and the use of 2SLS for binary treatments and non-continuous outcomes.

First, in the non-linear model setting, the estimand corresponding to the treatment effect will change by including not only confounders but also precision variables because of non-collapsibility.[43] Mathematically, because the covariates are encapsulated in a non-linear function (e.g. a link function), after adjusting for a precision variable, we cannot simply distribute the expected value such that we recover the "before adjustment" treatment effect.[86] A further consequence is that adjusting for any variable will increase coefficient standard errors of the variables already present. For instance, in logistic regression, the unexplained variance must stay fixed so the explained variance will increase upon adjustment for new variables, leading to coefficient values increasing in magnitude.[80, 105] Adjusting for precision variables will increase standard errors of the treatment effect but slightly increase the power to reject the null of no effect; this is due to the magnitude of the adjusted point estimate increasing relative to slight increases in the standard error of the estimate.[104]

The presence of non-collapsibility means that the interpretation of coefficients before and after adjusting for variables differs even in cases where there is independence between the adjustment variable and the treatment variable. Marginal estimates, without confounders, will therefore be different from conditional (on the confounders and precision variables) estimates. Comparing methodologies, IPTW will give a marginal estimate whereas regression will give a conditional estimate.[105] For example, after adjusting for the confounders, IPTW produces the odds ratio for the population characterized by the sample while a logistic regression model would compute the odds ratio for someone with the average value of the confounders.[64] In the linear case, the difference between conditional and marginal effects was only present under treatment effect heterogeneity.

Whether the marginal or conditional estimand is preferable depends on the scenario. One could argue that the conditional estimand is more applicable to settings where a physician is already conditioning upon knowledge of various patient characteristics like sex, age, and comorbidities. In addition, conditional estimands may be better transported to other populations such as future populations. On the other hand, marginal estimands can be more interpretable and comparable across studies.[118] They may also be preferred when the covariates potentially included in the model are not easily observed or measured in practice.

The issue with using IVs as confounders persists in the non-linear setting: adjusting for IVs as if they were confounders amplifies existing bias none and could even introduce bias where none previously existed.[87, 37] Amplification is for the same reasons as the linear setting while introducing bias occurs when the IVs are dependent on the outcome given the treatment.[89] Though Ding et al. do find specific situations where this bias was not present under their proposed monotonicity conditions for the treatment selection and outcome model.

A primary challenge for standard IVs methods is that 2SLS misspecifies a non-linear functional form in the case where we have a binary treatment. Nevertheless, analysts still may utilize 2SLS in non-linear settings such as through LPMs. Simulations have shown that LPM can produce low inconsistency in the estimates of the LATE.[13] A counterpart of 2SLS, two-stage residual inclusion (2SRI), where one takes the residuals from the first stage as a covariate in the second stage, did not perform nearly as well. Furthermore, claims that non-linear 2SRI, (using a probit model for example) are able to recover the ATE (as opposed to the LATE) are questionable.[26] Quantifying the effect of IV method misspeficiation relative to the confounder methods is an avenue of research.

Intuitively, using LPMs appears inappropriate as there is nothing preventing prediction of values that are outside of the interval $[0, 1]$, which leads to bias and inconsistency of LPM estimates. However, if the population of interest only has probabilities between a certain range not close to 0 or 1, then this will seldom be an issue because we will not have errant predicted values. Thus, LPMs will be consistent and unbiased.[54] One could also argue if there are true probabilities that are 0 or 1, then we have a positivity violation and propensity score methods, whether they use a logit, probit, or LPM to calculate propensity scores, will also run into issues. One solution for any method is to truncate probabilities but this will be at the cost of bias for the original causal effect.

We can rarely mimic the 2SLS substitution procedure with non-linear models, such as two-stage probit, and maintain consistency for the treatment effect (such an action is called "forbidden regression").[131] Instead, one can consider a three-step procedure: fit the first-stage with a non-linear model, regress the predicted values on the treatment in OLS excluding the IVs, and, lastly, fit the second stage with a linear or non-linear model.

2.3 Nonparametric Identification of Instrumental Variable Estimands

Recent developments in causal inference methodology have been mostly focused on relaxing assumptions on common approaches like OLS, IPTW, and 2SLS. These developments are predominately motivated by the wish to reduce the impacts of model misspecification and the view that many subtasks of estimation are primarily prediction-based, allowing for more flexible modeling using ML. Another motivation that we will not discuss is sparsity (e.g. incorporating more confounders or IVs than observations). Two prominent examples are targeted minimum loss-based estimation (TMLE) and double machine learning (DML).[114, 15, 28]

There are two central elements to nonparametric causal inference methods: influence functions (IF) and cross-fitting. Briefly, the IF framework is as follows. Suppose we wish to target some estimand $T(P)$ where P is an unknown distribution estimated with known distribution $\tilde{P}$. Usually, we use empirical distribution $\hat{P}$ for $\tilde{P}$. An IF-based approach takes a plug-in estimator $T(\tilde{P})$ and adds an extra term to correct for potential bias in non-parametric settings. This augmented estimator is called the "one-step estimator:"[40]

$$\hat{T}_{\text{1-step}} := T(\tilde{P}) + n^{-1}\sum_{i=1}^{n}\phi(z_i, \tilde{P}). \tag{2.7}$$

Notationally, $\phi(Z, P)$ represents the uncentered IF whereas $\mathbb{IF}(Z, P) = \phi(Z, P) - T(P)$ is the centered IF. A helpful connection is to visualize Eq. 2.7 as one step of a Netwon Raphson procedure with respect to the curve that maps the path of the space of possible distribution functions that link $\tilde{P}$ and P (see Fisher and Kennedy Figure 1).[40] Under some mild regularity conditions, this one-step estimator will be non-parametrically efficient and thus is often referred to as the "efficient IF."[65]

As a simple example of the IF, suppose we wanted to estimate $T(P) = E[E[Y|X, D = 1]] = E[\mu(X)]$. For propensity score $\pi(X) = P(D = 1|X)$, the centered IF is

$$\mathbb{IF}(Z, P) = \frac{D}{\pi(X)}\big\{Y - \mu(X)\big\} + \mu(X) - T(P). \tag{2.8}$$

This can be computed by taking the Gateux derivative of $T(P)$, which measures how an infinitesimal change in the distribution of P effects the value of of an estimator. Subsequently, we can estimate each nuisance function (π and μ) and plug the predicted values into the uncentered IF $\phi(Z, P)$ to obtain our point estimate. One may notice this has the same form as the augmented inverse probability weighting estimand for $T(P)$ yielding double-robustness. That is, we are consistent for $T(P)$ if either $\mu(X)$ or $\pi(X)$ is correctly specified.

We may estimate the uncentered IF by utilizing sample-splitting to prevent overfitting of our plug-in estimates. For sake of example, suppose we split our data $A = (X, D, Y, Z)$ with N_A observations in half: $A^n = (A_1, A_2, ..., A_n)$ and $A^N = (A_n + 1, ..., A_{N_A})$ with $n = \lceil \frac{N_A}{2} \rceil$. We would estimate all nuisance functions with A^n and then plug in predicted values resulting from inputting A^N to $\phi(Z, P)$. As to not discard data, we may additionally flip the roles of the partitions and average the two estimates – this is called cross-fitting.

The implication of using cross-fitting is that we may use ML to fit nuisance functions while avoiding verifying complicated empirical process conditions.[28] Because ML flexibly fits to the data, as the sample size grows so does the complexity of these algorithms (e.g. tree depth or number of trees). Increasing complexity translates to increasing entropy because the number of functions needed to cover the function class will need to increase. Thus, without strict limits on complexity, we are unable to meet Donsker conditions to ensure $\sqrt{n}$-convergence rates when we use ML. With cross-fitting, however, we are unconcerned with model complexity as we can compute $T(\hat{P})$ with the prediction on hold-out sets, essentially treating the nuisance functions as fixed. This means that in our estimator, we no longer need to account for the fact that we estimated the nuisance functions.[65]

From Kennedy (2023), the IF procedure to estimate the LATE for a binary treatment is as follows. Let $\mu_z(X) = E[Y|X, Z = z]$, $m_z(X) = E[D|X, Z = z]$, and $e(X) = P(Z = 1|X)$ where $e(X)$. We now have two uncentered IFs for the numerator and denominator:

$$\phi_{num} = \frac{Z}{e(X)}\{Y - \mu_1(X)\} - \frac{1 - Z}{1 - e(X)}\{Y - \mu_0(X)\} + \mu_1(X) - \hat{\mu}_0(X) \tag{2.9}$$

$$\phi_{denom} = \frac{Z}{e(X)}\{D - m_1(X)\} - \frac{1 - Z}{1 - e(X)}\{D - m_0(X)\} + m_1(X) - m_0(X) \tag{2.10}$$

Estimating the nuisance functions on a different sample, we may plug the values into Eqs. 4.4 and 4.6 to obtain the estimator for $\hat{\beta}_{LATE} = \mathbb{P}_n\hat{\phi}_{num}/\mathbb{P}_n\hat{\phi}_{denom}$. To determine the asymptotic distribution, we first write the following decomposition from Lee, Kennedy, and Mitra (2023) proof for Theorem 4.1.[70] Letting ψ_{num} and ψ_{denom} refer to the centered influence functions and $\hat{\phi}^{-1}_{denom} = \left[\mathbb{P}_n(m_1 - m_0)\right]^{-1}$, we have:

$$\hat{\beta}_{LATE} - \beta_{LATE} = \frac{\mathbb{P}_n\hat{\phi}_{num}}{\mathbb{P}_n\hat{\phi}_{denom}} - \frac{P\psi_{num}}{P\psi_{denom}} \tag{2.11}$$

$$= \hat{\phi}^{-1}_{denom}\left[\mathbb{P}_n\hat{\phi}_{num} - P\psi_{num} - \beta_{LATE}\left(\mathbb{P}_n\mathbb{P}_n\hat{\phi}_{num} - P\phi_{denom}\right)\right] \tag{2.12}$$

$$= \hat{\phi}^{-1}_{denom}\left[(\mathbb{P}_n - P)\left\{\phi_{num}(\eta) - \beta_{LATE}\phi_{denom}\right\}\right] \tag{2.13}$$

$$+ \hat{\phi}^{-1}_{denom}\left[(\mathbb{P}_n - P)\left\{\hat{\phi}_{num} - \phi_{num}\right\} - \beta_{LATE}\left(\mathbb{P}_n - P\right)\left\{\hat{\phi}_{denom} - \phi_{denom}\right\}\right] \tag{2.14}$$

$$+ \hat{\phi}^{-1}_{denom}\left[P\left\{\hat{\psi}_{num} - \psi_{num}\right\} - \beta_{LATE}P\left\{\hat{\psi}_{denom} - \psi_{denom}\right\}\right] \tag{2.15}$$

$$= T_1 + S^* + T_2 \tag{2.16}$$

We must also assume the following conditions:

C1: All nuisance functions belong to the Donsker class.

C2: $|Y|$ is bounded a.s. and $e < C_e$ for some constant C_e a.s.

C3: $\|\hat{e} - e\| = o_P(1), \|\hat{m}_z - m_z\| = o_P(1), \|\hat{\mu}_z - \mu_z\| = o_P(1)$ for $z \in \{0, 1\}$

C4: $\|\hat{e} - e\|\text{max}(\|\hat{m}_z - m_z\|\|\hat{\mu}_z - \mu_z\|) = o_P(n^{-1/2})$ for $z \in \{0, 1\}$

Because sample-splitting or cross-fitting circumvents the need to prove all nuisance functions are Donsker, we have that $T_1 + T_2 = o_P(n^{-1/2})$. Therefore, we have the following asymptotic distribution for $\hat{\beta}_{LATE}$:

$$n^{1/2}\left(\hat{\beta}_{LATE} - \beta_{LATE}\right) = n^{1/2}S^* + o_P(1) \xrightarrow{d} N\left(0, E[\Gamma^2]\right) \tag{2.17}$$

where $\Gamma = \left[\mathbb{P}_n(m_1 - m_0)\right]^{-1} \{\phi_{num} - \beta_{LATE}\phi_{denom}\}$.

(C4) gives us two important properties of IF-based estimators. First, it clearly defines the double robustness condition as the second-order error term, T_2. Second, it indicates that we can use estimators that converge at rate $o_P(n^{-1/4})$, such as machine learners, and still achieve $\sqrt{n}$ convergence.[65, 108] Key complexity conditions have been derived for the regularized regression, random forests, and neural networks.[28, 38] In general, by only requiring a $o_P(n^{-1/4})$ convergence rate for nuisance functions, these latter complexity conditions are generally weak and accommodate a wide array algorithms.

From these results, we may construct $(1-\alpha)$-level Wald confidence intervals for $\hat{\beta}_{LATE}$. Letting $q_{1-\alpha/2}$ be the $1-\alpha/2$ quantile of the standard normal distribution and $\hat{\Gamma}$ being the plug-in estimator for Γ, we have:

$$\hat{\beta}_{LATE} \pm q_{1-\alpha/2}\sqrt{\frac{\mathbb{P}_n\hat{\Gamma}^2}{n}}. \tag{2.18}$$

2.4 Causal Inference in Time to Event Settings

2.4.1 Survival Analysis and Confounder Adjustment

A common scenario in observational health research is isolating causal effects related to the impact of an exposure on the time until an event of interest such as disease progression or death. A primary challenge in this setting is that of right-censoring where individuals drop out of the dataset prior to their event meaning that their outcome is unknown. Discarding these observations typically leads to bias and inefficiency in causal estimates. The bias, in part, arises due to uncensored observations having, on average, shorter survival times than right-censored observations and potential covariate-dependent censoring, inducing selection bias. Therefore, any causal inference method in this setting must focus on mitigating the effects of right-censoring.

Consider a binary treatment $D \in \{0, 1\}$ and instrumental variable (IV) $Z \in \{0, 1\}$ with potential survival time denoted $T(d)$ and potential treatment assignment $D(z)$ where $T = DT(1) + (1 - D)T(0)$ and $A = ZD(1) + (1 - Z)D(0)$. Notationally, in time-to-event outcomes, for subject i, censoring prevents the observation of T_i if there exists censoring time C_i such that $T_i > C_i$. As such, let $\delta_i = I(T_i < C_i)$ and observed outcome $Y_i = min(T_i, C_i)$. We make the additional assumption that the censoring time is independent from the survival, given the covariates and treatment assignment: $T \perp\!\!\!\perp C|D, X$. This assumption is crucial to be able to treat uncensored observations as a simple random sample of the population.

Suppose for expository purposes, we wanted to estimate the ATE β_1. The most common censoring-robust method would be the Cox proportional-hazards model where the ATE the coefficient in front of D:

$$h(t \mid D, X) = h_0(t) \exp\left(\beta_1 D + \beta_2^T X\right). \tag{2.19}$$

where $h_0(t)$ is the baseline hazard at time t.

Under independent censoring, strong ignorability, positivity, and SUTVA, we may interpret $exp(\beta_1)$ as the causal hazard ratio (CHR), which tells us the relative increase or decrease of risk of an event over the study period due to the treatment. As discussed in previous sections, we may also use the propensity score to adjust for confounding, for example, via inverse probability weighting.[8]

One major concern with the CHR is its properties under the violation of the proportional hazards assumption, which states that the hazard ratio between the two treatment groups remains constant over time. For example, Hernan (2013) points out that since the CHRs are changing over time, a single number summary may be misleading and there is built-in selection bias over time as the subjects that remain at risk may be systematically different than those who have had the event.[49] To move away from the proportional hazards assumption, some have instead chosen to model time-to-event data via the accelerated failure time (AFT) wherein the survival time is regressed upon the treatment and covariates.[113, 125] The general form is log-linear:

$$log(T) = \beta_1 A + \beta_2^T X + \epsilon \tag{2.20}$$

where $\epsilon \sim N(0, 1)$. $exp(\beta_1)$ tells us the degree to which survival time is multiplicatively changed (i.e. accelerated or decelerated) in the treatment group compared to the control. Handling right-censoring in the AFT is important as T is not observed for all observations. One solution to this is to impute the survival time for the censored observations via the Buckley-James procedure.[24] In particular, we are imputing $log(T)$ with $Y^* = log(Y)\delta + E[log(T)|T > Y, A, X](1 - \delta)$ noting that $E[Y^*] = E[log(T)]$.

The focus of the procedure is on $E[log(T)|T > Y, X, A] = \mu(X) + \kappa$ where $\mu(X) = \beta_1 A + \beta_2^T X$ and $\kappa(X) = E[\epsilon|\epsilon > log(Y) - \mu(X), X]$. Since we cannot observe ϵ we instead use observed residuals $r_i = log(Y_i) - \mu(X)$. Ordering, $r_1 < r_2 < ... < r_{n_k}$, we can estimate κ nonparametrically via

$$\hat{\kappa} = \hat{S}(r_i)^{-1} \sum_{r_j > r_i} r_j \delta_j \Delta \hat{S}(r_j) \tag{2.21}$$

where $\hat{S}(r_i)$ is the Kaplan-Meier estimator of the survival function for residual r_i.[124] We fit $\mu(X)$ with $\beta = (\beta_1, \beta_2)$ iteratively as follows:

1. Set $M = 0$ and obtain initial estimate $\hat{\mu}^{(0)}$ using the uncensored data as well as $\hat{\kappa}$ from the residuals.

2. At the Mth iteration:

(a) $Y^* = \hat{\mu}^{(M-1)}(X_i) + e_i\delta_i + (1-\delta_i)\hat{\kappa}_{za}(X_i)$

(b) Fit $\hat{\mu}^{(M)}(X)$ using X and Y^*

3. Repeat Step 2 until a set number of iterations is reached or, for some constant $\alpha > 0$,

$$\frac{\left\|\hat{\beta}^{(M)}(X_i) - \hat{\beta}^{(M-1)}(X_i)\right\|}{\hat{\beta}^{(M-1)}(X_i)} < \alpha \tag{2.22}$$

2.4.2 Instrumental Variables in Time-to-Event Settings

As in with any causal setting, the aforementioned survival analysis methods are susceptible to bias for the ATE due to unmeasured confounding. Thus, a branch of the causal literature is focused on adapting survival methods to the IV settings. For example, the complier CHR can be estimated using 2SRI by taking the residuals from the first stage and incorporating them into a Weibull or Cox proportional hazards model.[112, 45, 109] The primary limitation of these methods is that non-collapsibility biases the estimate of the CCHR, which can be partially mitigated via including a frailty term in the second-stage.[76, 77] Another path to mitigate non-collapsibility, would be to assume an additive hazards model though considered by some to be biologically implausible.[111, 75]

Of course, targeting any causal HR, whether it be via the confounder or IV approach, is subject to the same limitations highlighted above. Thus, some have naturally extended AFT methods and, moreover, Buckley-James to the IV setting via two-stage predictor inclusion (2SPI).[81, 55] In this procedure, the predicted values from the first stage are put in place of the treatment in the outcome model. 2SRI may also be similarly used. Because the AFT is linear in nature, there is no issue with non-collapsibility. Lastly, methodology to nonparameterically estimate the difference in the survival probability at a certain time amongst compliers has been developed, which utilizes many of the IF properties previously described.[70]

Chapter 3

Choosing the Right Approach at the Right Time: A Comparative Analysis of Causal Effect Estimation using Confounder Adjustment and Instrumental Variables

3.1 Introduction

A common goal in observational studies is to estimate the average causal effect (ACE) of a treatment or exposure on a specific outcome such as the effect educational attainment on earnings. Due to the exposure not being randomized, the presence of confounders may bias estimates of the ACE. Confounders are factors that influence both the level of exposure (or treatment assignment) and the outcome. If uncontrolled for, confounders can create extraneous differences between the exposure groups that make it difficult to isolate the causal effect. The effectiveness of confounder adjustment for causality (CAC), however, is contingent on the untestable assumption that a sufficient set of confounders, or suitable proxies, is present in the dataset such that confounding can be accounted for appropriately. In other words, we cannot use observed data to prove that the CAC is consistent for the ACE. A simple example is if household income, a known confounder for educational attainment and earnings, and any proxies were missing from the dataset we would like to analyze. Without *a priori* knowledge that household income was a confounder, we would not be able to ascertain that the CAC assumptions were violated.

An alternative approach that may avoid concerns about unobserved confounders is the instrumental variable (IV) approach, or instrumental variable analysis for causation (IVAC). An IV is defined by three main conditions: (i) it influences the treatment assignment (relevance), (ii) it is not a cause of the outcome after conditioning on treatment assignment (exclusion restriction), and (iii) it is not associated with unobserved confounders (independence). In the event that we have a variable that satisfies these conditions, we could then use variation in the IV as a proxy for variation in the treatment and measure the effect on the outcome. Importantly, by (iii), the variation in the IV is independent from unobserved confounding and, therefore, unlike CAC, IVAC does not require appropriately accounting for all possible confounding to consistently estimate the ACE.[10]

When choosing to use CAC over the IVAC, and vice-versa, we trade one set of untestable assumptions for another. For CAC, we cannot prove that we have properly accounted for confounding while, for IVAC, we cannot prove we have a valid IV. For example, proving the exclusion restriction requires us to establish the lack of a direct relationship between the IV and outcome, or a null result, which is not possible with data. Furthermore, to be consistent for the ACE, the IVAC must meet certain untestable conditions surrounding treatment effect heterogeneity. We therefore have no guarantee that under either approach our produced estimate is consistent for the ACE. Despite this, we remain interested in estimating the ACE and subsequently direct our efforts towards attempting to estimate a parameter with the least possible distance from the ACE. In this vein, the key question and central premise of the current manuscript focuses on addressing the following question in practice: under the potential violation of the untestable assumptions, when is the parameter estimated by CAC closer to the ACE than that of IVAC, and vice versa?

In the literature, there exist general guidelines surrounding whether CAC is more appropriate than IVAC.[10, 135] Generally, we contend that analysts should weigh whether the potential degree of unobserved confounding outweighs the potential for violations in the IV assumptions. There is, however, little by way of theoretical research that directly compares the two approaches to assess these trade-offs. To examine these trade-offs between CAC and IVAC, we focus on the use of ordinary least squares (OLS) and two-stage least squares (2SLS), respectively, to estimate the causal effect in each paradigm for two main reasons. First, OLS and 2SLS are the most commonly used methodologies for each approach and, therefore, our findings would be readily applicable to a large group of analyses. Second, linear functional forms will allow us to give intuitive and tractable closed-form results for relative inconsistencies. Then, we will provide a sensitivity framework to guide analysts in determining whether the inconsistency of 2SLS is more than that of OLS, and vice versa.

Alternative estimators for CAC and IVAC include using non-linear machine learning (ML) models such as double machine learning or targeted minimum loss estimation approach.[28, 114] Though flexible modeling may protect against functional form misspecification, they are far from immune to inconsistency due to the assumption violations we will study. Yet, non-linear approaches will carry additional complexity that will make it difficult to analytically quantify the impact of potential violations in closed form. Thus, for our purposes, we will utilize OLS and 2SLS to estimate causal effects under a linear data generating mechanism with the general intuitions gleaned translating to non-linear settings.

In order to more succinctly express the relative performance of OLS and 2SLS under assumption violations, we focus on the scenario where, for a given variable, we must decide on whether the variable should be adjusted for as a confounder in OLS, used as an IV in 2SLS, or not incorporated into the analysis at all. Note that we allow confounders to be adjusted for in 2SLS – an IV is only used in the first stage whereas confounders would need to be present in both the first and second stage.

To our knowledge, there is no existing theoretical literature directly comparing the trade-offs of pursuing CAC versus IVAC, though authors have considered the impact of assumption violations in both settings. The first area of literature relevant to the themes of this chapter is bias amplification, which refers to the fact that using certain variables as confounders may increase pre-existing bias due to unobserved confounding. In the linear setting, an IV is a bias amplifier.[17, 132, 88, 107] In these papers, the authors compare the consistency for the ACE with and without adjusting for an IV. Pearl (2012) provides an extension where there is an imperfect IV (in that the exclusion restriction is violated) and shows that under certain conditions, adjusting for the imperfect IV may actually reduce bias. Similar to this work, he presents his results with both linear structural equations and directed acyclic graphs (DAG) edge-weights to aid understanding of the the trade-offs. Nevertheless, he does not address whether this variable may be more appropriately used in 2SLS.

The bias amplification literature and, by extension, our findings have important implications on applied practice. In particular, the rise of data-driven variable selection approaches for the propensity score, or probability of receiving the intervention, in confounder methods and first-stage for IV analysis. Though IVs, by definition, influence the treatment assignment, they may amplify bias if included in the model for the propensity score.[17] For IV methods, though a variable may not be a perfect IV it may still be worth using it as such. In addition, data-driven modeling of the first-stage may, for example, shrink to zero an important confounder used to achieve the IV assumptions. The gain in the strength of the first-stage may, however, offset the penalty incurred by omitted variable bias in 2SLS. As it stands, these intuitions are difficult to incorporate into variable selection procedures. With this work, we seek to elucidate these complicated scenarios.

In another area of the literature, there has been some work related to comparing OLS to 2SLS under the violation of IV assumptions. It is a well-known result that if an IV is poorly predictive of the intervention (i.e. weak), then small violations in the exclusion restriction and independence assumption can lead to large inconsistencies for the IV estimand.[19, 131]. In addition, to assess the independence assumption, one may compare the impact of intentionally omitting an observed confounder on OLS and 2SLS in order to compare the sensitivity of OLS to that of 2SLS in estimating the ACE.[21] The main assumption behind this procedure is that the impact of omitting an observed confounder on the consistency of OLS and 2SLS is similar to that of omitting a correlated unobserved confounder. A similar assumption and "benchmarking" procedure will be used in our sensitivity analysis.

There are several procedures in the literature regarding sensitivity analyses for violations in IV assumptions. For example, Cinelli and Hazlett (2022) provide a compelling framework and visualization scheme for omitted variable bias in 2SLS based on several partial R^2 measures.[32] Their framework addresses the question of how large the impact of an unobserved confounder would have to be in order to qualitatively change the inferential conclusions of a study, which covers both violations of the exclusion restriction and independence assumptions. A similar procedure to this is the E-value.[115] While we use many of the same tools – notably benchmarking unobserved R^2 measures with observed data from Cinelli and Hazlett (2022) – we do not focus on this sensitivity analysis paradigm of hypothesis testing but instead consider the relative inconsistencies of CAC and IVAC for the ACE.

Another complication to IVAC lies in treatment effect heterogeneity. In this setting, Imbens and Angrist (1994) state that, under monotonicity conditions, IVAC identifies the local average causal effect (LACE) or the causal effect of the "compliers" subpopulation (those whose treatment assignment varies with the IV).[61] If the factors that determine compliance also cause treatment effect heterogeneity, then the LACE may not be equal to the ACE. Hartwig et al. (2020) and Wang and Tchetgen Tchetgen (2018) give clear explanations of the assumptions needed for the LACE to equal the ACE.[48, 122] Essentially, the heterogeneity between the treatment and outcome should be independent of both the IV and the effect modification between the treatment and outcome. For the ease of parameterization in this chapter, we will use Wang and Tchetgen Tchetgen's notion that one of two conditions must be met: (i) no unmeasured confounders are additive effect modifiers of the relationship of both the instrument and treatment or (ii) no unmeasured confounders are additive effect modifiers the treatment and the outcome.

The rest of this chapter is organized as follows. First, we provide the general model setting of interest and introduce relative notation. We then present results regarding the consistency of no adjustment, OLS, and 2SLS for the scenarios of an exclusion restriction violation, independence violation, and treatment effect heterogeneity relevant to IV estimation with and without covariates. In all scenarios, we consider unobserved confounding and isolate the impact of individual assumption violations (e.g. both an exclusion restriction and independence violation). Following this, we present a sensitivity analysis procedure based on partial R^2 and benchmarking unobserved quantities with observed quantities. The goal of this procedure is to give the analyst relevant information to assess the plausibility of whether it may be more appropriate to adjust for a variable in OLS or 2SLS. Next, in simulations, we verify our closed-form results and demonstrate the use of the sensitivity analysis procedure in a variety of scenarios. Then, we apply the procedure to the analysis the effect of educational attainment on earnings conducted in Card (1993), which has been the subject of subsequent analyses over the validity of the IV utilized.[25, 32] We conclude with a discussion regarding the implications of our closed form results on the practice of causal inference and, additionally, provide further guidance on how to use our sensitivity analysis procedure.

3.2 Notation and Set-Up

Our goal is to estimate the ACE of some continuous treatment X on an outcome Y, denoted as $\beta_1 = \frac{\partial}{\partial x}E[Y|do(x)]$ in Pearl's notation.[86] We depict the causal relationships in Figure 3.1, a DAG with edge weights $c_0, \ldots, c_3$. We further consider the following structural equations where $E[\epsilon_1|U, Z] = 0$ and $E[\epsilon_2|X, U] = 0$:

$$X = \alpha_1 U + \alpha_2 Z + \epsilon_1, \tag{3.1}$$

$$Y = \beta_1 X + \beta_2 U + \epsilon_2. \tag{3.2}$$

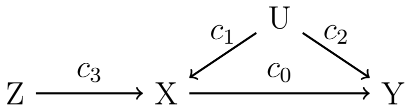


Figure 3.1: A directed acyclic graph with one confounder and one instrumental variable.

Given that all variables in Figure (3.1) are standardized to have mean 0 and variance 1, the edge weights are equivalent to the coefficients in Eqs. (3.1) and (3.2). For example, the ACE, β_1, is the same as the edge weight c_0. In addition, $\alpha_1 = c_1$, $\alpha_2 = c_3$, $\beta_1 = c_0$, and $\beta_2 = c_2$. This equivalence will be helpful in visually expressing assumptions surrounding different scenarios. Furthermore, the edge weights are correlations and are bounded between -1 and 1.

Throughout, we assume the stable unit treatment value assumption (SUTVA) of consistency and no interference. Because U is unobserved, we must estimate the the following reduced form regression where the subscript R indicates these are the values related to the reduced regression

$$Y = \beta_1^R X + \epsilon_2^R. \tag{3.3}$$

U is a confounder because it both influences X and Y, and, therefore, because we have failed to block the path through U, $\hat{\beta}_1^R$ is inconsistent for the ACE. Alternatively, to estimate the ACE, we may utilize Z, which is a valid IV if

1. $\alpha_2 > 0$ (relevance)

2. Z is not a cause of Y conditional on X (exclusion restriction)

3. Z is not influenced by any unaccounted for confounders such that $Z \not\perp\!\!\!\perp Y|X$ (independence).

Supposing that there is no treatment effect heterogeneity or that the conditions of Hartwig et al. (2020) or Wang and Tchetgen Tchetgen (2018) [48, 122] are met, we have $\beta_{IV} = \frac{Cov(Z,Y)}{Cov(Z,X)} = \frac{\beta_1 \alpha_1}{\alpha_1} = \beta_1$, and hence 2SLS will provide a consistent estimate of the ACE. The first equality comes from the definition of the LACE under a continuous treatment and outcome and, under the aforementioned conditions, the LACE is equal to the ACE.

We are interested in estimating and comparing the following estimands: the causal effect i.e. Eq. (3.4), one that omits Z i.e. Eq. (3.5), one that uses Z as a confounder i.e. Eq. (3.6), and one that utilizes Z as an IV i.e. Eq.(3.7):

$$A_1 = \frac{\partial}{\partial x} E[Y|do(x)] = c_0, \tag{3.4}$$

$$A_2 = \frac{\partial}{\partial x} E[Y|x], \tag{3.5}$$

$$A_3 = \frac{\partial}{\partial x} E[Y|x, z], \tag{3.6}$$

$$A_4 = \frac{Cov(Y, Z)}{Cov(X, Z)} = \frac{\frac{\partial}{\partial z} E[Y|z]}{\frac{\partial}{\partial z} E[X|z]}. \tag{3.7}$$

We define the degree of inconsistency for A_1 of estimates of A_2, A_3, and A_4 as a set of absolute differences: $\lambda_2 = |A_1 - A_2|$, $\lambda_3 = |A_1 - A_3|$, and $\lambda_4 = |A_1 - A_4|$. Our goal is to compare the magnitude of $lambda_2$, λ_3, and λ_4. One approach "performs better" than the other if the respective λ is smaller. For example, 2SLS performs better than OLS if $\lambda_4 > \lambda_3$.

As a simple example of the types of calculations and comparisons that we will do in the next section, we can use the conditions of Figure 3.1. From Pearl (2012), we have that $A_2 = c_0 + c_1 c_2$ by Wright's rules of path analysis and $A_3 = c_0 + c_2 \frac{\partial}{\partial x} E[U|x, z] = c_0 + \frac{c_1 c_2}{1 - c_3^2}$.[88] As a result of adjusting for Z as a confounder, we have bias amplification that increases with the strength of the IV (i.e. the magnitude of c_3). In addition, we have that $\hat{A}_4 \xrightarrow{p} c_0$. The proof of this is in the first section of the Appendix. Further, by Chebyshev's inequality and assuming finite variances hold, we also have that $\hat{A}_2 \xrightarrow{p} c_0 + c_1 c_2$ and $\hat{A}_3 \xrightarrow{p} c_0 + \frac{c_1 c_2}{1 - c_3^2}$. Thus, $\lambda_3 = \frac{c_1 c_2}{1 - c_3^2} > \lambda_2 = c_1 c_2 > \lambda_4 = 0$ or, in words, 2SLS performs better than no adjustment, which performs better than OLS.

3.3 Trade-offs Under Violations in Instrumental Variable Assumptions

In this section, we present results for the trade-offs between confounder and IV methods for three scenarios: (i) violation of the exclusion restriction assumption, (ii) violation of the independence assumption, and (iii) treatment effect heterogeneity. In all scenarios, U is unobserved, which provides the realistic setting where we may be motivated to use 2SLS due to the concern of unobserved confounding. For ease of exposition, we first derive the quantities of interest without adjustment for observed confounding. We then present the quantities with these observed covariates. For ease of comparison of the quantities of interest, we further assume all regression slope parameters are positive throughout this section and we will handle the general case in the sensitivity analysis portion. Unless otherwise stated, all proofs for the propositions in this section can be found in the Appendix.

3.3.1 Exclusion Restriction Violation

Figure 3.2 directly reproduces Figure 2 from Pearl (2012) and presents a violation of the exclusion restriction assumption for Z if $c_{ER} \neq 0$. We use this quantity to denote the degree of violation. By traditional logic, one would define Z as a confounder because it is a cause both of X and Y and use it as such. It is not, however, unequivocally true that one should use Z as a confounder. To see this, suppose we have the following structural equations:

$$X = c_1 U + c_3 Z + \epsilon_3 \tag{3.8}$$

$$Y = c_0 X + c_2 U + c_{ER} Z + \epsilon_4. \tag{3.9}$$

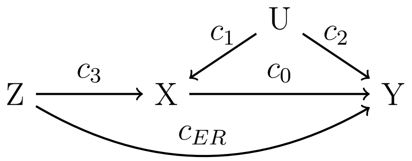


Figure 3.2: Exclusion Restriction Violation using Z as an IV.

Proposition 3.3.1. *Under the conditions of Eqs. (3.8) and (3.9),* $\hat{A}_2 \xrightarrow{p} c_0 + c_1 c_2 + c_3 c_{ER}$*,*

$\hat{A}_3 \xrightarrow{p} c_0 + \frac{c_1 c_2}{1-c_3^2}$, *and* $\hat{A}_4 \xrightarrow{p} c_0 + \frac{c_{ER}}{c_3}$.

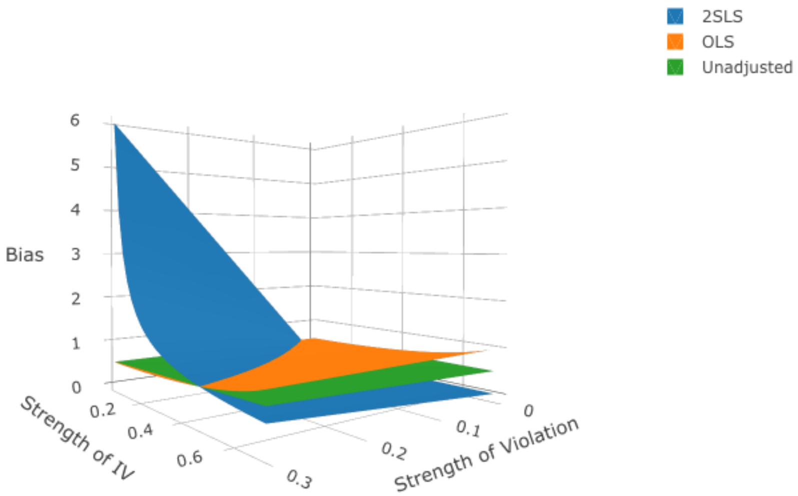
The convergence results for A_2 and A_3 can be found in Pearl (2012) so we omit them. The proof for A_4 can be found in the Appendix. We see that adjusting for Z decreases inconsistency compared to not adjusting for Z if $\frac{c_{ER}}{c_3} \geq \frac{c_1 c_2}{1-c_3^2}$. This inequality could be difficult to attain if the instrument is strong.[88] Interestingly, the left term of this inequality is the inconsistency of 2SLS and thus the IV being strong is relatively advantageous for the use of Z as an IV in 2SLS. We can re-arrange the inequality between A_3 and A_4 as $\frac{c_{ER}(1-c_3^2)}{c_3} \geq c_1 c_2$ where $c_1 c_2$ indicates the impact of unobserved confounding in the relationship between X and Y. Here, it becomes more clear that the strength of the IV can be large enough such that the degree of exclusion restriction violation (i.e. c_{ER}) is offset and is smaller than the impact of unobserved confounding.

The trade-offs between A_2, A_3, and A_4 can be visualized with a 3-D contour plot. In Figure 3.3, letting $c_0 = 0.3, c_1 = 0.7$, and $c_2 = 0.7$, we can vary the values of c_3 and c_{ER}. Note the plausible coefficient values for c_3 and c_{ER} are restricted due to the requirement that the variances for the variables to sum to one (see "Notes about Simulations" section in the Appendix). This image gives us the visual intuition that when the IV is stronger, moderate violations in the exclusion restriction violation do not preclude the use of Z as an IV. Furthermore, adjusting for Z will be inferior compared to not adjusting for Z. When the IV is weak, 2SLS predictably performs poorly in all cases.

3.3.2 Independence Violation

Figure 3.4 represents one violation of the independence assumption where the residual confounder U is a cause of Z. Here, c_I represents the degree of violation or, in this case, the effect of U on Z. Alternative specifications create a new confounder that is only associated with Z but not X. Nonetheless, our parameterization provides a useful case where both the independence assumption is violated and there is confounding in the relationship between

Figure 3.3: Analytical comparison of no adjustment, OLS adjusting for the IV, and 2SLS using the IV under varying IV strength and degree of exclusion restriction violation.



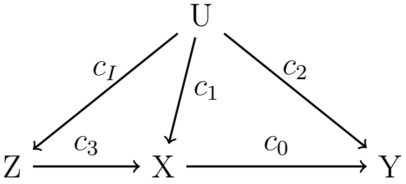


Figure 3.4: Z is Correlated with an Unobserved Confounder U

Z and X. For structural equations, we can re-use Eqs 3.1 and 3.2 as well as add an extra structural equation given by

$$Z = c_I U + \epsilon_5. \tag{3.10}$$

Proposition 3.3.2. *Under Eqs. (3.1), (3.2), and (3.10),* $\hat{A}_2 \xrightarrow{p} c_0 + c_1c_2 + c_2c_3c_I$, $\hat{A}_3 \xrightarrow{p} c_0 + \frac{c_1c_2(1-c_I^2)}{1-(c_3+c_1c_I)^2}$, *and* $\hat{A}_4 \xrightarrow{p} c_0 + \frac{c_2c_I}{c_3+c_1c_I}$.

Establishing the convergence result for A_2 is a straightforward application of Wright's path analysis so only the proofs for A_3 and A_4 are provided in the Appendix.[133] To better interpret these quantities, we can think about both paths on the DAG and remaining variance after orthogonalizing variables via the Frisch-Waugh-Lovell Theorem (FWL).[72] Looking at the pathways, c_1c_2 is the backdoor path from X to Y through U, $c_2c_3c_I$ is the backdoor path through Z, and c_2c_I is the path from Z to Y via U. $c_3 + c_1c_I$ is the unconditional correlation between Z and X, which includes both the direct and backdoor path via U. $1 - c_I^2$ depicts the remaining (stochastic) variation in Z after orthogonalizing U while $1 - (c_3 + c_1c_I)^2$ is the remaining variance in X after orthogonalizing Z.

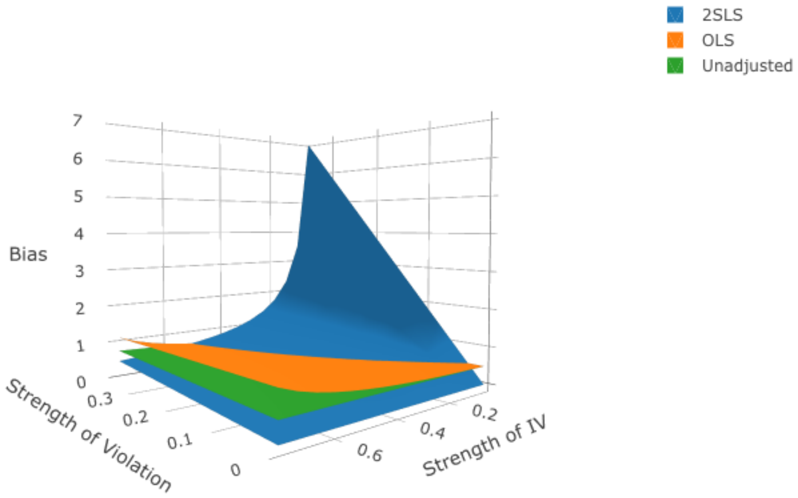
Of particular interest is the trade-off between using Z in 2SLS and using Z in OLS. We find that adjusting for Z is superior to using Z for 2SLS if $\frac{c_I}{c_3+c_1c_I} > \frac{c_1(1-c_I^2)}{1-(c_3+c_1c_I)^2}$. We can begin to interpret this inequality with the reoccurring theme that if the IV is strong then attaining this inequality is more difficult: a strong IV will cause $c_3 + c_1c_I$ to be large which inflates the inconsistency in A_3 while it decreases the consistency for A_4. The remaining variation in Z not caused by U, or c_I, is an important quantity because as c_I decreases, the inconsistency in A_3 will increase while the inconsistency in A_4 will decrease. In this sense, a simple sensitivity analysis procedure could be to benchmark the variation in the IV that is explained by the covariates. One could use this benchmark to conjecture how much of the variation in the IV is explained by an unobserved confounder. If this quantity is small then one could plausibly assume a fair amount of variation in Z free from unobserved confounding and, thus, c_I is small. These trade-offs can be visualized in Figure 3.5 where $c_0 = 0.3, c_1 = 0.7$, and $c_2 = 0.7$.

3.3.3 Treatment Effect Heterogeneity

We return the setting of Figure 3.1 where we have a perfect IV but now we omit the edge weights and introduce treatment effect heterogeneity. In this case, we do not know of any literature that gives us a salient way to to represent treatment heterogeneity using DAGs and edge weights. Therefore, the nuance is in the structural equations:

$$X = \alpha_1 Z + \alpha_2 U + \alpha_3 ZU + \epsilon_1, \tag{3.11}$$

Figure 3.5: Analytical comparison of no adjustment, OLS adjusting for the IV, and 2SLS using the IV under varying IV strength and degree of the independence violation.



$$Y = \beta_1 X + \beta_2 U + \beta_3 XU + \epsilon_2. \tag{3.12}$$

The estimand of interest remains the ACE or the average of the individual treatment effects, which is denoted by β_1. This parameterization is consistent with a violation of Wang and Tchetgen Tchetgen (2018) assumptions 5a and 5b if $\alpha_3 \neq 0$ and $\beta_3 \neq 0$. The key parameter for measuring the degree of "assumption violation" is α_3 because we aim to quantify how large an unobserved confounder needs to be in order to modify the effect on Z in the first-stage to render $\lambda_4 > \lambda_3$ and $\lambda_4 > \lambda_2$, or that IVAC is inferior to CAC.

We note that in this scenario, U is extended to be a composite variable that includes unobserved confounders but, additionally, unobserved effect modifiers. That is, a variable that only affects the outcome and thus contribute to the magnitude of β_2 and β_3 but not α_2 and α_3, and vice versa. This allows us to be more flexible in that we do not require all unobserved variables to be both confounders and effect modifiers but perhaps only effect modifiers.

Proposition 3.3.3. *Under the conditions of Eqs. (3.11) and (3.12) as well as assuming* $E[U^3] = 0$, $\hat{A}_2 \xrightarrow{p} \beta_1 + \alpha_2\beta_2 + 2\alpha_1\alpha_3\beta_3$, $\hat{A}_3 \xrightarrow{p} \beta_1 + \frac{\alpha_2\beta_2+\alpha_1\alpha_3\beta_3}{1-\alpha_1^2}$, *and* $\hat{A}_4 \xrightarrow{p} \beta_1 + \frac{\alpha_3\beta_3}{\alpha_1}$

When comparing the relative trade-off between not adjusting for Z and adjusting for Z, the latter is inferior if $\frac{\alpha_2\beta_2+\alpha_1\alpha_3\beta_3}{1-a_1^2} > \alpha_2\beta_2 + 2\alpha_1\alpha_3\beta_3$. Unlike the case of no effect modification, it is not always true that adjusting for an IV will amplify bias. Specifically, by rearranging terms we see the inequality holds if $2\alpha_1^2 + \frac{\alpha_2\beta_2}{\alpha_3\beta_3} > 1$. If $|\alpha_1| \gtrapprox 0.7$, or the IV is strongly associated with X, then this inequality will hold; however, this would not be the case if the IV is sufficiently weak and the multiplication of the coefficients for the interactions are larger than the multiplication of the coefficients representing the main effect.

In the more realistic case, if an analyst is aware Z is an IV and assuming that the IV was sufficiently strong, the main point of concern surrounds whether the LACE estimate is more inconsistent than the ACE estimate from the unadjusted OLS. This notion is true if $\frac{\alpha_3\beta_3}{\alpha_1} > \alpha_2\beta_2 + 2\alpha_1\alpha_3\beta_3$ or $\frac{\alpha_1\alpha_2\beta_2}{\alpha_3\beta_3} + 2\alpha_1^2 < 1$. Firstly, a strong IV, $|\alpha_1| \gtrapprox 0.7$, precludes this inequality from being attained. Alternatively, this inequality could fail to be attained if the ratio of main effects to interaction effects, scaled by the IV strength, is large. Setting $\alpha_2 = 0.15$, $\beta_1 = 0.1$, $\beta_2 = 0.2$, and $\beta_3 = 0.1$, we can visualize the above inequalities in Figure 3.6.

A final takeaway from these results is that the strength of an IV has implications on the the OLS inconsistencies even though it is independent from U (e.g. via the term $2\alpha_1\alpha_3\beta_3$ in $A2$). Therefore, accounting for any present IVs, even if never utilized, is important for understanding the degree of inconsistency present in confounder methodology. The results of the previous three Propositions are summarized in Table 3.1.

Table 3.1: Summary of Results

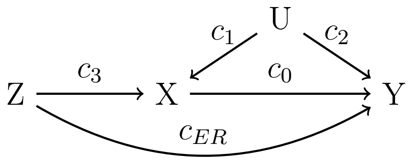
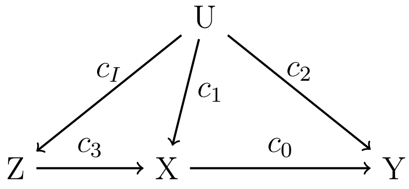
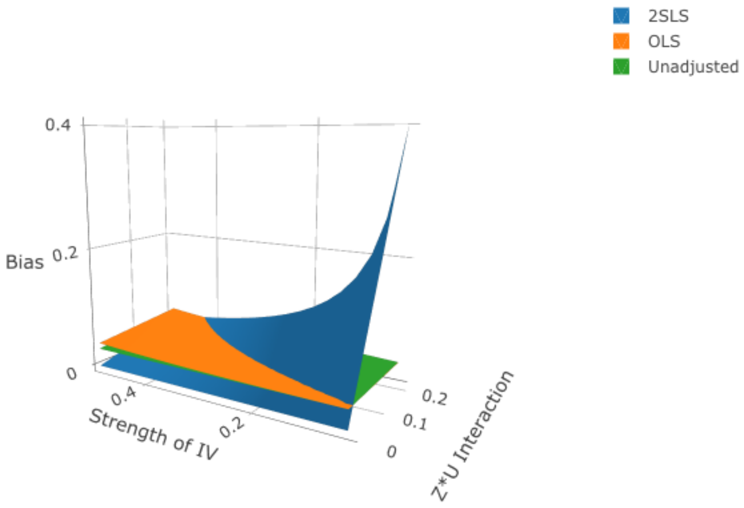
Scenario	A_2	A_3	A_4
Exclusion Restriction Violation  U c_1 c_2 Z c_3 X c_0 Y c_{ER} $X = c_1U + c_3Z + \epsilon_3$ $Y = c_0X + c_2U + c_{ER}Z + \epsilon_4.$	$c_0 + c_1c_2 + c_3c_{ER}$	$c_0 + \frac{c_1c_2}{1-c_3^2}$	$c_0 + \frac{c_{ER}}{c_3}$
Independence Assumption  U c_I c_1 c_2 Z c_3 X c_0 Y $X = c1_1U + c_3Z + \epsilon_1$ $Y = c_0X + c_2U + \epsilon_2$ $Z = c_IU + \epsilon_5.$	$c_0 + c_1c_2 + c_2c_3c_I$	$c_0 + \frac{c_1c_2(1-c_I^2)}{1-(c_3+c_1c_I)^2}$	$c_0 + \frac{c_2c_I}{c_3+c_1c_I}$
Treatment Effect Heterogeneity $X = \alpha_1Z + \alpha_2U + \alpha_3ZU + \epsilon_1,$ $Y = \beta_1X + \beta_2U + \beta_3XU + \epsilon_2.$	$\beta_1 + \alpha_2\beta_2 + 2\alpha_1\alpha_3\beta_3$	$\beta_1 + \frac{\alpha_2\beta_2+\alpha_1\alpha_3\beta_3}{1-a_1^2}$	$\beta_1 + \frac{\alpha_3\beta_3}{\alpha_1}$

Figure 3.6: Analytical comparison of no adjustment, OLS adjusting for the IV, and 2SLS using the IV under varying IV strength and degree of the treatment effect heterogeneity assumption.



3.3.4 Adding Additional Confounders

In the vast majority of data analyses, an analyst will have access to observed confounders that will be adjusted for to mitigate confounding both in OLS and 2SLS. Temporarily, we consider a single observed confounder, W, that is continuous and has mean 0 and variance 1. If we have multiple confounders, $\{W_1, W_2, ...W_J\}$, that are in the same form as W, we can simply redefine W as some function of the confounders, $f(W_1, W_2, ...W_J)$. For example, this might take a form linear combination derived from the first principal component of the W matrix. Therefore, the edge weights and regression coefficients will be the joint effect of the confounders and our updated results can accommodate multiple covariates. Depending on the which equation W is in, it can additionally represent the propensity score (first stage) or prognostic score (second stage) that can be estimated using ML models under a partial linear model that relaxes assumptions the functional form of covariates.[98, 28] In this work, for simplicity, we will not discuss such ways of modeling W.

Because we would like to use W to benchmark relationships involving U, we must update the DAGs and structural equations such that the W has a similar relationship to U in each assumption violation. As a consequence, our previous results will change. In all cases, W is orthogonal to U or, in other words, U represents confounding in ACE unrelated to U. With the exception of A_1, the quantities of interest are updated to

$$A_2 = \frac{\partial}{\partial x} E[Y|x, w] \tag{3.13}$$

$$A_3 = \frac{\partial}{\partial x} E[Y|x, z, w] \tag{3.14}$$

$$A_4 = \frac{Cov(Y, Z|W)}{Cov(X, Z|W)} = \frac{\frac{\partial}{\partial z} E[Y|z, w]}{\frac{\partial}{\partial z} E[X|z, w]}. \tag{3.15}$$

Exclusion Restriction

Figure 3.7 serves as an example of a covariate that has no direct impact on Z or U. Nevertheless, if we were to condition upon X and nothing else, U would no longer be independent of W or Z due to the collider effect. A collider effect induces an association between two variables that point to (i.e. cause) a single variable that has been conditioned.[86] However, conditioning upon W breaks this association.

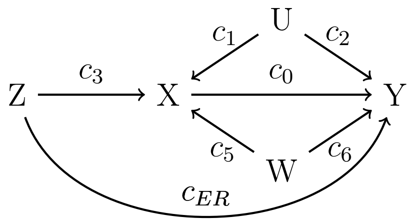


Figure 3.7: A DAG with observed confounder and violation of exclusion restriction (W).

The updated structural equations are

$$X = c_1 U + c_3 Z + c_5 W + \epsilon_3 \tag{3.16}$$

$$Y = c_0 X + c_2 U + c_{ER} Z + c_6 W + \epsilon_4. \tag{3.17}$$

Proposition 3.3.4. *Under Eqs. (3.16) and (3.17),* $\hat{A}_2 \xrightarrow{p} c_0 + \frac{c_1 c_2 + c_3 c_{ER}}{1 - c_5^2}$ *and* $\hat{A}_3 \xrightarrow{p} c_0 + \frac{c_1 c_2}{1 - c_3^2 - c_5^2}$

The proof is very similar to that of Proposition 3.3.1 so it is omitted. As expected, adjusting for W leads to decreased variance in X, which leads to a higher proportional contribution of U, at the benefit of eliminating the backdoor path via $c_5 c_6$. For 2SLS, the results for A_4 are not affected because the assumption violations vis-a-vis using Z as an IV are not influenced by W. Nevertheless, in practical settings, adjusting for W will usually increase precision of $\hat{A}_4$.

Independence

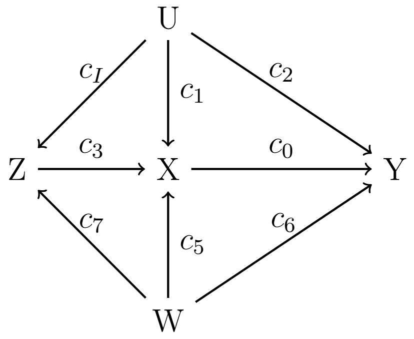


Figure 3.8: W mimics U in the DAG.

Besides introducing W as a confounder in the $X - Y$ relationship, Figure 3.8 extends W to be a confounder in the $Z - Y$ and $Z - X$ relationships. Therefore, the independence assumption of Z being met is contingent on conditioning upon both W and Z. Thus, the structural equations are updated to

$$Z = c_I U + c_7 W + \epsilon_5 \tag{3.18}$$

$$X = c_1 U + c_3 Z + c_5 W + \epsilon_1, \tag{3.19}$$

$$Y = c_0 X + c_2 U + c_6 W + \epsilon_2. \tag{3.20}$$

The results for all quantities, A_2, A_3, and A_4, can now be updated per the following proposition with the proof detailed in the Appendix:

Proposition 3.3.5. *Under Eqs. (3.19), (3.20), and (3.18):*

$$\hat{A}_2 \xrightarrow{p} c_0 + \frac{c_1 c_2 + c_2 c_3 c_I}{1 - (c_5 + c_3 c_7)^2}, \tag{3.21}$$

$$\hat{A}_3 \xrightarrow{p} c_0 + \frac{\frac{c_1 c_2 (1 - c_7^2 - c_I^2)}{1 - c_7^2}}{1 - (c_5 + c_3 c_7)^2 - (1 - c_7^2)(c_3 + \frac{c_1 c_I}{1 - c_7^2})^2}, \tag{3.22}$$

$$\hat{A}_4 \xrightarrow{p} c_0 + \frac{c_2 c_I}{(1 - c_7^2)(c_3 + \frac{c_1 c_I}{(1 - c_7^2)})}. \tag{3.23}$$

These results bear some resemblance to Proposition 3.3.2 when we did not have W present. For A_2, in the numerator, because W does not mitigate the influence of U in the DAG, we still have two backdoor paths from X to Y that go through U. Meanwhile, in the denominator the variance of X is reduced via controlling for W, which has a direct path to X as well as an indirect path via Z.

For A_3, the quantity is similar conceptually to our findings in Proposition 3.3.2 except that we must additionally account for controlling for W. In the numerator, $c_1 c_2(1 - c_7^2 - c_I^2)$ represents the magnitude of unobserved confounding reduced (i.e. multiplied) by the exogenous variance of Z due to there no longer being a backdoor path to Y via Z. We must account for the cost of adjusting W therefore this quantity is amplified (i.e. divided) by $1 - c_7^2$, the variance in Z free of W. In the denominator, we have the remaining variance of X after adjusting for W and Z. The first subtracting term represents the unconditional R^2 between X and Z while the second takes the variation of Z free of W and multiplies it by the partial R^2 between Z and X, adjusting for W. Because we haven't adjusted for U, the backdoor path via U remains and, furthermore, because W does not mitigate this, W essentially acts like a bias amplifier for $c_1 c_I$. Lastly, A_4 represents the association between Z and X via U as well as the variation exogenous from W directly from Z.

Treatment Effect Heterogeneity

Because we require our observed confounder to be of the same form of U for benchmarking purposes, we will set W as both an effect modifier of the treatment on the outcome and of the instrument on the treatment assignment with the following structural equations:

$$X = \alpha_1 Z + \alpha_2 U + \alpha_3 ZU + \alpha_4 W + \alpha_5 ZW + \epsilon_1, \tag{3.24}$$

$$Y = \beta_1 X + \beta_2 U + \beta_3 XU + \beta_4 W + \beta_5 XW + \epsilon_2. \tag{3.25}$$

The presence of XW, which is observed but still endogenous, means that in OLS, we must adjust for it and in 2SLS, must provide an an additional IV. In particular, we will choose

ZW. Therefore we modify the quantities of interest to

$$A_2 = \frac{\partial}{\partial x} E[Y|x, w, xw] \tag{3.26}$$

$$A_3 = \frac{\partial}{\partial x} E[Y|x, z, w, xw] \tag{3.27}$$

$$A_4 = \frac{\partial}{\partial \hat{x}} E[Y|\hat{x}, \widehat{xw}, w] \tag{3.28}$$

where $\hat{x}$ and $\widehat{xw}$ represent the fitted values from using Z, ZW, and W as regressors for X and XW, respectively. Note that now the main effect is obtained after orthogonalizing XW and W, or $\widehat{XW}$ and W in 2SLS, which we can find via FWL by treating XW as any other covariate. Therefore, the corresponding interpretation of the main effect is when $W = 0$, or we are at the average value of the covariate due to centering the variable.

Proposition 3.3.6. *Under Eqs. (3.24) and (3.25) as well as* $E[U^3] = 0$, $E[W^3] = 0$, $E[W^4] = 3$, *and* $E[U^4] = 3$ *we have* $\hat{A}_2 \xrightarrow{p} \beta_1 + \frac{\alpha_2\beta_2 + 2\alpha_1\alpha_3\beta_3}{1-\alpha_4^2 - \frac{4\alpha_1^2\alpha_5^2}{1+\alpha_4^2+2\alpha_5^2}}$ *and* $\hat{A}_4 \xrightarrow{p} \beta_1 + \frac{\alpha_1\alpha_3\beta_3 - \frac{2\alpha_1\alpha_3\alpha_5^2\beta_3}{\alpha_1^2+\alpha_5^2}}{\alpha_1^2+\alpha_5^2 - \frac{4\alpha_1^2\alpha_5^2}{(\alpha_1^2+\alpha_5^2})}$.

The proof for Proposition 3.3.6 is shown in the appendix. The interpretation of A_2 is consistent with Proposition 3.3.3 with the denominator reflecting the fact that we are adjusting for W and XW, which reduces the remaining variance of X. For A_4, because we are no longer computing the ratio of coefficients the interpretation is not directly comparable to Proposition 3.3.3. Nevertheless, we can see the influence of the $Z - U$ interaction on the inconsistency and observe that because XU is correlated with XW, adjusting for XW will reduce the influence of XU, hence the subtraction terms.

One may notice that we do not consider the convergence of A_3 and this is because we only wish to compare A_2 and A_4. The reason for this stems from our discussion of Proposition 3.3.3 where, assuming $sign(\alpha_2\beta_2) = sign(\alpha_3\beta_3)$ for sake of simplicity, bias amplification will hold in the effect modification case if $2\alpha_1^2 + \frac{\alpha_2\beta_2}{\alpha_3\beta_3} > 1$. When we have a strong IV, or $|\alpha_1| \gtrapprox 0.7$, then this inequality holds. If we instead lower the strength of the IV to where we must consider $\frac{\alpha_2\beta_2}{\alpha_3\beta_3} > 1 - 2\alpha_1^2$, we would require the multiplication interaction effects to be at least as large as the main effects with this requirement increasing as the IV strength increases. We argue that in this circumstance one should avoid using Z altogether because concern over this ratio would imply that the IV is weak and the LACE is likely to be far from the ACE due to large heterogeneity. Thus, a comparison between OLS adjusting for Z and 2SLS is not warranted.

3.4 Sensitivity Analysis

In this section, we use the closed form derivations of the previous section to develop a set of sensitivity analysis procedures that provide analysts with information surrounding whether OLS or 2SLS may be more appropriate given the observed data and hypothesized assumption violations. We focus on comparing A_3 and A_4, or using Z as a confounder versus as an IV, by graphically presenting the relative inconsistency $\phi = \frac{\lambda_4}{\lambda_3}$ as measured by the degree of IV assumption violations and unobserved confounding.

The graphical depiction of the relative inconsistencies is largely motivated by Cincelli and Hazlett (2020) [31] and Cincelli and Hazlett (2022)[32] where they use a set of partial R^2's to characterize how large the assumption violations in confounder and 2SLS analyses must be to render statistically significant results null. Instead of focusing on hypothesis testing, however, we examine how large unobserved confounding and IV assumption violations could be in order for move us away from ambivalence over the choice of methodology (i.e. $\phi = 1$) either towards 2SLS ($\phi > 1$) or OLS ($\phi < 1$). Quantifying this ϕ one IV assumption at a time, we build a detailed picture of the strengths and limitations relating to an analysis. Following this, like Cincelli and Hazlett, we use benchmarking to estimate the unobserved quantities contained within the closed-form derivations across a variety of scenarios. In the remainder of this section, we will focus on the sensitivity analysis procedure in the case of the exclusion restriction. Then, using the same principles, present the results for the independence and heterogeneity assumption violations more briefly.

3.4.1 Exclusion Restriction Violation

For the exclusion restriction, we have that

$$\phi = \left|\frac{c_{ER}}{c_3}\right|\left|\frac{c_1c_2}{1-c_3^2-c_5^2}\right|^{-1}. \tag{3.29}$$

Both c_3 and $1-c_3^2-c_5^2$ are observed quantities and can be directly estimated via OLS where $c_3 = \frac{\partial}{\partial z}E[X|z,w]$ and $1-c_3^2-c_5^2 = 1-R^2_{X\sim W+Z}$, or the residual variation in the model. c_{ER} measures the degree of the exclusion restriction violation and, crucially, we cannot directly identify it with the reduced model $\beta^R_{ER} = \frac{\partial}{\partial z}E[Y|x,w,z]$ because conditioning on X induces a collider effect between Z and U thus $\beta^R_{ER} = c_{ER} - \frac{c_1c_2c_3}{1-c_3^2-c_5^2}$ (see appendix). Therefore, we can substitute for c_{ER} resulting in $\phi = \left|\frac{\beta^R_{ER}+\frac{c_1c_2c_3}{1-c_3^2-c_5^2}}{c_3}\right|\left|\frac{c_1c_2}{1-c_3^2-c_5^2}\right|^{-1}$. The quantification of c_{ER} requires knowing $sign(c_1c_2)$, which is unobserved. As a result, we will calculate ϕ separately for the case when $sign(c_1c_2)=1$ and when when $sign(c_1c_2)=-1$

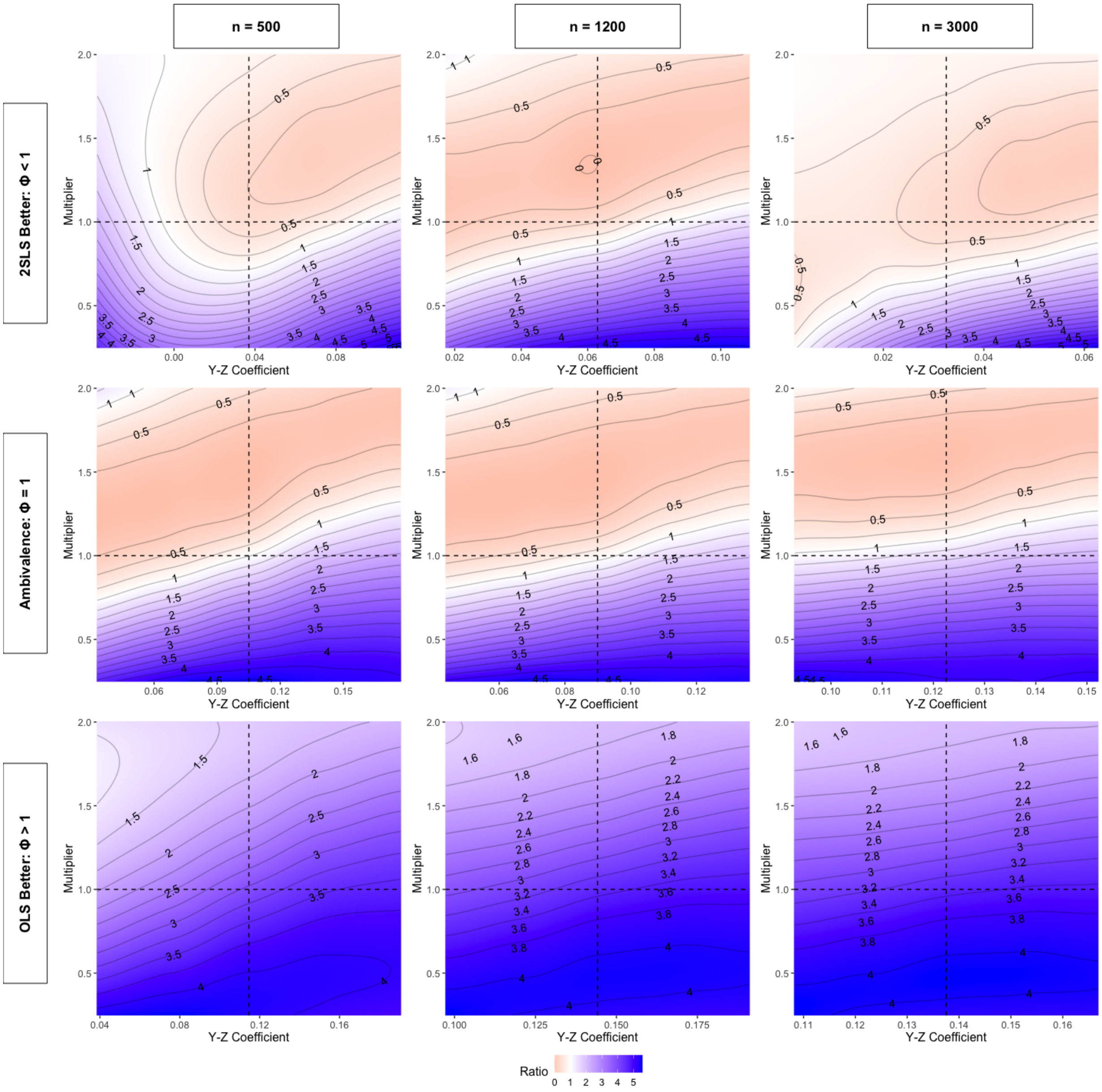
We now turn to c_1 and c_2, which together quantify the degree of unmeasured confounding. In order to reasonably benchmark these relationships, we must make the following assumption: for the set of observed confounders with cardinality J that make up the composite confounder W in our notation, the magnitude of the $W_j - X$ relationship and $W_j - Y$ relationship for $j \in \{1,2,...J\}$ are similar to the magnitude of the $U-X$ and $U-Y$ relationship, respectively. To be conservative, we take the largest such magnitude where we the benchmark c_1 via $c_1^B = \max_j \frac{\partial}{\partial w_j}E[X|z,w_1,w_2,...,w_J]$ and similar for c_2. Of course, the degree of unobserved confounding could be different than the observed benchmark. To mitigate this, we may additionally add a multiplier, for example $M \times c_1^B$, if we believe the observed confounders underestimate ($M<1$) or overestimate ($M>1$) the degree of unobserved confounding. For instance, if $M=0.5$, we are assuming that the unobserved confounding is half as much as the benchmarked confounding. Because the data is normalized, all quantities are on the scale of $[-1, 1]$ and can be effectively interpreted as partial correlations.

With the ability to calculate an estimate on ϕ, we can construct a graphical depiction of the relative trade-offs across a matrix of cases. That is, we can can calculate ϕ across an array of multipliers to account for several benchmark scenarios and, furthermore, across, for instance, the 90% confidence interval for β^R_{ER} to account for sampling variability in the estimation of the exclusion restriction violation. The final result is a contour plot colored by ϕ such as the one presented in Figure 3.9, which depicts the contour plots across a grid of example simulated scenarios where $\phi < 1$ (2SLS better), $\phi = 1$ (ambivalence), and $\phi > 1$ (OLS better) with sample sizes $n = 500, 1200, \text{and } 3000$ to show the impact of a widening confidence interval bounds on the decision-making.

Focusing on the plot in the upper-right, the dotted lines represent for the x-axis and y-axis the point estimate for β^R_{ER} and the multiplier of 1, respectively. The value of ϕ at the intersection of these dotted line indicates our decision if we take our data at face-value. The value of ϕ is about 0.5, indicating the inconsistency for the ACE of 2SLS due to an exclusion restriction violation is about half the inconsistency in OLS due to unmeasured confounding. This indicates that 2SLS may be the more appropriate method compared to OLS, which matches with our simulation. At a glance, referring to the legend, where the plot remains red (as opposed to white or blue) we maintain this conclusion. Nevertheless, we can see that as we move towards the upperbound of the confidence interval we need an increasingly larger multiplier on unmeasured confounding to cross into making the opposite conclusion due to $\phi > 1$. In the context of the current data analysis at hand, an analyst should consider the likelihood of both β^R_{ER}, which represents a partial correlation, nearly doubling and that the degree of unmeasured confounding is less than the available covariates to benchmark as it pertains to a certain multiplier.

As we move to the second row, ambivalence to methodology, we can see the intersection of the dotted lines lands on the white strip with translates to $\phi = 1$. Furthermore, roughly half

Figure 3.9: Contour Plots Across an Array of Exclusion Restriction Violation Scenarios



the scenarios are colored red while the other half is blue further indicating graphically that there is no clear choice of 2SLS or OLS. In this case, our sensitivity analysis for the exclusion restriction assumption is inconclusive and that other assumptions and factors regarding the analysis should be examined. In the third row, it is clear that across virtually all the scenarios observed, the exclusion restriction violation is large enough such that the 2SLS has a higher inconsistency for the ACE compare to the OLS, sometimes at a ratio of nearly 4 times.

3.4.2 Independence Violation

For the independence assumption, our ratio is now

$$\phi = \left|\frac{c_2 c_I}{c_3 + \frac{c_1 c_I}{1-c_7^2}}\right| \left|\frac{\frac{c_1 c_2 (1-c_7^2-c_I^2)}{1-c_7^2}}{1-(c_5+c_3 c_7)^2-(1-c_7^2)(c_3+\frac{c_1 c_I}{1-c_7^2})^2}\right|^{-1}.$$

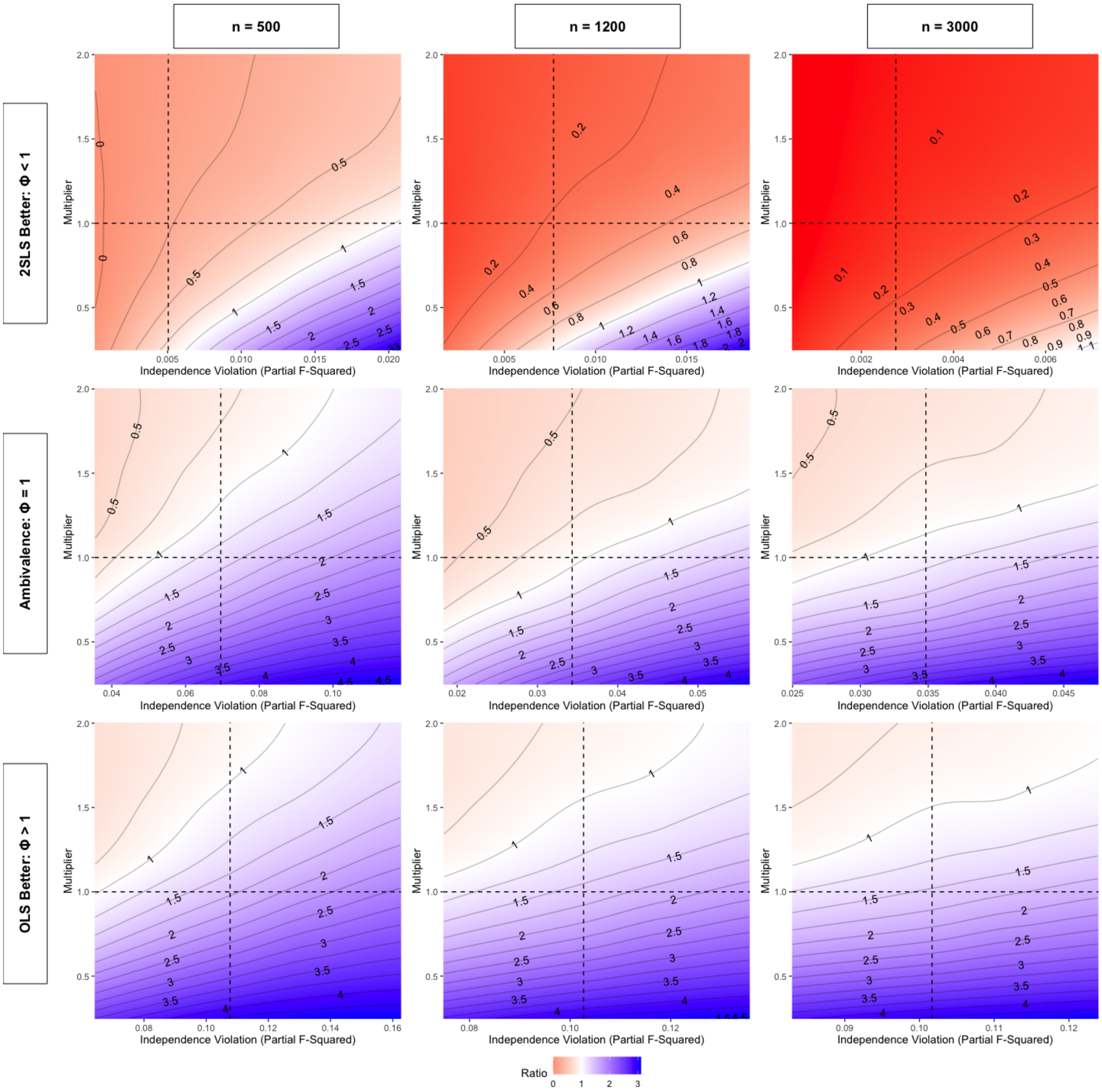
Involving partial R^2s for this assumption is more tractable to calculate and interpretable in this setting as opposed to the coefficients themselves. Furthermore, the key edge weight c_I plays a key role in both the inconsistency of 2SLS and OLS. Combining these both leads to a quantification of the degree of violation of the independence assumption using the partial Cohen's f^2 where $f^2_{U\sim Z|W} = \frac{R^2_{U\sim Z|W}}{1-R^2_{U\sim Z|W}}$. The partial Cohen's f^2 is a common measure for the general effect size of a relationship.[31] The notation in the super-script indicates we are interested in the relationship between U and Z conditioning on W. The inconsistency ratio can be translated to

$$\phi = f^2_{U\sim Z|W} \left(\frac{\sqrt{R^2_{X\sim U|Z,W}}\, sd(Z^{\perp W}) sd(X^{\perp Z,W}) R^2_{X\sim Z|W} \frac{Var(X^{\perp W})}{Var(Z^{\perp W})}}{1-R^2_{X\sim W}-(1-R^2_{X\sim W})(R^2_{X\sim Z|W}\frac{Var(X^{\perp W})}{Var(Z^{\perp W})})} \right)^{-2}$$

where, for example, $sd(Z^{\perp W})$ represents the standard deviation of the residuals produced from regressing Z on W (see appendix for the construction of this result).

There are only two quantities we cannot estimate directly: $R^2_{U\sim Z|W}$, the degree of violation, and $R^2_{X\sim U|Z,W}$, the degree of unmeasured confounding that influences X. We will benchmark the former with $R^2_{Z\sim W_j|\mathcal{W}_{-j}}$ and the latter with $\max_j R^2_{X\sim W_j|\mathcal{W}_{-j},Z}$. Similar to the exclusion restriction, we will compute a 90% confidence interval of the partial f^2 quantity (constructed via the bootstrap) and place a multiplier on the benchmarked unmeasured confounding. Using the same method to form the graphical representations, the contour plots across an array of scenarios are presented in Figure 3.10. The graphs have the same interpretation only that the x-axis directly measures the degree of violation. For example, the top left plot, at face value, the dotted lines intersect at a ϕ of 0.25 meaning that the inconsistency in 2SLS due to a potential violation is four times smaller than that of OLS due to unmeasured confounding. Thus, we favor 2SLS in this case. Where the white ambivalence line intersects the horizontal dotted line tells us that if we trust our benchmark at a multiplier of one, then the effect size will need to increase to about 0.02 to be ambivalent. The feasibility of this can be judged using subject matter expertise in a given analysis.

Figure 3.10: Contour Plots Across an Array of Independence Violation Scenarios



3.4.3 Treatment Effect Heterogeneity

As previously discussed, we will be comparing the inconsistency of OLS that does not adjust for Z (i.e. λ_2) to the estimate produced under 2SLS. We have that

$$\phi = \left| \frac{\alpha_1\alpha_3\beta_3 - \frac{2\alpha_1\alpha_3\alpha_5^2\beta_3}{\alpha_1^2+\alpha_5^2}}{\alpha_1^2 + \alpha_5^2 - \frac{4\alpha_1^2\alpha_5^2}{(\alpha_1^2+\alpha_5^2})} \right| \left| \frac{\alpha_2\beta_2 + 2\alpha_1\alpha_3\beta_3}{1 - \alpha_4^2 - \frac{4\alpha_1^2\alpha_5^2}{1+\alpha_4^2+2\alpha_5^2}} \right|^{-1}.$$

The key quantity to measure the heterogeneity violation is α_3, or the coefficient representing the $Z - U$ interaction. After re-arranging terms (see appendix), we can re-express the inconsistency ratio as

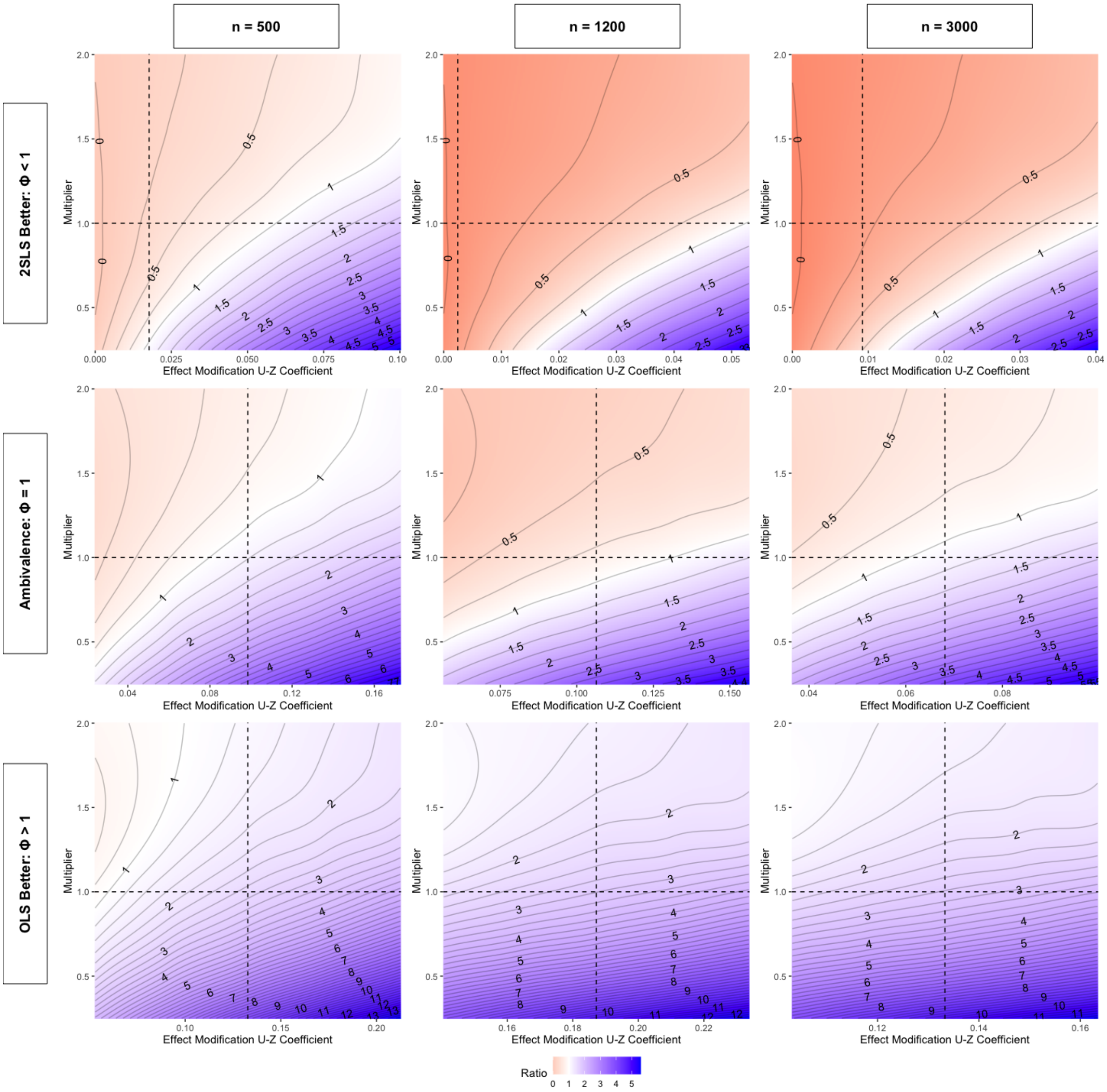
$$\phi = |\alpha_3| \left(\left| \frac{(\alpha_1^2 - \alpha_5^2)^2}{(\frac{1}{2}\alpha_1^2 - \frac{3}{2}\alpha_5^2)Var(X^{\perp W,XW}) - (\alpha_1^2 - \alpha_5^2)^2} \right| \left| \frac{\beta_2\alpha_2}{2\alpha_1\beta_3} \right| \right)^{-1}$$

We can benchmark α_3 by via the regression coefficient that the corresponds to the maximum absolute value of the coefficient between the observed covariates and Z in the model $E[X|z, w_1, w_2, ..., w_J, zw_1, zw_2, ..zw_J]$. Because this value comes directly from the regression model, we can easily derive 90% confidence intervals. β_3, or the $X - U$ interaction can be benchmarked in a similar fashion. Similar to c_1c_2 in the exclusion restriction and because all variables are centered, $\alpha_2\beta_2$ measures the degree of marginal unmeasured confounding, which we can benchmark in the same manner and include multipliers on these quantities. The only caveat is that we need to make sure to include the interaction terms in the benchmark model.

Figure 3.11 presents the contour plots across several scenarios where on the x-axis the absolute value α_3 is estimated at the vertical dotted line. Furthermore, the lowest and highest absolute values of the bounds of the confidence interval are plotted as the range in the plot. The bottom left plot shows that there is significant heterogeneity, obtaining a coefficient of 0.135 on the partial correlation scale, which produces a ϕ of about 2.5. This means that the heterogeneity pertaining to Z is large enough such that the LACE is 3 times more inconsistent for the

ACE than OLS despite unmeasured confounding. If we are at the lower end of the range and if the true degree of unmeasured confounding is twice that of our benchmark, then our ϕ is closer to 1 and we are more ambivalent in regards to this assumption violation causing inconsistency greater than that of OLS.

Figure 3.11: Contour Plots Across an Array of Treatment Effect Heterogeneity Scenarios



3.5 Simulation Results

In this section, we present results verifying the closed form derivations for all three assumption violation scenarios with and without covariates. Additionally, we further present results validating the accuracy of ϕ to capture the inconsistency. All variables were generated via the structural equations provided in the earlier sections with the error terms following a normal distribution with mean zero. When the variables were independent from all other variables in the system, such as U, they were standard normal. When the variable was determined by other variables in the system, such as X, it was still normal but with the variance of the error term being equal to one minus the variance of the other variables in the structural equation. This is so that the total variance of terms like $Var(X)$ will still equal one (see the "Notes about Simulations" section in the appendix). For example, in Eq 3.1, $\sigma_{\epsilon_1} = 1 - c_1^2 - c_3^2$ and, thus, $\epsilon_1 \sim N(0, 1 - c_1^2 - c_3^2)$.

To obtain the simulated numbers for OLS, we use the built-in `lm` from R function while for 2SLS, we use the `ivreg` function from the `ivreg` R package. In the exclusion restriction and independence setting, we present the Monte Carlo averages over 500 simulations of 500 observations generated. For treatment effect heterogeneity because the empirical results take more samples to converge, we used 500 simulations of 3000 observations. Note that for all simulations, for demonstration purposes, we set the IV to be sufficiently strong such that the estimates would converge on the population value within a reasonable sample size. Nevertheless, our results otherwise hold for weak IVs if the number of observations in each simulation increased significantly.

To demonstrate the sensitivity analysis plots, we examine an array of scenarios that scale the degree of violation from zero to a value where the inconsistency of 2SLS is far past that of OLS. At each degree of violation, we generate 500 samples from the data generating mechanism of sizes $n = 500, 1200, \text{and} 3000$ and record the proportion of calculated ϕ's that are above 1 given a multiplier of 1. As the violation increases, this proportion should increase from 0 to 1. Crucially, it should also equal 0.5 when we reach ambivalence between OLS and 2SLS. To simulate a set of covariates that we may use to benchmark the unobserved quantities, we generated three independent covariates (W_1, W_2, W_3) from a standard normal distribution and constructed W via the first principal component.

3.5.1 Exclusion Restriction

For the case with no covariates, we set the following structural parameters $c_0 = 0.3$, $c_1 = 0.5$, $c_2 = 0.5$, $c_3 = 0.5$, and $c_{ER} = 0.25$. For the case with covariates, we have $c_0 = 0.25$, $c_1 = 0.4$, $c_2 = 0.4$, $c_3 = 0.7$, $c_{ER} = 0.25$, $c_5 = 0.4$, and $c_6 = 0.4$. The results are presented in Table 3.2.

Table 3.2: Theoretical and Simulated Results for the Exclusion Restriction Violation

		Method		
		OLS without Z	OLS with Z	2SLS with Z
Without covariates	Closed Form Result	0.375	0.333	0.5
	Simulated Result	0.374	0.334	0.496
With covariates	Closed Form Result	0.399	0.457	0.357
	Simulated Result	0.398	0.459	0.354

Using the values $c_1 = c_2 = 0.35$, $c_3 = 0.3$, and $c_5 = c_6 = 0.385$ (to account for W being constructed via PCA for benchmarking), Figure 3.13 demonstrates the properties of ϕ as the exclusion violation grows in magnitude. As expected, when the violation is low, the average ϕ begins low and steadily increases before leveling off near one. The vertical dotted line represents magnitude of c_{ER} where the inconsistency of OLS is equal to that of 2SLS. For all sample sizes examined, we see that at this ambivalence point the average ϕ is about 0.5, indicating good performance of our procedure. Predictably, as the sample size increases, the behavior of ϕ at the ends of the range for the degrees violation (i.e. no violation and large) is more stable as there is less variability in the calculation of ϕ.

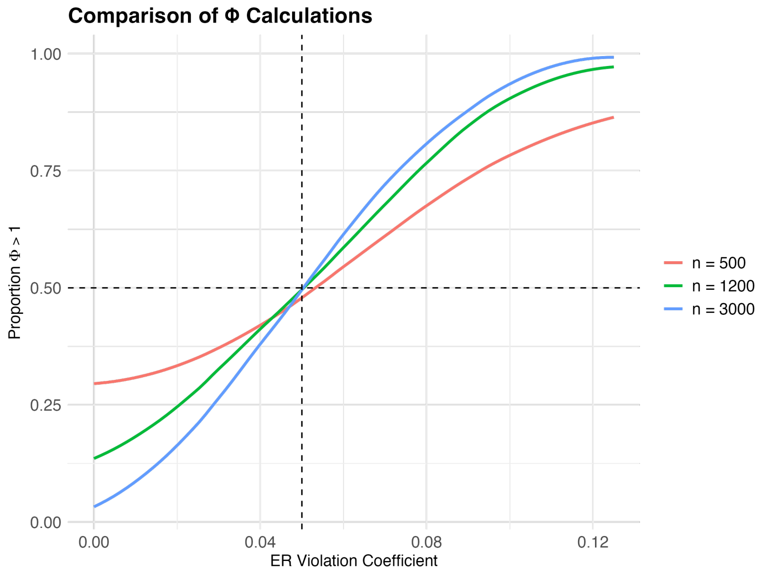


Figure 3.12: Average ϕ across 500 simulations of the given sample size over a variety of degree of exclusion restriction violations. The vertical dotted line represents the point of ambivalence.

3.5.2 Independence

For the case with no covariates, we set the following structural parameters $c_0 = 0.3$, $c_1 = 0.5$, $c_2 = 0.5$, $c_3 = 0.5$, $c_{ER} = 0.25$. For the case with covariates, we have $c_0 = 0.3$, $c_1 = 0.4$, $c_2 = 0.4$, $c_3 = 0.5$, $c_{ER} = 0.25$, $c_5 = 0.4$, $c_6 = 0.4$, and $c_7 = 0.25$. The results are presented in Table 3.3.

Table 3.3: Theoretical and Simulated Results for the Independence Violation

		Method		
		OLS without Z	OLS with Z	2SLS with Z
Without covariates	Closed Form Result	0.312	0.384	0.2
	Simulated Result	0.311	0.383	0.200
With covariates	Closed Form Result	0.290	0.391	0.176
	Simulated Result	0.290	0.394	0.178

In Figure 3.13, we present the results of increasing the independence violation with $c_1 =$

$c_2 = 0.25$, $c_3 = 0.2$, and $c_5 = c_6 = 0.35$ and c_7 scaling with the degree of violation to ensure accurate benchmarking. Based on the plotted proportions, the performance of ϕ matches our expectations.

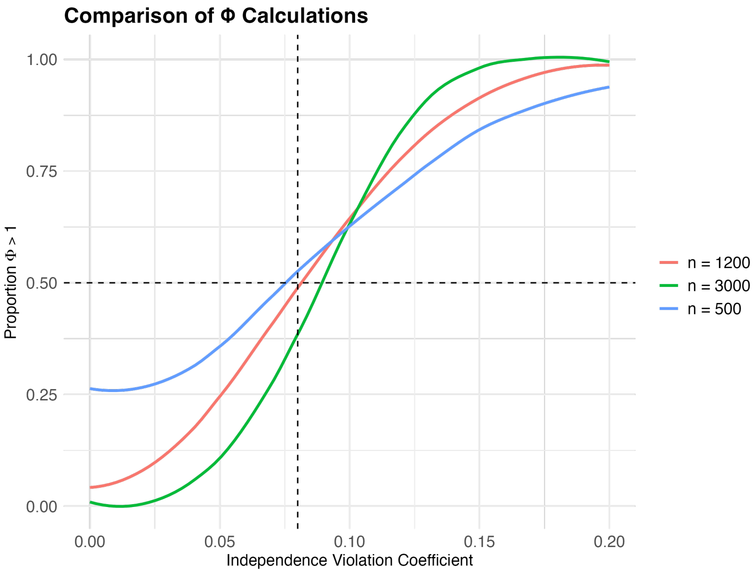


Figure 3.13: Average ϕ across 500 simulations of the given sample size over a variety of degrees of independence violations.

3.5.3 Treatment Effect Heterogeneity

For the case with no covariates, we set the following structural parameters $\beta_1 = 0.1$, $\beta_2 = 0.2$, $\beta_3 = 0.1$, $\alpha_1 = 0.45$, $\alpha_2 = 0.15$, and $\alpha_3 = 0.1$. For the case with covariates, we have $\beta_1 = 0.1$, $\beta_2 = 0.2$, $\beta_3 = 0.1$, β_4, β_5, $\alpha_1 = 0.45$, $\alpha_2 = 0.15$, $\alpha_3 = 0.1$, $\alpha_4 = 0.15$, and $\alpha_5 = 0.1$. The results are presented in Table 3.4. Figure 3.14 shows the performance of ϕ when $\beta_2 = 0.25$, $\beta_3 = 0.15$, $\beta_4 = 0.35$, $\beta_5 = 0.15$, $\alpha_1 = 0.4$, $\alpha_2 = 0.25$, and $\alpha_4 = 0.25$.

3.6 Applied Example

The causal effect of educational achievement on earnings have been the subject of several observational studies including the widely-cited results of Card (1993), which studies a sample of $n = 3,010$ men from the National Longitudinal Survey of Young Men (NLSYM).[25]

Table 3.4: Theoretical and Simulated Results for Treatment Effect Heterogeneity

		Method		
		OLS without Z	OLS with Z	2SLS with Z
Without covariates	Closed Form Result	0.039	0.044	0.022
	Simulated Result	0.039	0.044	0.021
With covariates	Closed Form Result	0.040	0.037	0.021
	Simulated Result	0.039	0.046	0.021

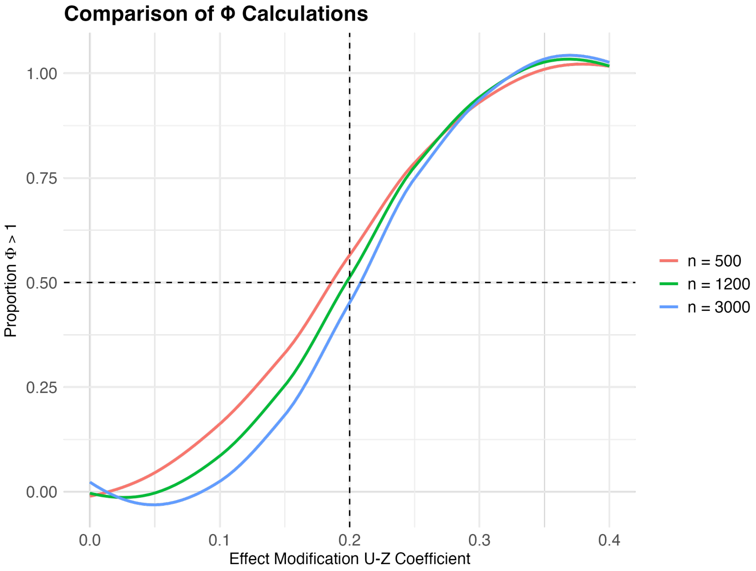


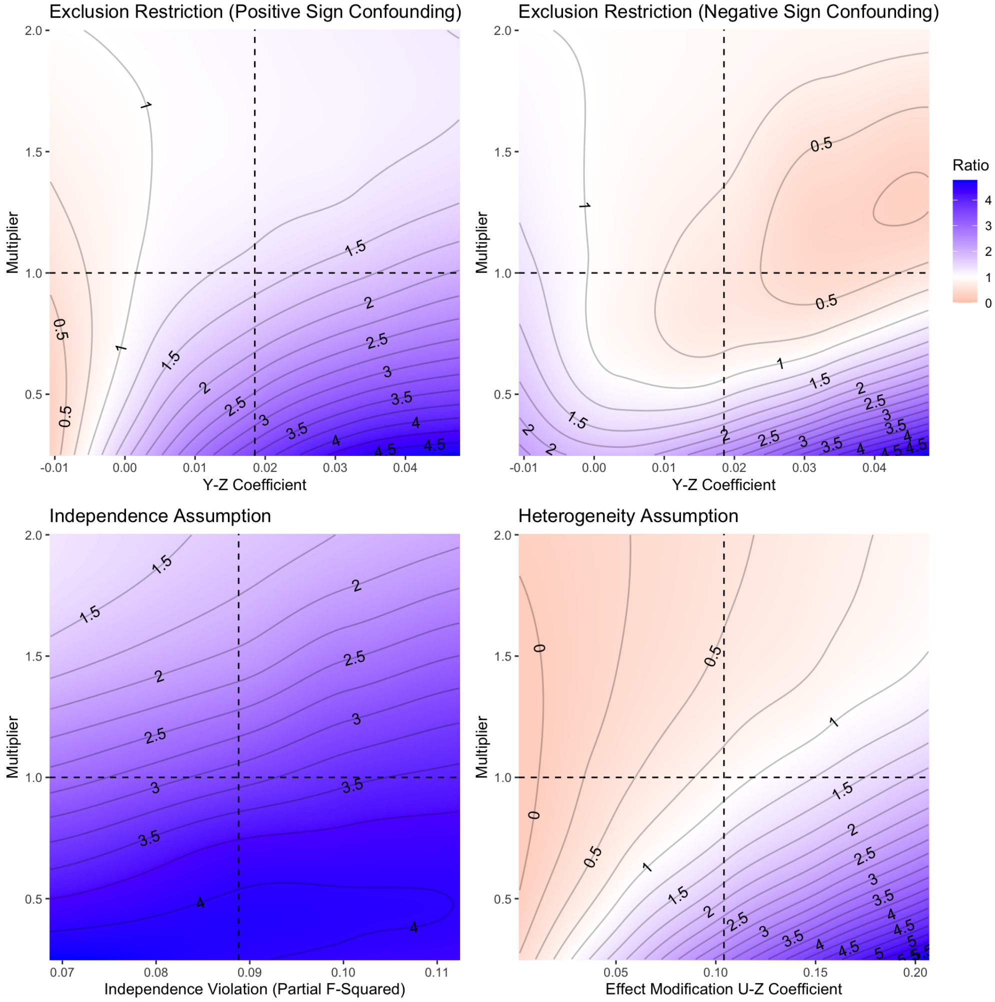
Figure 3.14: Average ϕ across 500 simulations of the given sample size over a variety of degrees of treatment effect heterogeneity interactions as measured by the effect modification via the interaction coefficient between U and Z in the first stage.

More details and access to the data can be found in `R` via `wooldridge` package under the `card` object. After log transforming earnings, Card (1993) found via OLS that there was a statistically significant 7.5% increase in earnings for each year of education. Nevertheless, there is large potential for residual confounding despite adjusting for race, experience, and regional indicators. As such, Card (1993) pursued an IV approach using an indicator variable for if the participant lived near a four-year college during their teenage years or not, which we will refer to as "Proximity." With this approach, 2SLS finds a significant 13.2% increase though with larger confidence intervals compared to OLS. Even so, as discussed in Cinelli and Hazlett (2022) there is concern that this IV may be invalid due to unmeasured geographical factors that are potentially associated with Proximity and Earnings.[32]

Though we use the same application, our sensitivity analysis addresses a different question than Cinelli and Hazlett (2022). Where as they aim to measure how large the strength of an an omitted variable that influences the IV and outcome must be to render 2SLS results null, we instead inquire whether for each assumption there exists a violation large enough to result in worse inconsistency than OLS. The findings of our procedure are presented in Figure 3.15. For sake of example, we interpret our findings based on the intersection of the dotted lines for each plot. The vertical dotted line represents the benchmarked assumption violation based on the data while the horizontal line denotes the multiplier of one, which assumes our data roughly quantifies the degree of unmeasured confounding.

With no prior knowledge on the sign of confounding, we may choose to average the results from the two exclusion restriction plots where ϕ takes the value of about 1.4 and 0.65 for the positive and negative sign respectively. This leads us to be ambivalent on whether the direct relationship between proximity and earnings is large enough such that the 2SLS estimate is more or less inconsistent than the OLS estimate. For heterogeneity, we have benchmarked a notable degree of interaction between an unmeasured confounder and proximity in the first stage, leading to a value of ϕ of approximately 0.9, slightly favoring 2SLS in this case. That is, there is likely some inconsistency in the LACE for the ACE but not to the amount where it renders the 2SLS estimate more inconsistent than the OLS estimate. Therefore, the deciding factor is in the independence assumption, which shows a rather large violation at a ϕ of nearly 3.25. This means that an unmeasured confounder with a partial f^2 of 0.09 has rendered the inconsistency of 2SLS to be 3.25 times larger than that of OLS. We thus conclude that using proximity as an instrument in 2SLS is sub-optimal compared to OLS due to a large potential violation in the independence assumption – a similar conclusion to

Figure 3.15: Examining the Three Assumptions for the Proximity IV



that of Cinelli and Hazlett (2022).

3.7 Discussion

In this work, we have investigated two predominant ways that analysts may isolate causal effects: CAC and IVAC via OLS and 2SLS, respectively. Considering these paradigms, we have based our study on the notion that each approach may work imperfectly due to assumption violations (as is most plausible in the vast majority of real world settings). Our closed-form results that capture interpretable rules of thumb based on DAG edge weights and coefficients, as well as our sensitivity analysis procedure, help guide analysts towards a practical philosophy on how one may execute observational studies. If the goal is to obtain an estimate of the ACE and there is a set of tools that one must choose from, such as the CAC and IVAC. Thus, one must ask: when is one tool more advantageous than the other? We have ultimately broken down this question by juxtaposing the degree of unobserved confounding to the degree IV assumption violations, providing results for analysts to offer evidence that the estimate computed was the least inconsistent it could have been given the scenario.

Our results suggest that properly defining confounders and IVs in the study paradigm is only a starting point. Whether a variable will be used in CAC or as an IV in IVAC is not necessarily congruent with its formal definition. In fact, we have presented analytically that, relative to OLS, there are scenarios where a variable should be used as an IV even though it does not meet the strict definition of an IV. For example, in an exclusion restriction violation the variable used as an IV is, by definition, a confounder and, although such a variable is not a "perfect IV," the relative performance of 2SLS would be better than that of OLS. Our sensitivity analysis procedure assists in detecting this by using the observed data.

Another relevant scenario lies in the fact that even though we may have a valid IV, the LACE estimand from 2SLS may be further away from the ACE than an estimand from OLS that is impacted by unobserved confounding. Through our closed form results and sensitivity analysis, we allow an analyst to judge how large heterogeneity would need to be in order for this to occur. Whether the resulting IV estimand remains scientifically useful is a subject of debate that we will not discuss here.[60] Rather, we are interested in directly targeting the ACE and will assume that treatment effect heterogeneity may provide a barrier to this goal. In this sense, the results from this work may provide case-by-case evidence for and against those who may argue IV analyses may still be useful for the ACE.

In our sensitivity analysis tool, we provide separate evaluations for the three distinct IV assumption violation scenarios. If all evaluations agree that either 2SLS is superior or OLS is superior then the suggested approach is clear to the analyst. However, one may encounter a scenario that the graphs disagree: for example, the exclusion restriction and independence show 2SLS is superior while the treatment effect heterogeneity graph does not. In this case, one should rank the importance of each assumption and consider the observed benchmarks for each assumption violation. Using the multipliers provided or inputting a custom multiplier, one could consider the degree of unobserved confounding across the board in order for all three graphs to agree and evaluate whether these findings are reasonable in the particular study scenario. Similarly, one may further use the implied assumption violations in the legend relative to the benchmarked assumption violations for these purposes. Ultimately, sensitivity analyses are matters of judgement and we aim to provide quantitative tools for analysts to navigate these scenarios.

Though we focus on inconsistency in this work, in practice, precision is important to consider as well. Even so, we believe that the first, and most important step, in any analysis is to verify that the correct estimand is being targeted, which requires reasoning about untestable causal assumptions. Our methods aim to assist analysts with this step. In the second step, the analyst will then judge which estimand is most efficient. This can be approximated by using the observed data to compare the standard error of each method. For instance, it would be clear that with a weak IV the confidence intervals will be comparatively wide. Of course, under unobserved confounding, there are implications on the first and second moment and, thus, inference. Examining the relative impacts on inference is a potential avenue of future research and could possibly be based upon the results of Cinelli and Hazlett.[31, 32]

Users of CAC often opt to capture confounding by adjusting for the propensity score (PS) using the fitted values of the exposure regressed on the confounders. Specifically, in the case of a continuous treatment, one will utilize the generalized propensity score (GPS).[59, 58] Like OLS, propensity scores estimates require conditional ignorability to be consistent so it falls in the purview of our scenarios (i.e. to adjust for Z or not). There are several ways to utilize the propensity score including matching, stratification, adjustment, and weighting. If the functional form of the propensity score is correct, then both the adjustment and propensity score methods would be consistent. In the case of unmeasured confounding, the inconsistency of each method would not be notably different because the propensity score would be incorrect by the same degree. Subsequently, in terms of judging consistency, the results we present for OLS would be identical to propensity score via adjustment for the propensity score as opposed to individual covariates. Because we encapsulate all confounders in a single covariate W, we could replace W and Z with a GPS containing a linear specification and achieve the same results. Though, as a result, the effect of Z in OLS would be mediated through the GPS, which would make our closed form derivations more complicated.

The results in this work also have implications on more sophisticated non-parametric, data-driven variable selection techniques. In order to reduce Type I error, one would model the PS away from the outcome and, thus, it appears reasonable to use penalization or machine learning (ML) techniques to optimize prediction error. There are two main issues with this automated procedure: (1) strong though imperfect IVs would be selected leading to possible bias amplification and (2) omitted variable bias may occur due to shrinkage to zero of important confounders. These points have been mentioned by others[88, 17, 29] and our results further provide analysts information to act in the face of such issues. As a starting point, one may use the general intuition gleaned from our results to quantify how strong the confounder or IV should be in order to avoid adjustment altogether or, instead, use it in an IVAC framework. Subsequently, future directions may include extending the framework developed in this chapter to techniques such as augmented inverse probability of treatment weighting, post-LASSO, targeted minimum loss estimation, and double machine learning to weigh the approaches against one another.[114, 15, 28]

Isolating causal effects in observational data presents many challenges, which foremost include the effect of unobserved confounding. The CAC and IVAC offer potential avenues to mitigate this confounding. Even so, under untestable assumptions, choosing the optimal approach for the problem at hand involves much conjecture. Our closed-form findings and sensitivity analysis approach helps analysts quantitatively justify the approach that they ultimately believe produces an estimate closest to the ACE. The upshot is that, with the information we provide, results from observational studies will both be more transparent and more useful in their interpretation.

Chapter 4

Nonparametric Replication of Instrumental Variable Estimates Across Studies

4.1 Introduction

In recent years, researchers have come to recognize that science faces a replicability crisis.[84, 63] That is, many findings produced by one study are unable to be repeated in subsequent similar studies. The end result is a potential reduction in the scientific community's confidence in published results. In the context of medical research, many have highlighted instances of limited replicability. [62, 14, 78] Epidemiologists frequently undertake questions for which controlled experimental data cannot be feasibly produced. Among other factors, this leads to a chief concern of spurious results due to unmeasured confounding, further highlighting the importance of replicability across multiple cohorts to increase confidence in conclusions. As an illustrative example, consider the effect of elevated blood lipid levels on cognitive decline. Though some work has found a that elevated lipid levels are associated with Alzheimer's

Disease (AD) incidence, others have reported null results.[106, 69, 110, 79]

One may hypothesize two natural reasons that explain the conflicting results in our example: unmeasured confounding and unaccounted for effect modification across studied cohorts. While much of the work on observational study methodology has centered around the former issue, there is comparatively less work that focuses on the latter. To relax the assumption that all confounders, or suitable proxies, were measured, we may turn to an instrumental variable (IV) approach via Mendelian randomization (MR). Briefly, MR involves using one or more single nucleotide polymorphism (SNP) that meets three main conditions: (i) it influences the level of triglycerides (relevance), (ii) it does not directly cause the cognitive decline (exclusion restriction), and (iii) there is no association any unobserved confounders (independence).[10]. For example, an analysis by Proitsi et al. (2014) found a null effect between triglycerides and late onset AD via MR.[93] Yet, considering the previous literature on this topic, it is natural to ask whether such findings can be replicated in other settings. This crucially involves a nuanced discussion surrounding how the conditions for replicability and, by extension, generalizability and transportability, interacts with the causal assumptions required for internal validity.

Suppose we wish to compare causal estimates across two independent studies or populations whether to replicate, transport, or generalize (herein encompassed by the verb replicate). This requires us to properly account for differences in the distribution of effect modifiers between each study such as medical history, concomitant medications, and lifestyle factors with many of these simultaneously being confounders. Traditionally, this may be done via survey weighting methodology. [34, 94, 16]. Nevertheless, there are three assumptions for such an approach to succeed: (i) all possible effect modifiers are measured, (ii) the distributions of these effect modifiers overlap (i.e. positivity), and (iii) the sampling weights are properly modeled. In many ways, these assumptions are similar to that of propensity score (PS) analyses in which we can require only balancing a correctly specified PS comprised of variables we are able to measure and that, furthermore, sufficiently overlap between treatment groups.[100, 23] Similar to the threats to internal validity posed by untestable causal assumptions, the threats to replicability are conceptually alike.

One path forward to relax the conditions required for sample weighting methods is by narrowing the scope of the target estimand from the whole population (i.e. the ATE) to a subset. By restricting the population of interest to be more homogeneous, the variability of all possible effect modifiers decreases, including those unobserved. Furthermore, the area of overlap is more likely to be proportionally larger. Through estimation of the local average treatment effect (LATE),[61] an IV approach with a valid instrument will achieve the same effect while mitigating unobserved confounding. The aforementioned "narrowing" occurs because the IV approach identifies the causal effect of a “complier” population, defined as those whose level of exposure monotonically varies with the value of the IV. This means that unless we assume treatment effect homogeneity or that treatment effect heterogeneity is unrelated to treatment assignment, the LATE does not equal the ATE.[122, 46] Even so, the IV approach could allow us to replicate causal estimates with increased confidence. Unlike the ATE, however, there has been little research regarding how the LATE may be used for replicability.

In the current work, assuming a binary IV and treatment, we extend LATE estimators to incorporate unknown sampling weights into a weighted LATE (WLATE). The WLATE seeks to generalize the results from one study to the target distribution of another, thus allowing for the basis of replication of effect estimates across studies. While existing WLATE methods could be employed to provide analysts valid point estimates, they require one to assume the sampling weights are known for valid inference and are further restrictive to which models can be used to estimate nuisance functions such as the weights.[30] Using the theory of influence functions (IFs) and sample-splitting,[65] our proposed method both accounts for the uncertainty in estimating the sample weights using a straightforward variance estimator and allows analysts to utilize machine learning (ML) to flexibly model nuisance functions in order to guard against potential misspecification. Furthermore, our estimator is doubly-robust with respect to within-study nuisance functions.

The remainder of this work is organized as follows. First, we describe the necessary conditions and key estimand of our methodology that we call the survey weighted LATE (SWLATE). Then, we develop an IF-based approach and describe a novel across-study cross-fitting procedure to estimate the SWLATE. To our knowledge, there is no result deriving the IF and showing the regularity conditions and asymptotic properties of a weighted LATE when using ML to estimate unknown sampling weights. Following this, to ameliorate concerns over the narrow scope of the LATE, we extend our developed method to estimate bounds on the weighted ATE (WATE) based on the approach of Kennedy et al. (2020).[66] Lastly, we demonstrate our estimators for the SWLATE and WATE both in simulation and in an MR analysis of the effect of triglycerides on cognitive decline.

4.2 Notation and Set-Up

Notationally, we assume a binary treatment $D \in \{0, 1\}$ and a binary IV $Z \in \{0, 1\}$. $Y(d)$ is the potential outcome under treatment assignment d and $D(z)$ as the potential treatment assignment under IV assignment of z and therefore $Y = DY(1) + (1 - D)Y(0)$ and $D = ZD(1) + (1 - Z)D(0)$. Denote $Y(d, z)$ as the potential outcome under treatment assignment d and IV assignment z. We will assume Y is continuous throughout this work. We additionally assume the standard assumption of the stable unit treatment value assumption (SUTVA) for all potential outcomes. Formally, the LATE and ATE are define as $\beta_{ATE} = E[Y(1) - Y(0)]$ and $\beta_{LATE} = E[Y(1) - Y(0)|D(1) > D(0)]$ where $D(1) > D(0)$ refers to the "principal strata" of compliers or those whose treatment status vary with the IV.[3] The other principal strata are always-takers, $D(1) = D(0) = 1$, never takers, $D(1) = D(0) = 0$, and defiers, $D(1) < D(0)$.

A set of covariates X may be required to satisfy the IV independence assumption.[41, 2] As such, we utilize the instrument propensity score (IPS) $e(X) = P(Z = 1|X)$. The conditions to identify the LATE via the IPS are similar to the that of ATE: positivity of the IPS and for the set of confounders to achieve strong ignorability. Additionally, conditioning on X means the standard IV assumptions of relevance and monotonicity are strata-specific. Formally, we may define the assumptions to identify the LATE conditional as follows:

Assumption 1 (Positivity of IPS): $0 < e(X) < 1$ a.s. for all $x \in \mathcal{S}_X$

Assumption 2 (Independence): $Z \perp\!\!\!\perp \{Y(0,0), Y(0,1), Y(1,0), Y(1,1), D(1), D(0)\} \mid X$

Assumption 3 (Exclusion Restriction): $Y(d,0) = Y(d,1)$ for $d \in \{0,1\}$

Assumption 4 (Relevance): $P[D(1) = 1|X] > P[D(0) = 1|X]$ a.s. for all $x \in \mathcal{S}_X$

Assumption 5 (Monotonicity): $P[D(1) \geq D(0)|X] = 1$ a.s. for all $x \in \mathcal{S}_X$

where $\mathcal{S}_X$ refers to the support of X. In order to formulate the SWLATE, we include weights $w(X)$ in the LATE estimand outlined in Frolich (2007) Theorem 1, which leads us to the following functional form

$$\beta_{SWLATE} = \frac{\int \big[\mu_1(x) - \mu_0(x)\big] w(x) f_X(x) dx}{\int \big[m_1(x) - m_0(x)\big] w(x) f_X(x) dx} \tag{4.1}$$

where $\mu_z(X) = E[Y|X, Z = z]$ and $m_z(X) = E[D|X, Z = z]$. To define $w(X)$ for the purposes of replicability, we must outline some further notation and conditions.

Let $A = (X^A, D^A, Y^A, Z^A)$ represent a sample of size N_A from study A with distribution P_A and $B = (X^B, D^B, Y^B, Z^B)$ a sample of size N_B from study B with distribution P_B. We assume that all variables were collected in the same manner and that Assumptions 1-5 are met across both studies. Assuming that $P_A \neq P_B$, for example due to differences in effect modifiers between the samples, the study-specific LATEs, β^A_{LATE} and β^B_{LATE}, are not equal.

For the purposes of replicability, this is the crux of the problem and may be ameliorated via modifying one of the LATEs (we arbitrarily choose study B) via sampling weights. When this is the case, our target estimand is β^A_{LATE} and we denote the SWLATE as β^B_{SWLATE} because it is estimated over P_B. Thus Eq. 4.1 can be rewritten as

$$\beta^B_{SWLATE} = \frac{E_{P_B}[w(X)\{\mu_1(X) - \mu_0(X)\}]}{E_{P_B}[w(X)\{m_1(X) - m_0(X)\}]}. \tag{4.2}$$

Let M_A and M_B represent indicators of membership for study A and B, respectively. We construct weights based upon the probability of being sampled into study B: $\eta(X) = P(M_B = 1|X)$. Typically, these probabilities are unknown and must be estimated by constructing and fitting a model for $E[M_B|X]$ using the data from both A and B. Consequentially, we must account for the uncertainty in estimating $\eta(X)$ in inference on β^B_{SWLATE}.

We assume the following conditions for our sampling weights: the subjects have a non-zero probability of being sampled into each study, both studies' conditional LATEs are equal within each possible strata of the covariates (i.e. no hidden strata-specific treatment effect heterogeneity), and that $\eta(X)$ is properly specified. Formally, letting $\mathcal{S}_X = \mathcal{S}^A_X \cup \mathcal{S}^B_X$ we have:

- **Assumption 6 (Positivity of Sampling Weights):** $0 < \eta(X) < 1$ a.s. for all $x \in \mathcal{S}_X$
- **Assumption 7 (Equality of Strata Specific LATEs):** $\beta(x)^A_{LATE} = \beta(x)^B_{LATE}$ a.s. for all $x \in \mathcal{S}_X$
- **Assumption 8 (Proper Specification of the Weights):** $|\hat{\eta}(X) - \eta(X)| = o_p(1)$

With these conditions, we may utilize $w(X) = \frac{1-\eta(X)}{\eta(X)}$ to re-weight β^B_{LATE} to replicate β^A_{LATE}. Assumption 7 encapsulates the previous argument that we may relax assumptions regarding unmeasured effect modification and positivity for the overall population by replacing them with similar assumptions for the complier population whose distribution of covariates varies less across studies.

4.3 Nonparametric Estimation of the SWLATE with Unknown Sampling Weights

To estimate Eq. 4.2, we extend the IF and cross-fitting framework for the LATE as detailed in Kennedy (2023)[65] to add survey weights $w(X)$. This allows us to non-parametrically identify and estimate the SWLATE with the use of ML (e.g. regularization, random forests, neural networks). Additionally, our estimator admits double-robustness for estimating the LATE within each study if either $m_1(X), m_0(X)$, $\mu_1(X)$, and $\mu_0(X)$ or $e(X)$ are correctly specified, which assists replicability by relaxing the assumptions for internal validity. When survey weights are unknown, our application of IF theory and sample-splitting both avoids complex derivations of asymptotic results using many simultaneous estimating equations and allows for flexible modeling of $w(X)$, protecting against potential misspecification.

In the remainder of the text, the empirical measure $\mathbb{P}_n$ refers to a sub-sample of B created by sample-splitting. For example, suppose the data has been split into two partitions, B^n and B^N, with $n = \lceil \frac{N_B}{2} \rceil$. Letting $\zeta_z = (\mu_z\{X\}, m_z\{X\}, e\{X\})$ represent the nuisance functions associated with study B, we can estimate ζ_z using B^N. Following this, we may fit $\eta(X)$ with both B^N and A. Computing the predicted values for each nuisance function using B^n, we may adapt the estimand presented in Mao et al. (2019) Appendix Theorem 2[74] to the following plug-in estimators for the uncentered IFs of the numerator and denominator of Eq. 4.1:

$$\phi_{num}(B; \hat{w}, \hat{\zeta_z}) = \frac{\hat{w}(X)}{\mathbb{P}_n\{\hat{w}(X)\}} \left[\frac{Z}{\hat{e}(X)} \{Y - \hat{\mu}_1(X)\} \right. \tag{4.3}$$

$$\left. - \frac{1 - Z}{1 - \hat{e}(X)} \{Y - \hat{\mu}_0(X)\} + \hat{\mu}_1(X) - \hat{\mu}_0(X) \right] \tag{4.4}$$

$$\phi_{denom}(B;\hat{w},\hat{\zeta}_z) = \frac{\hat{w}(X)}{\mathbb{P}_n\{\hat{w}(X)\}}\left[\frac{Z}{\hat{e}(X)}\{D-\hat{m}_1(X)\}\right. \tag{4.5}$$

$$\left. - \frac{1-Z}{1-\hat{e}(X)}\{D-\hat{m}_0(X)\} + \hat{m}_1(X) - \hat{m}_0(X)\right] \tag{4.6}$$

with point estimate $\hat{\beta}^B_{SWLATE} = \mathbb{P}_n\{\phi_{num}(B;\hat{w},\hat{\zeta}_z)\}/\mathbb{P}_n\{\phi_{denom}(B;\hat{w},\hat{\zeta}_z)\}$. We define the following cross-fitting procedure involving data from two studies and use this to formulate the main result of this chapter, Theorem .

Definition 4.3.1 (Cross-fitting Procedure for Two Studies)**.** *To estimate $\hat{\beta}^B_{SWLATE}$ via cross-fitting we execute the following procedure:*

1. *Randomly split B into K folds equally and $N_{B,k}$ sized. Let $k \in \{1, ...K\}$ be an arbitrary fold. Define $I_{B,k}$ as the indices in k and, similarly, $I^c_{B,k} := \{1...N_B\} \setminus I_{B,k}$.*

2. *For each k:*

   (a) Combine datasets into $C = (A, (B_i)_{i \in I^c_{B,k}})$.

   (b) Use C to estimate η_{-k} where $-k$ refers to omitting the fold k.

   (c) Use $(B_i)_{i \in I^c_{B,k}}$ to estimate the study-specific nuisance functions $\zeta_{z,-k}$

   (d) With the "hold-out" sample $(B_i)_{i \in I_{B,k}}$, compute the "predictions" $\hat{w}_k(X)$ and $\hat{\zeta}_{z,k}$.

   (e) Plug in these values into the estimators in Equations 4.4 and 4.6 to obtain $\hat{\beta}^B_{k,SWLATE}$.

3. *Average across the folds to obtain the final estimate: $\hat{\beta}^B_{SWLATE} = \frac{1}{K}\sum_{k=1}^K \hat{\beta}^B_{k,SWLATE}$.*

Theorem 4.3.1 (Inference for the SWLATE via the Influence Function)**.** *The following conditions must be met:*

1. *$N_A/N_B \xrightarrow{p} c > 0$ where c is some constant.*

2. *Assumption 10 holds as well as* $|\hat{\mu}_z - \mu_z| = o_P(1)$, $|\hat{m}_z - m_z| = o_P(1)$, *and* $|\hat{e}(X) - e(X)| = o_P(1)$ *for* $z \in \{0,1\}$.

3. $\|\hat{e} - e\| \max_{z\in\{0,1\}} (\|\hat{\mu_z} - \mu_z\|, \|\hat{m}_z - m_z\|) = o_p(N_B^{-1/2})$

4. $|Y| < C_Y$ *a.s. for some constant* C_Y

5. $|\mu_z| < C_{\mu_z}$ *and* $m_z < C_{m_z}$ *a.s. where* C_{μ_z} *and* C_{m_z} *are some constants for* $z \in \{0,1\}$

6. $\hat{e}, e > C_e$ *and* $\hat{\eta}, \eta > C_\eta$ *a.s. where* C_e *and* C_η *are some constants such that* $C_e, C_\eta > 0$

7. $\frac{\hat{w}(X)}{\mathbb{P}_n\{\hat{w}(X)\}} < C_w$ *a.s where* C_w *is some constant*

Using the estimator described in Definition 4.3.1, we have that $N_B^{1/2}(\hat{\beta}^B_{SWLATE} - \beta^A_{LATE}) \xrightarrow{d} N(0, E[\Gamma^2])$ *where*

$$\Gamma = \frac{w(X)}{E_{P_B}[w(X)\{m_1(X) - m_0(X)\}]}\Bigg\{\frac{2Z-1}{e(X,Z)}\Bigg[Y - \mu_Z(X) - \beta^A_{LATE}\{D - m_Z(X)\}\Bigg]$$
$$+ \mu_1(X) - \mu_0(X) - \beta^A_{LATE}\{m_1(X) - m_0(X)\}\Bigg\}$$

with $e(X,Z) = e(X)Z + \{1 - e(X)\}(1 - Z)$.

From the results of Theorem 4.3.1, we may construct $(1-\alpha)$-level Wald confidence intervals for $\hat{\beta}^B_{SWLATE}$. Letting $q_{1-\alpha/2}$ be the $1-\alpha/2$ quantile of the standard normal distribution and $\hat{\Gamma}$ being the plug-in estimator for Γ, we have:

$$\hat{\beta}^B_{SWLATE} \pm q_{1-\alpha/2}\sqrt{\frac{\mathbb{P}_{N_B}\hat{\Gamma}^2}{N_B}}. \tag{4.7}$$

A notable property of our estimator is that by taking $E_{P_B}[\Gamma^2]$, by iterated expectation on X and Z, we obtain the efficiency bound for the WLATE as given by Choi (2023) Theorem 1.[30] That is, our estimator achieves the lowest possible variance across all possible non-parametric

WLATE estimators.[65] In contrast with Choi (2023), we need not know the weights nor the IPS to achieve this lower bound. Altogether, this means that we may efficiently, non-parametrically, and flexibly compute a weighted LATE estimator to replicate findings from IV analyses across studies with all nuisance functions unknown.

4.4 Weighted ATE Bounds with Unknown Sampling Weights

Without restrictive and untestable assumptions regarding treatment effect heterogeneity, the LATE will not identify the ATE.[48, 122] Nevertheless, we may use the LATE to bound the ATE, which was first developed for non-compliance in clinical trials.[73, 96, 12] Kennedy et al. (2020) generalized this to tighter bounds that allows conditioning on covariates tha we will utilize.[66]

Let Y be bounded and scaled such that $Y \in [0, 1]$, then for $j \in \{l, u\}$, referring to the lower and upper bounds, Kennedy et al. (2020) defines the bound as

$$\beta^j_{ATE} = E[E[V_{j,1}|X, Z = 1] - E[V_{j,0}|X, Z = 0]] \tag{4.8}$$

for $V_{u,1} = YD + 1 - D$, $V_{l,1} = YD$, $V_{u,0} = Y(1 - D)$, and $V_{l,0} = Y(1 - D) + D$. The width of the interval $(\beta^l_{ATE}, \beta^u_{ATE})$ is inversely proportional to the strength of the instrument: stronger instruments will yield smaller intervals. Using the results from the previous section, we can straightforwardly extend this approach to incorporate survey weights.

Letting $\zeta_{j,z} = (v_{j,z}, e\{X\})$, the uncentered IF $\phi_{ATE,j}(B; w, \zeta_{j,z})$ takes the following form

$$\phi_{ATE,j}(B; \hat{w}, \hat{\zeta}_{j,z}) = \frac{\hat{w}(X)}{\mathbb{P}_n\{\hat{w}(X)\}}\Bigg[\frac{Z}{\hat{e}(X)}\{V_{j,1} - \hat{v}_{j,1}(X)\} \tag{4.9}$$

$$- \frac{1 - Z}{1 - \hat{e}(X)}\{V_{j,0} - \hat{v}_{j,0}(X)\} + \hat{v}_{j,1}(X) - \hat{v}_{j,0}(X)\Bigg]. \tag{4.10}$$

Denoting our study weighted estimate for the bounds ATE of study A as $\hat{\beta}^B_{SWATE,j}$, we denote $\hat{\beta}^B_{SWATE,l} = \mathbb{P}_n\{\phi_{ATE,l}(B;\hat{w},\hat{\zeta}_{l,z})\}$ and $\hat{\beta}^B_{SWATE,u} = \mathbb{P}_n\{\phi_{ATE,u}(B;\hat{w},\hat{\zeta}_{u,z})\}$ as the lower and upper bounds, respectively. Each weighted bound can be fit with the procedure described in Definition 4.3.1. Furthermore, inference follows from conditions identical to Theorem 4.3.1 since each bound is a essentially weighted ATE, we may use the results in the proof for the SWLATE numerator. Standard $1-\alpha$ level confidence intervals can be computed as

$$\hat{\beta}^B_{SWATE,j} \pm q_{1-\alpha/2}\sqrt{\frac{\mathbb{P}_{N_B}\hat{\Delta}_j^2}{N_B}} \tag{4.11}$$

where $\hat{\Delta}_j$ is the plug in estimator for the centered influence function, which takes the form

$$\Delta_j = \frac{w(X)}{E_{P_B}[w(X)]}\left\{\frac{2Z-1}{\hat{e}(X,Z)}[V_{j,z} - v_{j,z}(X)] + v_{j,1}(X) - v_{j,0}(X)\right\} - \beta^A_{ATE,j}. \tag{4.12}$$

From these results, analysts can compute bounds on the ATE based on the LATE that are weighted towards a target study population and, if desired, compute confidence intervals on these weighted bounds.

4.5 Simulation

We consider two general scenarios: a linear and non-linear data-generating mechanism (DGM). First, we generate six continuous covariates $X = (X_1, X_2, ..., X_6)$, from a multivariate normal distribution with $N = 1500$. Each variable has mean 0 and variance 1.5 with two correlated blocks, X_1, X_2, X_3 and X_4, X_5, X_6, with a covariance of 0.3. Table 5.1 defines all equations and coefficients for both the the linear and non-linear DGMs. First, we generate sampling probabilities of being in study A with $P(M_B = 1) = 0.5$ and use these probabilities to sample 1500 observations with replacement form study B's sample. Within each dataset, we assign the IV using the IPS with $P(Z = 1) = 0.5$.

Following this, we generate compliance status based on the "compliance score," or probability of being a complier. We set $P(D(1) > D(0))$ at 0.2, 0.5, and 0.8 to represent a "Weak IV","Moderate IV", and "Strong IV", respectively. Letting $\delta(X) = P(D(1) > D(0)|X)$, we calculate $P(D(1) = D(0) = 1|X) = P(D(1) = D(0) = 0|X) = \frac{1-\delta(X)}{2}$ for each subject, or the probabilities of being an always-taker and never-taker, respectively. From these probabilities, a subject's principal strata is selected via a multinomial distribution where $S = 0$ if a never-taker, $S = 1$ if an always-taker, and $S = 2$ if a complier. For subject i, the treatment assignment is $D_i = Z_i I(S_i = 2) + S_i I(S_i \neq 2)$.

In the outcome DGM, we set the Study A ATE (i.e. β_1) at 1. Our formulation in Table 5.1 allows us to write the LATE in closed form as $\beta^A_{LATE} = \beta_1 + \sum_{i=2}^{7} \beta_i E[X_i|S = 2]$ where we obtain the ground-truth LATEs via averaging across 5000 simulations to estimate $E[X_i|S = 2]$ where we know the complier status. Notationally, we let $\beta^A_{LATE,L}$ refer to the LATE in study A under the linear DGM while $\beta^A_{LATE,NL}$ refers to the non-linear DGM. The same interpretation applies to $\beta^B_{LATE,L}$ and $\beta^B_{LATE,NL}$.

To estimate the nuisance functions, we will compare utilizing generalized linear models (GLMs) to an ensemble of flexible models via the `SuperLearner` package.[92] For the ensemble, we will we use multivariate adaptive regression splines (EARTH), generalized additive models (GAMs), GLMs, random forest (RF), and recursive partitioning and regression trees (RPART). "SL" or "No SL" denotes whether the ensemble SuperLearner was used or not. Nuisance functions were estimated with cross-fitting with four splits of the data.

4.5.1 Results for the Survey Weighted LATE

Tables 4.2,4.3, and 4.4 detail the results over 1000 with the ground truth results in the footnotes. The unweighted LATE estimation and inference is equivalent to that detailed in Kennedy (2023). The simulation results show our SWLATE estimator can successfully replicate the target study LATE across a variety of scenarios as well as estimate in closed

Table 4.1: Summary of Data-Generating Mechanisms for Simulation

Model	**DGM**	**Equation**	**Coefficient Values**
Sampl. Prob	Linear	$logit[P(M_B = 1 \| X = x)] = \alpha_0 + \sum_{i=1}^{6} \alpha_i X_i$	$\alpha_0 = 0$, $\alpha_1 = \alpha_2 = \alpha_3 = 0.4$, $\alpha_4 = \alpha_5 = \alpha_6 = -0.4$
	Non-Linear	$logit[P(M_B = 1 \| X = x)] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2^2 + \alpha_3 X_3^2 + \alpha_4 \exp(X_4) + \alpha_5 \sin(X_5) + \alpha_6 X_6$	
IPS	Linear	$logit[P(D(1) > D(0) \| X = x)] = \theta_0 + \sum_{i=1}^{6} \theta_i X_i$	$\gamma_0 = 0$, $\gamma_1 = \gamma_2 = \gamma_3 = 0.1$, $\gamma_4 = \gamma_5 = \gamma_6 = -0.1$
	Non-Linear	$logit[P(Z = 1 \| X = x)] = \gamma_0 + \gamma_1 X_1 + \gamma_2 \exp(X_2) + \gamma_3 X_3^2 + \gamma_4 X_4 + \gamma_5 \cos(X_5) + \gamma_6 X_6$	
Compl. Score	Linear	$logit[P(D(1) > D(0) \| X = x)] = \theta_0 + \sum_{i=1}^{6} \theta_i X_i$	$\theta_0 = 0$, $\theta_1 = \theta_2 = \theta_3 = 0.4$, $\theta_4 = \theta_5 = \theta_6 = -0.8$
	Non-Linear	$logit[P(D(1) > D(0) \| X = x)] = \theta_0 + \theta_1 X_1^2 + \theta_2 \exp(X_2) + \theta_3 X_3 + \theta_4 X_4^3 + \theta_5 X_5 + \theta_6 X_6$	
Outcome	Linear	$Y = \beta_1 D + \sum_{i=2}^{7} \beta_i D X_i + f(X) + \epsilon$, where $f(X) = \sum_{i=1}^{6} \beta_i X_i$	$\beta_1 = 1$, $\beta_2 = \beta_3 = \beta_4 = 0.35$, $\beta_5 = \beta_6 = \beta_7 = -0.35$, $\beta_8 = \beta_9 = \beta_{10} = 1$, $\beta_{11} = \beta_{12} = \beta_{13} = 0.5$ $\epsilon \sim N(0, 1)$
	Non-Linear	where $f(X) = \beta_8 X_1^2 + \beta_9 X_2^2 + \beta_{10} \exp(X_3) + \beta_{11} X_4 + \beta_{12} \exp(X_5) + \beta_{13} \cos(X_6)$	

form the variance from estimating the weights with our cross-fitting procedure. In the linear setting, regardless of the instrument strength both no SL and SL have minimal absolute bias for the LATE of Study A as well as the desired coverage. In the non-linear setting, the SL approach is able to adapt to non-linearity in the DGM, with almost no bias and the desired coverage. On the other hand, not utilizing SL produces both substantial bias, almost as much as not weighting, and under-coverage. For example, in the moderate IV non linear setting, the absolute relative bias is 12.05%, almost as much as the unweighted SL. Nevertheless, the coverage of the weighted no SL estimator is better than the unweighted SL at 82.6% and 45.7%, respectively, which is largely a reflection of accounting for the estimation of the weights in our model-based estimator. These latter findings are mirrored the both the weak and strong instrument setting.

Table 4.2: Simulation Results for Linear and Non-Linear DGM Part I

Weak IV

			Point Estimate	**Bias (%)**	**Monte Carlo SE**	**Model-Based SE**	**Coverage (95% CI)**
Linear	Weighted	No SL	1.577	-2.14	0.251	0.233	0.93
		SL	1.605	-0.40	0.255	0.240	0.941
	Unweighted	No SL	1.790	11.02	0.172	0.167	0.811
		SL	1.787	10.86	0.173	0.167	0.822
Non-Linear	Weighted	No SL	1.762	19.80	0.300	0.287	0.843
		SL	1.469	-0.17	0.191	0.181	0.936
	Unweighted	No SL	1.880	27.80	0.216	0.214	0.501
		SL	1.636	11.23	0.141	0.137	0.773

Ground truth: $\beta^A_{LATE,L} = 1.612$, $\beta^B_{LATE,L} = 1.786$; $\beta^A_{LATE,NL} = 1.471$, $\beta^B_{LATE,NL} = 1.636$

Overall, the simulation results show our SWLATE estimator can successfully replicate the target study LATE across a variety of scenarios. Additionally, our efficient, model-based SE

Table 4.3: Simulation Results for Linear and Non-Linear DGM Part II

Moderate IV

			Point Estimate	**Bias (%)**	**Monte Carlo SE**	**Model-Based SE**	**Coverage (95% CI)**
Linear	Weighted	No SL	1.379	-1.78	0.128	0.124	0.946
		SL	1.400	-0.29	0.130	0.126	0.953
	Unweighted	No SL	1.603	14.17	0.100	0.095	0.431
		SL	1.601	14.03	0.100	0.094	0.435
Non-Linear	Weighted	No SL	1.431	12.05	0.166	0.162	0.826
		SL	1.273	0.08	0.105	0.101	0.943
	Unweighted	No SL	1.597	25.55	0.134	0.130	0.299
		SL	1.447	13.76	0.087	0.083	0.457

Ground truth: $\beta^A_{LATE,L} = 1.404$, $\beta^B_{LATE,L} = 1.604$; $\beta^A_{LATE,NL} = 1.272$, $\beta^B_{LATE,NL} = 1.447$

is aligns with the Monte Carlo results with and without SL. Clearly, the impact of having to estimate the weights is reflected in the difference between the weighted and unweighted SEs. In the linear DGM, SL incurs only a marginal increase in variance with a similar point estimate while in the non-linear DGM, SL is necessary to prevent bias and undercoverage.

4.5.2 Results for the Weighted ATE Bounds

For measuring the performance of our weighted ATE bounds, we again use 1000 simulations with the same DGM detailed above with a target ATE of 1. There are three main metrics measured across 1000 simulations: (i) the average of the lower bound estimate across simulations, (ii) the average of of the upper bound estimate across simulations, and (iii) the number of simulations whose intervals that cover the true ATE, which we will colloquially refer to as "coverage." Because the length of the ATE estimate interval is inversely proportional to IV strength, we present the results for strength ranging from 0.1 to 0.9. This also ensures

Table 4.4: Simulation Results for Linear and Non-Linear DGM Part III

Strong IV

			Point Estimate	**Bias (%)**	**Monte Carlo SE**	**Model-Based SE**	**Coverage (95% CI)**
Linear	Weighted	No SL	1.219	-0.51	0.085	0.084	0.951
		SL	1.234	0.72	0.087	0.085	0.943
	Unweighted	No SL	1.461	19.25	0.074	0.069	0.082
		SL	1.460	19.21	0.074	0.069	0.084
Non-Linear	Weighted	No SL	1.252	10.05	0.121	0.119	0.825
		SL	1.136	-0.15	0.075	0.074	0.949
	Unweighted	No SL	1.445	26.95	0.105	0.101	0.14
		SL	1.326	16.52	0.069	0.065	0.168

Ground truth: $\beta^A_{LATE,L} = 1.225$, $\beta^B_{LATE,L} = 1.461$; $\beta^A_{LATE,NL} = 1.138$, $\beta^B_{LATE,NL} = 1.327$

that our coverage results are informative since wider intervals are more likely to contain the true ATE.

Figure 4.1 shows how the average bounds across simulations vary with the IV strength, while Tables 4.5 and 4.6 detail those results along with the coverage. When the IV is of sufficient strength, the weighted bounds capture the true ATE for the target study at a far larger rate than the unweighted bounds. With a strong IV, the unweighted bounds almost always miss the true ATE while the weighted bounds increasingly center upon the target study ATE. At lower strengths, coverage was more a function of interval width than the result of a weighting. Similar to the LATE results, in the linear setting, SL had little to no impact on the results while in the non-linear setting, SL was crucial to ensuring that the bounds contained the target ATE at increasing strengths.

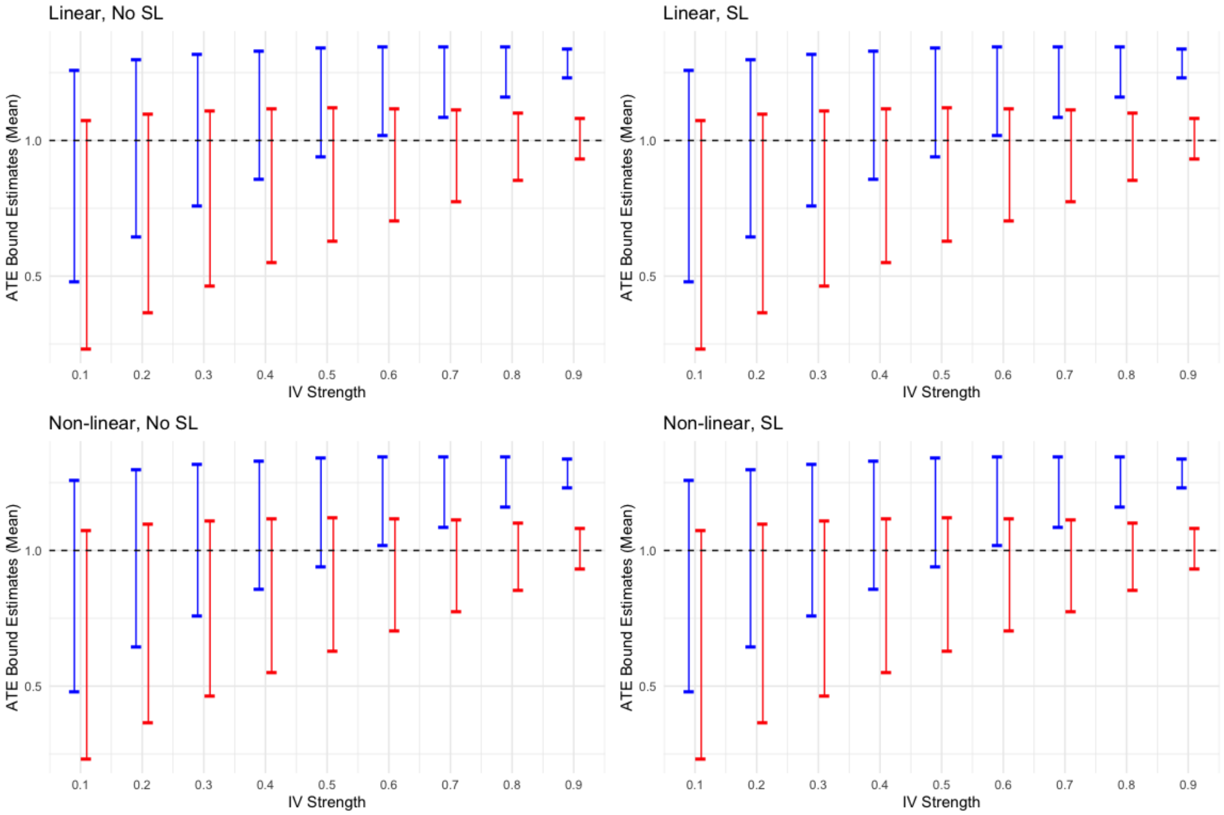


Figure 4.1: Simulation results from ATE bounds simulation. Blue lines are the unweighted bound estimators while red lines are the weighted bound estimators. The dotted black line indicates the target ATE ground truth of 1.

4.6 Applied Example

To demonstrate our methodology to replicate IV estimates across studies, we study the effect of triglycerides on cognitive decline by employing MR on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI is a multi-center, natural history longitudinal study that tracks cognitive performance among cognitively unimpaired, mild cognitively impaired, or cognitively impaired volunteers aged 55-90 who are in otherwise good health.[91] Here we consider two-year change from baseline in the clinical dementia rating sum of boxes (CDR-SB)[82] for those with mild cognitive impairment (MCI). Higher CDR-SB scores indicate more severe cognitive decline. The baseline CDR-SB is the earliest visit with recorded MCI. For the final measurement, there must be a CDR-SB measurement 1.75 to 2.5 years after their first measure. Our models both adjust for baseline CDR-SB and time from baseline.

Table 4.5: Simulation Results for ATE Bounds by Instrument Strength Part I

Linear, No SL									
Strength	**0.100**	**0.200**	**0.300**	**0.400**	**0.500**	**0.600**	**0.700**	**0.800**	**0.900**
Unweighted LB	0.483	0.646	0.762	0.857	0.941	1.018	1.088	1.160	1.234
Weighted LB	0.235	0.365	0.465	0.551	0.630	0.706	0.777	0.853	0.934
Unweighted UB	1.261	1.299	1.320	1.333	1.342	1.348	1.348	1.347	1.339
Weighted UB	1.077	1.100	1.111	1.118	1.121	1.120	1.114	1.104	1.082
Unweighted Coverage	0.999	1.000	0.996	0.965	0.779	0.385	0.108	0.009	0.000
Weighted Coverage	0.867	0.931	0.935	0.940	0.949	0.947	0.941	0.915	0.723
Linear, SL									
Unweighted LB	0.481	0.643	0.760	0.854	0.937	1.015	1.085	1.157	1.232
Weighted LB	0.244	0.373	0.471	0.557	0.634	0.710	0.780	0.856	0.937
Unweighted UB	1.258	1.296	1.316	1.330	1.339	1.346	1.345	1.345	1.339
Weighted UB	1.081	1.104	1.115	1.123	1.126	1.126	1.119	1.110	1.089
Unweighted Coverage	0.999	1.000	0.996	0.965	0.794	0.406	0.132	0.009	0.000
Weighted Coverage	0.867	0.928	0.937	0.939	0.955	0.948	0.937	0.904	0.721

Baseline covariates include medical history, age, body mass index (BMI), dementia family history, sex, APOE4 allele count, and years of education.

Using available genetic data, we constructed a binary IV by recording the presence of SNP *rs1260326*, which is associated with elevated triglyceride levels.[130] Our exposure, blood triglycerides levels, was binarized into "low" and "high" using the median level in our MCI cohort of 1.05 mmol/L. More details about how the genetic sequencing and lipid measurements were collected and about ADNI in general can be found at https://adni.loni.usc.edu/about/. Our final cohort had a total sample size of $n = 655$.

Table 4.6: Simulation Results for ATE Bounds by Instrument Strength Part II

				Non-linear, No SL					
Strength	**0.100**	**0.200**	**0.300**	**0.400**	**0.500**	**0.600**	**0.700**	**0.800**	**0.900**
Unweighted LB	0.625	0.804	0.920	1.011	1.086	1.151	1.209	1.263	1.315
Weighted LB	0.407	0.562	0.667	0.753	0.827	0.893	0.954	1.014	1.074
Unweighted UB	1.340	1.379	1.390	1.396	1.396	1.395	1.392	1.386	1.380
Weighted UB	1.184	1.208	1.209	1.207	1.200	1.194	1.184	1.173	1.161
Unweighted Coverage	0.999	0.959	0.769	0.445	0.215	0.071	0.021	0.007	0.005
Weighted Coverage	0.952	0.965	0.963	0.960	0.904	0.772	0.632	0.404	0.212
				Non-linear, SL					
Unweighted LB	0.531	0.703	0.815	0.904	0.976	1.040	1.095	1.147	1.198
Weighted LB	0.308	0.454	0.556	0.640	0.711	0.776	0.834	0.893	0.950
Unweighted UB	1.244	1.279	1.286	1.289	1.287	1.286	1.280	1.272	1.263
Weighted UB	1.078	1.102	1.101	1.098	1.091	1.083	1.073	1.059	1.043
Unweighted Coverage	0.994	0.995	0.980	0.860	0.613	0.309	0.105	0.016	0.001
Weighted Coverage	0.796	0.853	0.869	0.872	0.841	0.829	0.784	0.690	0.460

To demonstrate our methodology, we synthetically construct two cohorts by sampling the original data with replacement using sampling weights from the model

$$\begin{aligned} logit[P(M_B = 1|X = x)] &= \alpha_0 + \alpha_1 \mathrm{I}_{\text{Cardiovascular Disease}} + \alpha_2 \mathrm{I}_{\text{Neurological Disease}} \\ &\quad + \alpha_3 \mathrm{I}_{\text{Male}} + \alpha_4 \text{BMI} + \alpha_5 \text{Baseline CDR-SB} + \alpha_6 \text{Baseline Age} \end{aligned}$$

with $\alpha_0 = 0$ and $\alpha_1 = ... = \alpha_6 = 0.1$. Our model indicates those with higher values for the above covariates are more likely to be sampled in B. Henceforth, the original data is referred to as the target study and the sampled data the current study. Table 4.7 details the distributions of covariates across the cohorts. Binary IV strength is measured by the difference of the proportion of high triglyceride level subjects who do and do not have the gene, resulting in 0.13 for the current study and 0.11 for the target study, indicating an instrument of moderate strength.

Table 4.7: Baseline distributions of Covariates Stratified by Study

Variable	**Target Study (n = 655)**	**Current Study (n = 655)**	**Standardized Mean Difference**
Age (mean (SD))	72.96 (7.26)	71.30 (7.29)	0.228
Male (%)	394 (60.2)	400 (61.1)	0.019
Years Education (mean (SD))	15.98 (2.72)	15.94 (2.61)	0.014
BMI (mean (SD))	26.97 (4.77)	28.75 (5.38)	0.349
Family History (%)	371 (56.6)	394 (60.2)	0.071
Baseline CDR (mean (SD))	1.46 (0.88)	1.58 (0.97)	0.139
Time from Baseline (mean (SD))	2.03 (0.08)	2.03 (0.08)	0.042
APOE4 copies (%)			0.094
0	332 (50.7)	314 (47.9)	
1	256 (39.1)	255 (38.9)	
2	67 (10.2)	86 (13.1)	
Neurological Condition (%)	226 (34.5)	244 (37.3)	0.057
Cardiovascular Condition (%)	447 (68.2)	472 (72.1)	0.083
Metabolic Condition (%)	275 (0.42)	288 (0.44)	0.043

The results of our analysis are presented in 5.6. SL and no SL refer to the same estimators outlined in the previous section with five-fold cross-fitting for estimation. For a baseline ATE estimate, we ran OLS for both studies demonstrate a significant increase in CDR for those with higher triglycerides. To weight the current study OLS estimates, we computed survey weights via logistic regression and used 1000 bootstrap repetitions to calculate the variance. The weighted estimate does move towards the target study estimate but not notably so, indicating an opportunity to better model the weights.

Table 4.8: Estimated Effect of Triglycerides on Cognitive Decline with and without Weighting

Estimation Method	**Target Study**	**Current Study**	
		Weighted	**Unweighted**
OLS	0.34 (0.04, 0.65)	0.59 (0.29, 0.93)	0.63 (0.35, 0.93)
LATE			
No SL	2.13 (-0.93, 5.56)	3.53 (-2.46, 9.55)	5.36 (0.95, 9.77)
SL	1.44 (-1.7, 4.60)	1.73 (-4.20, 7.78)	2.73 (-3.50, 8.98)
ATE Bounds			
No SL	(-0.95, 0.16)	(-1.00, 0.08)	(-0.81, 0.25)
SL	(-1.11, 0.08)	(-0.93, 0.18)	(-0.77, 0.31)

The target study LATE, with and without SL, both have a positive point estimate but are not statistically significant. The unweighted LATE without SL in the current study shows a significant 5.36-point increase, but after weighting, the estimate attenuates toward the target LATE and becomes insignificant with our weighting method, reducing the relative difference from the target and current estimate from over 150% to 66%. SL for the current study LATE produced a lower point estimate and wider confidence intervals than the no SL equivalent (possibly a consequence of using flexible learners at lower sample-sizes) and, thus, was not statistically significant. After weighting the SL current study LATE, we see that the relative difference is reduced from 90% to 20%, showing notable attenuation towards the target SL point estimate.

Our ATE bounds in the target study for both no SL and SL suggest a null point estimate with the bounds crossing 0. The bounds for the unweighted current study estimates reach higher, which intuitively follows from the fact that the OLS and LATE estimates were higher. When we apply the weights, the bounds align well with the target study. For no SL, the percentage overlap between the ATE intervals increases from 87% to 93% while for SL, the percent overlap increases from 71% to 84%.

4.7 Discussion

Crucial to the trustworthiness of causal estimates from observational studies, including those found from MR, is whether they may be replicated across different populations. In this work, we have developed an approach to non-parametrically replicate the LATE across cohorts using possibly unknown survey weights. By focusing on the LATE as opposed to the ATE, our method has the inherent benefit of relaxing the untestable assumptions involved in replicability, including positivity and unmeasured effect modifiers. Explicitly, through the use of ML, our method can protect against functional misspecification of the survey weights. Our empirical results show that for all nuisance functions, using ensemble learning incurs little to no downside even when the DGM is linear, suggesting this approach is an acceptable default when the underlying DGM is unknown. Should we wish to target the ATE, our method performs well on providing bounds on the target ATE despite not making the accompanying causal assumptions, particularly in the case of a stronger IV.

Chapter 5

Nonparametric Instrumental Variable Estimation of Survival Times with Censored Data

5.1 Introduction

Survival analysis seeks to quantify the effect of an exposure on the time until an event of interest in the presence of censoring. Common event types include disease onset, disease progression or death. In the observational setting, many have focused on developing methods that can isolate causal effects while accounting for possible right-censoring.[51, 33, 8] For example, in studying the effect of blood lipid levels on the course of cognitive decline, older participants may become lost-to-followup for a variety of reasons. Due to potential violations in the assumption that confounding between exposure and outcome has been adequately accounted for, many have investigated the use of instrumental variables (IV) in survival analysis.[111, 75, 123]. Briefly, an IV must meet three main conditions: (i) it is associated with the exposure of interest, (ii) it is not associated with any unmeasured confounders, and

(iii) it is not directly associated with the outcome.[10] In the context of the previous example, if we wished to study the effect of high-density lipoprotein (HDL) levels on cognitive decline, we may choose to use one or more single nucleotide polymorphisms (SNPs) as IVs via a Mendelian Randomization (MR) approach.[93]

Under the setting of treatment effect heterogeneity related to the assignment of the IV, the IV approach identifies the local average treatment effect (LATE). That is, the effect of the treatment among the "compliers". Generally, the LATE is not equal to the population causal effect.[3, 122, 46] In MR, the compliers are those whose level of exposure changes monotonically with the presence of the selected SNPs. This work focuses on identifying the LATE in the survival setting where there is right-censoring.

Due to the ubiquity of the Cox model, some have developed IV estimators for the complier causal hazard ratio (CCHR). This typically involves a two-stage residual inclusion (2SRI) procedure wherein the residuals from the first stage are incorporated into a multiplicative hazards model to control for residual confounding.[112, 45, 109] Nevertheless, these methods have notable bias in estimating the CCHR, which can be attributed to non-collapsibility of the HR.[120, 121] One way to mitigate this bias to include a frailty term (i.e. random effect) in the second stage model.[76, 77] Others have developed estimators under the assumption of an additive hazards model, though some have criticized these approaches as "biologically implausible."[111, 75]

One salient disadvantage of the HR is that a clear causal interpretation heavily relies on the proportional hazards assumption being met.[50, 39] To avoid potential issues stemming from non-proportional hazards, other estimands have been investigated such as those based on the accelerated failure time (AFT) model in which the log-transformed survival time is regressed upon covariates of interest using a linear model.[113, 125] The AFT model yields an intuitive interpretation of regression coefficients as the multiplicative change in survival time. Due to the linear nature of AFT, traditional IV approaches such as 2SRI can be straightforwardly incorporated.[81, 55] One method to efficiently use both uncensored and censored observations to estimate the impact of a covariate on the survival time is by the Buckley-James method.[24] In this approach, the survival time of censored observations is iteratively imputed by incorporating a Kaplan-Meier estimator on the residuals.

Existing methods that extend the AFT model estimated via Buckley-James to the IV approach are limited to the linear setting[1]. While there are methods that incorporate machine learning (ML) to estimate non-linear AFTs, none have been developed in the IV setting to estimate the LATE.[124, 83] One path forward to non-parametrically identify the LATE in this setting is to build upon the recent developments in the "doubly debiased machine learning" literature.[28] Notably, Lee et al. (2023) develop an influence function (IF) approach to estimate the LATE of the survival probability.[70] Our work proposes a similarly derived estimator under the AFT model, incorporating the of use sample-splitting and machine learning to estimate nuisance functions, and yielding double robustness should some nuisance functions be misspecified. Additionally, we utilize a non-linear Buckley-James procedure as described by Wang and Wang (2013) to effectively include censored observations in the LATE estimation [124], thereby merging the AFT and modern causal inference literature. We specifically develop our method to estimate the relative difference in median survival times for compliers under the common scenario of a binary exposure and IV.

The remainder of this work is organized as follows. First, we introduce key notation and assumptions to identify the LATE using the AFT and BJ. In the next section, we describe our estimand and derive our nonparametric, IF-based estimator along with describing the asymptotic and double-robustness properties. Next, we present the procedure for estimation, including cross-fitting the BJ procedure and a modified variance estimator to account for the uncertainty in imputing of censored observations. We then compare the performance to the traditional two-stage predictor substitution approach across a variety of scenarios including potential misspecification. Finally, we apply our methodology to estimate the effect of HDL on progression of cognitive decline among those with mild cognitive impairment patients using a Mendelian Randomization approach and conclude by discussing the implications of our methodology.

5.2 Notation and Set-Up

Consider a binary treatment $A \in \{0, 1\}$ and instrumental variable (IV) $Z \in \{0, 1\}$ with log survival time denoted $T(a)$ and potential treatment assignment $A(z)$ where $T = AT(1) + (1 - A)T(0)$ and $A = ZA(1) + (1 - Z)A(0)$. We can define the local average treatment effect (LATE) of the log survival time among compliers as $\Psi_{LATE} = E[T(1) - T(0)|A(1) > A(0)]$. Assuming that the distribution of log-survival time is symmetric, exponentiating leads to the interpretation of $exp(\Psi_{LATE})$ as the ratio of the median survival times under exposure and non-exposure. Given a set of covariates X, we define the instrument propensity score (IPS) as $e(X) = P(Z = 1|X)$. To identify Ψ_{LATE}, we require the following assumptions conditional on confounders X:

A1 (Positivity of IPS): $0 < e(X) < 1$ a.s. for all $x \in \mathcal{S}_X$

A2 (Independence): $Z \perp\!\!\!\perp \{T(0,0), T(0,1), T(1,0), T(1,1), A(1), A(0)\} \,|\, X$

A3 (Exclusion Restriction): $T(a, 0) = T(a, 1)$ for $a \in \{0, 1\}$

A4 (Relevance): $P[A(1) = 1|X] > P[A(0) = 1|X]$ a.s. for all $x \in \mathcal{S}_X$

A5 (Monotonicity): $P[A(1) \geq A(0)|X] = 1$ a.s. for all $x \in \mathcal{S}_X$

With A1-A5, we may identify the LATE as[41]

$$\Psi_{LATE} = \frac{E[E[T|Z = 1, X] - E[T|Z = 0, X]]}{E[E[A|Z = 1, X] - E[A|Z = 0, X]]}. \tag{5.1}$$

In time-to-event outcomes, for individual i, censoring prevents the observation of T_i if there exists log censoring time C_i such that $T_i > C_i$. As such, let $\delta_i = I(T_i < C_i)$ and observed outcome $Y_i = min(T_i, C_i)$. We make the additional assumption that the censoring time is independent from the survival, given the covariates and treatment assignment.

A6 (Ignorable Censoring): $T \perp\!\!\!\perp C | A, X$

Thus, we will write

$$E[T|Z = z, X] = \sum_{a \in \{0,1\}} E[T|Z = z, A = a, X] P(A = a|X, Z = z) \text{ for } z \in \{0, 1\}. \quad (5.2)$$

Let us now focus on the estimation of the expectations of the log survival time such as $E[T|Z = 1, A = 1, X]$. Building off of the AFT literature, we posit a potentially non-linear relationship between covariates X and the log survival time as

$$T^{Z=1,A=1} = f^{Z=1,A=1}(X) + \epsilon^{Z=1,A=1} \quad (5.3)$$

such that $E[\epsilon^{Z=1,A=1}|Z = 1, A = 1, X] = 0$ and where the subscript refers to the particular strata where both $A = 1$ and $Z = 1$. We will drop the superscript notation henceforth, noting the remaining conditional expectations will take a similar form.

The primary goal of the Buckley-James approach is to incorporate the use of censored observations to estimate $f(X)$ by imputing T with $Y^* = Y\delta + E[T|T > Y, X](1 - \delta)$ noting that $E[Y^*] = E[T]$.[24]. Writing $E[T|T > Y, X] = f(X) + E[\epsilon|\epsilon > Y - f(X), X]$, we may use an iterative procedure to estimate $f(X)$ (i.e. the expectation of interest) via ML, compute the residuals, estimate $E[\epsilon|\epsilon > Y - f(X), X]$ via Kaplan-Meier, and then compute Y^* until convergence of $f(X)$. By incorporating the above decomposition of the log survival time into an IF-based estimator for Ψ_{LATE}, not only will we be able to utilize the Buckley-James procedure for estimation but we will completely non-parametric and doubly-robust for certain nuisance functions involved in Ψ_{LATE}.

5.3 Influence Function-based Buckley-James Estimation of the LATE

Define the following nuisance functions:

1. **Mean Survival Time Among Uncensored:** $m_{za}(X) = E[T \mid T < C, Z = z, A = a, X]$
2. **Mean Censoring Time Among Censored:** $\omega_{za}(X) = E[C|X, A = a, Z = z, T > C]$
3. **Mean Survival Time Adjustment:** $\lambda_{za}(X) = E[Y^* - C|X, A = a, Z = z, T > C]$
4. **Censoring Probability:** $\gamma^d_{za}(X) = P(\delta = d|Z = z, A = a, X)$ for $d \in \{0, 1\}$
5. **Treatment Propensity Score:** $\pi_z(X) = P(A = 1|Z = z, X)$
6. **IPS:** $e(X) = P(Z = 1|X)$
7. **Marginal Probability of X:** $p(x) = P(X = x)$.

Conditioning on A as in Eq. 5.2 the LATE can be written as

$$\Psi_{LATE} = \frac{\psi_{11} + \psi_{10} - \psi_{01} - \psi_{00}}{\phi_1 - \phi_0} \tag{5.4}$$

with $\psi_{za} = E\big[E[T|Z = z, A = a, X]P(A = a|Z = z, X)\big]$ and $\phi_z = E[\pi_z(X)]$. Focusing on the numerator, without loss of generality we will derive on the properties of $\psi_{11} = E\big[E[T|Z = 1, A = 1, X]P(A = 1|Z = 1, X)\big]$. In the following proposition, we may re-write $E[T|Z = 1, A = 1, X]$ in terms of the Buckley-James decomposition.

Proposition 5.3.1. *Via the Buckley-James decomposition,*

$$E[T \mid Z = 1, X, A = 1] = m_{11}(X)\gamma_{11}^{1}(X) + \omega_{11}(X)\gamma_{11}^{0}(X) + \lambda_{11}(X)\gamma_{11}^{0}(X)$$

Given the decomposition in Prop. 5.3.1, we may derive the IF for of the three terms separately using the properties outlined in Kennedy (2023) Section 3.[65]

Proposition 5.3.2. *The IF for $\psi_{1,1}$ is written as*

$$\begin{aligned}\mathbb{IF}(\psi_{1,1}) = &\left(\frac{Z}{e(X)}\left[\delta AT - m_{11}\gamma_{11}^{1}\pi_1\right] + m_{11}\gamma_{11}^{1}\pi_1\right) \\ &+ \left(\frac{Z}{e(X)}\left[(1-\delta)AC - \omega_{11}\gamma_{11}^{0}\pi_t\right] + \omega_{11}\gamma_{11}^{0}\pi_1\right) \\ &+ \left(\frac{Z}{e(X)}\left[(1-\delta)A(Y^* - C) - \lambda_{11}\gamma_{11}^{0}\pi_1\right] + \lambda_{11}\gamma_{11}^{0}\pi_1\right) - \psi_{11}\end{aligned}$$

We can generalize this quantity to any Z and A by replacing $e(X)$ with $e(X, Z) = P(Z|X)$ and $\pi_Z(X)$ with $\pi_Z(X, A) = P(A|Z, X)$. We denote this arbitrary estimand as $\psi_{Z,A}$ and can write the general IF of the numerator as

$$\varphi_{num} = \mathbb{IF}\big\{(\psi_{11} + \psi_{10}) - (\psi_{01} + \psi_{00})\big\} = \mathbb{IF}\{\psi_{11}\} + \mathbb{IF}\{\psi_{10}\} - \mathbb{IF}\{\psi_{01}\} - \mathbb{IF}\{\psi_{00}\}$$

The IF for the denominator is derived in Kennedy (2023) Example 6. That is, we have

$$\varphi_{denom} = \mathbb{IF}(\phi_1 - \phi_0) = \frac{2Z-1}{e(X,Z)}[A - \pi_Z(X)] + \pi_1(X) - \pi_0(X) - (\phi_1 - \phi_0)$$

Thus, the IF of the entire LATE is

$$\mathbb{IF}(\Psi_{LATE}) = \left[\mathbb{P}_n(\phi_1 - \phi_0)\right]^{-1} \left\{\varphi^U_{num}(\eta) - \Psi_{LATE}\varphi^U_{denom}(\eta)\right\}$$

Let φ^U_{num} and φ^U_{denom} refer to the uncentered IFs for the numerator and denominator, respectively. Assuming a known value for Y^*, we may utilize the plug-in estimator $\hat{\Psi}_{LATE} = \mathbb{P}_n\{\varphi^U_{num}(\hat{\eta})\}/\mathbb{P}_n\{\varphi^U_{denom}(\hat{\eta})\}$ where $\eta = \{e(X,Z), \pi_Z(X,A), \mu_{ZA}(X), \gamma_{ZA}(X), \kappa_{ZA}(X)\}$, the set of nuisance functions. We will discuss the case with an estimated Y^* in the next section. Once this estimate is obtained, we may exponentiate it to estimate the ratio of the median survival times between the treatment groups among compliers.

To establish the asymptotic and double robustness properties of $\hat{\Psi}_{LATE}$, we assume the following conditions:

C1: Each nuisance functions $\{g \in \eta : \|\hat{g} - g\| = o_P(1)\}$ belong to the Donsker class.

C2: There exist constants such that $|\hat{m}_{ZA}| < C_{m_{ZA}}, \hat{\gamma}^1_{ZA} > C_{\hat{\gamma}^1_{ZA}}, \gamma^1_{ZA} < C_{\gamma^1_{ZA}}$,

$\hat{\pi}_Z > C_{\pi_Z}, |\hat{\omega}_{ZA}| < C_{\hat{\omega}_{ZA}}$,

$\hat{\gamma}^0_{ZA} < C_{\hat{\gamma}^0_{ZA}}, \gamma^0_{ZA} < C_{\gamma^0_{ZA}}, |\hat{\lambda}_{ZA}| < C_{\hat{\lambda}_{ZA}},$ and $\hat{e} > C_{\hat{e}}$

, almost surely for $Z \in \{0,1\}$ and $A \in \{0,1\}$.

C3: $\{h \in \eta^* : \left\|\hat{h} - h\right\| \|\hat{e} - e\| = o_P(n^{-1/2})\}$ for $\eta^* = \{\pi_Z, m_{ZA}, \omega_{ZA}, \lambda_{ZA}, \gamma_{ZA}\}$

for $Z \in \{0,1\}$ and $A \in \{0,1\}$

Note that (C1) is not required if we use sample-splitting to estimate η[28, 65, 70], as proposed in the next section.

Theorem 5.3.1. *Given the following decomposition* $\hat{\Psi}_{LATE} - \Psi_{LATE} = S^* + T_1 + T_2$, *then* $T_1 + T_2 = o_P(n^{-1/2})$ *under (C1)-(C3) with*

$$n^{1/2}\left(\hat{\Psi}_{LATE} - \Psi_{LATE}\right) = n^{1/2}S^* + o_P(1) \xrightarrow{d} N\left(0, E[\Gamma^2]\right) \tag{5.5}$$

where $\Gamma = \left[\mathbb{P}_n(\phi_1 - \phi_0)\right]^{-1}\left\{\varphi^U_{num}(\eta) - \Psi_{LATE}\varphi^U_{denom}(\eta)\right\}$.

Using these results, constructing $(1-\alpha)$-level Wald-based confidence intervals for Ψ_{LATE} is straightforward. With $q_{1-\alpha/2}$ representing the $1-\alpha/2$ quantile of the standard Normal distribution, we can utilize a plug-in estimator for $\hat{\Gamma}$ to obtain

$$\hat{\Psi}_{LATE} \pm q_{1-\alpha/2}\sqrt{\frac{\mathbb{P}_n\hat{\Gamma}^2}{n}}.$$

which may then be exponentiated.

The double-robustness condition of our estimator is detailed by (C3), which is required for the term $T_2 = o_P(n^{-1/2})$. In summary, we jointly require that one of the following in each line to be correctly specified:

1. $\pi_Z(X)$ or $e(X)$ for $Z \in \{0,1\}$,
2. $m_{ZA}(X)$ or $e(X)$ for $Z \in \{0,1\}$ and $A \in \{0,1\}$,
3. $\gamma^1_{ZA}(X)$ or $e(X)$ for $Z \in \{0,1\}$ and $A \in \{0,1\}$,
4. $\omega_{ZA}(X)$ or $e(X)$ for $Z \in \{0,1\}$ and $A \in \{0,1\}$, and
5. $\lambda_{ZA}(X)$ or $e(X)$ for $Z \in \{0,1\}$ and $A \in \{0,1\}$.

5.4 Estimation

We propose estimating η, the set of nuisance functions in Ψ_{LATE} via sample-splitting.[65] Briefly, we randomly split i.i.d. data of size n into K mutually exclusive groups $\mathcal{Q}_n = q^{(1)} \cup q^{(2)} \cup ...q^{(K)}$. For $k \in \{1, 2, ...K\}$, using $\mathcal{R}_n \setminus q^{(k)}$, we estimate $\eta^{(k)}$ and then plug-in $q^{(k)}$ to obtain our estimate for $\hat{\Psi}^{(k)}_{LATE}$. With cross-fitting, we obtain a pooled final estimate

$$\hat{\Psi}_{LATE} = \frac{1}{K} \sum_{1}^{k} \hat{\Psi}^{(k)}_{LATE}.$$

By incorporating cross-fitting, we may non-parametrically estimate all the nuisance functions in $\mathbb{IF}(\Psi_{LATE})$ using flexible modeling (e.g. splines, random forests) while protecting against overfitting and straightforwardly avoiding the need to satisfy (C1). While we can estimate $e(X, Z)$,$\pi_Z(X, A)$, $m_{ZA}(X)$, $\gamma_{ZA}(X)$, and $\omega_{ZA}(X)$ by simply fitting them with training set of each fold in cross-fitting, $\lambda_{11}(X)$ will be fitted with $\hat{Y}^*$ obtained iteratively with the training set via the Buckley-James procedure.

The Buckley-James estimation procedure for fitting a non-linear function with sample-splitting for arbitrary Z and A is as follows. For training data of size n_k, we first define observed residuals $r_i = Y_i - \hat{\mu}_{za}(X_i)$ with ordering $r_1 < r_2 < ... < r_{n_k}$. The training data is additionally sorted in this order. Following the notion from Wang and Wang (2010), $\kappa_{za}(X) = E[\epsilon | \epsilon > Y - f(X), X, Z = z, A = a]$ can be estimated non-parametrically as follows

$$\hat{\kappa}_{za} = \hat{S}(r_i)^{-1} \sum_{r_j > r_i} r_j \delta_j \Delta \hat{S}(r_j) \tag{5.6}$$

where $\hat{S}(r_i)$ is the Kaplan-Meier estimator of the survival function for residual r_i.[124]

We can estimate both $\hat{f}_{za}$ and $\hat{\kappa}_{za}$ via the following iterative approach in Wang and Wang (2010).

1. Set $M = 0$ and obtain initial estimate $\hat{\mu}_{ZA}^{(0)}$ using the uncensored data as well as $\hat{\kappa}_{ZA}$ from the residuals.

2. At the Mth iteration:

   (a) $Y^* = \hat{f}_{za}^{(M-1)}(X_i) + e_i\delta_i + (1 - \delta_i)\hat{\kappa}_{za}(X_i)$

   (b) Fit $\hat{f}_{za}^{(M)}(X_i)$ using X and Y^*

3. Repeat Step 2 until a set number of iterations is reached or, for some constant $\alpha > 0$,

$$\frac{\left\| \hat{f}_{za}^{(M)}(X_i) - \hat{f}_{za}^{(M-1)}(X_i) \right\|}{\hat{f}_{za}^{(M-1)}(X_i)} < \alpha \tag{5.7}$$

With $\hat{\eta}$, we may now compute the relevant values using the hold-out sample and plug them into our estimator in Prop. 5.3.2. The first step is to compute residuals $r = Y - \hat{f}_{za}$. Then, using the fitted KM estimator $\hat{S}$ and training set cumulative sums in Eq. 5.6, we may compute $\hat{Y}^*$. In this case, the estimator in Theorem 5.3.1 underestimates the true variance because we need to account for the estimation of $\hat{Y}^*$. More specifically, the second-order bias term $\left\| \hat{Y}^* - Y^* \right\|$ in T_2 is likely not of order $o_P(n^{-1/2})$ for the IF of the right-hand side third term in Prop. 5.3.1 (see Appendix).

The second-order bias caused by imputation could, in theory, be accounted for by "higher-order" IFs.[95] That is, capturing additional terms beyond the one-step estimator in the von Mises expansion of the estimated IF around the true IF. Nevertheless, it is not known whether a higher-order de-biased estimator exists for our estimand, which additionally would may make restrictive structural assumptions about the underlying nuisance function space.[11] In traditional estimation settings, "Rubin's rules" (RR) is used to account for the additional variance of multiple-imputation where "within-imputed dataset" estimated variances are pooled and added to the "between-dataset" variance of the point-estimate.[?] We borrow this intuition to heuristically account for the remaining variance in T_2 in imputing Y^*, inducing randomness to measure between-dataset bootstrapping. This procedure is limited to the IF of the third term in Prop. 5.3.1 as the other two terms are not influenced by Y^*.

Without loss of generality, let φ_{11} represent the uncentered IF for the estimand ψ_{11}. Based on Prop. 5.3.1, we may decompose it into three terms $\varphi_{11} = \varphi_{11}^A + \varphi_{11}^B + \varphi_{11}^C$ where

$$\varphi_{11}^A(\eta^A) = \frac{Z}{e(X)}\left[\delta AT - m_{11}\gamma_{11}^1\pi_1\right] + m_{11}\gamma_{11}^1\pi_1,$$

$$\varphi_{11}^B(\eta^B) = \frac{Z}{e(X)}\left[(1-\delta)AC - \omega_{11}\gamma_{11}^0\pi_t\right] + \omega_{11}\gamma_{11}^0\pi_1,$$

$$\varphi_{11}^C(\eta^C) = \frac{Z}{e(X)}\left[(1-\delta)A(Y^* - C) - \lambda_{11}\gamma_{11}^0\pi_1\right] + \lambda_{11}\gamma_{11}^0\pi_1.$$

We may estimate $\varphi_{11}^A(\eta^A)$ and $\varphi_{11}^B(\eta^B)$ without the use of bootstrapping, resulting in two vectors of length n after plugging in the values of the hold-out set. For $\varphi_{11}^C(\eta^C)$, for a given train-test split, we take K bootstrap samples in the train data and fit $\{\hat{\eta}_{(1)}^C, \hat{\eta}_{(2)}^C, ...\hat{\eta}_{(K)}^C\}$. Crucially this gives us variability in calculation of Y^* and, subsequently, $\lambda_{11}(X)$. Then, using test set, we may compute $\{\varphi_{11}^C(\hat{\eta}_{(1)}^C), \varphi_{11}^C(\hat{\eta}_{(2)}), ...\varphi_{11}^C(\hat{\eta}_{(1)}^C)\}$, which forms a $n \times k$ matrix $\varphi_{11,boot}^C$ after plugging in the observed data. Thus, we have a matrix $\varphi_{11}^{pool} = \varphi_{11}^A + \varphi_{11}^B + \varphi_{11,boot}^C$ and we can take the column means $\overline{\varphi}_{11}^{pool}$ as well as the column means of the squared IF values $\overline{\varphi_{11,k}^{2,pool}}$. The point estimate is thus $\hat{\psi}_{11} = (K)^{-1}\sum_{k=1}^K \overline{\varphi}_{11,k}^{pool}$ with

$$Var(\hat{\psi}_{11}) = (K)^{-1}\sum_{k=1}^K n^{-1}\overline{\varphi_{11,k}^{2,pool}} + (1+\frac{1}{K})(\frac{1}{K-1})\sum_{k=1}^K\left(\hat{\psi}_{11} - \overline{\varphi_{11,k}^{2,pool}}\right)^2.$$

We generalize this to the entire IF for Ψ_{LATE} in Prop. 5.4.1.

Proposition 5.4.1. *Writing* $\varphi_{num} = \varphi_{num}^A + \varphi_{num}^B + \varphi_{num}^C$ *where* $\varphi_{num}^l = \varphi_{num,11}^l + \varphi_{num,10}^l - \varphi_{num,01}^l - \varphi_{num,00}^l$ *for* $l \in \{A, B, C\}$. *We estimate* $Var(\Psi_{LATE})$ *with the bootstrap as follows.*

1. *Compute $\hat{\varphi}^{A}_{num}$, $\hat{\varphi}^{B}_{num}$, and $\hat{\varphi}_{denom}$ using the standard cross-fitting procedure.*

2. *For $\hat{\varphi}^{C}_{num}$ in each fold's sample split we do have the following*

   (a) *Bootstrap sample the training data K times*

   (b) *For each bootstrap, compute the respective nuisance functions $\{\eta^{C}_{num,1}, \eta^{C}_{num,2}, ...\eta^{C}_{num,K}\}$*

   (c) *Using the test data, compute matrix $\hat{\varphi}^{C}_{num}$ with column vectors $(\hat{\varphi}^{C}_{num,(1)}), \hat{\varphi}^{C}_{num,(2)}), ...\hat{\varphi}^{C}_{num,(K)})$*

   (d) *Combine test set values across such that we have a $n \times K$ matrix.*

3. *Create $\hat{\varphi}^{pool}_{num}$ by adding $\hat{\varphi}^{A}_{num}$ $+\hat{\varphi}^{B}_{num}$ to each column vector $\hat{\varphi}^{C}_{num,(k)}$ for $k \in \{1, 2, ...K\}$.*

4. *Compute a K-length vector of point estimates, $\boldsymbol{\Psi_{LATE}}^{pool}$, where the kth entry is $\frac{\mathbb{P}_n \hat{\varphi}^{C}_{num,(k)}}{\mathbb{P}_n \hat{\varphi}_{denom}}$.*

5. *Compute pooled point estimate $\Psi^{pool}_{LATE} = K^{-1} \sum_{k=1}^{K} \frac{\mathbb{P}_n \hat{\varphi}^{C}_{num,k}}{\mathbb{P}_n \hat{\varphi}_{denom}}$*

6. *Compute a $n \times K$ matrix of LATE IF values as* $\Gamma^{pool} = \left[\mathbb{P}_n(\phi_1 - \phi_0)\right]^{-1} \left\{\hat{\varphi}^{pool}_{num} - \hat{\varphi}_{denom} \boldsymbol{\Psi_{LATE}}^{pool}(\eta)\right\}$

7. *Via Rubin's Rules:* $Var(\Psi_{LATE}) = (K)^{-1} \sum_{k=1}^{K} n^{-1} \overline{\left(\Gamma^{pool}_{k}\right)^{2}} + (1 + \frac{1}{K})(\frac{1}{K-1}) \sum_{k=1}^{K} (\Psi^{pool}_{LATE} - \boldsymbol{\Psi_{LATE,k}}^{pool})^2$

With this augmented variance estimator, we can account for the uncertainty introduced from having to impute Y^*, which is subsequently used when fitting λ.

5.5 Simulation

We study the properties of our non-parametric estimator under both linear and non-linear data generating mechanisms (DGMs). First, we generate six continuous covariates $X =$

$(X_1, X_2, X_3, X_4, X_5, X_6)$ from a multivariate normal distribution with $N = 4000$ and each variable having mean 0. The covariance matrix has two correlated blocks, X_1, X_2, X_3 and X_4, X_5, X_6 each with correlation 0.3. Furthermore, we have an uncorrelated unmeasured covariate U generated from the standard normal distribution.

Table 5.1 details the equations and coefficients for the linear and non-linear DGM. Briefly, once the six covariates and unmeasured covariate are generated we generate Z via the IPS with $P(Z = 1) = 0.5$. Then, we generate A from the treatment propensity score, where θ_1 represents the strength of the IV and $P(A = 1) = 0.5$. The censoring times are generated from the exponential distribution in which the rate parameter C_R is iteratively found such that the proportion of the censored data matches our simulation specification. Finally, the outcome is generated via a log-normal distribution with the treatment effect being $\beta_1 = -0.8$. Because there is no treatment effect heterogeneity the LATE is equal to the ATE.

In each DGM, as a baseline comparison method, we fit two-stage least squares (2SLS) via substituting Y^* for Y where the imputed values are derived from the traditional linear AFT BJ procedure. That is, we first use the `bj` function in the `rms` package to regress Y on the covariates and A to derive Y^*. Then, 2SLS is fit as usual upon Y^*. In our method, we fit all nuisance functions except κ with and without ensemble learning through the `SuperLearner` package.[92] We denote "No SL" to refer to the case where all nuisance functions are fit via generalized linear models (GLM) whereas "SL" refers to an ensemble of multivariate adaptive regression splines (EARTH), generalized additive models (GAMs), GLMs, random forest (RF), and recursive partitioning and regression trees (RPART). All estimates are computed via cross-fitting with four splits.

Within each DGM, we examine our estimator across three different scenarios varying the censoring rate from 10% to 60% with the ground truth being -0.8 on the log scale across all simulations. We examine five metrics: average point estimate, average relative bias, Monte Carlo variance, model-based variance, and the root mean squared error (MSE).For 2SLS,

Table 5.1: Summary of Data-Generating Mechanisms for Simulation

Model	**DGM**	**Equation**	**Coefficient Values**
IPS	Linear	$logit[P(Z=1\|X=x)] =$ $\gamma_0 + \sum_{i=1}^{6} \gamma_i X_i$	$\gamma_0 = 0,$ $\gamma_1 = \gamma_2 = \gamma_3 = 0.2,$ $\gamma_4 = \gamma_5 = \gamma_6 = -0.2$
	Non-Linear	$logit[P(Z=1\|X=x)] =$ $\gamma_0 + \gamma_1 X_1 + \gamma_2 \exp(X_2) + \gamma_3 \sin(X_3) +$ $\gamma_4 X_4 + \gamma_5 \cos(X_5) + \gamma_6 X_6^2$	
Treatment Propensity	Linear	$logit[P(A=1\|X=x)] =$ $\theta_0 + \theta_1 Z + \theta_2 U + \sum_{i=3}^{8} \theta_i$	$\theta_0 = 0,\ \theta_1 = 0.8,$ $\theta_2 = -0.6$ $\theta_3 = \theta_4 = \theta_5 = 0.4,$ $\theta_5 = \theta_6 = \theta_7 = -0.4$
	Non-Linear	$logit[P(A=1\|X=x)] = \theta_0 + \theta_1 Z +$ $\theta_2 exp(U) + \theta_3 \cos(X_1) + \theta_4 X_2^2 +$ $\theta_5 \sin(X_3) + \theta_6 X_4 + \theta_7 exp(X_5) + \theta_8 X$	
Censoring	Linear	$log(C) =$ $log(C_R) + \alpha_1 A + \sum_{i=2}^{7} \alpha_i X_i$	$\alpha_1 = -0.3,$ $\alpha_2 = \alpha_3 = \alpha_4 = 0.3,$ $\alpha_5 = \alpha_6 = \alpha_7 = -0.3$
	Non-Linear	$log(C) =$ $log(C_R) + \alpha_1 A + \alpha_2 exp(X_1)$ $+ \alpha_3 X_2^2 + \alpha_4 \sin(X_3) + \alpha_5 X_4 +$ $\alpha_6 X_5 + \alpha_7 exp(X_6)$	$\alpha_1 = -0.3,$ $\alpha_2 = \alpha_3 = \alpha_4 =$ $-0.15,$ $\alpha_5 = \alpha_6 = \alpha_7 = 0.3$
Outcome	Linear	$log(Y) =$ $\beta_0 + \beta_1 A + \beta_2 U + \sum_{i=3}^{8} \beta_i + \epsilon$	$\beta_0 = 1.5,\ \beta_1 = -0.8,$ $\beta_2 = 0.8,$ $\beta_3 = \beta_4 = \beta_5 = 0.4,$ $\beta_6 = \beta_7 = \beta_8 = -0.4,$ $\epsilon \sim N(0,1)$
	Non-Linear	$log(Y) = \beta_0 + \beta_1 A + \beta_2 exp(U)$ $+ \beta_3 \cos(X_1) + \beta_4 X_2^2 + \beta_5 \sin(X_3) +$ $\beta_6 exp(X_4) + \beta_7 X_5 + \beta_8 X_6 + \epsilon$	$\beta_0 = 1.5,\ \beta_1 = -0.8,$ $\beta_2 = 0.5,$ $\beta_3 = \beta_4 = \beta_5 = -0.4,$ $\beta_6 = \beta_7 = \beta_8 = 0.4,$ $\epsilon \sim N(0,1)$

Table 5.2: Simulation Results: Linear Data Generating Mechanism (n = 4000)

		Point Estimate	**Bias (%)**	**Monte Carlo SE**	**Model-Based SE**	**MSE**	**Coverage (%)**
10% Cens.	2SPI	-0.803	0.38	0.102	0.106	0.102	92.4
	Proposed Method	-0.779	2.26	0.102	0.109	0.102	95.8
20% Cens.	2SPI	-0.805	0.625	0.112	0.113	0.113	92.8
	Proposed Method	-0.777	2.88	0.113	0.113	0.114	95.0
30% Cens.	2SPI	-0.804	0.50	0.120	0.123	0.120	93.6
	Proposed Method	-0.783	2.12	0.117	0.123	0.117	95.8
40% Cens.	2SPI	-0.809	1.125	0.129	0.132	0.130	94.4
	Proposed Method	-0.793	0.80	0.149	0.136	0.149	94.4
50% Cens.	2SPI	-0.809	1.125	0.150	0.151	0.150	93.6
	Proposed Method	-0.802	0.01	0.184	0.160	0.184	94.4
60% Cens.	2SPI	-0.810	1.25	0.170	0.176	0.170	93.2
	Proposed Method	-0.788	1.50	0.269	0.206	0.269	91.0

the model-based variance is calculated via 500 bootstrap iterations; for our estimator, 10 bootstrap samples were used to calculate the variance estimator.

We present three sets of results with $n = 4000$. Table 5.2 details the results from 1000 simulations in the linear DGM. To examine the robustness to misspecification, we conducted another simulation in the linear setting that intentionally drops a covariate from the mean model for 2SPI (i.e. the `bj` function) and all nuisance functions besides the IPS. The results are presented in Table 5.3. Table 5.4 details the non-linear DGM results from 500 simulations (due to large computation times).

In the linear setting, as expected, both the 2SRI approach and the proposed method are

Table 5.3: Simulation Results: Misspecification (n = 4000)

		Point Estimate	**Bias (%)**	**Monte Carlo SE**	**Model-Based SE**	**MSE**	**Coverage (%)**
10% Cens.	2SPI	1.779	322.44	0.156	0.152	6.809	0
	Proposed Method	-0.766	4.22	0.087	0.091	0.088	95.4
20% Cens.	2SPI	1.756	319.56	0.155	0.155	6.690	0
	Proposed Method	-0.703	12.07	0.091	0.088	0.095	95.4
30% Cens.	2SPI	1.73	316.04	0.164	0.160	6.571	0
	Proposed Method	-0.714	10.69	0.106	0.103	0.113	94.4
40% Cens.	2SPI	1.70	312.31	0.179	0.167	6.42	0
	Proposed Method	-0.691	13.67	0.116	0.113	0.128	92.4
50% Cens.	2SPI	1.69	311.69	0.174	0.178	6.39	0
	Proposed Method	-0.667	16.67	0.146	0.132	0.163	92.8
60% Cens.	2SPI	1.66	307.63	0.206	0.191	6.26	0
	Proposed Method	-0.643	19.62	0.200	0.166	0.224	91.0

Table 5.4: Simulation Results: Non-linear Data Generating Mechanism (n=4000)

		Point Estimate	**Bias (%)**	**Monte Carlo SE**	**Model-Based SE**	**MSE**	**Coverage (%)**
10% Cens.	2SPI	-1.381	72.6	0.081	0.081	0.419	42.4
	Proposed Method	-0.838	4.75	0.065	0.072	0.066	96.0
20% Cens.	2SPI	-1.404	75.5	0.084	0.086	0.449	41.6
	Proposed Method	-0.848	6.00	0.064	0.078	0.066	94.8
30% Cens.	2SPI	-1.425	78.1	0.092	0.093	0.480	41.6
	Proposed Method	-0.867	8.38	0.085	0.087	0.088	95.2
40% Cens.	2SPI	-1.432	79.0	0.095	0.102	0.495	41.6
	Proposed Method	-0.869	8.62	0.102	0.102	0.105	94.8
50% Cens.	2SPI	-1.456	82.3	0.11	0.114	0.541	43.8
	Proposed Method	-0.917	14.62	0.128	0.125	0.140	91.6
60% Cens.	2SPI	-1.494	86.7	0.126	0.137	0.607	48.0
	Proposed Method	-0.918	14.8	0.213	0.177	0.226	90.4

approximately unbiased and have acceptable coverage. We note that our model-based variance estimator begins to underestimate the empirical variance when censoring is at 50% and beyond. When we misspecify key models, even though we remain the linear setting, 2SPI shows substantial bias while the proposed method shows minimal bias with this increasing bias increasing modestly with the censoring rate increasing due to the BJ imputations becoming increasingly incorrect (recall we are not doubly robust for $\hat{Y}^*$, see Appendix). This demonstrates the ability of the proposed flexible, nonparametric estimator to mitigate the effects of model misspecification. Lastly, in the non-linear setting, 2SPI contains large bias with poor coverage while the proposed method shows minimal bias, even at higher censoring rates where there are less uncensored observations to train complex machine learning models. Furthermore, the proposed variance estimator largely captures the empirical variance even in the non-linear setting.

We additionally examine the performance of our method in the linear and non-linear settings with a low sample size of $n = 1000$, which can be found in the Appendix. Across scenarios, though our method is approximately unbiased for the LATE, our computational variance estimator tends to overestimate the empirical variance, leading to overcoverage. This extra variability primarily stems from additional sparsity of strata induced by random sampling with replacement in the bootstrap. Indeed, there are some cases where we obtain extreme results or the BJ procedure simply fails to run, which particulary pronounced in the non-linear setting where adequately learning ML models is imperative. Typically, this occurs in the lower censoring rate scenarios (e.g. 10%) where we have only a few observations to fit a model on censored data.

5.6 Application to Dementia Progression

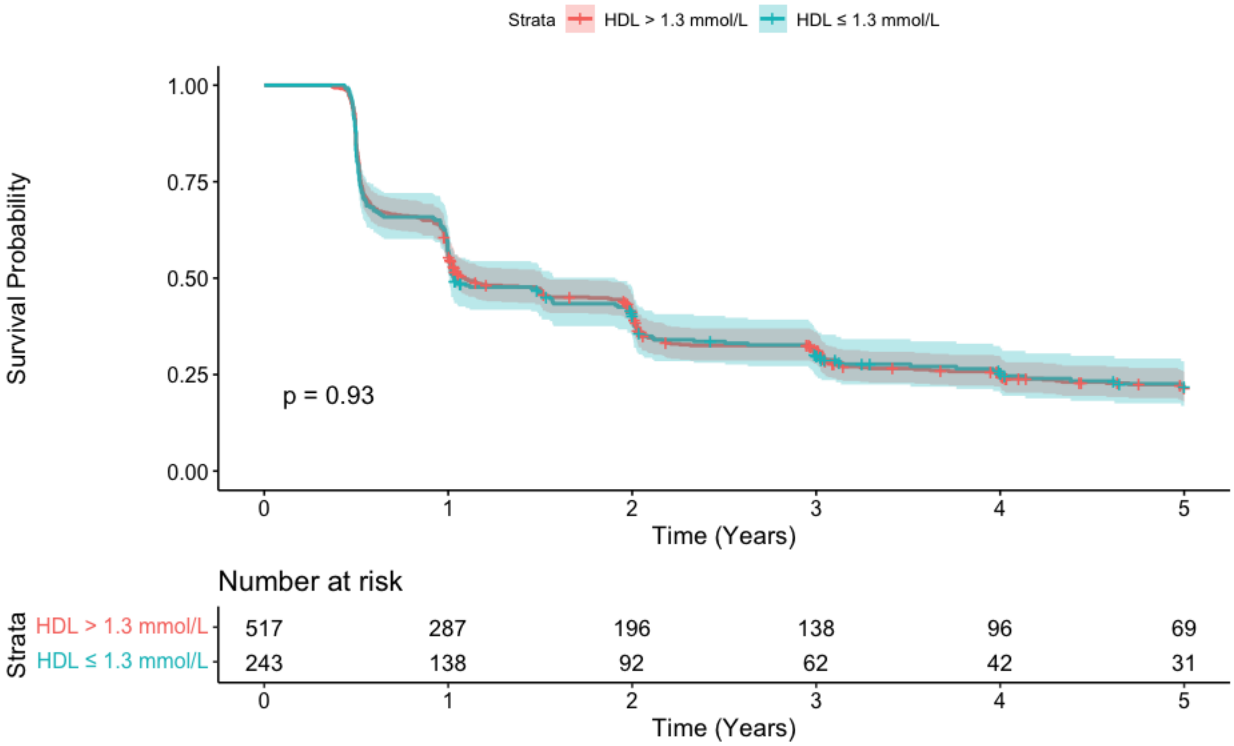
We apply our methodology to study the effect of HDL on cognitive decline from a Mendelian randomization analysis with data from the Alzheimer's Disease Neuroimaging Initiative

(ADNI). ADNI is a natural history that longitudinally tracks cognitively unimpaired, mild cognitively impaired, or cognitively impaired volunteers aged 55-90 who are in otherwise good health over time.[91] Our event of interest is the time until a half-point progression from baseline in the clinical dementia rating sum of boxes (CDR-SB)[82] for those with mild cognitive impairment (MCI). An increase of CDR indicates cognitive decline.

Our exposure is a participant's HDL levels at the earliest visit with recorded MCI, which we refer to as the "baseline visit." HDL was binarized into a "low" and "high" exposure group with a cut-off of 1.3 mmol/L. Using available genetic data, we constructed a binary IV that measures whether if a participant has two copies of the SNP *rs3764261*, which is associated with lower HDL levels.[130] More details regarding genetic and lipid data in ADNI can be found at https://adni.loni.usc.edu/about/.

Our final cohort contained $n = 760$ inidividuals and with administrative censoring at five years where 564 participants, or 74%, had an event. The Kaplan-Meier curve stratified by HDL level is shown in Figure 5.1. Baseline covariates include baseline CDR, include medical history, age, body mass index (BMI), dementia family history, sex, APOE4 allele count, and years of education (Table 5.5). The strength of the IV is estimated by the difference in the proportion of individuals with high HDL levels between those who have and do not have two copies of the gene. The proportion for the low HDL group was 0.56 while the proportion was 0.47 in the high HDL group, yielding a complier percentage of 9%, indicating instrument of moderate strength.

The results across a multiple estimation approaches are presented in Table 5.6. Using the traditional BJ confounder adjustment, we obtain a null and significant point estimate. With the MR approach, the complete case 2SPI yields a protective point estimate of high HDL on progression of cognitive decline with this estimate becoming statistically significant once we incorporate BJ. A point estimate of 0.17 indicates there is a 83% decrease in the median time until a 0.5 point progression in CDR in the high HDL group compared to the low HDL



Strata	0	1	2	3	4	5
HDL > 1.3 mmol/L	517	287	196	138	96	69
HDL ≤ 1.3 mmol/L	243	138	92	62	42	31

Figure 5.1: Kaplan Meier Curve of Time Until Half-Point Progression, Stratified by HDL Level

Table 5.5: Baseline distributions of Covariates Stratified by Study

Variable	**HDL $\leq$ 1.3**	**HDL $>$ 1.3**	**Standardized**
	(n = 243)	**(n = 517)**	**Mean Difference**
Age (mean (SD))	73.73 (7.03)	72.80 (7.40)	0.128
Male (%)	199 (81.9)	256 (49.5)	0.726
Years Education (mean (SD))	15.86 (2.85)	16.04 (2.71)	0.064
BMI (mean (SD))	28.42 (4.69)	26.21 (4.64)	0.474
Family History (%)	123 (50.6)	394 (52.4)	0.036
Baseline CDR (mean (SD))	1.43 (0.91)	1.50 (0.88)	0.083
APOE4 copies (%)			0.054
0	122 (50.2)	314 (49.7)	
1	99 (40.7)	255 (39.9)	
2	22 (9.1)	86 (10.6)	
Neurological Condition (%)	94 (38.7)	166 (32.1)	0.132
Cardiovascular Condition (%)	194 (79.8)	327 (63.2)	0.374
Metabolic Condition (%)	116 (47.7)	208 (40.2)	0.175

group. Our methodology approximately replicates this protective point estimate with the SL point estimate being slightly higher at 0.415. Nevertheless, the confidence intervals are notably wider due in part to sparsity and a lower performance of our at smaller sample sizes due to flexibility. Indeed, in some strata, such as those who have low HDL, do not have two copies of *rs3764261*, and do not experience the event, we have as little as 24 participants to learn flexible models.

5.7 Discussion

In this work, we have a developed a nonparametric estimator of the LATE stemming from the AFT and BJ framework. Our estimator allows model-agnostic estimation of all nuisance

Table 5.6: Treatment effect estimate by methodology

Approach	**Estimate**	**SE (log)**	**95% Conf Int.**
Confounder BJ	0.996	0.072	(0.869, 1.39)
Complete-Case 2SPI	0.179	1.37	(0.012, 1.63)
2SPI + BJ	0.17	0.682	(0.045, 0.644)
No SL	0.255	1.83	(0.02, 9.34)
SL	0.415	2.54	(0.003, 59.0)

functions, including those involved the BJ procedure. Through our IF-based estimator and cross-ftting, we may avoid complicated Donsker conditions making inference straightforward with additional double-robustness properties with all nuisance functions. We furthermore developed a computational procedure for accounting for the extra variability due to having to impute Y^*. In our simulations, we showed that inference for the LATE with our approach performs better than 2SPI in a variety of scenarios, including a non-linear DGM and misspecification of key functions.

Traditionally, the LATE can be boiled down to essentially a ratio of coefficients of Z from two linear models representing the propensity score and outcome. On the other hand, nonparametric approaches such as ours decompose the LATE into a set of nuisance functions that we may flexibly estimate. In practice, this translates to dividing our original sample into many subpopulations to estimate specific nuisance functions like, for example, $m_{1,1}$, which can only be estimated for only uncensored individual who both have $Z = 1$ and $A = 1$. At lower sample sizes, we risk sparsity and overfitting of ML models, resulting in high variability and overcoverage, which was demonstrated in our simulation results at $n = 1000$ and the applied example.

Chapter 6

Conclusion and Future Directions

In this dissertation, we have enhanced the utility of IVs in three important scenarios. First, in the linear settings, we analytically quantified the relative trade-offs between the confounder approach and the IV approach in the presence of key assumption violations. These results combined with our sensitivity analysis tool can be utilized by analysts wishing to examine whether the IV approach will improve upon the traditional regression adjustment or propensity score approach in consistently estimating the ATE. Shifting to the assumption of a valid IV, we extend recent developments in the nonparametric identification literature to accommodate unknown sampling weights for the purposes of replicating causal effects. Through flexible and doubly robust estimation of the LATE, we are able to guard not against misspecification within a specific population but enhance replicability across populations. We additionally apply this methodology to providing weighted bounds on the ATE. Lastly, we move to the time-to-event setting, once again extending the nonparametric literature as well as the AFT literature to flexibly estimate the percentage difference in the median survival time among compliers with a binary exposure. With this approach, we are able to avoid many of the causal complications stemming from using IVs to estimate the causal HR and, further, improve upon linear AFT and BJ methods.

For Chapter 3, there are several additional avenues for future research. As previously mentioned, we focus only on consistency but in estimation, we may also want to know whether CAC may be more efficient than IVAC or vice-versa. Additionally, the confidence interval overlap of OLS and 2SLS estimates could also cause ambivalence between the methods. Another future direction lies in moving beyond the setting of a continuous exposure and outcome. Nevertheless, if one believes linear probability models (LPMs) are appropriate for the study's context, then one may extend our results to a binary exposure and outcome. It has been shown that if the probabilities produced by the LPM are not outside of the range of $[0, 1]$, or if the probabilities of exposure or outcome are not extreme in the population, then OLS and 2SLS may still give consistent results.[54, 13] For the sensitivity analysis procedure, the results are only as useful as variables available and chosen for benchmarking, which is partially mitigated by using the multipliers. Certainly, other benchmarking quantities for the unobserved quantities than the ones we chose could be used. The aggregation of the covariates into one general confounder W could be done via other methods besides PCA, which is limited when there are categorical variables. Using non-continuous variables is also limited when capturing associations using R^2 due to the Frechet bounds on the correlation between non-continuous variables being potentially far narrower than $[-1, 1]$.[53]

There are a few future directions vis a vis Chapter 4. The ATE bounds could be further narrowed by incorporating sampling weights with developing work on "covariate-assisted" bounds.[71]. Despite double robustness on most nuisance functions, the survey weights do not share this property. Undoubtedly, the sensitivity to misspecification of the weights depends on the target estimand. Thus, future work could not only extend our work to estimands such as the local average treatment effect on the treated (LATT)[42] but furthermore assess scenarios where replicability may be more successful in one estimand compared to another. These results could be used to derive an "optimally weighted" estimator for replicability. In the IV literature, key conditions like monotonicity have been relaxed through similar approaches, trading off "localness" of the estimand for validity.[56] Finally, future work might consider extensions to binary outcomes, continuous IVs, and time-varying treatments.

For Chapter 5, the impacts of sparsity in finite samples is ultimately a limitation to our nonparametric approach and a trade-off for protecting against misspecification. These impacts can be partially alleviated if we need not condition on the treatment assignment for ignorable censoring as we effectively halve the number of expectations to estimate in the numerator. Another related aspect of our approach is that since we are training models directly on the censored subpopulation, we require a sufficient number of censored observations. In practice, censoring usually constitutes a significant proportion of the (e.g. 40%), which our estimator performs well on even at lower sample sizes. Future work may focus on maximizing sample size by effectively borrowing information across different subgroups to estimate nuisance functions but still remaining model agnostic. Furthermore, stratified sample-splitting and bootstrapping approaches could be investigated. Ultimately, analysts should ensure their data is able to sufficiently accommodate a nonparametric, flexible estimation approach before employing this general class of methods, not just ours.

In terms of inference, one challenge is properly accounting for the variability stemming from training nuisance functions on imputed outcomes such as what we did Y^*. While our bootstrap variance estimator captured most of this variability, it demonstrated bias when the censoring rate was high. As discussed earlier, a higher-order IF estimator could be utilized to potentially derive closed-form estimator, yet it is not known whether we may remain model-agnostic or if such a solution exists.[11] Lastly, we may extend our methodology to nonparametrically estimate several other causal quantities that involves time-to-event outcomes such as mediation effects with accommodation of continuous IVs and exposures.

Bibliography

[1] eTD Explore.

[2] A. Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, Apr. 2003.

[3] J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455, June 1996. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1996.10476902.

[4] J. D. Angrist and A. B. Krueger. Split-Sample Instrumental Variables Estimates of the Return to Schooling. *Journal of Business & Economic Statistics*, 13(2):225–235, Apr. 1995. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/07350015.1995.10524597.

[5] J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 1 edition, Jan. 2009.

[6] P. M. Aronow and A. Carnegie. Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable. *Political Analysis*, 21(4):492–506, 2013. Publisher: Cambridge University Press.

[7] P. M. Aronow and C. Samii. Does Regression Produce Representative Estimates of Causal Effects? *American Journal of Political Science*, 60(1):250–267, 2016. Publisher: [Midwest Political Science Association, Wiley].

[8] P. C. Austin. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, 33(7):1242–1258, Mar. 2014.

[9] P. C. Austin and E. A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28):3661–3679, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6607.

[10] M. Baiocchi, J. Cheng, and D. S. Small. Tutorial in Biostatistics: Instrumental Variable Methods for Causal Inference*. *Statistics in medicine*, 33(13):2297–2340, June 2014.

[11] S. Balakrishnan, E. H. Kennedy, and L. Wasserman. The Fundamental Limits of Structure-Agnostic Functional Estimation, May 2023.

[12] A. Balke and J. Pearl. Bounds on Treatment Effects From Studies With Imperfect Compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[13] A. Basu, N. B. Coe, and C. G. Chapman. 2SLS versus 2SRI: Appropriate methods for rare outcomes and/or rare exposures. *Health Economics*, 27(6):937–955, June 2018.

[14] C. G. Begley and L. M. Ellis. Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, Mar. 2012. Number: 7391 Publisher: Nature Publishing Group.

[15] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica*, 80(6):2369–2429, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA9626.

[16] O. M. Bernstein, B. G. Vegetabile, C. R. Salazar, J. D. Grill, and D. L. Gillen. Adjustment for biased sampling using NHANES derived propensity weights. *Health Services and Outcomes Research Methodology*, 23(1):21–44, Mar. 2023.

[17] J. Bhattacharya and W. Vogt. Do Instrumental Variables Belong in Propensity Scores? Technical Report t0343, National Bureau of Economic Research, Cambridge, MA, Sept. 2007.

[18] C. Blandhol, J. Bonney, M. Mogstad, and A. Torgovitsky. When is TSLS Actually LATE? *SSRN*, page 69, Feb. 2022.

[19] J. Bound, D. A. Jaeger, and R. M. Baker. Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, 90(430):443–450, 1995. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[20] M. Brookhart, T. Stürmer, R. Glynn, J. Rassen, and S. Schneeweiss. Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, 48(6 0):S114–S120, June 2010.

[21] M. A. Brookhart and S. Schneeweiss. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The international journal of biostatistics*, 3(1):14, 2007.

[22] M. A. Brookhart, S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. Variable Selection for Propensity Score Models. *American Journal of Epidemiology*, 163(12):1149–1156, June 2006.

[23] J. M. Brooks and R. L. Ohsfeldt. Squeezing the balloon: propensity scores and unmeasured covariate balance. *Health Services Research*, 48(4):1487–1507, Aug. 2013.

[24] J. BUCKLEY and I. JAMES. Linear regression with censored data. *Biometrika*, 66(3):429–436, Dec. 1979.

[25] D. Card. Using Geographic Variation in College Proximity to Estimate the Return to Schooling, Oct. 1993.

[26] C. G. Chapman and J. M. Brooks. Treatment Effect Estimation Using Nonlinear Two-Stage Instrumental Variable Estimators: Another Cautionary Note. *Health Services Research*, 51(6):2375–2394, Dec. 2016.

[27] Y. Chen and B. A. Briesacher. Use of Instrumental Variable in Prescription Drug Research with Observational Data: A Systematic Review. *Journal of clinical epidemiology*, 64(6):687–700, June 2011.

[28] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, Feb. 2018.

[29] V. Chernozhukov, C. Hansen, and M. Spindler. Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments. *American Economic Review*, 105(5):486–490, May 2015.

[30] B. Y. Choi. Instrumental variable estimation of weighted local average treatment effects. *Statistical Papers*, Mar. 2023.

[31] C. Cinelli and C. Hazlett. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12348.

[32] C. Cinelli and C. Hazlett. An Omitted Variable Bias Framework for Sensitivity Analysis of Instrumental Variables, Sept. 2022.

[33] S. R. Cole and M. A. Hernán. Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine*, 75(1):45–49, July 2004.

[34] S. R. Cole and E. A. Stuart. Generalizing Evidence From Randomized Clinical Trials to Target Populations. *American Journal of Epidemiology*, 172(1):107–115, July 2010.

[35] A. N. Corallo, R. Croxford, D. C. Goodman, E. L. Bryan, D. Srivastava, and T. A. Stukel. A systematic review of medical practice variation in OECD countries. *Health Policy*, 114(1):5–14, Jan. 2014.

[36] N. M. Davies, K. H. Thomas, A. E. Taylor, G. M. Taylor, R. M. Martin, M. R. Munafò, and F. Windmeijer. How to compare instrumental variable and conventional regression analyses using negative controls and bias plots. *International Journal of Epidemiology*, 46(6):2067–2077, Dec. 2017.

[37] P. Ding, T. Vanderweele, and J. M. Robins. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, 104(2):291–302, June 2017.

[38] M. H. Farrell, T. Liang, and S. Misra. Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1):181–213, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA16901.

[39] M. P. Fay and F. Li. Causal interpretation of the hazard ratio in randomized clinical trials. *Clinical Trials*, page 17407745241243308, Apr. 2024. Publisher: SAGE Publications.

[40] A. Fisher and E. H. Kennedy. Visually Communicating and Teaching Intuition for Influence Functions. *The American Statistician*, 75(2):162–172, May 2021. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2020.1717620.

[41] M. Frölich. Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75, July 2007.

[42] M. Frölich and B. Melly. Identification of Treatment Effects on the Treated with One-Sided Non-Compliance. *Econometric Reviews*, Mar. 2013. Publisher: Taylor & Francis Group.

[43] S. Greenland, J. Pearl, and J. M. Robins. Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29–46, Feb. 1999. Publisher: Institute of Mathematical Statistics.

[44] P. Guggenberger. ON THE ASYMPTOTIC SIZE DISTORTION OF TESTS WHEN INSTRUMENTS LOCALLY VIOLATE THE EXOGENEITY ASSUMPTION. *Econometric Theory*, 28(2):387–421, 2012. Publisher: Cambridge University Press.

[45] J. Hadley, K. R. Yabroff, M. J. Barrett, D. F. Penson, C. S. Saigal, and A. L. Potosky. Comparative Effectiveness of Prostate Cancer Treatments: Evaluating Statistical Adjustments for Confounding in Observational Data. *JNCI Journal of the National Cancer Institute*, 102(23):1780–1793, Dec. 2010.

[46] F. P. Hartwig, L. Wang, G. Davey Smith, and N. M. Davies. Average Causal Effect Estimation Via Instrumental Variables: the No Simultaneous Heterogeneity Assumption. *Epidemiology*, 34(3):325, May 2023.

[47] F. P. Hartwig, L. Wang, G. D. Smith, and N. M. Davies. Average causal effect estimation via instrumental variables: the no simultaneous heterogeneity assumption. Oct. 2020.

[48] F. P. Hartwig, L. Wang, G. D. Smith, and N. M. Davies. Average causal effect estimation via instrumental variables: the no simultaneous heterogeneity assumption, Sept. 2022. arXiv:2010.10017 [stat].

[49] M. A. Hernán. The Hazards of Hazard Ratios. *Epidemiology (Cambridge, Mass.)*, 21(1):13–15, Jan. 2010.

[50] M. A. Hernán. The Hazards of Hazard Ratios. *Epidemiology (Cambridge, Mass.)*, 21(1):13–15, Jan. 2010.

[51] M. A. Hernán, B. Brumback, and J. M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology (Cambridge, Mass.)*, 11(5):561–570, Sept. 2000.

[52] K. Hirano and G. W. Imbens. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology*, 2(3):259–278, Dec. 2001.

[53] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[54] W. C. Horrace and R. L. Oaxaca. Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3):321–327, Mar. 2006.

[55] J. D. Huling, M. Yu, and A. J. O'Malley. Instrumental variable based estimation under the semiparametric accelerated failure time model. *Biometrics*, 75(2):516–527, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12985.

[56] N. Huntington-Klein. Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness. *Journal of Causal Inference*, 8(1):182–208, Jan. 2020. Publisher: De Gruyter.

[57] K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12027.

[58] K. Imai and D. A. van Dyk. Causal Inference With General Treatment Regimes. *Journal of the American Statistical Association*, 99(467):854–866, Sept. 2004. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214504000001187.

[59] G. Imbens and K. Hirano. The Propensity Score with Continuous Treatments. 2004.

[60] G. W. Imbens. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48(2):399–423, June 2010.

[61] G. W. Imbens and J. D. Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475, 1994. Publisher: [Wiley, Econometric Society].

[62] J. P. A. Ioannidis. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *JAMA*, 294(2):218–228, July 2005.

[63] J. P. A. Ioannidis. Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8):e124, Aug. 2005. Publisher: Public Library of Science.

[64] K. B. Karlson, F. Popham, and A. Holm. Marginal and Conditional Confounding Using Logits. *Sociological Methods & Research*, page 0049124121995548, Apr. 2021. Publisher: SAGE Publications Inc.

[65] E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review, Jan. 2023. arXiv:2203.06469 [stat].

[66] E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*, 48(4):2008–2030, Aug. 2020. Publisher: Institute of Mathematical Statistics.

[67] T. W. Kinal. The Existence of Moments of k-Class Estimators. *Econometrica*, 48(1):241–249, 1980. Publisher: [Wiley, Econometric Society].

[68] J. F. Kiviet and S. Kripfganz. Instrument approval by the Sargan test and its consequences for coefficient estimation. *Economics Letters*, 205:109935, Aug. 2021.

[69] M. Kivipelto, E.-L. Helkala, M. P. Laakso, T. Hänninen, M. Hallikainen, K. Alhainen, S. Iivonen, A. Mannermaa, J. Tuomilehto, A. Nissinen, and H. Soininen. Apolipoprotein E e4 Allele, Elevated Midlife Total Cholesterol Level, and High Midlife Systolic Blood Pressure Are Independent Risk Factors for Late-Life Alzheimer Disease. *Annals of Internal Medicine*, 137(3):149–155, Aug. 2002. Publisher: American College of Physicians.

[70] Y. Lee, E. H. Kennedy, and N. Mitra. Doubly robust nonparametric instrumental variable estimators for survival outcomes. *Biostatistics*, 24(2):518–537, Apr. 2023.

[71] A. W. Levis, M. Bonvini, Z. Zeng, L. Keele, and E. H. Kennedy. Covariate-assisted bounds on causal effects with instrumental variables, Sept. 2023. arXiv:2301.12106 [stat].

[72] M. C. Lovell. A Simple Proof of the FWL Theorem. *The Journal of Economic Education*, 39(1):88–91, Jan. 2008. Publisher: Routledge _eprint: https://doi.org/10.3200/JECE.39.1.88-91.

[73] C. F. Manski. Nonparametric Bounds on Treatment Effects. *The American Economic Review*, 80(2):319–323, 1990. Publisher: American Economic Association.

[74] H. Mao, L. Li, and T. Greene. Propensity score weighting analysis and treatment effect discovery. *Statistical Methods in Medical Research*, 28(8):2439–2454, Aug. 2019. Publisher: SAGE Publications Ltd STM.

[75] T. Martinussen, D. Nørbo Sørensen, and S. Vansteelandt. Instrumental variables estimation under a structural Cox model. *Biostatistics*, 20(1):65–79, Jan. 2019.

[76] P. Martínez-Camblor, T. Mackenzie, D. O. Staiger, P. P. Goodney, and A. J. O'Malley. Adjusting for bias introduced by instrumental variable estimation in the Cox proportional hazards model. *Biostatistics*, 20(1):80–96, Jan. 2019.

[77] P. Martínez-Camblor, T. A. MacKenzie, D. O. Staiger, P. P. Goodney, and A. James O'Malley. An Instrumental Variable Procedure for Estimating Cox Models with Non-Proportional Hazards in the Presence Of Unmeasured Confounding. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 68(4):985–1005, Aug. 2019.

[78] M. B. Mathur and M. P. Fox. Toward Open and Reproducible Epidemiology. *American Journal of Epidemiology*, 192(4):658–664, Apr. 2023.

[79] M. Mielke, P. Zandi, H. Shao, M. Waern, S. Östling, X. Guo, C. Björkelund, L. Lissner, I. Skoog, and D. Gustafson. The 32-year relationship between cholesterol and dementia from midlife to late life. *Neurology*, 75(21):1888–1895, Nov. 2010. Publisher: Wolters Kluwer.

[80] C. Mood. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review*, 26(1):67–82, Feb. 2010.

[81] Muhammad Atiyat. Instrumental Variable Modeling in a Survival Analysis Framework. June 2011.

[82] S. E. O'Bryant, S. C. Waring, C. M. Cullum, J. Hall, L. Lacritz, P. J. Massman, P. J. Lupo, J. S. Reisch, and R. Doody. Staging Dementia Using Clinical Dementia Rating Scale Sum of Boxes Scores. *Archives of neurology*, 65(8):1091–1095, Aug. 2008.

[83] M. Pang, R. W. Platt, T. Schuster, and M. Abrahamowicz. Flexible extension of the accelerated failure time model to account for nonlinear and time-dependent effects of covariates on the hazard. *Statistical Methods in Medical Research*, 30(11):2526–2542, Nov. 2021. Publisher: SAGE Publications Ltd STM.

[84] H. Pashler and C. R. Harris. Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, 7(6):531–536, Nov. 2012. Publisher: SAGE Publications Inc.

[85] J. PEARL. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, Dec. 1995.

[86] J. Pearl. *Causality*. Cambridge University Press, Cambridge, 2 edition, 2009.

[87] J. Pearl. On a Class of Bias-Amplifying Variables that Endanger Effect Estimates, 2010.

[88] J. Pearl. On a Class of Bias-Amplifying Variables that Endanger Effect Estimates. *arXiv:1203.3503 [cs, stat]*, Mar. 2012. arXiv: 1203.3503.

[89] J. Pearl and A. Paz. Confounding Equivalence in Causal Inference. *Journal of Causal Inference*, 2(1):75–93, Mar. 2014. Publisher: De Gruyter.

[90] M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, Feb. 2012.

[91] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner. Alzheimer's Disease Neuroimaging Initiative (ADNI). *Neurology*, 74(3):201–209, Jan. 2010.

[92] E. Polley, E. LeDell, C. Kennedy, S. Lendle, and M. v. d. Laan. SuperLearner: Super Learner Prediction, Feb. 2024.

[93] P. Proitsi, M. K. Lupton, L. Velayudhan, S. Newhouse, I. Fogh, M. Tsolaki, M. Daniilidou, M. Pritchard, I. Kloszewska, H. Soininen, P. Mecocci, B. Vellas, f. t. A. D. N. Initiative, J. Williams, f. t. G. Consortium, R. Stewart, P. Sham, S. Lovestone, and J. F. Powell. Genetic Predisposition to Increased Blood Cholesterol and Triglyceride Lipid Levels and Risk of Alzheimer Disease: A Mendelian Randomization Analysis. *PLOS Medicine*, 11(9):e1001713, Sept. 2014. Publisher: Public Library of Science.

[94] G. Ridgeway, S. A. Kovalchik, B. A. Griffin, and M. U. Kabeto. Propensity Score Analysis with Survey Weighted Data. *Journal of causal inference*, 3(2):237–249, Sept. 2015.

[95] J. Robins, L. Li, E. Tchetgen, and A. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman*, 2:335–422, Jan. 2008.

[96] J. M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412, Jan. 1994. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610929408831393.

[97] J. M. Robins, M. A. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, 11(5):550–560, Sept. 2000.

[98] P. M. Robinson. Root-N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931–954, 1988. Publisher: [Wiley, Econometric Society].

[99] D. Roodman. A Note on the Theme of Too Many Instruments*. *Oxford Bulletin of Economics and Statistics*, 71(1):135–158, 2009. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0084.2008.00542.x.

[100] P. R. ROSENBAUM and D. B. RUBIN. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, Apr. 1983.

[101] D. B. Rubin. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[102] E. Sanderson, M. M. Glymour, M. V. Holmes, H. Kang, J. Morrison, M. R. Munafò, T. Palmer, C. M. Schooling, C. Wallace, Q. Zhao, and G. Davey Smith. Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):1–21, Feb. 2022. Number: 1 Publisher: Nature Publishing Group.

[103] J. D. Sargan. The Estimation of Economic Relationships using Instrumental Variables. *Econometrica*, 26(3):393–415, 1958. Publisher: [Wiley, Econometric Society].

[104] E. F. Schisterman, S. R. Cole, and R. W. Platt. Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. *Epidemiology*, 20(4):488–495, July 2009.

[105] N. A. Schuster, J. W. R. Twisk, G. ter Riet, M. W. Heymans, and J. J. M. Rijnhart. Noncollapsibility and its role in quantifying confounding bias in logistic regression. *BMC Medical Research Methodology*, 21(1):136, July 2021.

[106] A. Solomon, M. Kivipelto, B. Wolozin, J. Zhou, and R. A. Whitmer. Midlife Serum Cholesterol and Increased Risk of Alzheimer's and Vascular Dementia Three Decades Later. *Dementia and Geriatric Cognitive Disorders*, 28(1):75–80, Aug. 2009.

[107] T. Stokes, R. Steele, and I. Shrier. Causal simulation experiments: Lessons from bias amplification. *Statistical Methods in Medical Research*, 31(1):3–46, Jan. 2022. Publisher: SAGE Publications Ltd STM.

[108] K. Takatsu, A. W. Levis, E. Kennedy, R. Kelz, and L. Keele. Doubly robust machine learning for an instrumental variable study of surgical care for cholecystitis, July 2023. arXiv:2307.06269 [stat].

[109] H.-J. Tan, E. C. Norton, Z. Ye, K. S. Hafez, J. L. Gore, and D. C. Miller. Long-term survival following partial versus radical nephrectomy among older patients with early-stage kidney cancer. *JAMA : the journal of the American Medical Association*, 307(15):10.1001/jama.2012.475, Apr. 2012.

[110] Z. S. Tan, S. Seshadri, A. Beiser, P. W. F. Wilson, D. P. Kiel, M. Tocco, R. B. D'Agostino, and P. A. Wolf. Plasma Total Cholesterol Level as a Risk Factor for Alzheimer Disease: The Framingham Study. *Archives of Internal Medicine*, 163(9):1053–1057, May 2003.

[111] E. J. Tchetgen Tchetgen, S. Walter, S. Vansteelandt, T. Martinussen, and M. Glymour. Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.)*, 26(3):402–410, May 2015.

[112] J. V. Terza, A. Basu, and P. J. Rathouz. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3):531–543, May 2008.

[113] A. A. Tsiatis. Estimating Regression Parameters Using Linear Rank Tests for Censored Data. *The Annals of Statistics*, 18(1):354–372, Mar. 1990. Publisher: Institute of Mathematical Statistics.

[114] M. J. van der Laan. Targeted Maximum Likelihood Based Causal Inference: Part I. *The International Journal of Biostatistics*, 6(2):2, Feb. 2010.

[115] T. J. VanderWeele and P. Ding. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*, 167(4):268–274, Aug. 2017.

[116] T. J. VanderWeele and I. Shpitser. On the definition of a confounder. *Annals of statistics*, 41(1):196–220, Feb. 2013.

[117] T. J. VanderWeele, E. J. Tchetgen Tchetgen, M. Cornelis, and P. Kraft. Methodological challenges in Mendelian randomization. *Epidemiology (Cambridge, Mass.)*, 25(3):427–435, May 2014.

[118] S. Vansteelandt and N. Keiding. Invited Commentary: G-Computation–Lost in Translation? *American Journal of Epidemiology*, 173(7):739–742, Apr. 2011.

[119] B. G. Vegetabile, D. L. Gillen, and H. S. Stern. Optimally Balanced Gaussian Process Propensity Scores for Estimating Treatment Effects. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 183(1):355–377, Jan. 2020.

[120] F. Wan, D. Small, J. E. Bekelman, and N. Mitra. Bias in estimating the causal hazard ratio when using two-stage instrumental variable methods. *Statistics in medicine*, 34(14):2235–2265, June 2015.

[121] F. Wan, D. Small, and N. Mitra. A general approach to evaluating the bias of 2-stage instrumental variable estimators. *Statistics in Medicine*, 37(12):1997–2015, May 2018.

[122] L. Wang and E. Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 80(3):531–550, June 2018.

[123] L. Wang, E. Tchetgen Tchetgen, T. Martinussen, and S. Vansteelandt. Instrumental variable estimation of the causal hazard ratio. *Biometrics*, 79(2):539–550, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13792.

[124] Z. Wang and C. Wang. Buckley-James Boosting for Survival Analysis with High-Dimensional Biomarker Data. *Statistical Applications in Genetics and Molecular Biology*, 9(1):24, June 2010.

[125] L. J. Wei. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879, 1992. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780111409.

[126] S. Weitzen, K. L. Lapane, A. Y. Toledano, A. L. Hume, and V. Mor. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiology and Drug Safety*, 14(4):227–238, 2005. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pds.986.

[127] J. Wennberg and n. Gittelsohn. Small area variations in health care delivery. *Science (New York, N.Y.)*, 182(4117):1102–1108, Dec. 1973.

[128] J. E. Wennberg. Dealing With Medical Practice Variations: A Proposal for Action. *Health Affairs*, 3(2):6–33, Jan. 1984. Publisher: Health Affairs.

[129] D. Westreich and S. R. Cole. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, 171(6):674–677, Mar. 2010.

[130] C. J. Willer, E. M. Schmidt, S. Sengupta, M. L. Buchkovich, S. Mora, Jacques S. Beckmann, J. L. Bragg-Gresham, H.-Y. Chang, P. Fontanillas, R. M. Fraser, D. F. Freitag, D. Gurdasani, K. Heikkilä, E. Hyppönen, A. Isaacs, A. U. Jackson, Johansson, M. Kaakinen, J. Kettunen, M. E. Kleber, X. Li, J. Luan, L.-P. Lyytikäinen, P. K. E. Magnusson, E. Mihailov, M. E. Montasser, M. Müller-Nurasyid, I. M. Nolte, J. R. O'Connell, C. D. Palmer, M. Perola, A.-K. Petersen, S. Sanna, R. Saxena, S. K. Service, S. Shah, D. Shungin, C. Sidore, C. Song, R. J. Strawbridge, I. Surakka, T. Tanaka, T. M. Teslovich, G. Thorleifsson, E. G. Van den Herik, B. F. Voight, K. A. Volcik, L. L. Waite, A. Wong, Y. Wu, W. Zhang, D. Absher, G. Asiki, I. Barroso, L. F. Been, J. L. Bolton, L. L. Bonnycastle, P. Brambilla, M. S. Burnett, G. Cesana, M. Dimitriou, A. Döring, P. Elliott, S. E. Epstein, G. Ingi Eyjolfsson, B. Gigante, H. Grallert, M. L. Gravito, C. J. Groves, G. Hallmans, A.-L. Hartikainen, C. Hayward, D. Hernandez, A. A. Hicks, H. Holm, Y.-J. Hung, T. Illig, M. R. Jones, P. Kaleebu, K.-T. Khaw, E. Kim, N. Klopp, P. Komulainen, M. Kumari, C. Langenberg, T. Lehtimäki, S.-Y. Lin, J. Lindström, R. J. F. Loos, F. Mach, W. L. McArdle, C. Meisinger, B. D. Mitchell, G. Müller, R. Nagaraja, N. Narisu, T. V. M. Nieminen, R. N. Nsubuga, I. Olafsson, K. K. Ong, A. Palotie, T. Papamarkou, C. Pomilla, A. Pouta, D. J. Rader, M. P. Reilly, P. M. Ridker, F. Rivadeneira, I. Rudan, A. Ruokonen, N. Samani, H. Scharnagl, J. Seeley, K. Silander, A. Stančáková, K. Stirrups, A. J. Swift, L. Tiret, A. G. Uitterlinden, L. J. van Pelt, S. Vedantam, N. Wainwright, C. Wijmenga, S. H. Wild, G. Willemsen, T. Wilsgaard, J. F. Wilson, E. H. Young, J. H. Zhao, L. S. Adair, D. Arveiler, T. L. Assimes, S. Bandinelli, F. Bennett, M. Bochud, B. O. Boehm, D. I. Boomsma, I. B. Borecki, S. R. Bornstein, P. Bovet, M. Burnier, H. Campbell, A. Chakravarti, J. C. Chambers, Y.-D. I. Chen, F. S. Collins, R. S. Cooper, J. Danesh, G. Dedoussis, U. de Faire, A. B. Feranil, J. Ferrières, L. Ferrucci, N. B. Freimer, C. Gieger, V. Gudnason, U. Gyllensten, A. Hamsten, T. B. Harris, A. Hingorani, J. N. Hirschhorn, A. Hofman, G. K. Hovingh, C. A. Hsiung, S. E. Humphries, S. C. Hunt, C. Iribarren, M.-R. Järvelin, A. Jula, M. Kähönen, J. Kaprio, A. Kesäniemi, M. Kivimaki, J. S. Kooner, P. J. Koudstaal, R. M. Krauss, D. Kuh, J. Kuusisto, K. O. Kyvik, M. Laakso, T. A. Lakka, L. Lind, C. M. Lindgren, N. G. Martin, W. März, M. I. McCarthy, C. A. McKenzie, P. Meneton, A. Metspalu, L. Moilanen, A. D. Morris, P. B. Munroe, I. Njølstad, N. L. Pedersen, C. Power, P. P. Pramstaller, J. F. Price, B. M. Psaty, T. Quertermous, R. Rauramaa, D. Saleheen, V. Salomaa, D. K. Sanghera, J. Saramies, P. E. H. Schwarz, W. H.-H. Sheu, A. R. Shuldiner, A. Siegbahn, T. D. Spector, K. Stefansson, D. P. Strachan, B. O. Tayo, E. Tremoli, J. Tuomilehto, M. Uusitupa, C. M. van Duijn, P. Vollenweider, L. Wallentin, N. J. Wareham, J. B. Whitfield, B. H. R. Wolffenbuttel, J. M. Ordovas, E. Boerwinkle, C. N. A. Palmer, U. Thorsteinsdottir, D. I. Chasman, J. I. Rotter, P. W. Franks, S. Ripatti, L. A. Cupples, M. S. Sandhu, S. S. Rich, M. Boehnke, P. Deloukas, S. Kathiresan, K. L. Mohlke, E. Ingelsson, and G. R. Abecasis. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45(11):1274–1283, Nov. 2013.

[131] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, USA, 2 edition, Oct. 2010.

[132] J. M. Wooldridge. Should instrumental variables be used as matching variables?

Research in Economics, 70(2):232–237, June 2016.

[133] S. Wright. The Method of Path Coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, Sept. 1934. Publisher: Institute of Mathematical Statistics.

[134] A. Young. Consistency without Inference: Instrumental Variables in Practical Application. *European Economic Review*, 147:104112, Aug. 2022.

[135] R. S. Zawadzki, J. D. Grill, D. L. Gillen, and and for the Alzheimer's Disease Neuroimaging Initiative. Frameworks for estimating causal effects in observational settings: comparing confounder adjustment and instrumental variables. *BMC Medical Research Methodology*, 23(1):122, May 2023.

Appendix A

Appendix

A.1 Appendices for Chapter 3

A.1.1 Proof for the Consistency $\hat{A}_4$ under Figure 3.1

Proof. We begin by defining following system of structural equations under Figure 3.1 where $E[\epsilon_1|U, Z] = 0$ and $E[\epsilon_2|X, U] = 0$.

$$X = \alpha_1 U + \alpha_2 Z + \epsilon_1 \tag{A.1}$$

$$Y = \beta_1 X + \beta_2 U + \epsilon_2 \tag{A.2}$$

By the standard definition of the IV estimator, $\hat{A}_4 = \frac{\widehat{Cov}(Y,Z)}{\widehat{Cov}(X,Z)}$. Where n is the sample size, by finite variance and Slutsky's theorem we can write:

$$\frac{\widehat{Cov}(Y,Z)}{\widehat{Cov}(X,Z)} = \frac{n^{-1}\sum_{i=1}^{n} y_i z_i}{n^{-1}\sum_{i=1}^{n} x_i z_i} \xrightarrow{p} \frac{E[YZ]}{E[XZ]} = \frac{Cov(Y,Z)}{Cov(X,Z)} \tag{A.3}$$

Clearly, $Cov(X,Z) = \alpha_2 = c_3$ because $Z \perp\!\!\!\perp U$. Moving onto $Cov(Y,Z)$ we write:

$$E[Y|z] = \sum_u E[Y|z,u]P(u|z) \tag{A.4}$$

$$= \sum_u [c_0 c_3 z + c_2 u]P(u) \tag{A.5}$$

$$= c_0 c_3 z \sum_u u + c_2 \sum_u uP(u) \tag{A.6}$$

$$= c_0 c_3 z \tag{A.7}$$

□

A.1.2 Proof for Proposition 3.3.1

Proof. Using an equivalent definition of the IV estimator, $\hat{A}_4 = (Z^T X)^{-1}(Z^T Y)$ we can substitute 3.9 for Y

$$(Z^T X)^{-1}(Z^T Y) = (Z^T X)^{-1}(Z^T(c_0 X + c_2 U + c_{ER} Z + \epsilon_3)) \tag{A.8}$$

$$= c_0 + c_2 (Z^T X)^{-1} Z^T U + c_{ER}(Z^T X)^{-1} Z^T Z + (Z^T X)^{-1} Z^T \epsilon_3 \tag{A.9}$$

$$= c_0 + c_2 \frac{n^{-1}\sum_{i=1}^{n} z_i u_i}{n^{-1}\sum_{i=1}^{n} z_i x_i} + c_{ER} \frac{n^{-1}\sum_{i=1}^{n} z_i^2}{n^{-1}\sum_{i=1}^{n} z_i x_i} + \frac{n^{-1}\sum_{i=1}^{n} \epsilon_{3i} x_i}{n^{-1}\sum_{i=1}^{n} z_i x_i} \tag{A.10}$$

$$\xrightarrow{p} c_0 + c_2\frac{Cov(Z,U)}{Cov(Z,X)} + c_{ER}\frac{Var(Z)}{Cov(Z,X)} + \frac{Cov(Z,\epsilon_3)}{Cov(Z,X} \tag{A.11}$$

$$= c_0 + \frac{c_{ER}}{c_3} \tag{A.12}$$

$$(Z \perp\!\!\!\perp U \text{ and } Cov(X,Z) = c_3) \tag{A.13}$$

The last line follows from the fact that $Z \perp\!\!\!\perp U$ and $E[\epsilon_3 Z] = 0$ by iterated expectation. □

A.1.3 Proofs for Proposition 3.3.2

Proof for the Consistency of $\hat{A}_3$

Proof. We need to find the value of $\frac{\partial}{\partial x}E[Y|x,z] = c_0 + c_2\frac{\partial}{\partial x}E[U|x,z]$. We will use the FWL theorem to find $\frac{\partial}{\partial x}E[U|x,z]$ by orthogonalizing Z and using consistency. Assuming no intercept, we can write the quantity as $\frac{X^TM_ZU}{X^TM_ZX}$ where $M_Z = I - Z^T(Z^TZ)^{-1}Z^T$ (residual making matrix). Noting that $Cov(X,Z) = c_3 + c_1c_I$ and $Cov(U,Z) = c_I$, we have:

$$\frac{X^TM_ZU}{X^TM_ZX} = \frac{\sum_{i=1}^n [x_i - (c_3 + c_1c_I)z_i][u_i - c_Iz_i]}{\sum_{i=1}^n [x_i - (c_3 + c_1c_I)z_i]^2} \tag{A.14}$$

$$\xrightarrow{p} \frac{Cov[X - (c_3 + c_1c_I)Z, U - c_IZ]}{Var[X - (c_3 + c_1c_I)Z]} \tag{A.15}$$

$$= \frac{c_1 + c_3c_I - c_I(c_3 + c_1c_I)}{1 - (c_3 + c_1c_I)^2} = \frac{c_1(1 - c_I^2)}{1 - (c_3 + c_1c_I)^2} \tag{A.16}$$

□

Proof for the Consistency of $\hat{A}_4$

Proof. In a similar derivation Proposition 3.1 but plugging in $c_0X + c_2U + \epsilon_3$ for the structural equation of Y, we obtain

$$\hat{A}_4 \xrightarrow{p} c_0 + c_2\frac{Cov(Z,U)}{Cov(Z,X)} + \frac{Cov(Z,\epsilon_3)}{Cov(Z,X)} \tag{A.17}$$

$$= c_0 + c_2\frac{Cov(Z,U)}{Cov(Z,X)} \tag{A.18}$$

$$= c_0 + \frac{c_2c_I}{c_3 + c_1c_I} \tag{A.19}$$

Note that the unconditional association between X and Z goes has two paths: the direct path $X \rightarrow Z$ and the indirect path $X \leftarrow U \rightarrow Z$. □

A.1.4 Proofs for Proposition 3.3.3

Proof for the consistency of $\hat{A}_2$

Proof. We simply will compute the regression of Y on X where $\hat{A}_2 \xrightarrow{p} Cov(Y,X)$.

$$Cov(Y,X) = Cov(\beta_1X + \beta_2U + \beta_3XU + \epsilon_2, X) = \beta_1 + \beta_2\alpha_2 + \beta_3Cov(XU,X) \tag{A.20}$$

$$= \beta_1 + \alpha_2\beta_2 + 2\alpha_1\alpha_3\beta_3 \tag{A.21}$$

Where $Cov(XU,X) = 2\alpha_1\alpha_3$ because

$$Cov(XU,X) = Cov(U(\alpha_1Z + \alpha_2U + \alpha_3ZU + \epsilon_1), X) \tag{A.22}$$

$$= \alpha_1Cov(UZ,X) + \alpha_2Cov(U^2,X) + \alpha_3Cov(ZU^2,X) \tag{A.23}$$

$$= \underbrace{\alpha_1E[UZX]}_{B_1} + \underbrace{\alpha_2E[U^2X]}_{B_2} + \underbrace{\alpha_3E[ZU^2X]}_{B_3} \tag{A.24}$$

$B_1 = \alpha_1\alpha_3$ because

$$E[XUZ] = E[ZE[XU|Z]] = E[ZE[(\alpha_1 Z + \alpha_2 U + \alpha_3 ZU + \epsilon_1)U|Z]] \tag{A.25}$$

$$= E[Z(\alpha_1 E[UZ|Z] + \alpha_2 E[U^2|Z] + \alpha_3 E[ZU^{|}Z] + E[U\epsilon_1|Z)] \tag{A.26}$$

$$= E[\alpha_1 Z^2 E[U] + \alpha_2 ZE[U^2] + \alpha_3 ZE[U^2] + E[U\epsilon_1]] \tag{A.27}$$

$$= E[\alpha_2 Z + \alpha_3 Z^2] = \alpha_3 \tag{A.28}$$

$B_2 = 0$ because $E[U^2X] = E[U^2E[X|U]] = \alpha_2 E[U^3] = 0$ because $E[U^3] = 0$.

$B_3 = \alpha_1\alpha_3$ because

$$E[ZU^2X] = E[ZU^2(\alpha_1 Z + \alpha_2 U + \alpha_3 ZU + \epsilon_1)] \tag{A.29}$$

$$= \alpha_1 E[Z^2U^2] + \alpha_2 E[ZU^3] + \alpha_3 E[Z^2U^3] \tag{A.30}$$

$$= \alpha_1 \tag{A.31}$$

Where the last equality follows because $Z \perp\!\!\!\perp U$ and $E[U^3] = 0$. □

Proof for the consistency of $\hat{A}_3$

Proof. Similar to proposition 3.2, we will proceed via FWL to find the value of $A_3 = \frac{X^T M_Z Y}{X^T M_Z X}$. First, we need to find $Cov(Z, X)$ and $Cov(Z, Y)$:

$$Cov(Z, X) = Cov(Z, \alpha_1 Z + \alpha_2 U + \alpha_3 ZU + \epsilon_1) = \alpha_1 \tag{A.32}$$

$$\text{and} \tag{A.33}$$

$$Cov(Z, Y) = Cov(Z, \beta_1 X + \beta_2 U + \beta_X U + \epsilon_2) = E[Z(\beta_1 X + \beta_2 U + \beta_X U + \epsilon_2)] \tag{A.34}$$

$$= E[\beta_1 XZ + \beta_2 UZ + \beta_3 XUZ + Z\epsilon_2] = \alpha_1\beta_1 + \alpha_3\beta_3 \tag{A.35}$$

Going back to FWL, letting $\lambda = \alpha_1\beta_1 + \alpha_3\beta_3$, we now have

$$\frac{X^T M_Z Y}{X^T M_Z X} = \frac{(X - \alpha_1 Z)^T(Y - \lambda Z)}{(X - \alpha_1 Z)^T(X - \alpha_1 Z)} = \frac{\sum_{i=1}^n X_i Y_i - \lambda X_i Z_i - \alpha_1 Z_i Y_i + \alpha_1 \lambda Z_i^2}{\sum_{i=1}^n (X_i - \alpha_1 Z_i)^2} \tag{A.36}$$

$$\xrightarrow{p} \frac{E[X_i Y_i - \lambda X_i Z_i - \alpha_1 Z_i Y_i + \alpha_1 \lambda Z^2]}{E[(X_i - \alpha_1 Z_i)^2]} \tag{A.37}$$

$$= \frac{Cov(X,Y) - \lambda Cov(X,Z) - \alpha_1 Cov(Z,Y) + \alpha_1 \lambda Var(X)}{1 - \alpha_1^2} \tag{A.38}$$

$$= \frac{(\beta_1 + \alpha_2\beta_2 + 2\alpha_1\alpha_3\beta_3) - \lambda\alpha_1 - \alpha_1\lambda + \alpha_1\lambda}{1 - \alpha_1^2} \tag{A.39}$$

$$= \frac{\beta_1 + \alpha_2\beta_2 + 2\alpha_1\alpha_3\beta_3 - \alpha_1^2\beta_1 + \alpha_3\beta_3\alpha_1}{1 - \alpha_1^2} = \beta_1 + \frac{\alpha_2\beta_2 + \alpha_1\alpha_3\beta_3}{1 - \alpha_1^2} \tag{A.40}$$

□

Proof for the consistency of $\hat{A}_4$

Proof. We can compute the Wald estimator from the quantities computed in the proof for $\hat{A}_3$. In particular, $Cov(Z,Y)$ and $Cov(Z,X)$.

$$\hat{A}_4 \xrightarrow{p} \frac{Cov(Z,Y)}{Cov(Z,X)} = \frac{\alpha_1\beta_1 + \alpha_3\beta_3}{\alpha_1} = \beta_1 + \frac{\alpha_3\beta_3}{\alpha_1} \tag{A.41}$$

□

A.1.5 Proofs for Proposition 3.3.5

Proof for the consistency of $\hat{A}_2$

Proof. Our goal is to find $\frac{\partial}{\partial x} E[Y|x,w]$, which, by FWL, is equivalent to the convergence in probability of $\frac{X^T M_W Y}{X^T M_W X}$. Because we have already demonstrated the use of FWL, we will skip several intermediate steps. Letting $\eta = c_6 + c_5 c_0$

$$\frac{X^T M_W Y}{X^T M_W X} = \frac{(X - c_5 W)^T (Y - \eta W}{((X - c_5 W))} \xrightarrow{p} \frac{Cov(X,Y) - \eta Cov(W,Y) + c_5 \eta}{1 - c_5^2} \tag{A.42}$$

$$= \frac{c_0 + c_1 c_2 + c_2 c_3 c_I - \eta c_5 - \eta c_5 + c_5 \eta)}{1 - c_5^2} = c_0 + \frac{c_1 c_2 + c_2 c_3 c_I}{1 - c_5^2} \tag{A.43}$$

□

Proof for the consistency of $\hat{A}_3$

Proof. Our goal is to find $\frac{\partial}{\partial x} E[Y|x, w, z]$, which, by FWL, is equivalent to the limit of $\frac{X^T M_{\mathbf{B}} Y}{X^T M_{\mathbf{B}} X}$ where $\mathbf{B} = [W, Z]$ or the residuals of regressing both W and Z. To simplify matters, we can expand Y

$$\frac{X^T M_{\mathbf{B}} Y}{X^T M_{\mathbf{B}} X} = \frac{X^T M_{\mathbf{B}} (c_0 + c_2 U + c_6 W + \epsilon_2)}{X^T M_{\mathbf{B}} X} \tag{A.44}$$

$$= c_0 + c_2 \frac{X^T M_{\mathbf{B}} U}{X^T M_{\mathbf{B}} X} \tag{A.45}$$

Where the last line follows from the orthogonality of ϵ_2 and the result of W being regressed upon itself being that the residuals are 0. So we are focused on two regressions: $E[X|w, z] = \gamma_1 W + \gamma_2 Z$ and $E[U|w, z] = \pi_1 W + \pi_2 Z$. Using FWL, we find the value of the coefficients to be

$$\gamma_1 = \frac{W^T M_Z X}{W^T M_Z W} \xrightarrow{p} c_5 - \frac{c_1 c_I c_7}{1 - c_7^2}, \tag{A.46}$$

$$\gamma_2 = \frac{Z^T M_W X}{Z^T M_W Z} \xrightarrow{p} c_3 + \frac{c_1 c_I}{1 - c_7^2}, \tag{A.47}$$

$$\pi_1 = \frac{W^T M_Z U}{W^T M_Z W} \xrightarrow{p} \frac{-c_I c_7}{1 - c_7^2}, \tag{A.48}$$

$$\pi_2 = \frac{Z^T M_W U}{Z^T M_W Z} \xrightarrow{p} \frac{c_I}{1 - c_7^2} \tag{A.49}$$

Putting this altogether, we write $A_3 = c_0 + c_2 \frac{Cov(X-\gamma_1 W-\gamma_2 Z, U-\pi_1 W-\pi_2 Z)}{Var(X-\gamma_1 W-\gamma_2 Z)}$ (note that we are now in the limit). Expanding the numerator we get

$$c_2 \Bigg[Cov(X,U) - \pi_1 Cov(X,W) - \pi_2 Cov(X,Z) - \gamma_1 Cov(W,U) + \gamma_1 \pi_1 \tag{A.50}$$

$$+ \gamma_1 \pi_2 Cov(W,Z) - \gamma_2 Cov(Z,U) + \gamma_2 \pi_1 Cov(W,Z) + \gamma_2 \pi_2 \Bigg] \tag{A.51}$$

Noting that $Cov(X,U) = c_1 + c_3 c_I$, $Cov(X,W) = c_5 + c_3 c_7$, $Cov(X,Z) = c_3 + c_1 c_I + c_5 c_7$, $Cov(W,U) = 0$, $Cov(W,Z) = c_7$, and $Cov(Z,U) = c_I$, we can simplify the above expression to

$$c_1 c_2 + \frac{c_1 c_2 c_I^2 c_7}{1-c_7^2} - \frac{c_1 c_2 c_I^2 c_7}{(1-c_7^2)^2} + \frac{c_1 c_2 c_I c_7^3}{(1-c_7^2)^2} - \frac{c_1 c_2 c_I^2}{1-c_7^2} \tag{A.52}$$

$$= c_1 c_2 + \frac{c_1 c_2 c_I^2 c_7 - c_1 c_2 c_I^2}{1-c_7^2} + \frac{c_1 c_2 c_I^2 c_7 (c_7^2 - 1}{(1-c_7^2)^2} = c_1 c_2 + \frac{c_1 c_2 c_I^2 c_7 - c_1 c_2 c_I^2 - c_1 c_2 c_I^2 c_7}{1-c_7^2} \tag{A.53}$$

$$= c_1 c_2 - \frac{c_1 c_2 c_I^2}{1-c_7^2} \tag{A.54}$$

For the denominator

$$Var(X - \gamma_1 W - \gamma_2 Z) = 1 + \gamma_1^2 + \gamma_2^2 - 2\gamma_1 Cov(X,W) \tag{A.55}$$

$$- 2\gamma_2 Cov(X,Z) + 2\gamma_1 \gamma_2 Cov(W,Z) \tag{A.56}$$

$$\tag{A.57}$$

After combining like terms we obtain

$$1 - c_5^2 - c_3^2 + \frac{c_1^2 c_I^2}{(1-c_7^2)^2} - 2c_3 c_5 c_7 - 2c_1 c_3 c_I - 2\frac{c_1^2 c_I^2}{1-c_7^2} - \frac{c_1^2 c_I^2 c_7^2}{(1-c_7^2)^2} \tag{A.58}$$

$$= 1 - c_5^2 - c_3^2 - 2c_3c_5c_7 - 2c_1c_3c_I - \frac{c_1^2c_I^2}{1-c_7^2} + [c_3^2c_7^2 - c_3^2c_7^2] \tag{A.59}$$

$$= 1 - (c_5 + c_3c_7)^2 + c_3^2c_7^2 - c_3^2 - 2c_1c_3c_I - \frac{c_1^2c_I^2}{1-c_7^2} \tag{A.60}$$

$$= 1 - (c_5 + c_3c_7)^2 - (1-c_7^2)(c_3^2 + \frac{2c_1c_3c_I}{1-c_7^2} + \frac{c_1^2c_I^2}{1-c_7^2}) \tag{A.61}$$

$$= 1 - (c_5 + c_3c_7)^2 - (1-c_7^2)(c_3 + \frac{c_1c_I}{1-c_7^2})^2 \tag{A.62}$$

□

Proof for the consistency of $\hat{A}_4$

Proof. Via FWL, $\frac{\frac{\partial}{\partial z}E[Y|z,w]}{\frac{\partial}{\partial z}E[X|z,w]} = \ = \frac{Z^TM_WY}{Z^TM_WX}$. Re-using quantities from the proof for A_3 and, additionally $Cov(Z,Y) = c_0c_3 + c_2c_I + c_6c_7 + c_0c_5c_7 + c_0c_1c_I$ and $Cov(Y,W) = c_0c_5 + c_6 + c_0c_3c_7$ we have

$$\frac{Z^TM_WY}{Z^TM_WX} \xrightarrow{p} \frac{Cov(Z,Y) - Cov(Z,W)Cov(Y,W)}{Cov(X,Z) - Cov(Z,W)Cov(X,W)} \tag{A.63}$$

$$= \frac{c_0c_3 + c_2c_I + c_6c_7 + c_0c_5c_7 + c_0c_1c_I - c7(c_0c_5 + c_6 + c_0c_3c_7)}{c_3 + c_1c_I + c_5c_7 - c_7(c_5 + c_3c_7)} \tag{A.64}$$

$$= \frac{c_0(c_3 + c_1c_I - c_7^2c_3) + c_2c_I}{c_3 + c_1c_I - c_7^2c_3} = \frac{c_2c_I}{c_3(1-c_7^2) + c_1c_I} \tag{A.65}$$

□

A.1.6 Proof for Proposition 3.3.6

Results for A_2

Proof. Our goal is to find the value of $\frac{\partial}{\partial x}\frac{X^T M_{\mathbf{C}} Y}{X^T M_{\mathbf{C}} X}$ where $\mathbf{C} = [W, XW]$. Because $Cov(W, XW) = 0$, we can write the following linear conditional expectations

$$E[X|W, XW] = Cov(X, W)W + \frac{Cov(X, XW)}{Var(XW)}XW = \alpha_4 W + \frac{2\alpha_1\alpha_5}{\alpha_4^2 + 2\alpha_5^2 + 1}XW, \tag{A.66}$$

$$E[Y|W, XW] = Cov(Y, W)W + \frac{Cov(Y, XW)}{Var(XW)}XW \tag{A.67}$$

$$= (\beta_1\alpha_4 + \beta_4)W + \frac{2\alpha_1\alpha_5\beta_1 + \beta_5(\alpha_4^2 + 2\alpha_5^2 + 1)}{\alpha_4^2 + 2\alpha_5^2 + 1}XW. \tag{A.68}$$

Returning back to the FWL expression, we now plug in the above conditional expectations. Focusing on the numerator, which is rewritten as $Cov(X - E[X|W, XW], Y - E[X|W, XW])$, we have

$$Cov(X - \alpha_4 W - \frac{2\alpha_1\alpha_5}{\alpha_4^2 + 2\alpha_5^2 + 1}XW, Y - \beta_1\alpha_4 + \beta_4 W \tag{A.69}$$

$$- \frac{2\alpha_1\alpha_5\beta_1 + \beta_5(\alpha_4^2 + 2\alpha_5^2 + 1)}{\alpha_4^2 + 2\alpha_5^2 + 1}XW \tag{A.70}$$

$$= Cov(X, Y) - \frac{\frac{2\alpha_1\alpha_5\beta_1 + \beta_5(\alpha_4^2 + 2\alpha_5^2 + 1)}{\alpha_4^2 + 2\alpha_5^2 + 1}Cov(X, XW)}{\alpha_4^2 + 2\alpha_5^2 + 1} - \alpha_4 Cov(W, Y) - \tag{A.71}$$

$$\frac{2\alpha_4\alpha_5 Cov(XW, Y)}{\alpha_4^2 + 2\alpha_5^2 + 1} + \frac{\frac{2\alpha_1\alpha_5\beta_1 + \beta_5(\alpha_4^2 + 2\alpha_5^2 + 1)}{\alpha_4^2 + 2\alpha_5^2 + 1}2\alpha_1\alpha_5}{\alpha_4^2 + 2\alpha_5^2 + 1} \tag{A.72}$$

$$= \beta_1 + \beta_2\alpha_2 + 2\alpha_1\alpha_3\beta_3 + \beta_4\alpha_4 + 2\alpha_1\alpha_5\beta_5 - \alpha_4(\beta_1\alpha_4 + \beta_4) \tag{A.73}$$

$$- \frac{2\alpha_1\alpha_5\frac{2\alpha_1\alpha_5\beta_1 + \beta_5(\alpha_4^2 + 2\alpha_5^2 + 1)}{\alpha_4^2 + 2\alpha_5^2 + 1}}{\alpha_4^2 + 2\alpha_5^2 + 1} \tag{A.74}$$

$$= \beta_1 + \beta_2\alpha_2 + 2\alpha_1\alpha_3\beta_3 - \alpha_4^2\beta_1 - \frac{4\alpha_1^2\alpha_5^2\beta_1}{\alpha_4^2 + 2\alpha_5^2 + 1}. \tag{A.75}$$

Focusing on the denominator:

$$Var(X - \alpha_4 W - \frac{2\alpha_1\alpha_5}{\alpha_4^2 + 2\alpha_5^2 + 1} XW) = 1 + \alpha^4 + \frac{4\alpha_1^2\alpha_5^2}{\alpha_4^2 + 2\alpha_5^2 + 1} - 2\alpha_4^2 - \frac{8\alpha_1^2\alpha_5^2}{\alpha_4^2 + 2\alpha_5^2 + 1} \tag{A.76}$$

$$= 1 - \alpha_4 - \frac{4\alpha_1^2\alpha_5^2}{\alpha_4^2 + 2\alpha_5^2 + 1}. \tag{A.77}$$

Thus, we obtain the final result $\beta_1 + \frac{\beta_2\alpha_2 + 2\alpha_1\alpha_3\beta_3}{1 - \alpha_4 - \frac{4\alpha_1^2\alpha_5^2}{\alpha_4^2 + 2\alpha_5^2 + 1}}$. □

Results for A_4

Due to there being two endogenous variables, X and XW, we will need to utilize two instruments, which are Z and ZW, respectively. We will have the following series of linear conditional expectations, noting that $Cov(XW, W) = 0$

$$E[X|W, Z, ZW] = Cov(X, W)W + Cov(Z, X)Z + Cov(ZW, X)ZW, \tag{A.78}$$

$$E[XW|Z, ZW] = Cov(XW, Z)Z + Cov(XW, ZW)ZW, \tag{A.79}$$

$$E[Y|\hat{X}, X\hat{W}, W] = \frac{Cov(Y, \hat{X})}{Var(\hat{X})}\hat{X} + \frac{Cov(Y, X\hat{W})}{Var(X\hat{W})}X\hat{W} + Cov(Y, W)W \tag{A.80}$$

where $\hat{X}$ and $X\hat{W}$ are the fitted values from the $E[X|W, Z, ZW]$ and $E[XW|Z, ZW]$, respectively. The coefficient of interest from 2SLS is thus $\frac{Cov(Y,\hat{X})}{Var(\hat{X})}$ or, returning to FWL, denoted as $\frac{\partial}{\partial \hat{x}} \frac{\hat{X} M_{\mathbf{D}} Y}{\hat{X} M_{\mathbf{D}} \hat{X}}$ where $\mathbf{D} = [W, X\hat{W}]$.

First, we will write out the relevant covariances and variances:

$$Cov(W,\hat{X}) = \alpha_4 \tag{A.81}$$

$$Cov(\hat{XW},\hat{X}) = 2\alpha_1\alpha_5 \tag{A.82}$$

$$Cov(W,Y) = \beta_1\alpha_4 + \beta_4 \tag{A.83}$$

$$Cov(\hat{XW},Y) = \alpha_5\beta_1 Cov(Z,X) + \alpha_5\beta_3 Cov(Z,XU) + \beta_5\alpha_5 Cov(Z,XW) \tag{A.84}$$

$$+\alpha_1\beta_1 Cov(ZW,X) + \alpha_1\beta_3 Cov(ZW,XU) + \alpha_1\beta_5 Cov(ZW,XW) \tag{A.85}$$

$$= 2\alpha_1\alpha_5\beta_1 + \alpha_3\alpha_5\beta_3 + \beta_5\alpha_5^2 + \alpha_1^2\beta_5 \tag{A.86}$$

$$Var(\hat{X}) = \alpha_4^2 + \alpha_1^2 + \alpha_5^2 \tag{A.87}$$

$$Var(\hat{XW}) = \alpha_5^2 + \alpha_1^2 \tag{A.88}$$

$$Cov(\hat{X},Y) = \alpha_4 Cov(W,Y) + \alpha_1 Cov(Z,Y) + \alpha_5 Cov(ZW,Y) \tag{A.89}$$

$$= \alpha_4(\beta_1\alpha_4 + \beta_4) + \alpha_1(\alpha_1\beta_1 + \alpha_3\beta_3 + \alpha_5\beta_5) + \alpha_5(\beta_1\alpha_5 + \beta_5\alpha_1) \tag{A.90}$$

$$Cov(ZW,Y) = \beta_1\alpha_5 + \beta_5\alpha_1 \tag{A.91}$$

Because $Cov(\hat{XW},W) = 0$ we can simply substitute the covariances in the FWL expression as such

$$\frac{Cov(\hat{X} - \alpha_4 W - \frac{2\alpha_1\alpha_5}{\alpha^5+\alpha_1^2}\hat{XW}, Y - (\beta_1\alpha_4 + \beta_4)W - Cov(\hat{XW},Y)\hat{XW})}{\hat{X} - \alpha_4 W - \frac{2\alpha_1\alpha_5}{\alpha^5+\alpha_1^2}\hat{XW}} \tag{A.92}$$

Now focusing on the numerator, can further simplify to

$$Cov(\hat{X},Y) - (\beta_1\alpha_4 + \beta_4)Cov(\hat{X}W) - \frac{Cov(\hat{XW},Y)}{\alpha_5^2 + \alpha_1^2} - \frac{2\alpha_1\alpha_5 Cov(\hat{XW},Y)}{\alpha^5 + \alpha_1^2} \tag{A.93}$$

$$+ \frac{2\alpha_1\alpha_5 Cov(X\hat{W}, Y) Var(X\hat{W})}{(\alpha^5 + \alpha_1^2)^2} \tag{A.94}$$

$$= \alpha_1(\alpha_1\beta_1 + \alpha_3\beta_3 + \alpha_5\beta_5) + \alpha_5(\beta_1\alpha_5 + \beta_5\alpha_1) - \frac{2\alpha_1\alpha_5 Cov(X\hat{W}, Y)}{\alpha^5 + \alpha_1^2} \tag{A.95}$$

$$= \alpha_1\beta_1 + \alpha_1\alpha_3\beta_3 + 2\alpha_1\alpha_5\beta_5 + \beta_1\alpha_5^2 \tag{A.96}$$

$$- \frac{4\alpha_1\alpha_5^2\beta_1 + 2\alpha_1\alpha_3\alpha_5^2\beta_3 + 2\alpha_1\alpha_5^3\beta_5 + 2\alpha_1^3\alpha_5\beta_5}{\alpha_5^2 + \alpha_1^2} \tag{A.97}$$

$$= \beta_1\left[\alpha_1^2 + \alpha_5^2 - \frac{4\alpha_1^2\alpha_5\beta_1}{\alpha^5 + \alpha_1^2}\right] + \alpha_1\alpha_3\beta_3 - \frac{2\alpha_1\alpha_3\alpha_5^2\beta_3}{\alpha_1^2 + \alpha_5^2}. \tag{A.98}$$

For the denominator, we have

$$Var(\hat{X} - \alpha_4 W - \frac{2\alpha_1^2\alpha_5^2}{\alpha_1^2 + \alpha_5^2} X\hat{W}) = \alpha_4^2 + \alpha_1^2 + \alpha_5^2 + \alpha_4^2 + \frac{4\alpha_1^2\alpha_5^2}{\alpha_1^2 + \alpha_5^2} \tag{A.99}$$

$$- 2\alpha_4^2 - \frac{8\alpha_1^2\alpha_5^2}{\alpha_1^2 + \alpha_5^2} \tag{A.100}$$

$$= \alpha_1^2 + \alpha_5^2 - \frac{4\alpha_1^2\alpha_5^2}{\alpha_1^2 + \alpha_5^2}. \tag{A.101}$$

Therefore, all together we obtain the final result of

$$\beta_1 + \frac{\alpha_1\alpha_3\beta_3 - \frac{2\alpha_1\alpha_3\alpha_5^2\beta_3}{\alpha_1^2 + \alpha_5^2}}{\alpha_1^2 + \alpha_5^2 - \frac{4\alpha_1^2\alpha_5^2}{(\alpha_1^2 + \alpha_5^2})} \tag{A.102}$$

A.1.7 Derivation for Reduced Form Coefficient in Exclusion Restriction β_{ER}^R

Proof. We will first derive the proof in the case with no covariates in the DAG (i.e. $E[Y|X, Z|)$ as it is more tractable to show and describe how it applies when W is present. The coefficient takes the form $\frac{\text{Cov}(Y - E[Y|X], Z - E[Z|X])}{\text{Var}(Z - E[Z|X])}$. As shown in the previous proof, the denominator is

$1 - c^3$ so we will focus on the numerator.

$$\begin{aligned}
\text{Cov}(Y - E[Y|X], Z - E[Z|X])\text{Var}(Z - E[Z|X]) &= \text{Cov}(Y, Z) - \text{Cov}(Y, \text{Cov}(Z, X)X) \\
&\quad - \text{Cov}(\text{Cov}(Y, X)X, Z) + \text{Cov}(Y, X) \cdot \text{Cov}(Z, X) \\
&= c_{ER} + c_3c_0 - c_3(c_0 + c_1c_2 + c_3c_{ER}) \\
&\quad - c_3(c_0 + c_1c_2 + c_3c_{ER}) + c_3(c_0 + c_1c_2 + c_3c_{ER}) \\
&= c_{ER} + c_3c_0 - c_3c_0 - c_1c_2c_3 - c_3^2c_{ER} \\
&= c_{ER}(1 - c_3^2) - c_1c_2c_3
\end{aligned}$$

Thus, the quantity is equal to $c_{ER} - \frac{c_1c_2c_3}{1-c^3}$. Adding W to the DAG, because we are conditioning for it, there is no impact besides a portion of variation in X being explained away, leading us the final quantity of $c_{ER} - \frac{c_1c_2c_3}{1-c_3^2-c_5^2}$ □

A.1.8 Proof for Independence ϕ Result

Proof. We will first translate the edge weights or the linear combination of the edge weights as R^2 quantities. From the main text, the initial quantity is

$$\phi = \left|\frac{c_2c_I}{c_3 + \frac{c_1c_I}{1-c_7^2}}\right|\left|\frac{\frac{c_1c_2(1-c_7^2-c_I^2)}{1-c_7^2}}{1 - (c_5 + c_3c_7)^2 - (1 - c_7^2)(c_3 + \frac{c_1c_I}{1-c_7^2})^2}\right|^{-1}. \tag{A.103}$$

We can re-write the different terms as follows [31]:

$$|c_1| = \sqrt{R^2_{X\sim U|Z,W}}\frac{sd(X^{\perp Z,W})}{sd(U^{\perp Z,W})} \tag{A.104}$$

$$|c_I| = \sqrt{R^2_{Z\sim U|W}}\frac{sd(U^{\perp W})}{sd(Z^{\perp W})} \tag{A.105}$$

$$|c_7| = \sqrt{R^2_{Z\sim W}} \tag{A.106}$$

$$(c_5 + c_3c_7)^2 = R^2_{X\sim W} \tag{A.107}$$

$$(c_3 + \frac{c_1c_I}{1-c_7^2})^2 = R^2_{X\sim Z|W}\frac{Var(X^{\perp W})}{Var(Z^{\perp W})} \tag{A.108}$$

$$\frac{1-c_I^2-c_7^2}{1-c_7^2} = \frac{1-R^2_{Z\sim W+U}}{1-R^2_{Z\sim W}} = 1 - R^2_{Z\sim U|W}. \tag{A.109}$$

Canceling out c_2 on both sides and substituting these quantities, we obtain

$$\frac{\sqrt{R^2_{Z\sim U|W}\frac{sd(U^{\perp W})}{sd(Z^{\perp W})}}}{R^2_{X\sim Z|W}\frac{Var(X^{\perp W})}{Var(Z^{\perp W})}} = \frac{\sqrt{R^2_{X\sim U|Z,W}\frac{sd(X^{\perp Z,W})}{sd(U^{\perp Z,W})}}}{1-R^2_{X\sim W}-(1-R^2_{X\sim W})(R^2_{X\sim Z|W}\frac{Var(X^{\perp W})}{Var(Z^{\perp W})})}. \tag{A.110}$$

Noting that $\frac{sd(U^{\perp W,Z})}{sd(U^{\perp W})} = \sqrt{1-R^2_{Z\sim U|W}}$, we can re-arrange terms such that

$$\frac{\sqrt{R^2_{Z\sim U|W}}}{\sqrt{1-R^2_{Z\sim U|W}}} = \frac{\sqrt{R^2_{X\sim U|Z,W}}sd(Z^{\perp W})sd(X^{\perp Z,W})R^2_{X\sim Z|W}\frac{Var(X^{\perp W})}{Var(Z^{\perp W})}}{1-R^2_{X\sim W}-(1-R^2_{X\sim W})(R^2_{X\sim Z|W}\frac{Var(X^{\perp W})}{Var(Z^{\perp W})})}. \tag{A.111}$$

Squaring both sides obtains the final result. □

A.1.9 Proof for Heterogeneity ϕ Result

Proof. From the main text, the original quantity is

$$\phi = \left|\frac{\alpha_1\alpha_3\beta_3 - \frac{2\alpha_1\alpha_3\alpha_5^2\beta_3}{\alpha_1^2+\alpha_5^2}}{\alpha_1^2+\alpha_5^2-\frac{4\alpha_1^2\alpha_5^2}{(\alpha_1^2+\alpha_5^2})}\right|\left|\frac{\alpha_2\beta_2+2\alpha_1\alpha_3\beta_3}{1-\alpha_4^2-\frac{4\alpha_1^2\alpha_5^2}{1+\alpha_4^2+2\alpha_5^2}}\right|^{-1}. \tag{A.112}$$

Noting that $1-\alpha_4^2-\frac{4\alpha_1^2\alpha_5^2}{1+\alpha_4^2+2\alpha_5^2} = Var(X^{\perp W,XW})$, we may simplify as follows

$$\frac{\frac{\beta_2\alpha_2}{2\alpha_1\alpha_3\beta_3}+1}{Var(X^{\perp W,XW})}\left(\frac{\frac{1}{2}-\frac{2a_5^2}{a_1^2+a_5^2}}{a_1^2+a_5^2-\frac{4a_1^2a_5^2}{a_1^2+a_5^2}}\right)^{-1} \tag{A.113}$$

$$\frac{\frac{\beta_2\alpha_2}{2\alpha_1\alpha_3\beta_3}+1}{Var(X^{\perp W,XW})}\left(\frac{\frac{1}{2}(a_1^2+a_5^2)-2a_5^2}{(a_1^2+a_5^2)^2-4a_1^2a_5^2}\right)^{-1} \tag{A.114}$$

$$\frac{\frac{\beta_2\alpha_2}{2\alpha_1\alpha_3\beta_3}+1}{Var(X^{\perp W,XW})}\left(\frac{\frac{1}{2}a_1^2-\frac{3}{2}a_5^2}{(a_1^2-a_5^2)^2}\right)^{-1} \tag{A.115}$$

$$\left|\frac{\beta_2\alpha_2}{2\alpha_1\alpha_3\beta_3}\right|\left|\frac{(\frac{1}{2}a_1^2-\frac{3}{2}a_5^2)Var(X^{\perp W,XW})-(a_1^2-a_5^2)^2}{(a_1^2-a_5^2)^2}\right|^{-1} \tag{A.116}$$

$$\left|\frac{2\alpha_1\alpha_3\beta_3}{\beta_2\alpha_2}\right|\left|\frac{(a_1^2-a_5^2)^2}{(\frac{1}{2}a_1^2-\frac{3}{2}a_5^2)Var(X^{\perp W,XW})-(a_1^2-a_5^2)^2}\right|^{-1} \tag{A.117}$$

$$|\alpha_3|\left(\left|\frac{(a_1^2-a_5^2)^2}{(\frac{1}{2}a_1^2-\frac{3}{2}a_5^2)Var(X^{\perp W,XW})-(a_1^2-a_5^2)^2}\right|\left|\frac{\beta_2\alpha_2}{2\alpha_1\beta_3}\right|\right)^{-1} \tag{A.118}$$

$$\tag{A.119}$$

This gives us the final result in the text. □

A.1.10 Notes about Simulations

Specifically, we have restricted the variance for all variables to 1. Therefore, choosing values of the structural coefficients for simulations to verify our derivations is subject to constraints. Importantly, the variance of the stochastic error (e.g. σ_{ϵ_1} in Eq 3.1) must be chosen carefully such that we simulate the random variables properly. We detail these constraints for each scenario below, which are calculated by taking variances of X and Y and computing the relevant covariances.

Perfect IV

We are subject to the following constraints:

1. $\sigma_X^2 = 1 = c_3^2 + c_1^2 + \sigma_{\epsilon_1}^2$

2. $\sigma_Y^2 = 1 = c_0^2 + c_2^2 + 2c_0c_1c_2 + \sigma_{\epsilon_2}^2$

Exclusion Restriction

Without Covariates

1. $\sigma_X^2 = 1 = c_3^2 + c_1^2 + \sigma_{\epsilon_1}^2$

2. $\sigma_Y^2 = 1 = c_0^2 + c_2^2 + c_{ER}^2 + 2c_0c_1c_2 + 2c_0c_3c_{ER} + \sigma_{\epsilon_2}^2$

With Covariates

1. $\sigma_X^2 = 1 = c_1^2 + c_3^2 + c_5^2 + \sigma_{\epsilon_3}^2$

2. $\sigma_Y^2 = 1 = c_0^2 + c_2^2 + c_{ER}^2 + c_6^2 + 2c_0c_1c_2 + 2c_0c_3c_{ER} + 2c_0c_5c_6 + \sigma_{\epsilon_4}^2$

U is a cause of Z

Without Covariates

1. $\sigma_X^2 = 1 = c_1^2 + c_3^2 + 2c_1c_3c_I + \sigma_{\epsilon_1}^2$

2. $\sigma_Y^2 = 1 = c_0^2 + c_2^2 + 2c_0c_2(c_1 + c_3c_I) + \sigma_{\epsilon_2}^2$

3. $\sigma_Z^2 = 1 = c_I^2 + \sigma_{\epsilon_3}^2$

With Covariates

1. $\sigma_X^2 = 1 = c_1^2 + c_3^2 + 2c_1c_3c_I + 2c_3c_5c_7 + \sigma_{\epsilon_4}^2$

2. $\sigma_Y^2 = 1 = c_0^2 + c_2^2 + 2c_0c_2(c_1 + c_3c_I) + 2c_0c_6(c_5 + c_3c_7) + \sigma_{\epsilon_5}^2$

3. $\sigma_Z^2 = 1 = c_I^2 + c_7^2 + \sigma_{\epsilon_6}^2$

Treatment Effect Heterogeneity

Without Covariates

1. $\sigma_X^2 = 1 = \alpha_1^2 + \alpha_2^2 + \alpha_3^2 + \sigma_{\epsilon_1}^2$

2. $\sigma_Y^2 = 1 = \beta_1^2 + \beta_2^2 + \beta_3^2 + 2\beta_1\beta_2\alpha_2 + 4\beta_1\beta_2\beta_3\alpha_1\alpha_3 + \sigma_{\epsilon_2}^2$

With Covariates

1. $\sigma_X^2 = 1 = \alpha_1^2 + \alpha_2^2 + \alpha_3^2 + \alpha_4^2 + \alpha_5^2 + \sigma_{\epsilon_3}^2$

2. $\sigma_Y^2 = 1 = \beta_1^2 + \beta_2^2 + \beta_3^2 + 2\beta_1\beta_2\alpha_2 + 4\beta_1\beta_2\beta_3\alpha_1\alpha_3 + 2\beta_1\alpha_4\beta_4 + 4(\beta_1\beta_5\alpha_1\alpha_5) + 2\beta_3\beta_5(\alpha_2\alpha_4 + 2\alpha_3\alpha_5) + \sigma_{\epsilon_4}^2$

A.2 Appendices for Chapter 4

A.2.1 Proof of Theorem 4.3.1

Proof. The overall structure of the proof is as follows. We largely follow the strategy outlined in Kennedy (2023) where we must show that $\hat{\beta}_{LATE}^A - \beta_{LATE}^A = S^* + T_1 + T_2$ takes a asymptotically linear form of $(\mathbb{P}_n - P)(\phi(B; w, \zeta_z)) + o_P(n^{-1/2})$.[65] In other words, we have $S^* = (\mathbb{P}_n - P)\{\phi(B; w, \zeta_z)$ and show $T_1 + T_2 = o_P(n^{-1/2})$. We will establish this result for the numerator, whose influence function will be denoted by ϕ, and skip proof for the denominator as it is similarly a weighted difference of conditional expectations. Once we have done this, we can apply Lemma S.1 from Takatsu et al. (2023), which states the ratio of two asymptotically linear estimators is also asymptotically linear.[108]

For simplicity of the proof, we assume that our estimation uses sample-splitting with $K = 2$, but not cross-fitting, as described in Proposition 4.3.1. Suppose we had an i.i.d. sample from distribution P_B: $(B_1, B_2, ..., B_{N_B})$ and that $n = \lceil \frac{N_B}{2} \rceil$, then we fit the nuisance functions with

$B^N = (B_n + 1, ..., B_{N_B})$ and computed the predicted values over $(B_1, ...B_n)$ where empirical measure $\mathbb{P}_n$ is over this independent partition. The fact that we are using sample-splitting means that our weights estimator only needs to be consistent for $\frac{w(X)}{E_{P_B}[w(X)]}$. Nevertheless, because study A is involved in estimating the weights, we will take the expectations over P, the joint distribution of P_A and P_B.

Lemma A.2.1. $|\hat{\eta} - \eta| = o_P(1)$ *implies* $|\hat{w} - w| = o_P(1)$

Proof. By positivity and the bounding conditions outlined, $\eta, \hat{\eta} > C_\eta$ so by a simple arithmetic arrangement we have $|\hat{w} - w| = \left|\frac{\eta - \hat{\eta}}{\eta\hat{\eta}}\right| \leq \frac{1}{C_\eta^2}|\hat{\eta} - \eta|$. □

Lemma A.2.2. $T_1 = o_P(n^{-1/2})$

Proof. Proving $T_1 = o_P(n^{-1/2})$ is an application of results of Kennedy et al (2020) Lemma 2, which states $T_1 = O_P\left(\frac{\|\hat{\phi} - \phi\|}{n^{-1/2}}\right)$ due to sample-splitting.[66] Therefore, we meet the condition for T_1 as long as $\left\|\hat{\phi} - \phi\right\| = o_P(1)$. We can investigate this condition in more detail for $T(P_B) = \frac{E_{P_B}[w(X)E[Y|Z=1,X]]}{E_{P_B}[w(X)]}$, which naturally extends to our estimand:

$$\hat{\phi} - \phi = \frac{\hat{w}}{\mathbb{P}_n\hat{w}}\left\{\hat{\mu}_1 + \frac{Z(Y - \hat{\mu}_1)}{\hat{e}}\right\} - \frac{w}{E_{P_B}[w]}\left\{\mu_1 + \frac{Z(Y - \mu_1)}{e}\right\} \tag{A.120}$$

$$= \frac{\hat{w}\hat{\mu}_1}{\mathbb{P}_n\hat{w}}(1 - \frac{Z}{\hat{e}}) - \frac{w\mu_1}{E_{P_B}[w]}(1 - \frac{Z}{e}) + \frac{ZY}{\hat{e}e}(\frac{\hat{w}\hat{e}}{\mathbb{P}_n\hat{w}} - \frac{we}{E_{P_B}[w]}) \tag{A.121}$$

$$\leq (\frac{\hat{w}\hat{\mu}_1}{\mathbb{P}_n\hat{w}} - \frac{w\mu_1}{E_{P_B}[w]})(1 + \frac{1}{C_e}) + \frac{C_Y}{{C_e}^2}(\frac{\hat{w}\hat{e}}{\mathbb{P}_n\hat{w}} - \frac{we}{E_{P_B}[w]}) \tag{A.122}$$

Therefore, taking the L_2 norm, we will have

$$\left\|\hat{\phi} - \phi\right\| \leq \left\|\frac{\hat{w}\hat{\mu}_1}{\mathbb{P}_n\hat{w}} - \frac{w\mu_1}{E_{P_B}[w]}\right\|(1 + \frac{1}{C_e}) + \frac{C_Y}{{C_e}^2}\left\|\frac{\hat{w}\hat{e}}{\mathbb{P}_n\hat{w}} - \frac{we}{E_{P_B}[w]}\right\| \tag{A.123}$$

Thus, it is sufficient that $\left\|\frac{\hat{w}\hat{\mu}_1}{\mathbb{P}_n\hat{w}} - \frac{w\mu_1}{E_{P_B}[w]}\right\| = o_P(1)$ and $\left\|\frac{\hat{w}\hat{e}}{\mathbb{P}_n\hat{w}} - \frac{we}{E_{P_B}[w]}\right\| = o_P(1)$ for the whole term to be $o_P(1)$. This is achieved as long as we have consistency of the first term of each L_2 norm to the second term of each L_2 norm and smoothness of the various nuisance functions.[28, 65] Using sample splitting, we can completely avoid Donsker conditions regarding the complexity of the nuisance functions. For $\hat{\mu}_1$, this is implied by assumption. In the following steps, we will show consistency for $\hat{w}$ and $\mathbb{P}_n\hat{w}$. □

To show $|\mathbb{P}_n\hat{w} - E_{P_B}[w]| = o_P(1)$, we will proceed by Markov's inequality, noting that we have estimated w via sample splitting and, therefore, conditioning on B^N and A, the data from study A, will yield $\hat{w}$ fixed.

$$P(|\mathbb{P}_n\hat{w} - E_{P_B}[w]| \geq \epsilon) = E[P(|\mathbb{P}_n\hat{w} - E_{P_B}[w]| \geq \epsilon|\big|B^N, A)] \tag{A.124}$$

$$\leq \epsilon^{-1}E[|\mathbb{P}_n\hat{w} - E_{P_B}[w]|\big|B^N, A] \tag{A.125}$$

$$= \epsilon^{-1}E[|\mathbb{P}_n\hat{w} - E_{P_B}[w] + \mathbb{P}_n w - \mathbb{P}_n w|\big|B^N, A] \tag{A.126}$$

$$\leq \epsilon^{-1}E[|\hat{w} - w|] + E[|\mathbb{P}_n w - E_{P_B}[w]|] = o(1) \tag{A.127}$$

where the last inequality follows by a combination of triangle inequality and the fact that

$$E[|\mathbb{P}_n(\hat{w} - w)|\big|B^N, A] \leq n^{-1}\sum_i E\big[|\hat{w}(x_i) - w(X)|\big|B^N, A\big],$$

an i.i.d. sum, because w is now fixed, similar to the proof of Kennedy et al (2020) Lemma 2.[66] Now, the first term goes to 0 by $|\hat{w} - w| = o_P(1)$ and uniform integrability because $\hat{w}$ and w are bounded. The second term goes to 0 by weak law of large numbers and and uniform integrability due to boundedness.

Now that we have established all the terms in A.123 are consistent, the whole term is consistent by Slutsky's theorem. Thus, we we will have L_2 convergence due to boundedness and, consequentially, $T_1 = O_P\left(\frac{\|\hat{\phi}-\hat{\phi}\|}{n^{-1/2}}\right) = o_P(n^{-1/2})$.

Lemma A.2.3. $T_2 = o_P(n^{-1/2})$

Proof. For T_2, we must derive the remainder term $R_2(P,\hat{P}) = E_P[\hat{T}_{\text{1-step}}] - T(P_B)$ for the numerator. We will begin first with $T(P_B) = \frac{E_{P_B}[w(X)E[Y|Z=1,X]]}{E_{P_B}[w(X)]}$:

$$R_2(P,\hat{P}) = \int \frac{\hat{w}}{\mathbb{P}_n\hat{w}}\left[\frac{Z}{\hat{e}}(Y-\hat{\mu}_1)+\hat{\mu}_1\right]dP - \frac{E_{P_B}[w\mu_1]}{E_{P_B}[w]} \tag{A.128}$$

$$= \int \frac{\hat{w}}{\mathbb{P}_n\hat{w}}\left[\frac{\mu_1 e}{\hat{e}} - \frac{\hat{\mu}_1}{\hat{e}} + \hat{\mu}_1\right]dP - \frac{E_{P_B}[w\mu_1]}{E_{P_B}[w]} \tag{A.129}$$

$$= \int \frac{\hat{w}}{\mathbb{P}_n\hat{w}}\left[\frac{-\hat{e}}{e}(\hat{\mu}_1-\mu_1)+\hat{\mu}_1+(\mu_1-\mu_1)\right]dP - \frac{E_{P_B}[w\mu_1]}{E_{P_B}[w]} \tag{A.130}$$

$$= \int \frac{\hat{w}}{\mathbb{P}_n\hat{w}}\left[\frac{1}{\hat{e}}(\hat{e}-e)(\hat{\mu}_1-\mu_1)\right]dP + \int \mu_1\left[\frac{\hat{w}}{\mathbb{P}_n\hat{w}} - \frac{w}{E_{P_B}[w]}\right]dP \tag{A.131}$$

where the second equality follows by iterated expectation and the last follows because P is a joint distribution that includes P_B. Thus, taking the absolute value by bounding assumptions, triangle equality, and Cauchy-Schwarz, we have

$$|R_2(P,\hat{P})| \le \frac{1}{C_{\hat{e}}}\int |\frac{\hat{w}}{\mathbb{P}_n\hat{w}}||\hat{e}-e||\hat{\mu}_1-\mu_1|dP + C_{\mu_1}\int |\frac{\hat{w}}{\mathbb{P}_n\hat{w}} - \frac{w}{E_{P_B}[w]}|dP \tag{A.132}$$

$$\le \frac{C_w}{C_{\hat{e}}}\|\hat{e}-e\|\|\hat{\mu}_1-\mu_1\| + C_{\mu_1}\int\left|\frac{\hat{w}}{\mathbb{P}_n\hat{w}} - \frac{w}{E_{P_B}[w]}\right|dP \tag{A.133}$$

$$= o_P(n^{-1/2}) + o_P(1). \tag{A.134}$$

The first term is $o_P(n^{-1/2})$ by boundedness of $\left\|\frac{\hat{w}}{\mathbb{P}_n\hat{w}}\right\|$ and the product $\|\hat{e}-e\|\|\hat{\mu}_1-\mu_1\| = o_P(n^{-1/2})$, which holds if $\|\hat{e}-e\| = o_P(n^{-1/4})$ and $\|\hat{\mu}_1-\mu_1\| = o_P(n^{-1/4})$. Now we will focus on the second term, where we will show L_1 convergence.

Firstly, we have $|\hat{w}-w| = o_P(1)$ by Lemma A.2.1. Furthermore, when proving the rate of T_1, we showed that $|\mathbb{P}_n\hat{w} - E_{P_B}[w]| = o_P(1)$. Thus, by Slutsky's theorem we have that $|\frac{\hat{w}}{\mathbb{P}_n\hat{w}} - \frac{w}{E_{P_B}[w]}| = o_P(1)$. Given that we have uniform integrability due to boundedness of $\frac{\hat{w}}{\mathbb{P}_n\hat{w}}$ so we have L_1 convergence or, in other words, $E\left[|\frac{\hat{w}}{\mathbb{P}_n\hat{w}} - \frac{w}{E_{P_B}[w]}|\right] = o_P(1)$. □

The proof for $T(P_B) = \frac{E_{P_B}[w(X)\{\mu_1 - \mu_0\}]}{E_{P_B}[w(X)]}$ is the same as the one above where we repeat the process for μ_1 and μ_0 because it takes the form of a WATE. The proof for the denominator mirrors that of the numerator. Now we are ready to prove convergence and derive the asymptotic variance of our estimator.

From the result of Takatsu Lemma S.1, assuming there exists $\epsilon > 0$ such that $|\phi_{num}(\mathbf{B}; \hat{w}, \hat{\zeta}_z)| \wedge |\phi_{denom}(\mathbf{B}; \hat{w}, \hat{\zeta}_z)| > \epsilon$, then we have the following asymptotically linear form

$$
\begin{aligned}
&\frac{\mathbb{P}_n \hat{\phi}_{num}}{\mathbb{P}_n \hat{\phi}_{num}} - \frac{P\phi_{num}}{P\phi_{denom}} \\
&= \mathbb{P}_n \Bigg\{ \Bigg(\frac{E_{P_B}[w\{m_1(X) - m_0(X)\}]}{E[w(X)]} \Bigg)^{-1} \Bigg(\mathbb{IF}_{num} - \beta^A_{LATE} \mathbb{IF}_{denom} \Bigg) \Bigg\} + o_P(n^{-1/2}) \\
&= \mathbb{P}_n \Bigg\{ \Bigg(\frac{E_{P_B}[w\{m_1(X) - m_0(X)\}]}{E_{P_B}[w(X)]} \Bigg)^{-1} \\
&\Bigg(\frac{2Z-1}{e(X,Z)} \frac{w(X)}{E[w(X)]} \{Y - \mu_z(X)\} + \frac{w\{\mu_1(X) - \mu_0(X)\}}{E_{P_B}[w(X)]} \\
&- \beta^A_{LATE} \Bigg[\frac{2Z-1}{e(X,Z)} \frac{w(X)}{E_{P_B}[w(X)]} \{D - m_z(X)\} + \frac{w\{m_1(X) - m_0(X)\}}{E_{P_B}[w(X)]} \Bigg] \Bigg) \Bigg\} + o_P(n^{-1/2}) \\
&= \mathbb{P}_n \Bigg(\frac{w(X)}{E_{P_B}[w(X)\{m_1(X) - m_0(X)\}]} \Bigg\{ \frac{2Z-1}{e(X,Z)} \Bigg[Y - \mu_Z(X) - \beta^A_{LATE}\{D - m_Z(X)\} \Bigg] \\
&+ \mu_1(X) - \mu_0(X) - \beta^A_{LATE}\{m_1(X) - m_0(X)\} \Bigg) + o_P(n^{-1/2})
\end{aligned}
$$

where $e(X,Z) = e(X)Z + \{1 - e(X)\}(1 - Z)$. We can observe this is a sample mean of the influence function for the weighted LATE and, thus, after multiplying each side by $n^{-1/2}$, we have that $n^{-1/2}(\hat{\beta}^A_{LATE} - \beta^A_{LATE}) \xrightarrow{d} N(0, E_{P_B}[\Gamma^2])$ where

$$
\begin{aligned}
\Gamma = &\frac{w(X)}{E_{P_B}[w(X)\{m_1(X) - m_0(X)\}]} \Bigg\{ \frac{2Z-1}{e(X,Z)} \Bigg[Y - \mu_Z(X) - \beta^A_{LATE}\{D - m_Z(X)\} \Bigg] \\
&+ \mu_1(X) - \mu_0(X) - \beta^A_{LATE}\{m_1(X) - m_0(X)\} \Bigg\}
\end{aligned}
$$

via Slutsky's theorem and the standard central limit theorem, giving us $\sqrt{n}$-convergence. □

A.3 Appendices for Chapter 5

A.3.1 Proof of Proposition 5.3.1

Given $\delta_i = 0$, we know that $T_i = C_i + \alpha_i$ for some $\alpha_i > 0$. The crux of the Buckley-James procedure is essentially computing the BJ adjustment for the censored observations. Thus we may re-write the expected survival time of each strata as

$$E[T \mid Z = 1, A = 1, X] = E[T\delta \mid Z = 1, A = 1, X] + E[(1-\delta)T \mid Z = 1, A = 1, X]$$

$$= E[T \mid Z = 1, A = 1, X, T \leq C]\, P(T \leq C \mid Z = 1, A = 1, X)$$

$$+ E[\delta + \alpha \mid Z = 1, A = 1, X, T > C]\, [1 - P(T \leq C \mid Z = 1, A = 1, X)]$$

$$= E[T \mid Z = 1, A = 1, X, T \leq C]\, P(T \leq C \mid Z = 1, A = 1, X)$$

$$+ E[C \mid Z = 1, A = 1, X, T > C]\, [1 - P(T \leq C \mid Z = 1, A = 1, X)]$$

$$+ E[\alpha \mid Z = 1, A = 1, X, T > C]\, [1 - P(T \leq C \mid Z = 1, A = 1, X)]$$

$$= E[T \mid Z = 1, A = 1, X, T \leq C]\, P(T \leq C \mid Z = 1, A = 1, X)$$

$$+ E[C \mid Z = 1, A = 1, X, T > C]\, [1 - P(T \leq C \mid Z = 1, A = 1, X)]$$

$$+ E[Y^* - C \mid Z = 1, A = 1, X, T > C]\, [1 - P(T \leq C \mid Z = 1, A = 1, X)].$$

Where the last equality follows from that fact that

$$E[Y^* - C \mid Z = 1, A = 1, X, T > C]$$

$$= E[Y\delta + (1-\delta)E[T|Z=1, A=1, X, T>C] \mid Z=1, A=1, X, T>C]$$

$$- E[C \mid Z=1, A=1, X, T>C]$$

$$= E[E[T|Z=1, A=1, X, T>C] \mid Z=1, A=1, X, T>C]$$

$$- E[C \mid Z=1, A=1, X, T>C]$$

$$= E[T|Z=1, A=1, X, T>C] - E[C \mid Z=1, A=1, X, T>C]$$

$$= E[\alpha \mid Z=1, A=1, X, T>C].$$

A.3.2 Proof of Proposition 5.3.2

$$\mathbb{IF}\{\psi_{11}\} = \mathbb{IF}\left\{m_{11}(x)\gamma_{11}^{1}(x)\pi_1(x)p(x)\right\} + \mathbb{IF}\left\{\omega_{11}(x)\gamma_{11}^{0}(x)\pi_1(x)p(x)\right\} \tag{A.135}$$

$$+ \mathbb{IF}\left\{\lambda_{11}(X)\gamma_{11}^{0}(X)\pi_1(x)p(x)\right\} \tag{A.136}$$

Focusing on the first term, we may re-write it as

$$\mathbb{IF}\Bigg\{\sum_x m_{11}(x)\gamma_{11}^{1}(X)\pi_1(x)\gamma_{11}^{1}(x)p(x)\Bigg\} = \sum_x \Big[\mathbb{IF}\{m_{11}(x)\}\gamma_{11}^{1}(x)\pi_1(x)p(x)$$

$$+ m_{11}(x)\mathbb{IF}\{\gamma_{11}^{1}(X)\}\pi_1(x)p(x) + m_{11}(x)\gamma_{11}^{1}(x)\mathbb{IF}\{\pi_1(x)\}p(x) + m_{11}(x)\gamma_{11}^{1}(x)\pi_1(x)\mathbb{IF}\{p(x)\}\Big]$$

$$= (A) + (B) + (C) + (D)$$

Beginning with (A), we have

$$(A) = \sum_x \frac{I(X=x, T<C, Z=1, A=1)}{p(X=x, T<C, Z=1, A=1)}[T - m_{11}(x)]\gamma_{11}^{1}(x)\pi_1(x)p(x) \tag{A.137}$$

$$= \sum_x \frac{\delta Z A}{p(Z=1|X=x)}[T - m_{11}(x)] = \frac{\delta Z A}{e(X)}[T - m_{11}(X)]. \tag{A.138}$$

Similarly, we can write (B), (C), and (D) as

$$(B) = \frac{m_{11}(X) A Z}{e(X)}[\delta - \gamma_{11}^1(X)], \tag{A.139}$$

$$(C) = \frac{m_{11}(X)\gamma_{11}^1(X) Z}{e(X)}[A - \pi_1(X)], \tag{A.140}$$

$$(D) = m_{11}(X)\gamma_{11}^1(X)\pi_1(X) - E[m_{11}(X)\gamma_{11}^1(X)\pi_1(X)\pi_1(X)]. \tag{A.141}$$

Thus, combining these terms, we are left with

$$(A) + (B) + (C) + (D) = \frac{Z}{e(X)}\Big\{\delta A[T - m_{11}(X)] + m_{11}(X) A[\delta - \gamma_{11}^1(X)] \tag{A.142}$$

$$+ m_{11}(X)\gamma_{11}^1(X)[A - \pi_1(X)]\Big\} \tag{A.143}$$

$$+ m_{11}(X)\gamma_{11}^1(X)\pi_1(X) - E[m_{11}(X)\gamma_{11}^1(X)\pi_1(X)] \tag{A.144}$$

$$= \frac{Z}{e(X)}\Big\{\delta A T - m_{11}(X)\gamma_{11}^1(X)\pi_1(X)\Big\} \tag{A.145}$$

$$+ m_{11}(X)\gamma_{11}^1(X)\pi_1(X) - E[m_{11}(X)\gamma_{11}^1(X)\pi_1(X)]. \tag{A.146}$$

We may similarly write out the second term into four terms $(E) + (F) + (G) + (H)$ where

$$(E) = \frac{A Z \gamma_{11}^0(X)}{e(X)}[C - \omega_{11}(X)], \tag{A.147}$$

$$(F) = \frac{A Z \omega_{11}(X)}{e(X)}[(1 - \delta) - \gamma_{11}^0(X)], \tag{A.148}$$

$$(G) = \frac{Z \omega_{11}(X)\gamma_{11}^0(X)}{e(X)}[A - \pi(X)], \tag{A.149}$$

$$(H) = \omega_{11}(X)\gamma_{11}^{0}(X)\pi_1(X) - E[\omega_{11}(X)\gamma_{11}^{0}(X)\pi_1(X)]. \tag{A.150}$$

Combining and simplifying we are left with

$$(E) + (F) + (G) + (H) = \frac{Z}{e(X)}\left\{(1-\delta)AC - \omega_{11}(X)\gamma_{11}^{1}(X)\pi_1(X)\right\} \tag{A.151}$$

$$+\,\omega_{11}(X)\gamma_{11}^{1}(X)\pi_1(X) - E[\mu_{11}(X)\gamma_{11}^{0}(X)\pi_1(X)] \tag{A.152}$$

For the third term, we write $(I) + (J) + (K) + (L)$. For (I) we can write

$$(I) = \frac{\delta ZA}{e(X)}[(Y^* - C) - \lambda_{11}(X)] \tag{A.153}$$

$$(J) = \frac{\lambda_{11}(X)ZA}{e(X)}[(1-\delta) - \gamma_{11}(X)], \tag{A.154}$$

$$(K) = \frac{Z\lambda_{11}(X)\gamma_{11}^{0}(X)}{e(X)}[A - \pi_1(X)], \tag{A.155}$$

$$(L) = \kappa_{11}(X)\gamma_{11}^{0}(X)\pi_1(X) - E[\lambda_{11}(X)(X)\gamma_{11}^{0}(X)\pi_1(X)]. \tag{A.156}$$

Combining the results together and simplifying, we obtain

$$(I) + (J) + (K) + (L) = \frac{Z}{e(X)}\left\{A(1-\delta)(Y^* - C) - \lambda_{11}(X)\gamma_{11}^{0}(X)\pi_1(X)\right\} \tag{A.157}$$

$$+\,\lambda_{11}(X)\gamma_{11}^{0}(X)\pi_1(X) - E[\lambda_{11}(X)\gamma_{11}^{0}(X)\pi_1(X)]. \tag{A.158}$$

Therefore, the final influence function is denoted as

$$\mathbb{IF}(\psi_{1,1}) = \frac{Z}{e(X)}\left[\delta AT - m_{11}\gamma_{11}^{1}\pi_1\right] + m_{11}\gamma_{11}^{1}\pi_1 + \frac{Z}{e(X)}\left[(1-\delta)AC - \omega_{11}\gamma_{11}^{0}\pi_1\right] \tag{A.159}$$

$$+\,\omega_{11}\gamma_{11}^{0}\pi_1 + \frac{Z}{e(X)}\left[(1-\delta)A(Y^*-C) - \lambda_{11}\gamma_{11}^{0}\pi_1\right] + \lambda_{11}\gamma_{11}^{0}\pi_1 - \psi_{11} \tag{A.160}$$

A.3.3 Proof of Theorem 5.3.1

Letting $\psi_{num} = \psi_{11} + \psi_{10} - \psi_{01} - \psi_{00}$ and $\phi_{denom} = \phi_1 - \phi_0$ from Lee, Kennedy, and Mitra (2023) proof for Theorem 4.1 we can write the following:[70]

$$\hat{\Psi}_{LATE} - \Psi_{LATE} = \frac{\mathbb{P}_n\{\varphi^U_{num}(\hat{\eta})\}}{\mathbb{P}_n\{\varphi^U_{denom}(\hat{\eta})\}} - \frac{P\psi_{num}}{P\psi_{denom}} \tag{A.161}$$

$$= (\mathbb{P}_n\hat{\phi}_{denom})^{-1}\left[\mathbb{P}_n\{\varphi^U_{num}(\hat{\eta})\} - P\psi_{num} - \Psi_{LATE}\left(\mathbb{P}_n\{\varphi^U_{denom}(\hat{\eta})\} - P\phi_{denom}\right)\right] \tag{A.162}$$

$$= (\mathbb{P}_n\hat{\phi}_{denom})^{-1}\left[(\mathbb{P}_n - P)\left\{\varphi^U_{num}(\eta) - \Psi_{LATE}\varphi^U_{denom}(\eta)\right\}\right] \tag{A.163}$$

$$+\,(\mathbb{P}_n\hat{\phi}_{denom})^{-1} \tag{A.164}$$

$$\left[(\mathbb{P}_n - P)\left\{\varphi^U_{num}(\hat{\eta}) - \varphi^U_{num}(\eta)\right\} - \Psi_{LATE}\,(\mathbb{P}_n - P)\left\{\varphi^U_{denom}(\hat{\eta}) - \varphi^U_{denom}(\eta)\right\}\right] \tag{A.165}$$

$$+\,(\mathbb{P}_n\hat{\phi}_{denom})^{-1}\left[P\left\{\varphi^U_{num}(\hat{\eta}) - \psi_{num}\right\} - \Psi_{LATE}P\left\{\varphi^U_{denom}(\hat{\eta}) - \psi_{denom}\right\}\right] \tag{A.166}$$

$$= S^* + T_1 + T_2 \tag{A.167}$$

The S^* is asymptotically normal by the central limit theorem while $T_1 = o_P(n^{-1/2})$ under the condition that each estimator of the nuisance function are apart of the Donsker class, which is achieved by estimation via sample splitting.[28, 65, 70]. Thus, we must show that the remainder term $T_2 = o_P(n^{-1/2})$, which will additionally reveal the double robustness conditions of our estimators. As the numerator is composed of four alike terms, we will derive the bounds only for ψ_{11}, which itself is comprised of three terms that we derived in the previous section and can study the remainder of. For simplicity of notation, we omit X from the nuisance functions assuming so unless otherwise stated.

Beginning with the first term, we have

$$\mathbb{E}\left[\frac{Z}{\hat{e}}\left(\delta AT - \hat{m}_{11}\hat{\gamma}^1_{11}\hat{\pi}_1\right) + \hat{m}_{11}\hat{\gamma}^1_{11}\hat{\pi}_1 - m_{11}\gamma^1_{11}\pi_1\right] \tag{A.168}$$

$$= \mathbb{E}\left[\frac{Z}{\hat{e}}\left(\delta AT - \hat{m}_{11}\hat{\gamma}^1_{11}\hat{\pi}_1\right) + \mathbb{E}\left[\hat{m}_{11}\hat{\gamma}^1_{11}\hat{\pi}_1 - m_{11}\gamma^1_{11}\pi_1\right]\right] \tag{A.169}$$

Via iterated expectation on X, we have (A.170)

$$= \mathbb{E}\left[\frac{1}{\hat{e}}\mathbb{E}\left[Z\delta AT|X\right] - \hat{m}_{11}\hat{\gamma}^1_{11}\hat{\pi}_1\left(\frac{e}{\hat{e}}\right)\right] + \mathbb{E}\left[\hat{m}_{11}\hat{\gamma}^1_{11}\hat{\pi}_1 - m_{11}\gamma^1_{11}\pi_1\right] \tag{A.171}$$

$$= \mathbb{E}\left[\frac{e}{\hat{e}}m_{11}\gamma^1_{11}\pi_1 - \frac{e}{\hat{e}}\hat{m}_{11}\hat{\gamma}^1_{11}\hat{\pi}_1\right] + \mathbb{E}\left[\hat{m}_{11}\hat{\gamma}^1_{11}\hat{\pi}_1 - m_{11}\gamma^1_{11}\pi_1\right] \tag{A.172}$$

$$= \mathbb{E}\left[\left(\hat{m}_{11}\hat{\gamma}^1_{11}\hat{\pi}_1 - m_{11}\gamma^1_{11}\pi_1\right)\left(\frac{\hat{e}-e}{\hat{e}}\right)\right] \tag{A.173}$$

To better express the remainder term, we may further manipulate the term (A.174)

inside the parenthesis (A.175)

$$= \mathbb{E}\left[\left(\hat{m}_{11}\hat{\gamma}^1_{11}\hat{\pi}_1 - m_{11}\gamma^1_{11}\pi_1 + \hat{m}_{11}\hat{\gamma}^1_{11}\pi_1 - \hat{m}_{11}\hat{\gamma}^1_{11}\pi_1 + \hat{m}_{11}\gamma^1_{11}\pi_1 - \hat{m}_{11}\gamma^1_{11}\pi_1\right)\left(\frac{\hat{e}-e}{\hat{e}}\right)\right] \tag{A.176}$$

$$= \mathbb{E}\left[\hat{m}_{11}\hat{\gamma}^1_{11}\left(\hat{\pi}_1 - \pi_1\right)\left(\frac{\hat{e}-e}{\hat{e}}\right)\right] + \mathbb{E}\left[\pi_1\hat{m}_{11}\left(\hat{\gamma}^1_{11} - \gamma^1_{11}\right)\left(\frac{\hat{e}-e}{\hat{e}}\right)\right] \tag{A.177}$$

$$+ \mathbb{E}\left[\pi_1\gamma^1_{11}\left(\hat{m}_{11} - m_{11}\right)\left(\frac{\hat{e}-e}{\hat{e}}\right)\right] \tag{A.178}$$

By the Cauchy-Schwarz inequality (A.179)

$$\leq \frac{C_{\hat{m}_{11}} C_{\hat{\gamma}^1_{11}}}{C_{\hat{e}}} \|\hat{\pi}_1 - \pi_1\| \|\hat{e} - e\| + \frac{C_{\pi_1} C_{\hat{m}_{11}}}{C_{\hat{e}}} \|\hat{\gamma}^1_{11} - \gamma^1_{11}\| \|\hat{e} - e\| \tag{A.180}$$

$$+ \frac{C_{\pi_1} C_{\gamma^1_{11}}}{C_{\hat{e}}} \|\hat{m}_{11} - m_{11}\| \|\hat{e} - e\|. \tag{A.181}$$

The form of the second term is identical to that of the first term, thus we can omit intermediate steps:

$$\mathbb{E}\left[\frac{Z}{\hat{e}(X)}\left[(1-\delta)AC - \hat{\omega}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1\right] + \hat{\omega}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1 - \omega_{11}\gamma^0_{11}\pi_1\right] \tag{A.182}$$

$$= \mathbb{E}\left[(\hat{\omega}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1 - \omega_{11}\gamma^0_{11}\pi_1)(\frac{\hat{e} - e}{\hat{e}})\right] \tag{A.183}$$

$$\leq \frac{C_{\hat{\omega}_{11}} C_{\hat{\gamma}^0_{11}}}{C_{\hat{e}}} \|\hat{\pi}_1 - \pi_1\| \|\hat{e} - e\| + \frac{C_{\pi_1} C_{\hat{\omega}_{11}}}{C_{\hat{e}}} \|\gamma^1_{11} - \hat{\gamma}^1_{11}\| \|\hat{e} - e\| \tag{A.184}$$

$$+ \frac{C_{\pi_1} C_{\gamma^0_{11}}}{C_{\hat{e}}} \|\hat{\omega}_{11} - \omega_{11}\| \|\hat{e} - e\|. \tag{A.185}$$

For the third term, assuming a known Y^*, the form of the remainder term is as follows:

$$\mathbb{E}\left\{\frac{Z}{\hat{e}(X)}\left[(1-\delta)A(Y^* - C) - \hat{\lambda}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1\right] + \hat{\lambda}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1 - \lambda_{11}\gamma^0_{11}\pi_1\right\} \tag{A.186}$$

$$\leq \frac{C_{\hat{\lambda}_{11}} C_{\hat{\gamma}^0_{11}}}{C_{\hat{e}}} \|\hat{\pi}_1 - \pi_1\| \|\hat{e} - e\| + \frac{C_{\pi_1} C_{\hat{\lambda}_{11}}}{C_{\hat{e}}} \|\gamma^1_{11} - \hat{\gamma}^1_{11}\| \|\hat{e} - e\| \frac{C_{\pi_1} C_{\gamma^0_{11}}}{C_{\hat{e}}} \|\hat{\lambda}_{11} - \lambda_{11}\| \|\hat{e} - e\| \tag{A.187}$$

Thus, for the entire numerator, substituting $1 - \gamma^1_{11}$ for γ^0_{11} we can factor the final bounding on the remainder term as

$$\left(\frac{C_{\hat{m}_{11}} C_{\hat{\gamma}^1_{11}} + C_{\hat{\omega}_{11}} C_{\gamma^0_{11}} + C_{\hat{\lambda}_{11}} C_{\hat{\gamma}^0_{11}}}{C_{\hat{e}}}\right) \|\pi_1 - \hat{\pi}_1\| \|e - \hat{e}\| \tag{A.188}$$

$$+ \left(\frac{C_{\pi_1} C_{\hat{m}_{11}} + C_{\pi_1} C_{\hat{\omega}_{11}} + C_{\pi_1} C_{\hat{\lambda}_{11}}}{C_{\hat{e}}}\right) \|\gamma^1_{11} - \hat{\gamma}^1_{11}\| \|e - \hat{e}\| \tag{A.189}$$

$$+ \left(\frac{C_{\pi_1} C_{\gamma^1_{11}}}{C_{\hat{e}}} \right) \|m_{11} - \hat{m}_{11}\| \|e - \hat{e}\| \tag{A.190}$$

$$+ \left(\frac{C_{\pi_1} C_{\gamma^0_{11}}}{C_{\hat{e}}} \right) \|\omega_{11} - \hat{\omega}_{11}\| \|e - \hat{e}\| \tag{A.191}$$

which can be generalized to any combination of Z and A. Thus, we have double robustness in the following sense that one of the two must be correctly specified:

1. $\pi_Z(X)$ or $e(X)$ for $Z \in \{0, 1\}$
2. $m_{ZA}(X)$ or $e(X)$ for $Z \in \{0, 1\}$ and $A \in \{0, 1\}$
3. $\gamma^1_{ZA}(X)$ or $e(X)$ for $Z \in \{0, 1\}$ and $A \in \{0, 1\}$
4. $\omega_{ZA}(X)$ or $e(X)$ for $Z \in \{0, 1\}$ and $A \in \{0, 1\}$
5. $\lambda_{ZA}(X)$ or $e(X)$ for $Z \in \{0, 1\}$ and $A \in \{0, 1\}$

Therefore, if the estimation of each nuisance function has a rate of $o_P(n^{-1/4})$ or better (e.g. $\|m_{ZA} - \hat{m}_{ZA}\| = o_P(n^{-1/4})$) in each of these sets, then we have $T_2 = o_P(n^{-1/2})$.

Bounding the denominator is the same as any derivation for the ATE and is detailed in Kennedy (2023) Section 4.3 Example 2. It is sufficient to show, without loss of generality, the result for ϕ_1.

$$\mathbb{E}\left[\frac{Z}{\hat{e}}(A - \hat{\pi}_1) + \hat{\pi}_1 - \pi_1 \right] \leq \frac{1}{C_{\hat{e}}} \|\hat{e} - e\| \|\hat{\pi}_1 - \pi_1\| \tag{A.192}$$

This reveals that either the propensity scores, π_1 and π_0, or the instrument propensity score must be correctly specified. The term is $o_P(n^{-1/2})$ if $\|\hat{\pi}_Z - \pi_Z\| = o_P(n^{-1/4})$ and $|\hat{e} - e\| = o_P(n^{-1/4})$ or better.

Now that we have shown under which conditions $T_2 = o_P(n^{-1/2})$, we may now state the asymptotic distribution of $\hat{\Psi}_{LATE}$, which is defined by the remaining term S^*:

$$n^{1/2}\left(\hat{\Psi}_{LATE} - \Psi_{LATE}\right) \xrightarrow{d} N\left(0, E[\Gamma^2]\right) \tag{A.193}$$

where $\Gamma = \varphi^U_{num}(\eta) - \Psi_{LATE}\varphi^U_{denom}(\eta)$.

A.3.4 Remaining Bias When Plugging in for Y^*

For the third term, we must account for the fact that we are estimating Y^*, which we denote as $\hat{Y}^*$ in our plug-in estimator. The form of the remainder term is as follows:

$$\mathbb{E}\left\{\frac{Z}{\hat{e}(X)}\left[(1-\delta)A(\hat{Y}^* - C) - \hat{\lambda}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1\right] + \hat{\lambda}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1 - \lambda_{11}\gamma^0_{11}\pi_1\right\} \tag{A.194}$$

$$= \mathbb{E}\left\{\frac{Z}{\hat{e}(X)}\left[(1-\delta)A([\hat{Y}^* - Y^*] + [Y^* - C] - \hat{\lambda}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1\right] + \hat{\lambda}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1 - \lambda_{11}\gamma^0_{11}\pi_1\right\} \tag{A.195}$$

$$= \mathbb{E}\left[\frac{e\gamma^0_{11}\pi_1}{\hat{e}}(\hat{Y}^* - Y^*) + \frac{\hat{\lambda}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1 - \lambda_{11}\gamma^0_{11}\pi_1}{\hat{e}} + \hat{\lambda}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1 - \lambda_{11}\gamma^0_{11}\pi_1\right] \tag{A.196}$$

$$= \mathbb{E}\left[\frac{e\gamma^0_{11}\pi_1}{\hat{e}}(\hat{Y}^* - Y^*)\right] + \mathbb{E}\left[\frac{(\hat{e} - e)}{\hat{e}}(\hat{\lambda}_{11}\hat{\gamma}^0_{11}\hat{\pi}_1 - \lambda_{11}\gamma^0_{11}\pi_1)\right] \tag{A.197}$$

$$\leq \frac{C_e C_{\gamma^0_{11}} C_{\pi_1}}{C_{\hat{e}}}\left\|\hat{Y}^* - Y^*\right\| + \frac{C_{\hat{\lambda}_{11}} C_{\hat{\gamma}^0_{11}}}{C_{\hat{e}}}\|\hat{\pi}_1 - \pi_1\|\|\hat{e} - e\| + \frac{C_{\pi_1} C_{\hat{\lambda}_{11}}}{C_{\hat{e}}}\|\gamma^1_{11} - \hat{\gamma}^1_{11}\|\|\hat{e} - e\| \tag{A.198}$$

$$+ \frac{C_{\pi_1} C_{\gamma^0_{11}}}{C_{\hat{e}}}\|\hat{\lambda}_{11} - \lambda_{11}\|\|\hat{e} - e\| \tag{A.199}$$

with the first term resulting from the fact that we must impute Y^* with $\hat{Y}^*$.

A.3.5 Extended Simulation Results

For each DGM (i.e. linear and non-linear) and each of the 500 simulations, we generated 1000 observations and estimated nuisance functions with two-fold sample splitting, to mitigate sparsity. We further restricted the algorithms to exclude RF and RPART. Results that failed to run or produced extreme results, defined as a closed form variance of less than 3 on the log scale (exponentiated is 20 on the multiplicative scale, or 2000%), were excluded from the results with the total number of such iterations reported in the footnote of each DGM's table.

Table A.1: Simulation Results: Linear Data Generating Mechanism (n = 1000)

		Point Estimate	**Bias (%)**	**Monte Carlo SE**	**Model-Based SE**	**RMSE**	**Coverage (%)**
10% Cens.	2SPI	-0.737	7.84	0.359	0.506	0.361	92.8
	Proposed Method*	-0.778	2.75	0.578	1.341	0.578	99.5
20% Cens.	2SPI	-0.732	8.51	0.394	0.544	0.397	94.0
	Proposed Method	-0.747	6.52	0.500	0.806	0.501	99.6
30% Cens.	2SPI	-0.722	9.63	0.411	0.596	0.415	94.8
	Proposed Method	-0.746	6.75	0.592	0.768	0.597	98.6
40% Cens.	2SPI	-0.689	13.81	0.505	0.648	0.505	93.2
	Proposed Method	-0.746	6.75	0.744	0.876	0.746	99.2
50% Cens.	2SPI	-0.722	9.64	0.530	0.720	0.534	92.0
	Proposed Method	-0.825	3.13	0.967	1.071	0.966	98.0
60% Cens.	2SPI	-0.737	7.91	0.644	0.837	0.645	92.8
	Proposed Method	-0.896	12.00	1.551	1.516	1.558	97.0

*291 excluded iterations

Table A.2: Simulation Results: Non-linear Data Generating Mechanism (n = 1000)

		Point Estimate	**Bias (%)**	**Monte Carlo SE**	**Model-Based SE**	**RMSE**	**Coverage (%)**
10% Cens.	2SPI	-1.361	56.14	0.323	0.458	0.637	77.6
	Proposed Method*	-0.878	9.69	0.489	1.009	0.493	98.0
20% Cens.	2SPI	-1.365	56.79	0.347	0.486	0.661	80.3
	Proposed Method	-0.853	6.58	0.432	0.861	0.434	99.6
30% Cens.	2SPI	-1.404	60.38	0.408	0.546	0.771	83.6
	Proposed Method	-0.862	7.81	0.480	0.793	0.482	99.4
40% Cens.	2SPI	-1.409	61.25	0.431	0.604	0.798	83.5
	Proposed Method	-0.829	3.62	0.475	0.629	0.475	98.3
50% Cens.	2SPI	-1.468	66.83	0.470	0.667	0.915	84.4
	Proposed Method	-0.819	2.38	0.610	0.827	0.609	98.7
60% Cens.	2SPI	-1.500	70.73	0.531	0.757	1.015	85.6
	Proposed Method	-0.727	9.12	0.864	1.151	0.867	98.4

*Iterations excluded: 10% Cens. 296, 20% Cens. 192, 30% Cens. 168, 40% Cens. 141, 50% Cens. 122, 60% Cens. 112