Title
Towards Credible Causal Inference under Real-World Complications

Author
Huang, Melody

Publication Date
2023

Peer reviewed|Thesis/dissertation

Towards Credible Causal Inference under Real-World Complications

by

Melody Huang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Erin Hartman, Co-chair
Professor Avi Feller, Co-chair
Professor Samuel Pimentel
Professor Peng Ding
Professor Chad Hazlett

Spring 2023

Towards Credible Causal Inference under Real-World Complications

Abstract

Towards Credible Causal Inference under Real-World Complications

by

Melody Huang

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Erin Hartman, Co-chair

Professor Avi Feller, Co-chair

Causal inference provides tools for researchers to answer scientific and policy questions. The validity of estimated causal effects depends on many factors, from research design to the credibility of the underlying assumptions. The following dissertation addresses three aspects of causal inference: credibility, generalizability, and utility. Each chapter of the dissertation addresses the intersection of these three aspects of causality.

The first chapter examines credibility and generalizability, and introduces a sensitivity analysis framework for estimating externally valid causal effects. When estimating externally valid causal effects, researchers must leverage a conditional ignorability assumption to account for confounding effects from selection into the experimental sample. This assumption allows researchers to theoretically identify generalized (or transported) causal effects; however, like many assumptions in causal inference, this assumption is not testable, and in practice, can be untenable. The proposed framework allows researchers to quantify how much bias there can be in generalizing or transporting causal effects before the estimated effect substantively changes. The contributions in this chapter are three-fold. First, I show that the sensitivity parameters are scale-invariant and standardized, and introduce an estimation approach for researchers to simultaneously account for the bias in their estimates from omitting a moderator, as well as potential changes to their inference. Second, I propose several tools researchers can use to perform sensitivity analysis: (1) graphical and numerical summaries for researchers to assess how robust an estimated effect is to changes in magnitude as well as statistical significance; (2) a formal benchmarking approach for researchers to estimate potential sensitivity parameter values using existing data; and (3) an extreme scenario analysis. Finally, I demonstrate that the proposed framework can be easily extended to the class of doubly robust, augmented weighted estimators. The sensitivity analysis framework is applied to a set of Jobs Training Program experiments.

The second chapter focuses on utility and generalizability. While recent papers developed various weighting estimators for the population average treatment effect (PATE), many of these methods result in large variance because the experimental sample often differs substantially from the target population, and estimated sampling weights are extreme. In the following chapter, we propose *post-residualized weighting*, in which we use the outcome measured in the observational population data to build a flexible predictive model (e.g., machine learning methods) and residualize the outcome in the experimental data before using conventional weighting methods. We show that the proposed PATE estimator is consistent under the same assumptions required for existing weighting methods, importantly without assuming the correct specification of the predictive model. We examine the efficiency gains in the context of a set of jobs training program experiments, and find that using post-residualized weighting can result between a 5 - 25% reduction in variance over standard approaches.

The final chapter addresses credibility and utility. I introduce a new set of sensitivity models called the "variance-based sensitivity model". The variance-based sensitivity model characterizes the bias from omitting a confounder by bounding distributional differences that arise in the weights from omitting a confounder, with several notable innovations over existing approaches. First, the variance-based sensitivity model can be parameterized by an R^2 parameter that is both standardized and bounded. We introduce a formal benchmarking procedure that allows researchers to use observed covariates to reason about plausible parameter values in an interpretable and transparent way. Second, we show that researchers can estimate valid confidence intervals under the variance-based sensitivity model, and provide extensions for incorporating substantive knowledge about the confounder to help tighten the intervals. Last, we demonstrate, both empirically and theoretically, that the variance-based sensitivity model can provide improvements on both the stability and tightness of the estimated confidence intervals over existing methods. We illustrate our proposed approach on a study examining blood mercury levels using the National Health and Nutrition Examination Survey (NHANES).

The results from the dissertation collectively provide a broad range of methods for researchers to estimate causal effects more transparently and robustly.

To my parents.

Contents

List of Figures

List of Tables

Acknowledgments

They say it takes a village to raise a child, and a village and then some to raise an academic. The number of people in my village I owe thank yous to from the past five years is too many to possibly enumerate. This is in no way meant to be an exhaustive list.

First and foremost, I thank my advisor, Erin Hartman. I met Erin during my first quarter of graduate school, when I discovered the world of causal inference for the first time, and proceeded to spend most of my time in her office hours asking non-sequitur questions. Since then, Erin has helped me turn harebrained ideas into fully formed projects, taught me the importance of precise, mathematical thinking, and continues to inspire me to work on impactful and interesting research problems. Erin's mentorship through every stage of my Ph.D. has been instrumental in shaping not only who I am as a researcher, but also who I am as a person, and she has set a high bar for the type of advisor and professor I can only aspire to be. While counterfactual outcomes are neither knowable nor identifiable, I am certain that the counterfactual world in which I did not have a chance to work with Erin is a worse one, and I am forever indebted and grateful to her.

I would like to also thank my other committee members and collaborators. Sam Pimentel provided countless hours of advice and guidance during the last two years of my Ph.D. Getting to work with Sam and learning from his methodical approach to both research and writing has undoubtedly been a highlight of moving to Berkeley. Thank you to Avi Feller, Chad Hazlett, and Peng Ding, for the continual advice and guidance, not only about research, but also about navigating life and academia. Thank you to Naoki Egami and Luke Miratrix for numerous enlightening conversations on external validity. I look forward to collaborating again in the future. Thank you to Mark Handcock and Jennie Brand, who served as my committee members when I was still at UCLA. Both of them provided invaluable feedback on early stages of the work in this dissertation.

I have had the unique experience of spending my Ph.D. at two different institutions, where I have met some of the most wonderful friends and colleagues. To my office mates, Tiffany Tang and Austin Zane, thank you for always taking time to celebrate the highs and commiserate the lows of graduate school with me. I will miss our afternoon walks to every dessert place within a one mile radius of campus. Thank you to Dan Soriano, Benji Lu, Arisa Sadeghpour, Sizhu Lu, Yaxuan Huang, Lauren Liao, Emily Flanagan, and everyone else in the causal community at Berkeley. I am immensely lucky to get to be a part of such a wonderful group. Ciara Sterbenz, Sydney Kahmann, Leonard Wainstein, and many others in the UCLA Causal Inference Reading Group read multiple early iterations of my work. The work in this dissertation would not have been possible without their critical and detailed feedback. Furthermore, thank you to Jimmy Butler, Ashley Chiu, Tiffany Ding, Jeremy Goldwasser, Shuchi Goyal, Ella Hiesmayr, Ana Kenney, Conor Kresin, Drew Nguyen, Andy Shen, Tanvi Shinkre, Eric Xia, and so many others for making the Ph.D. not only survivable, but dare I say, also a fun experience.

My Ph.D. was generously funded by the National Science Foundation's Graduate Research Fellowship program. I am grateful for the support and opportunities that this pro-

vided. Any opinion, findings, and conclusions or recommendations expressed in this dissertation are those of the authors and do not necessarily reflect the views of the National Science Foundation. Furthermore, I was lucky to receive financial support from Erin Hartman, the San Francisco Bay Area Chapter of the American Statistical Association, and UCLA's Department of Statistics.

I am grateful to my support system outside of the immediate Berkeley and UCLA bubbles. Thank you to Addison Hu for a particularly memorable last year at Berkeley, and to Colleen Chan for dedicating time each day to work on the New York Times Crossword with me. My friends at home–Karen Chou, Ariela Feitelberg, Heather Lee, Amy Gimlin, Kyungna Kim–cheered me on when I first applied to Ph.D. programs, and have continued to cheer me on as I continue onward in academia.

Finally, I thank my parents, without whom, this would not have been possible. It is not lost upon me that doing a Ph.D. is an act of incredible privilege, and I would not have had this privilege without their hard work and sacrifice. Everything I have been able to accomplish has been possible because of them.

Chapter 1

Introduction

The credibility revolution has pushed for careful causal identification in the biomedical and social sciences. The move towards valid and transparent causal inference has highlighted the importance of research design and randomized control trials. Experiments are often considered the gold standard for estimating causal effects–within an experimental setting, researchers have full knowledge and control over the treatment assignment mechanism, thereby allowing for causal effects to be identified with relatively few assumptions. However, even in the context of a fully randomized experiment, challenges persist.

The proposed dissertation thus addresses three aspects of causal inference: *credibility*, *generalizability*, and *utility*. Estimating causal effects relies on an underlying set of identifying assumptions. When identifying assumptions are not met, the resulting estimates can be biased and lead researchers to erroneous conclusions. Thus, the *credibility* of estimated causal effects depend on these assumptions being met. *Generalizability* refers to the external validity of our causal estimates. Causal effects estimated across experiments have high degrees of internal validity. However, experiments are often conducted across convenience samples, which leads to the question of how well these effects generalize to larger (or different) target populations. Finally, even when causal effects are correctly identified and generalizable, issues such as instability and variance inflation limit the *utility* of the estimated causal quantities.

An overview of the prospectus follows. The first chapter introduces a sensitivity analysis for generalizability estimators. The second chapter, based on work with Naoki Egami, Erin Hartman, and Luke Miratrix, proposes a method to improve the efficiency of generalizability estimators by leveraging observational population-level data. The last chapter introduces a set of modified, variance-based sensitivity models that allow researchers to obtain more informative bounds under unobserved confounding, and is based on work with Sam Pimentel. The three chapters provide new methods that address the intersection of the three aforementioned aspects of causality to help researchers conduct more robust and generalizable causal inference under real-world complications.

1.1 Outline of Contributions

Sensitivity Analysis for Generalizing Experimental Results

The first chapter of the prospectus addresses *generalizability* and *credibility* when estimating causal effects. More specifically, external validity focuses on generalizing or transporting causal effects beyond an experimental sample. To account for the confounding effects of selection into the experimental sample, a common estimation approach is to re-weight the experimental data to control for distributional shifts in treatment effect moderators–i.e., pre-treatment covariates that drive propensity of selection into the experimental sample, as well as treatment effect heterogeneity–across the experimental sample and the population. In this chapter, I show that the bias for a weighted estimator from omitting a variable can be decomposed into a function of three different components: (1) a correlation term, which represents how the omitted confounder is related to the individual-level treatment effect; (2) an R^2 measure, which represents how imbalanced the omitted confounder is; (3) an observable scaling factor that increases or decreases the inherent sensitivity in the estimated effect to any omitted variable bias. The correlation measure and the R^2 value represent the omitted confounder's relationship with the outcome and the selection process, and serve as the sensitivity parameters. Furthermore, I introduce a percentile bootstrap approach for researchers to estimate valid confidence intervals under unobserved confounding, which allows researchers to assess the impact of an omitted variable on their statistical inference. Unlike existing methods, the sensitivity framework simultaneously accounts for both changes in point estimates and uncertainty estimates when omitting a variable.

To help researchers perform the sensitivity analysis in practice, I propose a set of sensitivity tools. While sensitivity tools have become more prevalent in outcome-based sensitivity analyses, such approaches do not currently exist for weighted estimators. More specifically, I introduce a suite of sensitivity summary measures, which includes a numerical summary measure (the robustness value), relative measures of sensitivity (minimum relative confounding strength), and graphical summaries (bias contour plots). Furthermore, I introduce formal benchmarking for weighted estimators. This allows researchers to estimate the parameter values for an omitted confounder with equivalent confounding strength to an observed covariate. In settings when researchers have a strong substantive understanding of important covariates, this is a powerful tool to argue about the plausibility of different sensitivity parameters, and provides much needed interpretability to conducting the sensitivity analysis. When used collectively, the proposed sensitivity tools allow researchers to not only quantitatively reason about sensitivity, but also transparently incorporate their substantive expertise into assessing how sensitive their estimates are to potential bias and changes in statistical significance.

Improving the Generalization of Experimental Results

The second chapter of the prospectus focuses on *generalizability* and *utility*. Under the aforementioned conditional ignorability assumption, recent literature has proposed the use of different estimation approaches to consistently estimate the PATE. However, because the experimental sample often differs substantially from the population, many of the generalization methods result in large variance and little statistical power. The resulting variance inflation often raises the question of whether the bias-variance trade-off is "worth" it (Miratrix et al., 2018), and limits the value of the PATE inference.

This chapter, which is based on joint work with Naoki Egami, Erin Hartman, and Luke Miratrix, introduces a novel method known as *post-residualized weighting* to improve efficiency in PATE estimation. Post-residualized weighting allows researchers to use outcomes measured in the observational population data to build a flexible predictive model and residualize the outcome in the experimental data, before using conventional weighting methods to recover the PATE. We show that post-residualized weighting results in consistent estimates of PATE, under the same assumptions required for existing weighting methods. This is true *regardless* of how the the predictive model is specified. Furthermore, we formalize the efficiency gains from post-residualized weighting, both theoretically and through simulations, and propose a diagnostic measure that allows researchers to check when post-residualized weighting will result in efficiency gain.

Improved Bounds for Sensitivity Models

Finally, the third chapter of the dissertation addresses *credibility* and *utility*. Recent developments in sensitivity analysis has seen the introduction of a variety of different approaches to assessing sensitivity to omitted confounding. However, many of these existing approaches rely on bounding a worst-case error. As a result, the constructed bounds on the potential bias for a sensitivity model are extremely conservative, limiting the utility of conducting the analysis.

In this chapter, based on joint work with Sam Pimentel, we introduce a one-parameter sensitivity model for weighted estimators. The contributions of this chapter are two-fold. First, we simplify the sensitivity framework introduced in Chapter 1 to a one-parameter setting, allowing researchers to conduct the sensitivity analysis using only the R^2 measure. We derive a closed form solution for the maximum bias that can occur for a fixed R^2 value, which we refer to as *optimal bias bounds*. Second, we formalize the theoretical relationship between our proposed method and existing sensitivity analyses in the literature. More formally, the proposed sensitivity analysis can be formulated as a bias maximization problem, with a constraint on the weighted average error from omitting a confounder. In contrast, existing approaches are constraining a global error. Using this dual formulation, we show that by moving away from a worst-case, global error, our proposed method results in more informative and stable bounds over current state-of-the-art sensitivity analyses.

Chapter 2

Sensitivity Analysis for Generalizing Experimental Results

2.1 Introduction

Randomized controlled trials (RCT's) provide researchers with a rich understanding of the treatment effect within an experimental sample. Because researchers have the ability to eliminate confounding by randomly assigning treatment in a controlled environment, experiments have a high degree of internal validity. However, problematically, a causal effect estimated from an RCT may not directly generalize to populations of interest when the experimental sample is not representative of the larger population. One prominent source of bias arises from distributional differences in treatment effect moderators—i.e., covariates that drive propensity of selection into the experimental sample, as well as treatment effect heterogeneity—between the experimental sample and the population (i.e., Imai et al. (2008); Cole and Stuart (2010); Olsen et al. (2013); see Egami and Hartman (2022) for discussion on alternative sources of bias). To properly generalize or transport the results from an experiment into a target population, researchers must either re-weight the experimental sample to be representative of the target population, or successfully model the treatment effect heterogeneity (Stuart et al. (2011); Kern et al. (2016)).

In practice, it is impossible to know whether the set of treatment effect moderators has been correctly identified. Researchers rely on the measured variables that are available in the sample and the population, and often assume that the observed covariates sufficiently capture the confounding effect. However, when moderators are omitted from estimation, the resulting point estimates will be biased. Existing sensitivity analyses in generalizability and transportability allow researchers to assess how robust their point estimates are to omitted confounders. However, many of the existing approaches require researchers to justify sensitivity parameters that may be arbitrarily large or small, and/or invoke parametric assumptions used to model the estimated bias from moderators (i.e., Nguyen et al. (2017); Nie et al. (2021); Dahabreh et al. (2019)).

In the following paper, we introduce a sensitivity analysis framework for unobserved moderators when using a weighted estimator for generalizing or transporting a causal effect. We focus on developing a sensitivity analysis for assessing bias in the point estimate of a causal effect, with discussions for how researchers may address changes in uncertainty from omitting a confounder in Section 2.6. The proposed framework builds on the sensitivity analysis literature from observational studies (Hong et al., 2021; Cinelli and Hazlett, 2020; Shen et al., 2011), as well as existing sensitivity analysis approaches for generalizing or transporting an estimated treatment effect (Nguyen et al., 2017; Dahabreh et al., 2019)), with several important innovations.

The paper provides several key contributions. First, we demonstrate that the bias of a weighted estimator may be decomposed into three bounded components, which serve as the sensitivity parameters in the proposed framework. We show that two of the components are standardized representations of the omitted variable's relationship with (1) the individual-level treatment effect and (2) the selection mechanism. The last component is related to how much inherent treatment effect heterogeneity and imbalance there is in the data. To help researchers account for estimation uncertainty, we introduce a bootstrapping-based approach for researchers to simultaneously consider not only potential bias that would occur from omitting a variable, but also changes in statistical inference.

Second, we introduce several sensitivity tools to help researchers conduct their sensitivity analysis in transparent and interpretable ways. While sensitivity tools have become more prevalent in outcome-based sensitivity analyses (i.e., Cinelli and Hazlett (2020); Zheng et al. (2021)), such approaches do not currently exist for weighted estimators, and are important in helping researchers interpret and reason about the potential bias from omitting a variable. The first approach is a graphical summary of sensitivity in the form of bias contour plots. The second is a numerical summary of sensitivity, and extends the robustness value from Cinelli and Hazlett (2020) for the weighted estimator setting. The robustness value serves as a summary measure for how much uncertainty there is in an estimate due to confounding from selection. Finally, we propose a formal benchmarking procedure that leverages observed covariates to posit plausible parameter values, and allows researchers to incorporate their substantive knowledge for the relative strength of moderators. We provide extensions of the sensitivity analysis and sensitivity tools for the class of augmented weighted estimators.

The paper is organized as follows. Section 2.2 introduces the notational framework, identifying assumptions, related literature, and the running example. Section 2.3 formalizes the proposed sensitivity analysis framework. In Section 2.4, we discuss three different tools that researchers can use to conduct the sensitivity analysis. Section 2.5 extends the framework for the class of augmented weighted estimators. Section 2.6 concludes. Proofs and extensions are provided in the Appendix.

2.2 Background

Notation and Set-Up

To begin, we define an infinite super-population from which the target population and the experimental sample are drawn. We define the target population as a sample of N units, drawn i.i.d. randomly from the target population. Following Buchanan et al. (2018), we define the experimental sample of n units as a potentially biased i.id. sample from the infinite super-population. Define S_i as an indicator for whether the unit is in the experimental sample (i.e., $S_i = 1$ when unit i is in the experiment, and $S_i = 0$ otherwise), and let $\mathcal{S}$ denote the set of indices for units included in the experimental sample.

Let T_i be a binary treatment assignment variable, where $T_i = 1$ for units assigned to treatment, and $T_i = 0$ for control. We assume full compliance, such that treatment assigned implies treatment received, and following the potential outcomes framework, define $Y_i(t)$ to be the potential outcome when unit i receives treatment $T_i = t$ where $t \in \{0, 1\}$ (Neyman, 1923; Rubin, 1974). Throughout the paper, we make the standard assumptions of no interference and that treatments are identically administered across all units (i.e., SUTVA, defined in Rubin (1980)). We assume a set of pre-treatment covariates $\mathcal{X}_i$ exists across both the experimental sample and the target population. Finally, we define the individual-level treatment effect τ_i as the difference between the potential outcomes of unit i:

$$\tau_i = Y_i(1) - Y_i(0)$$

Because we can never observe both potential outcomes of a specific unit, the individual-level treatment effect can never be observed in practice (Holland, 1986). To formalize, we assume that both $\{\tau_i, T_i, \mathcal{X}_i \mid S_i = 1\}_{i=1}^{n}$ and $\{\tau_i, T_i, \mathcal{X}_i \mid S_i = 0\}_{i=1}^{N}$ are drawn i.i.d. from an infinite super-population. When the experimental sample is a biased sample from the super-population, the sampling distributions for the experimental sample and the target population will not be the same (i.e., $\mathbb{P}(\tau_i, T_i, \mathcal{X}_i \mid S_i = 1) \neq \mathbb{P}(\tau_i, T_i, \mathcal{X}_i \mid S_i = 0)$).

The sample average treatment effect (SATE) is defined as as the average treatment effect across the experimental sample (i.e., $\tau_{\mathcal{S}} \equiv \mathbb{E}\{\tau_i \mid S_i = 1\}$). Assuming equal probability of treatment assignment, a simple difference-in-means estimator can be used to estimate the SATE:

$$\widehat{\tau}_{\mathcal{S}} \equiv \frac{1}{\sum_{i \in \mathcal{S}} T_i} \sum_{i \in \mathcal{S}} T_i Y_i - \frac{1}{\sum_{i \in \mathcal{S}} 1 - T_i} \sum_{i \in \mathcal{S}} (1 - T_i) Y_i, \tag{2.1}$$

where $\mathcal{S}$ represents the set of indices that correspond to units in the experimental sample (i.e., $\mathcal{S} = \{i : \ S_i = 1\}$). The population (or target) average treatment effect (PATE) is the causal quantity of interest, formally defined as:

$$\tau \equiv \mathbb{E}\{\tau_i \mid S_i = 0\} \tag{2.2}$$

where the expectation is taken over the realized target population.[1]

If the experimental sample is randomly drawn from the super-population, then $\widehat{\tau}_{\mathcal{S}}$ is an unbiased estimator for the PATE. However, in most settings, the experimental sample is not representative of the target population, and experimental results cannot be directly extrapolated to the population (Cole and Stuart, 2010; Olsen et al., 2013; Nguyen et al., 2017). In these settings, an additional identifying assumption is necessary to recover the PATE from the experimental sample:

Assumption 1 (Conditional Ignorability of Sampling)

$$\tau_i \perp\!\!\!\perp S_i \mid \mathcal{X}_i \tag{2.3}$$

Assumption 4 states that there exists some set of pre-treatment covariates $\mathcal{X}_i$ for which, conditioned on the set $\mathcal{X}$, the distribution of the individual-level treatment effects in the sample will be equivalent to the distribution of individual-level treatment effects in the population (Kern et al., 2016).[2] Egami and Hartman (2019) formally define the set of covariates $\mathcal{X}_i$ that allow the sampling mechanism to be conditionally independent from the treatment effect heterogeneity as the *separating set*.

In addition to Assumption 4, we assume positivity–conditional on $\mathcal{X}$, the probability of being included in the sample is non-zero (Rosenbaum and Rubin, 1983).

Assumption 2 (Positivity)

$$0 < \mathbb{P}(S_i = 1 \mid \mathcal{X}_i) < 1 \tag{2.4}$$

Violations of the positivity assumption result in attempting to generalize beyond the support of the data (see Stuart et al. (2011) and Tipton (2014) as two examples).

The most common approach to estimating the PATE is through a weighted estimator, where the observations in the experimental sample are re-weighted to resemble that of the target population (Stuart et al., 2011; Olsen et al., 2013):

$$\hat{\tau}_W = \frac{1}{n_1} \sum_{i \in \mathcal{S}} w_i T_i Y_i - \frac{1}{n_0} \sum_{i \in \mathcal{S}} w_i (1 - T_i) Y_i,$$

where the weights are defined as the sampling weights (i.e., $w_i \propto \mathbb{P}(S_i = 0 \mid \mathcal{X}_i)/\mathbb{P}(S_i = 1 \mid \mathcal{X}_i)$), n_1 and n_0 are the number of units in the treatment and control groups, respectively. Weights are often estimated using logistic regression (Cole and Stuart, 2010; Stuart et al., 2011; Buchanan et al., 2018). Recently, alternative weighting methods have been proposed, including more general balancing methods, such as entropy balancing, which adjust for distributional differences between the experimental sample and population observations without

[1] Researchers may instead, treat the estimand of interest as the average treatment effect, across the infinite super-population, instead of the realized population. The proposed sensitivity analysis will extend for both cases. We refer readers to Huang et al. (2021) for more discussion of this setting.

[2] For PATE identification, Assumption 4 can be relaxed for mean exchangeability. See Hartman et al. (2021) for more discussion.

explicitly modeling the underlying probability function (Särndal et al., 2003; Hainmueller, 2012; Josey et al., 2021; Lu et al., 2021).

In practice, researchers estimate the PATE under the assumption that they have correctly identified the full separating set $\mathcal{X}_i$. When Assumption 4 holds, the weighted estimators will be consistent estimators for PATE. However, violations of this assumption can result in biased estimation. The goal of this paper is to formalize a framework for assessing the sensitivity of the PATE estimates to a variable $\mathbf{U}_i$ being omitted from the separating set $\mathcal{X}_i$, which we refer to as a *confounder* (i.e., a variable missing from the separating set necessary for Assumption 4 to hold).[3]

Running Example: Jobs Training Partnership Act

To enrich our discussion of the sensitivity analysis, we will use a set of experiments conducted on the Jobs Training Partnership Act (JTPA) as a running example throughout the paper. The national JTPA study ran from 1987 to 1989, and assessed the effectiveness of the jobs training programs in helping individuals in the study find employment and increase their earnings. The original study was conducted across 16 different experimental sites. Individuals were first interviewed to determine whether or not they were eligible for JTPA services; those deemed eligible were assigned randomly to treatment and control using a 2:1 ratio. Individuals assigned to treatment were given access to JTPA services, while those assigned to control were told they were ineligible for the program. Following treatment assignment, a follow-up survey was conducted 18 months later, in which individuals were asked about their earnings (Bloom et al., 1993). We focus our analysis on the subset of adult women, the largest target group within the JTPA study.[4]

We leverage the nature of the original multi-site experiment to perform a benchmarking exercise for the sensitivity analysis. More specifically, we pick one of the 16 experimental sites and generalize the estimated effect of JTPA access on earnings from this site to the remaining 15 sites. The benchmark PATE is defined as the average treatment effect across the units in the other 15 experimental sites. This allows us to evaluate the actual error that is incurred from generalizing. To estimate the sample selection weights, we use entropy balancing across a set of pre-treatment covariates measured in the baseline survey (Hainmueller, 2012; Josey et al., 2021). Entropy balancing directly optimizes on covariate balance (i.e., the average covariate value in the experimental sample, versus the average covariate value in the target population) to estimate the weights, instead of first estimating the probabilities of selection into sample.[5] We weight on previous earnings, age, hourly wage, years of education, whether

[3]With some abuse of terminology, we use the term *confounder* instead of moderator to be consistent with other sensitivity frameworks. This idea is consistent with the notion that the set-up can be thought of as a missing data problem, in which the individual-level treatment effect τ_i is the 'outcome'.

[4]The estimated impacts of JTPA for the other target groups were not found to be statistically significant in the original study.

[5]The sensitivity analysis are agnostic to whether we use inverse-propensity score weights, or probability-like balancing weights. Zhao and Percival (2016) demonstrated that entropy balancing weights are implicitly

or not the individual graduated high school (or has a GED), whether or not the individual is married, and indicators for whether the individual is black or Hispanic.

To illustrate the sensitivity analysis, we examine the site of Coosa Valley, Georgia, which consists of 788 individuals, 519 of whom were assigned to treatment, with the remainder in control. The target population (i.e., the other 15 experimental sites) consists of 5,314 individuals. To showcase the performance of the sensitivity analysis across alternative experimental sites, we also conduct the sensitivity analysis on the other 15 experimental sites from JTPA. The results are provided in Appendix A.4.

	Unweighted	Weighted
Impact of JTPA access on earnings*	1.63 (0.95)	2.81 (1.21)

*-Estimates reported in thousands of USD

Table 2.1: Estimates of impact of JTPA access on earnings, generalizing the estimated effect from the site of Coosa Valley, Georgia to the other 15 experimental sites. Standard errors are reported in the parentheses.

The within-site estimated impact of JTPA access on earnings in Coosa Valley, Georgia is $\$1,630$. After weighting, the estimated impact of JTPA access earnings is $\$2,810$. In the following sections, we will introduce a sensitivity framework that allow researchers to assess how robust the estimate is to unobserved confounders.

2.3 Sensitivity Analysis for Weighted Estimators

In the following section, we will introduce a sensitivity analysis for weighted estimators when omitting a confounder from the weight estimation.

Bias of a Weighted Estimator when Omitting a Confounder

We consider the sensitivity of a weighted estimator to a confounder that has been omitted in the estimation of the weights. We formally define the minimum separating set as $\mathcal{X}_i = \{\mathbf{X}_i, \mathbf{U}_i\}$. In other words, for the weighted estimator to be unbiased, we would have had to estimate the weights using both $\mathbf{X}_i$ and $\mathbf{U}_i$; however, we omit $\mathbf{U}_i$. We write the weights estimated using just $\mathbf{X}_i$ as w_i, and the *ideal* weights that would have been estimated, had we included both $\mathbf{X}_i$ and $\mathbf{U}_i$, as w_i^*. We note that in defining the estimated and ideal weights as such, the proposed sensitivity framework will not account for settings in which researchers

estimating propensity score weights, with a modified loss function. See Wang and Zubizarreta (2020), Soriano et al. (2021), and Ben-Michael et al. (2021) for more discussion on the connection between balancing weights and inverse-propensity score weighting.

naively use uniform weights (i.e., $w_i = 1$ for all units), or settings in which the ideal weights are uniform (i.e., $w_i^* = 1$ for all units). Finally, we define ε_i as the linear error in the weights from omitting $\mathbf{U}_i$:

$$\varepsilon_i := w_i - w_i^*. \tag{2.5}$$

In the following sections, we will assume that researchers are estimating inverse propensity score weights. This allows us to examine a closed-form solution for the error term, which can help provide intuition. We provide extensions for balancing weights in Appendix A.1. Furthermore, we will assume that had researchers included $\mathbf{U}_i$, they would have been able to consistently estimate the weights.[6] Throughout, consistent with Shen et al. (2011) and Hong et al. (2021), we will refer to bias as the expectation of estimator minus the true value (i.e., true statistical bias).

The bias of a weighted estimator from omitting a confounder $\mathbf{U}_i$ is a function of ε_i and the degree to which this error term is related to treatment effect heterogeneity. We formalize this in the following theorem:

Theorem 2.3.1 (Bias of a Weighted Estimator from Omitting a Confounder) *Assume $Y_i(1) - Y_i(0) \perp\!\!\!\perp S_i \mid \{\mathbf{X}_i, \mathbf{U}_i\}$. Let w_i be the weights estimated using only $\mathbf{X}_i$, and let w_i^* be the (correct) weights, obtained using $\{\mathbf{X}_i, \mathbf{U}_i\}$. The bias of a weighted estimator from using w_i instead of w_i^* is given as:*[7]

$$Bias(\hat{\tau}_W) = \begin{cases} \rho_{\varepsilon,\tau} \sqrt{\dfrac{R_\varepsilon^2}{1 - R_\varepsilon^2} \cdot var_{\mathcal{S}}(w_i) \cdot \sigma_\tau^2} & \text{if } R_\varepsilon^2 < 1 \qquad (2.6) \\ \rho_{\varepsilon,\tau} \sqrt{var_{\mathcal{S}}(w_i^*) \cdot \sigma_\tau^2} & \text{if } R_\varepsilon^2 = 1, \qquad (2.7) \end{cases}$$

where $\rho_{\varepsilon,\tau}$ is the correlation between ε_i and τ_i (i.e., $\rho_{\varepsilon,\tau} := cor_{\mathcal{S}}(\varepsilon_i, \tau_i)$), R_ε^2 is the ratio of variances between ε_i and w_i^* (i.e., $R_\varepsilon^2 := var_{\mathcal{S}}(\varepsilon_i)/var_{\mathcal{S}}(w_i^*)$), and σ_τ^2 is the variance of τ_i. Derivation is provided in Appendix A.2.

Theorem 2.3.1 identifies the three drivers of bias in a weighted estimator when a confounder is omitted in the weight estimation: (1) the remaining imbalance in the omitted confounder (i.e., R_ε^2), (2) the correlation between ε_i and the individual-level treatment effect (i.e., $\rho_{\varepsilon,\tau}$), and (3) a scaling factor, represented by the product between the variance in the estimated weights and the amount of treatment effect heterogeneity (i.e., $\text{var}_{\mathcal{S}}(w_i) \cdot \sigma_\tau^2$). Theorem 2.3.1 provides a natural foundation for a sensitivity analysis. In particular, R_ε^2 and

[6] Misspecification concerns can also be addressed with the sensitivity analysis if researchers can write the error as an omitted variable problem. For example, if a linear probability model is used, $\mathbf{U}_i$ can include non-linear functions of $\mathbf{X}_i$ that matter for modeling selection.

[7] The derived bias expression will be the *exact* bias when researchers are using a Horvitz-Thompson style weighted estimator. In cases when researchers are using a stabilized weighted estimator, there will be finite-sample bias of order $o_{\mathbb{P}}(1/n)$. However, the finite-sample bias will be dominated by the bias incurred from omitting a confounder from the weights (see Miratrix et al. (2013), Rosenbaum (2010a), Lunceford and Davidian (2004) for more discussion).

$\rho_{\varepsilon,\tau}$ will serve as our sensitivity parameters, while the scaling factor can be conservatively bounded using observed data. We show in Appendix A.1 that a similar bias decomposition holds for augmented weighted estimators, and provide an extension of the sensitivity analysis framework.

Remark. For the setting in which $R^2_\varepsilon = 1$, the bias decomposition in Equation 2.6 will be undefined, and researchers will have to use an alternative decomposition, given in Equation 2.7. However, we note that in order for the R^2_ε parameter to be equal to 1, researchers would have to include covariates $\mathbf{X}_i$ that are exactly orthogonal to the confounder $\mathbf{U}_i$, *and* completely unrelated to the selection process. Thus, while it is mathematically possible to be in this setting, it is practically implausible.[8]

Interpreting the Parameters

In the following subsection, we discuss the interpretation of each of the sensitivity parameters. Instead of relying on unbounded sensitivity parameters (i.e., Shen et al. (2011); Hong et al. (2021)), the proposed sensitivity analysis uses a correlation value and an R^2 measure to represent how related the confounder is to the individual-level treatment effect and the selection mechanism. Both of these parameters are scale invariant, which can make it easier for researchers to reason about plausible sensitivity parameters, especially when paired with the sensitivity tools introduced in Section 2.4.

Variation in Ideal Weights Explained by ε_i (R^2_ε)

The R^2_ε term is defined as the ratio of variances between the error term and the ideal weights. In the following lemma, we show that the variation in the true weights can be decomposed into two components: variation explained by the estimated weights, and the variation explained by the error term ε_i; therefore, R^2_ε is bounded on the interval of 0 and 1. As such, we can interpret R^2_ε as the proportion of variation in the true weights explained by the error term ε_i.

Lemma 2.3.1 (Variance Decomposition of w^*_i)
For inverse propensity score weights, the variance of the true weights w^*_i can be decomposed linearly into two components:

$$var_{\mathcal{S}}(w^*_i) = var_{\mathcal{S}}(w_i) + var_{\mathcal{S}}(\varepsilon_i) \implies \frac{var_{\mathcal{S}}(w_i)}{var_{\mathcal{S}}(w^*_i)} + \underbrace{\frac{var_{\mathcal{S}}(\varepsilon_i)}{var_{\mathcal{S}}(w^*_i)}}_{:=R^2_\varepsilon} = 1$$

Therefore, R^2_ε is bound between 0 and 1.

[8] We also note that an alternative setting in which R^2_ε could equal 1 is if researchers posit naive, uniform weights. However, our definition for the estimated and ideal weights rules out this scenario.

The results of Lemma 2.3.1 follow from the fact that we may recover the estimated weights from projecting the ideal weights onto the space of the observed covariates $\mathbf{X}$. This is a general property of inverse propensity score weights. In Appendix A.1, we provide an extension of this result for a class of balancing weights.

As the amount of residual imbalance in the omitted confounder increases, R^2_ε will increase. If the residual imbalance of the omitted confounder (i.e., imbalance in $\mathbf{U}_i$, conditional on $\mathbf{X}_i$) is relatively small, then the estimated weights will be close to the true weights. As a result, R^2_ε will be close to 0. In contrast, if the residual imbalance of the omitted confounder is large, then much of the variation in w_i^* will be driven by ε_i, and R^2_ε will be large, approaching 1. Consider our running example. The original study cited the latent variable of motivation as a potential confounder (Bloom et al., 1993). While we cannot include motivation directly in the weights, we have included variables such as education and previous earnings, which are likely correlated to motivation. If, by controlling for variables such as education and previous earnings, we have accounted for much of the imbalance in motivation, then including motivation into the weight estimation should result in weights w_i^* similar to the estimated weights w_i, and R^2_ε will be relatively small (i.e., R^2_ε is close to zero).

Correlation between ε_i and τ_i ($\rho_{\varepsilon,\tau}$)

The correlation between ε_i and the individual-level treatment effect is a standardized measure for how much treatment effect heterogeneity $\mathbf{U}_i$ explains. When $\rho_{\varepsilon,\tau}$ is very high (i.e., $\rho_{\varepsilon,\tau} \approx 1$), then units with a large τ_i are overweighted ($w_i > w_i^*$ corresponds to large τ_i). Thus, in these settings, there will be positive bias. Conversely, if $\rho_{\varepsilon,\tau} \approx -1$, the opposite would be true—we underweight units with a large individual-level treatment effect, which results in a negatively biased estimated PATE. If the correlation between the error term and the individual-level treatment effect were close to zero, then the imbalance in the omitted confounder $\mathbf{U}_i$ is not related to treatment effect heterogeneity, and as such, omitting $\mathbf{U}_i$ would not result in much bias.

While $\rho_{\varepsilon,\tau}$ is inherently bounded on the interval $[-1, 1]$, we can decompose $\rho_{\varepsilon,\tau}$ as a function of R^2_ε to restrict the set of feasible correlation values to a tighter range.

Lemma 2.3.2 (Correlation Decomposition)
The correlation between ε_i and the individual-level treatment effects is bound on the following range:

$$-\sqrt{1 - cor^2_{\mathcal{S}}(w_i, \tau_i)} \leq \rho_{\varepsilon,\tau} \leq \sqrt{1 - cor^2_{\mathcal{S}}(w_i, \tau_i)}$$

Lemma 2.3.2 demonstrates that $\rho_{\varepsilon,\tau}$ will be bounded between $\pm\sqrt{1 - \text{cor}^2_{\mathcal{S}}(w_i, \tau_i)}$. If the estimated weights w_i can explain most of the variation in treatment effect heterogeneity, the additional variation that can be explained by adding in the omitted confounder must be small.

The correlation between the estimated weights and τ_i will take on large values when (1) the covariates contained in w_i explain much of the treatment effect heterogeneity, *and* (2) the covariates that explain the treatment effect heterogeneity are imbalanced across the population and the experimental sample. To help provide intuition for this, consider our running example. If access to JTPA services was only effective for women who graduated high school, then if educational attainment were imbalanced across the experimental sample and the population, estimating weights on educational attainment would result in a large $|\text{cor}_{\mathcal{S}}(w_i, \tau_i)|$ value. However, if educational attainment were not very imbalanced across the experimental sample and population, even though educational attainment explains much of the variation in the treatment effect heterogeneity, $\text{cor}_{\mathcal{S}}(w_i, \tau_i)$ will low. In such a scenario, the true $\rho_{\varepsilon,\tau}$ value should also be small; however, this would not be reflected in the bound.

Remark on Estimating $\text{cor}_{\mathcal{S}}(w_i, \tau_i)$: In practice, it is not possible to directly calculate $\text{cor}_{\mathcal{S}}(w_i, \tau_i)$, since τ_i is unidentified. Researchers may conservatively estimate the correlation of w_i and τ_i by using $\text{cov}_{\mathcal{S}}(w_i, Y_i(1))$ and $\text{cov}_{\mathcal{S}}(w_i, Y_i(0))$, which is identified by randomization. More specifically:

$$\widehat{\text{cor}}_{\mathcal{S}}(w_i, \tau_i) = \frac{\widehat{\text{cov}}_{\mathcal{S}}(w_i, Y_i(1)) - \widehat{\text{cov}}_{\mathcal{S}}(w_i, Y_i(0))}{\sqrt{\sigma^2_\tau \cdot \widehat{\text{var}}_{\mathcal{S}}(w_i)}}$$

Because $\widehat{\text{cor}}_{\mathcal{S}}(w_i, \tau_i)$ is a function of the variation in the individual-level treatment effect (i.e., σ^2_τ), if researchers use a more conservative estimate of σ^2_τ, this will subsequently lead to a more conservative estimate on $\text{cor}_{\mathcal{S}}(w_i, \tau_i)$, and by extension, a more conservative estimate for the bounds on $\rho_{\varepsilon,\tau}$. See Section 2.3 for details on specifying σ^2_τ.

Scaling Factor ($\text{var}_{\mathcal{S}}(w_i) \cdot \sigma^2_\tau$)

The last term in the bias decomposition is a scaling factor, made up of the product of the variance of the estimated weights and the variance in the individual-level treatment effect (i.e., σ^2_τ). This term is not related to the confounder, and is instead, intrinsic to the inherent data generating process. However, it does increase or decrease our exposure to bias from omitting a confounder.

We consider both terms in the scaling factor. The first term, $\text{var}_{\mathcal{S}}(w_i)$, corresponds to how much inherent imbalance there is in the observed covariates between the experimental sample and the target population. As the variance of our estimated weights increases, this implies that the weights are accounting for larger distributional differences between the experimental sample and the target, and the potential for bias also increases.

The second term is the magnitude of treatment effect heterogeneity (σ^2_τ). This is related to in Meng (2018)'s 'problem difficulty.' More specifically, when there exists a large degree of treatment effect heterogeneity, the task of recovering the PATE becomes harder, and even small imbalances in the confounders can result in a large degree of bias. When there is less treatment effect heterogeneity, we have more leeway in mis-specifying the weights without

incurring large amounts of bias. In the most extreme case of no treatment effect heterogeneity, we need not adjust for any confounders to have unbiased estimation. Because treatment effect heterogeneity is inherent to the underlying data generating process, regardless of what variables are included in the weights, σ_τ^2 is fixed. We apply the results from Ding et al. (2019) to show that σ_τ^2, while unidentifiable, can be bounded using Fréchet-Hoeffding bounds (Hoeffding, 1941; Fréchet, 1951) using the observed data, with opportunities for tighter bounds in cases when researchers are willing to invoke additional assumptions about the potential outcomes (see Appendix A.1 for more details).

In general, to estimate a conservative upper bound for the scaling factor, researchers can directly estimate $\text{var}_{\mathcal{S}}(w_i)$ and an upper bound for σ_τ^2 (which we denote as $\sigma_{\tau,\max}^2$).

Accounting for Changes in Inference

In practice, researchers are concerned about not only the resulting bias from omitting a confounder, but also potential changes to their inference. In particular, not only will the estimated effect change in magnitude due to bias from omitting a variable, but the estimated uncertainty associated with an estimate will also change. In particular, weighting on additionally imbalanced variables can result in an inflation in variance. However, estimating the variance inflation factor in the weighted estimator setting is challenging. In particular, there exists higher-order dependencies between the error term ε_i and the individual-level treatment effect τ_i that are not represented by the existing sensitivity parameters.

Instead, we leverage the results from Huang and Pimentel (2022) to estimate confidence intervals for a specified set $\{R_\varepsilon^2, \rho_{\varepsilon,\tau}, \sigma_\tau^2\}$ using a percentile bootstrap. More formally, for any set of $\{R_\varepsilon^2, \rho_{\varepsilon,Y}, \sigma_\tau^2\}$ values, researchers can compute the associated confidence intervals of the adjusted point estimate. This approach allows researchers to simultaneously account for the bias that occurs from omitting a confounder, as well as the changes in uncertainty, without introducing additional sensitivity parameters. We provide details in Appendix A.4. Researchers can compute the confidence intervals for the adjusted point estimates for a grid of $\{R_\varepsilon^2, \rho_{\varepsilon,Y}, \sigma_\tau^2\}$ values. Then, using the estimated confidence intervals, researchers can find the minimum bias that can occur before the intervals around the adjusted point estimate contain the null estimate, which would imply that omitting a confounder resulted in a change in the statistical significance of an estimated effect.

Summary of the Sensitivity Framework

To summarize the sensitivity analysis framework thus far, we have parameterized the bias of a weighted estimator when omitting a confounder in the estimation of the weights with the following components: (1) an R^2 measure that is bounded between 0 and 1 (i.e., R_ε^2), (2) the correlation between the error term ε_i and the individual-level treatment effect (i.e., $\rho_{\varepsilon,\tau}$), and (3) variation in the individual-level treatment effect (i.e., σ_τ^2). We summarize this below.

Summary of Sensitivity Framework for Weighted Estimators

Step 1. Estimate an upper bound for σ_τ^2 (i.e., $\sigma_{\tau,\max}^2$).

Step 2. Using $\sigma_{\tau,\max}^2$, estimate $\widehat{\text{cor}}_{\mathcal{S}}^2(w_i, \tau_i)$ as a bound for $\text{cor}_{\mathcal{S}}^2(w_i, \tau_i)$.

Step 3. Vary $\rho_{\varepsilon,\tau}$ from $-\sqrt{1 - \widehat{\text{cor}}_{\mathcal{S}}^2(w_i, \tau_i)}$ to $\sqrt{1 - \widehat{\text{cor}}_{\mathcal{S}}^2(w_i, \tau_i)}$.

Step 4. Vary R_ε^2 from the range of $[0, 1)$.

Step 5. Evaluate the bias (Theorem 2.3.1) and uncertainty (Table A.1 in Appendix A.1).

2.4 Tools for Sensitivity Analysis

In the following section, we provide different tools that researchers can use to help understand the degree of sensitivity associated with their estimated effects. We introduce two summary measures: (1) a graphical representation of sensitivity, in the form of bias contour plots, and (2) a numerical measure, referred to as a robustness value, which summarizes how much confounding must be present for an omitted confounder to result in change in the estimated effect. To assess the plausibility of parameter values, we introduce a formal benchmarking approach that allows researchers to use observed covariates to calibrate their understanding of potential sensitivity parameters. In Appendix A.1, we provide an extreme scenario analysis that evaluates an upper bound for the bias that in the extreme case that ε_i is maximally correlated with the individual-level treatment effect.

Summary Measures of Sensitivity

We provide two approaches for researchers to summarize the sensitivity in their point estimates. The first approach is graphical, while the second is a numerical measure.

Graphical Summary: Contour Plots

A simple way to summarize and visualize the sensitivity of the point estimates is through bias contour plots (see Figure 2.1). To generate the plots, the y-axis represents values that the correlation term can take on (i.e., the estimated range from Lemma 2.3.2), and the x-axis represents values of R_ε^2 across the interval of $[0, 1)$.

Furthermore, we recommend researchers shade in the "killer confounder" region. The killer confounder region represents the set of $\{R_\varepsilon^2, \rho_{\varepsilon,\tau}\}$ values for which we expect, given an omitted confounder in this set, the bias is large enough to substantively alter the estimated effect. Throughout the paper, we consider two different types of killer confounders: (1) a confounder that is strong enough to result in a change in the directional sign of a treatment

effect, or bring the treatment effect to zero; and (2) a confounder that alters the statistical significance of our estimated effect. If the killer confounder region is large, then there exists a greater degree of sensitivity to violations to the conditional ignorability assumption. If the region is small, there is less sensitivity.

Numerical Summary: Robustness Value

In practice, justifying whether the killer confounder region is large or small can be challenging. As such, we propose the robustness value as a standardized, numerical summary of how sensitive a point estimate is to confounders that may change the substantive interpretation of an estimated treatment effect. This extends the robustness value proposed by Cinelli and Hazlett (2020) for weighted estimators.

The robustness value measures how strong a confounder must be in order for the bias to equal $100 \times q\%$ of the estimated effect:

$$RV_q = \frac{1}{2}\left(\sqrt{a_q^2 + 4a_q} - a_q\right), \quad \text{where } a_q = \frac{q^2 \cdot \hat{\tau}_W^2}{\sigma^2_{\tau,\max} \cdot \text{var}_{\mathcal{S}}(w_i)} \tag{2.8}$$

Evaluating the robustness value at $q = 1$ provides a measure for minimum confounding strength in order for the bias to equal the point estimate, which would result in the point estimate being equal to zero. RV_q is interpreted as the minimum amount of variation in treatment effect heterogeneity *and* the true sample selection weights w_i^*, that the error term ε_i must explain (i.e., $\rho^2_{\varepsilon,\tau} = R^2_\varepsilon \geq RV_q$) for the bias to be $q \times 100\%$ that of the point estimate. Similarly, we may evaluate the robustness value associated with the minimum confounding strength of a confounder that results in the estimated effect changing its statistical significance. We denote this as RV_α. More details and derivations are provided in Appendix A.3.

A key property of the robustness value is that it exists on a scale from 0 to 1. When the robustness value is close to 1, then this implies that ε_i must explain close to 100% of the variation in both τ_i and w_i^* for the confounder to be a killer confounder. In contrast, if the robustness value is close to zero, then if ε_i is able to explain a small amount of variation in both τ_i and w_i^*, the error from omitting a confounder will be strong enough to result in a killer confounder. While the robustness value cannot rule out the possibility of a killer confounder, it can help researchers discuss the plausibility of such a confounder. Like standard error, which summarizes our uncertainty due to sampling error, the robustness value serves as a summary measure of our uncertainty due to systematic bias.

Geometric Connection to Bias Contour Plots. The robustness value is connected to the boundary of the killer confounder region. For example, if researchers are considering just changes to their point estimates, the killer confounder region would be defined by the part of the plot in which the bias is large enough to reduce the estimate to zero. The point on the boundary for which $\rho^2_{\varepsilon,\tau} = R^2_\varepsilon$ is representative of the robustness value $RV_{q=1}$. The same

interpretation applies if researchers define the killer confounder region with respect to the minimum bias associated with a change in the statistical significance of their estimated effect. The boundary of the killer confounder region represents the set of *all* potential parameter values associated with a killer confounder. As such, we recommend researchers report both the robustness value and the bias contour plots when performing sensitivity analysis.

Example: Sensitivity Summary Measures in JTPA

We illustrate the proposed sensitivity summary measures in our running example. To conduct the sensitivity analysis, we use an estimated bound of 29.01 for σ^2_τ. (Details on how $\hat{\sigma}^2_{\tau,\max}$ was chosen is provided in Appendix A.4.) Table 2.2 provides the different sensitivity statistics:

	Unweighted	Weighted	$RV_{q=1}$	$RV_{\alpha=0.05}$
Impact of JTPA access on earnings*	1.63 (0.95)	2.81 (1.21)	0.56	0.08

$\hat{\sigma}^2_{\tau,\max} = 29.01$; $\widehat{\text{cor}}_{\mathcal{S}}(w_i, \tau_i) = 0.37$, *-Estimates reported in thousands of USD

Table 2.2: Summary of point estimates and sensitivity statistics.

We see that the estimated robustness value is $RV_{q=1} = 0.56$, which implies that the error in the weights for omitting a confounder (i.e., ε_i) must explain 56% of the variation in the individual-level treatment effect, as well as 56% of the variation in the ideal weights in order for the treatment effect to be brought down to 0. Whether or not the robustness value is large or small depends on whether researchers believe that it is plausible for the error in omitting a confounder to explain 56% of the variation in both the ideal weights and the treatment effect heterogeneity. The estimated robustness value for a confounder that alters the statistical significance of an estimated effect is $RV_{\alpha=0.05} = 0.08$, which is much lower. As such, if the error from omitting a confounder explains 8% of the variation in the ideal weights and the treatment effect heterogeneity, then the estimated effect will no longer be statistically significant.

We also examine a bias contour plot, in which we shade in blue the part of the plot for which the bias is large enough to reduce the estimated impact of JTPA access on earnings to zero or negative (see Figure 2.1). The boundary of this region visualizes the full set of $\{R^2_\varepsilon, \rho_{\varepsilon,\tau}\}$ that corresponds to a confounder strong enough to reduce the estimated treatment effect to zero. For example, an omitted confounder that results in an error term that explains less than half the variation in the ideal weights (i.e., $R^2_\varepsilon = 0.46$), but explains a large amount of variation in the individual-level treatment effect (i.e., $\rho_{\varepsilon,\tau} = 0.86$) would reduce our estimate to zero. Similarly, a confounder that results in an error term that explains a large amount of the variation in the ideal weights (i.e., $R^2_\varepsilon = 0.91$), but a small portion of the variation in the individual-level treatment effect (i.e., $\rho_{\varepsilon,\tau} = 0.25$) would also reduce our estimate to zero.

We additionally shade in light gray the region of the plot in which the point estimate will still be positive, but the estimated effect will no longer be statistically significant. We see that this dominates a much larger part of the plot.

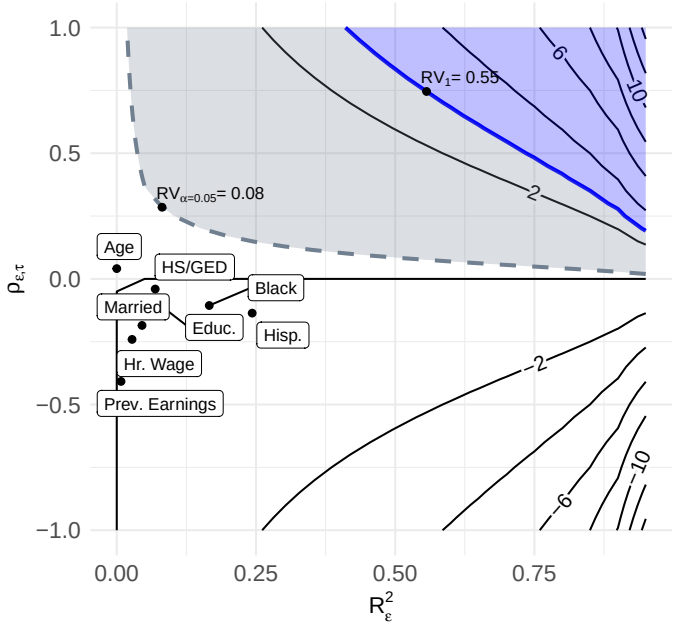

Figure 2.1: Bias Contour Plot for Coosa Valley, Georgia. The blue region represents the region for which the estimated effect will be equal to zero, or become negative. The gray region represents the part of the plot in which the estimated effect will no longer be statistically significant. To aid in our discussion, we use formal benchmarking (introduced in Section 2.4) to estimate the parameter values for an omitted confounder with similar confounding strength as an observed covariate.

Formal Benchmarking to Infer Reasonable Parameters

A challenge in sensitivity analysis is positing reasonable values for the sensitivity parameters to take on. Furthermore, justifying whether the killer confounder region of a bias contour plot, or the robustness value, is large or small can be challenging in practice. In the following subsection, we introduce a formal benchmarking approach for researchers to use observed covariates to calibrate their understanding of plausible parameter values using relative strength.

To begin, let $\mathbf{X}^{(j)}$ be an observed covariate (i.e., $\mathbf{X}^{(j)} \in \{\mathbf{X}\}$, and $\mathbf{X} \in \mathbb{R}^{n \times p}$, $j \in \{1, ..., p\}$). Define $\varepsilon_i^{-(j)}$ as the error term that compares the weights estimated using all

covariates $\mathbf{X}_i$ with the weights estimated using all the covariates, except for $\mathbf{X}_i^{(j)}$:

$$\varepsilon_i^{-(j)} := w_i^{-(j)} - w_i, \tag{2.9}$$

where $w_i^{-(j)}$ is the set of weights estimated using all the covariates $\mathbf{X}_i$, except for $\mathbf{X}_i^{(j)}$, and w_i is the set of weights estimated using all available covariates $\mathbf{X}_i$. We define the amount of confounding strength an omitted confounder has by how much variation ε_i explains in the ideal weights w_i^* and the individual-level treatment effect τ_i. Thus, to obtain formal benchmarks, we posit the amount of variation explained in w_i^* and τ_i by ε_i, in comparison to $\varepsilon_i^{-(j)}$. More formally, define:

$$k_\sigma = \frac{\text{var}_\mathcal{S}(\varepsilon_i)/\text{var}_\mathcal{S}(w_i^*)}{\text{var}_\mathcal{S}(\varepsilon_i^{-(j)})/\text{var}_\mathcal{S}(w_i^*)}, \qquad k_\rho = \frac{\text{cor}_\mathcal{S}(\varepsilon_i, \tau_i)}{\text{cor}_\mathcal{S}(\varepsilon_i^{-(j)}, \tau_i)}, \tag{2.10}$$

where the numerators (i.e., $\text{var}_\mathcal{S}(\varepsilon_i)/\text{var}_\mathcal{S}(w_i^*)$ and $\text{cor}_\mathcal{S}(\varepsilon_i, \tau_i)$) correspond to the sensitivity parameters introduced in Section 2.3. k_σ represents how much relative variation in the true sample selection weights w_i^* the residual imbalance in $\mathbf{U}$ (i.e., ε_i) explains, relative to the residual imbalance in $\mathbf{X}_i^{-(j)}$ (i.e., $\varepsilon_i^{-(j)}$). If the residual imbalance in the omitted confounder $\mathbf{U}_i$ is greater than the observed residual imbalance in the covariate $\mathbf{X}_i^{(j)}$, then we expect $k_\sigma > 1$. k_ρ represents how correlated the individual-level treatment effect and the error term ε_i are, relative to $\varepsilon_i^{-(j)}$. k_σ and k_ρ intuitively represent the relative confounding strength of an observed covariate. When $k_\sigma = k_\rho = 1$, then we say that an omitted confounder has *equivalent confounding strength* to an observed covariate.

With a researcher-specified k_σ and k_ρ, we obtain the formally benchmarked sensitivity parameters. Theorem 2.4.1 formalizes this.

Theorem 2.4.1 (Formal Benchmarking for Sensitivity Parameters) *Let k_σ and k_ρ be defined as in Equation (2.10). Let $R_\varepsilon^{2-(j)} := var_\mathcal{S}(\varepsilon_i^{-(j)})/var_\mathcal{S}(w_i)$, and $\rho_{\varepsilon,\tau}^{-(j)} := cor_\mathcal{S}(\varepsilon_i^{-(j)}, \tau_i)$. The sensitivity parameters R_ε^2 and $\rho_{\varepsilon,\tau}$ can be written as a function of k_σ and k_ρ:*

$$R_\varepsilon^2 = \frac{k_\sigma \cdot R_\varepsilon^{2-(j)}}{1 + k_\sigma \cdot R_\varepsilon^{2-(j)}}, \qquad \rho_{\varepsilon,\tau} = k_\rho \cdot \rho_{\varepsilon,\tau}^{-(j)}$$

Theorem 2.4.1 provides a way for researchers to estimate parameter values for an omitted confounder, after specifying the confounding strength, relative to an observed covariate. There are several key takeaways to highlight. First, in addition to providing a better understanding of potential parameter values, formal benchmarking can be used to assess the plausibility of killer confounders. We elaborate on this point in the following subsection. Secondly, because both R_ε^2 and $\rho_{\varepsilon,\tau}$ are inherently bounded, k_σ and k_ρ will also be bounded. As such, researchers can estimate the maximum confounding strength of an omitted confounder, relative to an observed covariate. Finally, we note that Theorem 2.4.1 can be extended for

a subset of covariates. This is helpful if researchers believe that a subset of observed covariates (or interactions) is particularly important to explaining the sample selection process or treatment effect heterogeneity, and wish to assess the effect of omitting a confounder with similar strength to the entire group of covariates.

Using Benchmarking to Understand Killer Confounders

We will now detail how benchmarking can be employed to help researchers reason about the plausibility of a killer confounder. We can do this in two different ways. The first is to compare the benchmarked bias with either the estimate (to assess sensitivity to a confounder reducing the estimated effect to zero), or the minimum bias estimated that can result in a statistically insignificant effect. The second approach is to compare the benchmarking results with the robustness value (either RV_q or RV_α).

Minimum Relative Confounding Strength: Benchmarking the sensitivity parameters allows researchers to estimate the resulting bias from omitting a confounder with fixed relative confounding strength as a covariate. We propose a natural summary measure, referred to *minimum relative confounding strength* (MRCS) for how much relative confounding strength an omitted variable must have to result in a killer confounder. If researchers define a killer confounder as a confounder strong enough to reduce their estimated effect to zero, the MRCS can be simply solved by dividing the point estimate with the estimated bias when $k_\rho = k_\sigma = 1$:

$$\text{MRCS}(\mathbf{X}_i^{-(j)}) = \frac{\hat{\tau}_W}{\widehat{\text{Bias}}(\varepsilon_i^{-(j)}, k_\rho = 1, k_\sigma = 1)}. \tag{2.11}$$

Similarly, if researchers are interested in killer confounders that would alter the statistical significance of their estimated effects, they can evaluate Equation 2.11 using the estimated minimum bias threshold instead of the point estimate.

If the estimated MRCS is small (i.e., MRCS < 1), then this implies that an omitted confounder, with weak confounding strength, relative to the covariate $\mathbf{X}_i^{-(j)}$, could lead to a killer confounder, if MRCS is large (i.e., MRCS > 1), then this indicates that an omitted confounder must be stronger than the observed covariate to result in a killer confounder. MRCS is an especially helpful measure when researchers have strong substantive priors for what may be important covariates.

Comparing benchmarking results with the robustness value: From benchmarking, researchers can estimate the necessary k_ρ and k_σ values in order for $R^2_\varepsilon = \rho^2_{\varepsilon,\tau} = RV_q$ (or RV_α). We denote these values as $\{k_\rho^{min}, k_\sigma^{min}\}$. The interpretation of k_ρ^{min} and k_σ^{min} is similar to that of the MRCS; however, researchers can now look at the drivers of bias with respect to the confounder's relationship to the sample selection process and treatment effect heterogeneity separately.

Example: Applying Formal Benchmarking in JTPA

To help assess plausible sensitivity parameters in the JTPA application, we perform formal benchmarking. Table 2.3 presents the results. For each of the covariates included in the weights, we estimate R^2_ε and $\rho_{\varepsilon,\tau}$, the MRCS, and $\{k^{min}_\sigma, k^{min}_\rho\}$. To account for estimation uncertainty in our benchmarking results, we perform benchmarking across repeated bootstrap iterations, and estimate the percentage of benchmarked results that result in enough bias to either (1) reduce the estimated effect to zero, or (2) change the statistical significance of the estimated effect. We provide the results in Appendix A.4.

				Estimated Effect = 0			Changes in Signif.		
Variable	$\hat{R}^2_\varepsilon$	$\hat{\rho}_{\varepsilon,\tau}$	$\widehat{\text{Bias}}$	MRCS	k^{min}_σ	k^{min}_ρ	MRCS	k^{min}_σ	k^{min}_ρ
Prev. Earnings	0.01	-0.41	-0.12	-23.4	72.9	-1.8	-2.4	10.7	-0.7
Age	0.00	0.04	0.00	—	—	18.2	—	—	7.0
Married	0.05	-0.19	-0.14	-20.7	12.3	-4.0	-2.1	1.8	-1.5
Hourly Wage	0.03	-0.24	-0.14	-20.6	20.2	-3.1	-2.1	3.0	-1.2
Black	0.17	-0.11	-0.16	-17.7	3.4	-7.1	-1.8	0.5	-2.7
Hispanic	0.24	-0.14	-0.26	-10.8	2.3	-5.5	-1.1	0.3	-2.1
HS/GED	0.07	-0.04	-0.04	-76.1	8.1	-18.5	-7.7	1.2	-7.1
Education	0.07	-0.10	-0.09	-30.0	7.5	-7.6	-3.0	1.1	-2.9

Point Estimate ($\hat{\tau}_W$): 2.81; $\hat{\sigma}^2_{\tau,\max} = 29.01$; $RV_1 = 0.56$; $RV_{\alpha=0.05} = 0.08$

Table 2.3: Formal benchmarking results for Coosa Valley, Georgia. The estimated bias is reported in thousands of USD.

From the benchmarking results, we see that omitting a confounder with equivalent confounding strength to the covariates previous earnings, whether or not the individual is Hispanic or Black, or hourly wage will result in the largest amount of bias. This is consistent with the substantive findings from the original study, which reported strong subgroup effects when looking at race and previous earnings (Bloom et al., 1993). However, the magnitude of the biases from omitting these covariates is relatively low, with most ranging from 0.12 to 0.26.

There are several takeaways to highlight from formal benchmarking. First, we see that considering the two dimensions associated with an omitted confounder matter—i.e., its relationship with the individual-level treatment effect, and its relationship with the selection mechanism. In particular, omitting a confounder like previous earnings results in a relatively "large" correlation values of $\hat{\rho}_{\varepsilon,\tau} = -0.41$; however, the benchmarked R^2_ε value associated with previous earnings is relatively low, at 0.01. As such, the overall bias from omitting a variable like previous earnings is relatively low, at -0.12. In contrast, omitting a covariate like whether or not an individual is black results in a relatively large benchmarked $\hat{R}^2_\varepsilon$ value of 0.17, but a smaller benchmarked correlation value of -0.11. As a result, the bias from

omitting a variable like whether or not an individual is Black is also relatively low, at -0.16. By looking at both the $\hat{R}^2_{\varepsilon}$ and $\hat{\rho}_{\varepsilon,\tau}$ measures, we are able to obtain a more holistic view of the types of confounders that may lead to potential changes in our analysis.

Second, we see that there is a large degree of robustness to an omitted confounder being strong enough to reduce to estimated effect to zero. In particular, a confounder would have to be 10 to 20 times stronger than an observed covariate to reduce the estimated effect to zero. However, when accounting for uncertainty, we see that an omitted confounder 1.1 times as strong as whether or not an individual is Hispanic, or twice as strong as hourly wage would be strong enough to result in a statistically insignificant effect. As such, we conclude that while there is a large degree of robustness to a confounder reducing the point estimate to 0, there is some sensitivity to potential changes in the statistical significance of our estimated effect. Finally, we highlight that while the running example throughout this paper focused on one experimental site in the JTPA study, we provide an illustration the sensitivity analysis for all 16 experimental sites in Appendix A.4.

2.5 Sensitivity Analysis for Augmented Weighted Estimators

In the following section, we extend the proposed sensitivity analysis for the class of augmented weighted, doubly robust estimators. Doubly robust estimators are a popular approach used to help improve the robustness of estimators to potential misspecfications (Dahabreh et al., 2019; Tan, 2007; Bang and Robins, 2005). There are many different doubly robust estimators (Kang et al., 2007), but we will focus on the augmented weighted estimator:

Definition 2.5.1 (Augmented Weighted Estimator)

$$\hat{\tau}_W^{Aug} = \hat{\tau}_W - \underbrace{\frac{1}{n}\sum_{i\in\mathcal{S}} w_i\hat{\tau}(\mathbf{X}_i) + \frac{1}{N}\sum_{i\in\mathcal{P}}\hat{\tau}(\mathbf{X}_i)}_{Augmented\ Component}$$

where $\mathcal{P}$ represents the set of indices of the units in the target population, $\hat{\tau}(\mathbf{X}_i)$ is the estimated individual-level treatment effect, and the weights are defined in the same manner as before. Doubly robust estimators, like the augmented weighted estimator, allow practitioners to model both the probability of sample selection and the treatment effect heterogeneity simultaneously. When one of these processes is specified correctly, then the estimator will be unbiased and asymptotically consistent.

In the following section, we introduce a sensitivity analysis for the augmented weighted estimator when omitting a confounder from the minimum separating set. We show that there are strong parallels between the sensitivity analysis for the augmented weighted estimator and the sensitivity analysis for the weighted estimator.

Bias Formula

To begin, we show that the bias of an augmented weighted estimator when omitting a variable from the minimum separating set can be written as a function of three components.

Theorem 2.5.1 (Bias of Augmented Weighted Estimator)
The bias of an augmented weighted estimator when a variable has been omitted from the minimum separating set:

$$Bias(\hat{\tau}_W^{Aug}) = \rho_{\varepsilon,\xi} \cdot \sqrt{\frac{R_\varepsilon^2}{1 - R_\varepsilon^2} \cdot var(w_i) \cdot \sigma_\xi^2} \tag{2.12}$$

where ξ_i represents the difference between the true individual-level treatment effect and estimated treatment effect (i.e., $\xi_i = \tau_i - \hat{\tau}(\mathbf{X}_i)$).

There are several key takeaways from Theorem 2.5.1. First, the double robustness of the augmented weighted estimator is apparent from Theorem 2.5.1 by noting that if there is no error in the estimated weights (i.e., $\varepsilon_i = 0$), or there is no error in estimating the treatment effect heterogeneity (i.e., $\hat{\tau}(\mathbf{X}_i)$ is a consistent model for τ_i), then ξ_i will be made up of random noise, and the correlation between ξ_i and ε_i will be zero (i.e., $\rho_{\varepsilon,\xi} = 0$). Second, Theorem 2.5.1 highlights that the bias of an augmented weighted estimator from omitting a confounder is very similar to the bias of a weighted estimator (i.e., Equation (2.6)). The primary difference is that instead of the individual-level treatment effect τ_i, we are interested in ξ_i, which is the residual component of τ_i that cannot be explained by $\hat{\tau}(\mathbf{X}_i)$.

Remark. Researchers can adapt Theorem 2.5.1 to the case where they are not re-weighting the data at hand, and are focused solely on modeling the individual-level treatment effect $\hat{\tau}(\mathbf{X}_i)$. If we assume that the individual-level treatment effect follows a linear model, then we recover the results from Nguyen et al. (2017) (see Appendix A.1 for more details). In other words, previously proposed sensitivity analysis frameworks that rely on parametric assumptions are special cases of our proposed bias decomposition. In cases when researchers do not wish to impose parametric assumptions, Theorem 2.5.1 provides a flexible approach for sensitivity analysis.

Sensitivity Analysis for Augmented Weighted Estimators

In the previous subsection, we showed that the primary differentiation between the bias formula for the augmented weighted estimator and the weighted estimator is ξ_i (i.e., the residuals in the treatment effect model). This results in two new components in the augmented weighted estimator setting: $\rho_{\varepsilon,\xi}$ and σ_ξ^2. The third component in the bias decomposition is R_ε^2, which is identical in both the weighted and augmented weighted estimator setting. We show in Appendix A.1 that similar bounds to the ones derived in Section 2.3 apply to this setting. As such, after estimating an adequate upper bound for σ_ξ^2, researchers may vary

both R^2_ε and $\rho_{\varepsilon,\xi}$ across bounded ranges to assess the sensitivity of an augmented weighted estimator to omitted confounders. Similarly, the sensitivity tools in Section 2.4 can also be extended for the augmented weighted estimator case. Details are provided in Appendix A.1, and Appendix A.4 illustrates the sensitivity analysis using JTPA.

2.6 Conclusion

Generalizing or transporting causal effects from an experiment to a different, or larger, population requires researchers to correctly identify a separating set of pre-treatment covariates that allow the confounding effect of sample selection to be conditionally ignorable. When this separating set is not correctly identified, PATE estimation will be biased.

In this paper, we formalize a sensitivity analysis framework for weighted estimators in the generalization or transportability setting, with extensions for augmented weighted estimators. We demonstrate that the proposed framework is a more general version of previously proposed sensitivity analysis frameworks. The proposed framework has several advantages to existing approaches. First, it allows researchers to bound both the magnitude of the imbalance in an omitted confounder, as well as the relationship between the omitted confounder and the individual-level treatment effect. Furthermore, the framework allows researchers to simultaneously consider bias and changes in inference from omitting a variable. Second, the sensitivity analysis allows researchers to work with standardized, scale-invariant parameters, and introduces benchmarking for researchers to use observed covariates to reason about the plausibility of parameter values. Third, we propose a set of sensitivity analysis tools to help researchers understand and summarize the degree of sensitivity that is present in their estimation. We introduce two summary measures, and demonstrate that the proposed sensitivity parameters can be bounded in an extreme scenario analysis, allowing researchers to quantify worst-case scenarios for their estimates. These tools collectively allow researchers to encode their substantive knowledge to quantitatively reason about sensitivity in their estimated effects.

Finally, in concluding this paper, it is important to emphasize the limits of the sensitivity tools. The proposed sensitivity framework provides researchers with different quantitative and graphical measures to assess the degree of robustness that is present in their point estimate. However, these tools cannot be used to *eliminate* the possibility of killer confounders, and akin to Cinelli and Hazlett (2020), we do not provide cutoff measures for measures such as the robustness value or the minimum relative confounding strength. We caution researchers from using these tools without also considering substantive judgment. The sensitivity framework provides a strong foundation for researchers to discuss the plausibility of killer confounders, but should not be used in lieu of substantive understanding of the underlying covariates and context.

Chapter 3

Leveraging Population Outcomes to Improve the Generalization of Experimental Results

3.1 Introduction

The Job Training Partnership Act (JTPA) was introduced by the U.S. Congress in 1982 to help provide employment and training programs to economically disadvantaged adults and youths. To assess its effectiveness, the national JTPA study evaluated the impact of the program across a diverse set of sixteen experimental sites between 1987 and 1989. Eligible individuals assigned to treatment were given access to the JTPA services, while those assigned to control were told that the services were not available. Eighteen months later, researchers checked on whether these study participants were employed, and measured their recent earnings (Bloom et al., 1993). The hope is that those offered the program would be more often employed, and would generally be earning higher wages.

Each site can be considered a stand-alone randomized trial. Each site has a different collection of individuals from the other sites. If a policymaker had only run their experiment in one specific population, how representative would their results have been for the other populations? This question is the essence of a current and serious critique of large-scale randomized evaluations: does a rigorous and robust finding regarding a program evaluated in a specific population actually shed light on wider questions of a program's effectiveness for a "real-world" population? Originally, the "credibility revolution" elevated the role of randomized, controlled trials (RCTs), generally praised for their strong internal validity (Banerjee and Duflo, 2009; Falk and Heckman, 2009; Baldassarri and Abascal, 2017). RCTs are attractive in that they allow researchers to draw causal inferences about treatment effects with only minimal assumptions, but only for the experimental sample. And perhaps this last clause is too great a cost; perhaps the emphasis on causality has led researchers to overly narrow the scope of their inquiry (Huber, 2013; Deaton and Cartwright, 2018). Especially

with a policy finding, if one cannot generalize, what should one make of a found result? Concerns about generalizability span the social and biomedical sciences, and are related to discussions about participant recruitment in pragmatic study designs (Ford and Norrie, 2016).

This critique has inspired a robust literature on methods for how to generalize an experimental results to broader populations of interest. In our case, for example, one could imagine extending the results found for a specific population in one site to populations living in the other sites, adjusting the impact estimate to account for differences in populations served. The generalizability literature has provided clear outlines for the necessary assumptions for such generalization, providing tools to identify the population average treatment effect (PATE), i.e., the effect of the experimental treatment in a clearly defined target population that differs from the experimental sample (Cole and Stuart, 2010; Bareinboim and Pearl, 2016; Egami and Hartman, 2022). In practice, the most common approaches model the experimental sample inclusion probability, with the PATE then estimated using weighting estimators (Stuart et al., 2011; Tipton, 2013; Hartman et al., 2015; Buchanan et al., 2018). Alternative estimators focus on modeling treatment effect heterogeneity (Kern et al., 2016; Nguyen et al., 2017) or doubly robust estimation (Dahabreh et al., 2019).

Generalizing, however, can be prohibitively costly. In practice, weighted estimators are often far more imprecise than unweighted estimators, especially when the experimental sample differs substantially from the target population. This makes it difficult for policymakers and practitioners to draw conclusions about the impact of treatment in the target population to guide their policy recommendations. Indeed, researchers empirically find that weighted estimators often increase the mean squared error for the PATE compared to a biased estimator that ignores sampling weights, due to paying for a smaller bias with much larger standard errors (Miratrix et al., 2018). More generally, considering the bias-variance tradeoff, the cost of large precision loss associated with the conventional weighting methods makes it unclear if it is "worth weighting," and questions the applicability of these weighting methods that researchers are advocating for.

This provides a quandary: the more the target population differs from the sample, the greater the cost of generalizing, due to more extreme weights, but the greater the need to generalize to keep the findings of the original experiment relevant. In this work, we seek to mitigate this tradeoff by exploiting a valuable resource commonly left on the table: the outcome data measured in the population. In particular, we aim to incorporate observational population data to reduce the noise from generalizing an experimental result. Population data often have larger sample sizes and therefore provide an opportunity to model complex covariate-outcome relationships with more flexible modeling approaches. It is this opportunity — to incorporate large population data sets that contain outcome data to improve precision — that serves as the foundation of our method.

The multisite design of our JPTA experiment serves as an ideal test bed for our method. We generalize the results of each site individually to a target population defined by the units in the other fifteen sites, allowing us to benchmark our estimates against the experimentally identified causal estimate of the excluded sites. We can then evaluate any precision gains as

compared to other generalization approaches as well as to no adjustment. We can also, for each site in turn, assess whether one should generalize, based on a diagnostic test. Ultimately, using this within study comparison approach (LaLonde, 1986), we find between a 5% and 25% reduction in variance from exploiting population data and outcomes, for those sites where we determine that our methods are applicable.

Our method is post-residualized weighting, where we leverage outcome data measured in the population to improve precision in estimation of the PATE. We begin by constructing a predictive model of the outcome using the population data. We then use this to residualize the experimental outcome data, and these residuals replace the experimental outcome in the standard inverse probability weighting estimators used for generalization. Identification of the PATE proceeds under the same assumptions required for existing inverse probability weighting methods, namely that the sampling weights are correctly specified. We show that this estimator is consistent, regardless of the residualizing model constructed in the population data. Therefore, we can safely use machine learning methods to build a predictive model. We then establish under what conditions the proposed post-residualized weighting estimator is more efficient than existing methods.

We also extend our estimator to the weighted least squares framework, which has three advantages: (1) it incorporates the well-known benefits of stabilized weighting estimators (i.e. Hàjek estimators), (2) it allows for additional precision gains from prognostic variables measured only within the experiment, and (3) it addresses concerns about scaling differences between the outcomes measured in the experiment and the population data. Importantly, we also provide a diagnostic that allows researchers to assess when the post-residualized weighting method is likely to result in efficiency gains.

As far as we know, using covariates and outcome data in this manner has not been investigated. While inverse probability weighting methods do leverage population data about pre-treatment covariates when modeling the sampling weights, use of outcome data has primarily been limited to use in placebo tests (Cole and Stuart, 2010; Hartman et al., 2015). Recently, the data fusion literature proposed using experimental data to help aid the estimation of causal effects in observational studies (e.g., see Athey et al. (2020, 2019); Kallus and Mao (2020)), which bears some similarity to our problem.

We proceed by further introducing our empirical application. We then introduce notation and existing methods for estimating the population average treatment effect from experimental data in Section 3.2. In Section 3.3 we introduce post-residualized weighting, prove its statistical properties, and introduce a diagnostic to assess whether researchers should expect efficiency gains in their applications. We consider both a weighted estimator (a.k.a., Hàjek estimator) and a weighted least squares estimator. We extend results to a case in which we include the predicted outcome as a covariate in Section 3.4. Finally, we provide simulation evidence supporting the performance of post-residualized weighting estimators and diagnostic tools in Section 3.5 and apply them to the Job Training Partnership Act in Section 3.6.

Background and Data

The Job Training Partnership Act (JTPA) was a large study with a 2:1 treatment to control ratio. A variety of outcomes were measured with a follow-up survey 18 months after assignment (Bloom et al., 1993). We use the 16 experimental sites from the national JTPA study as the basis for our analysis. While the original study focused on four target groups: adult women and men (categorized formally as ages 22 and older), and female and male out-of-school youths (ages 16-21), we focus our analysis on adult women, the largest target group within the JTPA study.[1] We consider two different outcomes: employment status (binary outcome) and total earnings (zero-inflated, continuous outcome). Across the 16 sites, the average effect on earnings was \$1240 and employment was 1.63%, but point estimates across sites ranged from -\$5210 in Butte, MT to \$3030 in Providence, RI for earnings and -7% in Butte, MT and Marion, OH to 7% in Heartland, FL and Providence, RI. Had a policymaker only run their experiment in Providence, RI, they may have concluded that the treatment was effective, but not so in Butte, MT. Weighted estimators can adjust for demographic differences across sites, but many of the sites, such as Butte, MT, contain few units, emphasizing the need for precise estimators when generalizing results to other populations.

Using a within study comparison approach, we generalize the results of each site individually to a target population defined by the units in the other 15 sites, allowing us to benchmark our estimator against the experimentally identified causal estimate of the excluded sites and evaluate precision gains from post-residualized weighting. A summary of the JTPA experimental set up is provided in Appendix B.5.

3.2 Existing Estimators for Generalization

Setup

We begin by defining the target population as an infinite super-population $\mathcal{P}$ with probability distribution F and probability density dF, for which we wish to infer the effectiveness of treatment. Following Buchanan et al. (2018), suppose we observe n units as the "experimental sample," but, as with most experiments in practice, the selection into the experiment from the target population is biased. Let $\mathcal{S}$ represent the random set of n indices for the units in the experimental sample.

Units in our experimental sample are treated, or not, with treatment indicator $T_i = 1$ for units assigned to treatment, and $T_i = 0$ for control. Using the potential outcomes framework (Neyman, 1923; Rubin, 1974), we define $Y_i(t)$ to be the potential outcome of unit i that would realize if unit i receives treatment $T_i = t$, where $t \in \{0, 1\}$. Our primary causal quantity of interest is the population average treatment effect (PATE), which is formally defined as:

$$\tau := \mathbb{E}_F\{Y_i(1) - Y_i(0)\}, \tag{3.1}$$

[1]The estimated impact of JTPA for the other target groups were not found to be statistically significant in the original study.

where the expectation is taken over the target population distribution F. This is in contrast to the sample average treatment effect (SATE) of

$$\tau_{\mathcal{S}} := \mathbb{E}_{\tilde{F}}\{Y_i(1) - Y_i(0)\},$$

where the expectation is taken over the experimental sample distribution $\tilde{F}$.

For each unit in the experiment, only one of the potential outcome variables can be observed, and the realized outcome variable for unit i is denoted by $Y_i = T_i Y_i(1) + (1-T_i)Y_i(0)$. We also observe pre-treatment covariates $\mathbf{X}_i$ for units in the experiment. We use $\tilde{F}$ to represent the sampling distribution for the experimental sample, i.e., $\{Y_i(1), Y_i(0), T_i, \mathbf{X}_i\}_{i=1}^n \overset{iid}{\sim} \tilde{F}$ with density $d\tilde{F}$. Because we consider settings where the selection into the experiment from the target population $\mathcal{P}$ is biased, $F \neq \tilde{F}$.

We assume that the treatment assignment is randomized within the experiment.

Assumption 3 (Randomization within Experiment)

$$d\tilde{F}(Y_i(1), Y_i(0), T_i, \mathbf{X}_i) = d\tilde{F}(Y_i(1), Y_i(0), \mathbf{X}_i) \cdot d\tilde{F}(T_i) \tag{3.2}$$

In other words, the treatment assignment T_i is independent of the tuple $\{Y_i(1), Y_i(0), \mathbf{X}_i\}$. Under this assumption, the SATE can be estimated without bias using a difference-in-means estimator of

$$\widehat{\tau}_{\mathcal{S}} = \frac{1}{\sum_{i\in\mathcal{S}} T_i} \sum_{i\in\mathcal{S}} T_i Y_i - \frac{1}{\sum_{i\in\mathcal{S}}(1-T_i)} \sum_{i\in\mathcal{S}} (1-T_i) Y_i. \tag{3.3}$$

The SATE is important for evaluating the effectiveness of treatment. However, researchers often want to know to what extent the internally valid findings of an experiment are externally valid to the target population (Cole and Stuart, 2010; Miratrix et al., 2018; Egami and Hartman, 2022). When the experimental sample is randomly drawn from the target population $F = \tilde{F}$, $\widehat{\tau}_{\mathcal{S}}$ can be used as an unbiased estimator for τ. However, in most settings, experimental units are not randomly drawn from the target population with equal probability.

To estimate the PATE, we also assume we observe an *i.i.d.* sample of N units from the target super-population $\mathcal{P}$ as the "population data," which is separate from the experimental sample. This design is most common in the social sciences, and is called the non-nested design in that the experimental sample is not a subset of the population data (Colnet et al., 2020).[2] Typically, the size of the population data is much larger than the experimental data, i.e., $N \gg n$. In the conventional setup, researchers only observe pre-treatment covariates $\mathbf{X}_i$ for each unit i in the population data. In the next subsection, we review assumptions and estimators for the PATE under this conventional setup. In Section 3.3, we then consider our setting in which researchers also observe an outcome measure in addition to pre-treatment covariates in the population data. Importantly, because the treatment is not randomized in the population data, we cannot identify the PATE just using the population data.

[2]While we focus on the non-nested design in this paper, the same proposed approach is useful for the nested design where the experimental sample is a subset of the population data. The main difference arises in the analytical expressions of the efficiency gain from our proposed approach.

Assumptions

We make the standard assumptions of no interference and that treatments are identically administered across all units (i.e., SUTVA, defined in Rubin (1980)). In order to identify the PATE using experimental data, we require additional assumptions about the sampling of the experimental units. First, we assume that, conditional on a set of pre-treatment covariates $\mathbf{X}_i$, the sample selection mechanism is ignorable. More formally,

Assumption 4 (Ignorability of Sampling and Potential Outcomes)

$$dF(Y_i(1), Y_i(0) \mid \mathbf{X}_i = \mathbf{x}) = d\tilde{F}(Y_i(1), Y_i(0) \mid \mathbf{X}_i = \mathbf{x}) \tag{3.4}$$

Assumption 4 states that, conditional on $\mathbf{X}_i$, the distribution of the potential outcomes $\{Y_i(1), Y_i(0)\}$ is the same across the experimental sample and the target population (Stuart et al., 2011; Pearl and Bareinboim, 2014; Kern et al., 2016).[3] We also assume that, for any pre-treatment covariate profile $\mathbf{X}_i = x$ we might see in the population, we have a non-zero chance of seeing it in the sample as well (Westreich and Cole, 2010):

Assumption 5 (Positivity)
For all $\mathbf{x}$ *with* $dF(\mathbf{X}_i = \mathbf{x}) > 0$, *we have*

$$dF(\mathbf{X}_i = \mathbf{x}) > 0 \Rightarrow d\tilde{F}(\mathbf{X}_i = \mathbf{x}) > 0. \tag{3.5}$$

Estimation of PATE

There is a robust, and growing, literature on methods for estimating the PATE. The most common approach is the inverse probability weighting estimator (IPW) (Cole and Stuart, 2010). The IPW estimator relies on sampling weights usually defined as an inverse of the probability of being sampled into the experiment. In our case, given the infinite superpopulation defined by F, this translates to, for each unit i,

$$w_i \propto \frac{1}{\pi(\mathbf{X}_i)},$$

with $\pi(\mathbf{X}_i)$ the relative density of

$$\pi(\mathbf{X}_i) = \frac{d\tilde{F}(\mathbf{X}_i)}{dF(\mathbf{X}_i)}. \tag{3.6}$$

Weights are typically estimated using a binary outcome model, such as logistic regression, by exploiting the fact that weights are proportional to the relative probability of being in the

[3]For identification of the PATE, a weaker assumption of conditional ignorability of sampling and treatment effect heterogeneity may be invoked instead. However, our variance derivations rely on the conditional ignorability of sampling and potential outcomes.

observed population data to the probability of being in the experimental sample, conditional on being in either set:

$$w_i \propto \frac{\Pr(S_i = 0 \mid \mathbf{X}_i)}{\Pr(S_i = 1 \mid \mathbf{X}_i)},$$

where S_i takes on a value of 1 if the unit belongs to the experimental sample, and 0 if the unit belongs to the observed population data.

Researchers can estimate $\Pr(S_i = 1 \mid \mathbf{X}_i)$ and $\Pr(S_i = 0 \mid \mathbf{X}_i)$ using a binary outcome model, regressing S_i on $\mathbf{X}_i$ using the stacked dataset of both the experimental and population data (Stuart et al., 2011; Buchanan et al., 2018; Egami and Hartman, 2019; O'Muircheartaigh and Hedges, 2014). Alternatively, researchers can use balancing methods, such as entropy balancing, which estimates weights such that weighted moments (e.g., means of each pre-treatment covariate $\mathbf{X}_i$) of the experimental data equal the corresponding moments of the observed population data (Deville and Särndal, 1992; Hainmueller, 2012; Hartman et al., 2015).

Once researchers have estimated the sampling weights, the PATE can be estimated using a weighted estimator, also known as the Hàjek estimator:

$$\hat{\tau}_W := \frac{\sum_{i \in \mathcal{S}} \hat{w}_i T_i Y_i}{\sum_{i \in \mathcal{S}} \hat{w}_i T_i} - \frac{\sum_{i \in \mathcal{S}} \hat{w}_i (1 - T_i) Y_i}{\sum_{i \in \mathcal{S}} \hat{w}_i (1 - T_i)}. \tag{3.7}$$

As with estimation of the SATE, researchers can also include covariate adjustment to increase efficiency. This approach is popular because, while the estimation of the weights requires covariates to be measured across both the population and the experimental data, covariate adjustment can leverage covariates that are only measured in the experimental data (Stuart and Rhodes, 2017).

The weighted least squares estimator $\hat{\tau}_{wLS}$ for the PATE can be computed via a weighted regression of the outcome on an intercept, the treatment indicator and pre-treatment covariates with estimated weights. Formally,

$$(\hat{\tau}_{wLS}, \hat{\alpha}, \hat{\gamma}) = \underset{\tau,\alpha,\gamma}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in \mathcal{S}} \hat{w}_i \left(Y_i - (\tau T_i + \alpha + \widetilde{\mathbf{X}}_i^\top \gamma) \right)^2 \tag{3.8}$$

where $\widetilde{\mathbf{X}}_i$ are experimental pre-treatment covariates included in the covariate adjustment. Covariates $\widetilde{\mathbf{X}}_i$ can differ from the $\mathbf{X}_i$ required for Assumptions 4–5. The weighted estimator (equation (3.7)) is a special case of this weighted least squares estimator (equation (3.8)) because it is numerically equivalent to the estimated coefficient of the treatment indicator when no covariate is included, i.e., $\widetilde{\mathbf{X}}_i = \emptyset$. Because the weighted estimator is a special case of the weighted least squares estimator, we focus on the weighted least squares estimator in this paper, but use the simpler weighted estimator to illustrate intuitions when appropriate. Under Assumption 3–5 and the consistent estimation of the sampling weights, the weighted estimator $\hat{\tau}_W$ and the weighted least squares estimator $\hat{\tau}_{wLS}$ are both consistent for the PATE, regardless of what covariates $\widetilde{\mathbf{X}}$ we include as covariate adjustment (Buchanan et al., 2018; Dahabreh et al., 2019).

In practice, weighted estimators can suffer from large variance due to extreme weights, which in this case depends on how much the individual unit-level probabilities of inclusion in the experimental sample varies relative to their average probability of inclusion. This problem has been highlighted in the observational causal inference literature with respect to inverse propensity score weighted estimators, in which large imbalances between treatment and control groups can result in extreme weights (Kang et al., 2007; Stuart, 2010). This issue is often exacerbated in the generalization setting, where imbalances between a convenience experimental sample and target population can be relatively large. As a result, losses in precision from weighting can be challenging to overcome when generalizing from the SATE to the PATE (Miratrix et al., 2018).

3.3 Post-Residualized Weighting

Existing methods, such as the weighted estimator and weighted least squares estimator described above, require pre-treatment covariate data, measured in both the experimental sample and target population, for estimating the sampling weights. However, researchers often have access to an outcome variable in the observational population data as well. Our proposed method, *post-residualized weighting*, aims to improve precision in estimation of the PATE by leveraging this outcome variable measured in the observational population data. See Figure 3.1 for a visualization of the difference in settings from conventional methods.

In addition to our JTPA application, which inspires our method, we next describe two canonical social science examples below that motivate the data settings that underpin our method. We return to these examples, in addition to the JTPA application, for conceptual clarity. We describe our benchmark analysis of the JTPA data in Section 3.6.

Example: Get-Out-the-Vote (GOTV) Experiments Political scientists have conducted a number of field experiments to evaluate the impact of canvassing efforts, including door-to-door, phone, and mail, on voter turnout. Such GOTV experiments typically rely on administrative data to measure the outcome, namely voter turnout data from the Secretary of State. These experiments are often conducted in a small geographic region (e.g., New Haven, Connecticut in Gerber and Green (2000)), but scholars are often interested in generalizing the effect to broader populations, such as for a statewide election. Importantly, when considering generalization, the outcome variable of voter turnout is available not only for the experimental data but also for the broader target population of interest. In our framework, we use this information about voter turnout measured in the observational population data to improve precision in the estimation of the PATE.

Example: Education Experiments Education research also relies on experiments to evaluate the performance of classroom interventions, such as the impact of smaller class size on curriculum-based and standardized tests (e.g., Word et al., 1990). These experiments are often done in partnership with school systems. For example, the Tennessee STAR experiment

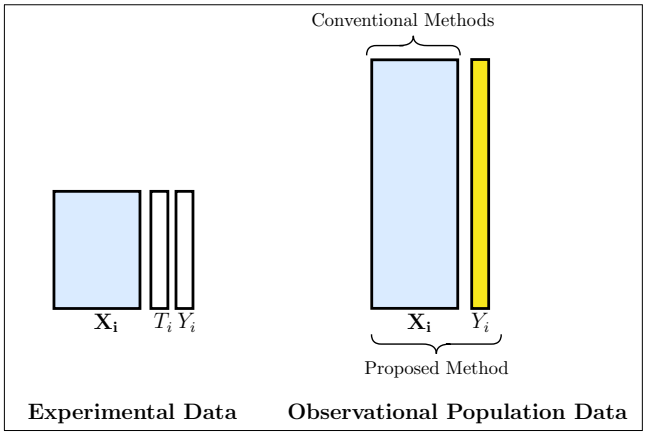

Figure 3.1: Data Requirements. Conventional estimation methods only use the covariate data $\mathbf{X}_i$ (in light gray). Our proposed approach leverages the outcome data, in addition to the covariate data at the population level (as highlighted in dark gray).

was conducted in classrooms across Tennessee. However, researchers are interested in the broader impact of such interventions. For example, a researcher may ask what the long term impact of small class sizes in primary school is on standardized test scores, such as the SAT, for all public schools in the United States. To estimate the PATE, existing methods use demographic variables from a random sample of public school students to construct sampling weights. In our framework, we can additionally use SAT scores measured for a random sample of public school students, which improves estimation accuracy.

Remark We emphasize that the outcome variable available in the population data can be either the potential outcomes under treatment $Y(1)$, the potential outcomes under control $Y(0)$, or their mix. Indeed, researchers do not need to know the treatment condition of units in the target population. This is because consistency of our proposed approach does not depend on the correct specification of a predictive model we will build with the outcome variable available in the population data (see Theorem 3.3.1). More generally, the outcome variable available in the population data can be a proxy of the outcome variable in the experimental data (i.e., not equal to either the potential outcomes under treatment or control), and we consider this case in Section 3.4. □

Post-residualized Weighted Estimators

Our proposed post-residualized weighting approach exploits the outcome measured in the population data to improve precision in estimation of the PATE. The key idea is that we estimate a predictive model with the outcome measured in the population data and then

Post-residualized Weighting for the PATE estimation:	
Step 1:	Estimate sampling weights, w_i, for units in the experimental sample.
Step 2:	Choose a residualizing model $g(\mathbf{X}_i)$: $\mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is the support of $\mathbf{X}_i$. Using the population data, estimate $\hat{g}(\mathbf{X}_i)$ that predict the population outcomes using pre-treatment covariates $\mathbf{X}_i$.
Step 3:	Predict $\hat{Y}_i = \hat{g}(\mathbf{X}_i)$ for each unit in the experimental data, and compute residual $\hat{e}_i = Y_i - \hat{Y}_i$ for units in the experimental sample.
Step 4:	Estimate the PATE using residuals $\hat{e}_i$ and estimated sampling weights $\hat{w}_i$. *No covariate adjustment within the experimental data* ↳ See post-residualized weighted estimator $\hat{\tau}_W^{res}$ in equation (3.10). *With covariate adjustment within the experimental data* ↳ See post-residualized weighted least squares estimator $\hat{\tau}_{wLS}^{res}$ (Definition 3.3.1).

Table 3.1: Summary of Post-Residualized Weighting.

use this estimated predictive model to *residualize* outcomes in the experimental data, before using conventional weighting estimators for the PATE. For example, in our JTPA application, we predict earnings or employment across the target sites (i.e., the 'population'), which we use to residualize the outcomes in the experimental site.

In total, post-residualized weighting has four steps. The first step is to estimate sampling weights w_i, which is the same as the conventional weighting approach. In the second step, we fit a flexible model in the population data to predict the outcome variable Y_i using pre-treatment $\mathbf{X}_i$. We refer to this predictive model fit in the population data as a *residualizing model*, and formally denote it as $g(\mathbf{X}_i)$: $\mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is the support of $\mathbf{X}_i$. In the third step, we use the estimated residualizing model to predict outcomes $\widehat{Y_i}$ in the experimental data, which is separate from the population data used to estimate the residualizing model. In the fourth and final step, we apply the weighted least squares estimator (equation (3.8)) using the residuals from this prediction, (denoted by $\hat{e}_i = Y_i - \hat{Y}_i$) as outcomes (instead of Y_i used in the conventional weighted least squares estimator).

We summarize our proposed approach in Table 3.1. In the following section, we directly extend the weighted estimator and the weighted least squares estimator discussed in Section 3.2.

Definition 3.3.1 (Post-Residualized Weighted Least Squares Estimator) *Given a residualizing model estimated as $\hat{g}(\cdot)$, the post-residualized weighted least squares*

estimator $\hat{\tau}^{res}_{wLS}$ *for the PATE is defined as,*

$$(\hat{\tau}^{res}_{wLS}, \hat{\alpha}^{res}, \hat{\gamma}^{res}) = \underset{\tau,\alpha^{res},\gamma^{res}}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in \mathcal{S}} \hat{w}_i (\hat{e}_i - \tau T_i - \alpha^{res} - \widetilde{\mathbf{X}}_i^\top \gamma^{res})^2 \tag{3.9}$$

where $\hat{e}_i = Y_i - \hat{g}(\mathbf{X}_i)$ *and* $\widetilde{\mathbf{X}}_i$ *are experimental pre-treatment covariates included in the covariate adjustment. We allow* $\widetilde{\mathbf{X}}_i$ *to differ from* $\mathbf{X}_i$ *used to calculate* $\hat{g}(\mathbf{X}_i)$*.*

In practice, the post-residualized weighted least squares estimator can be estimated by running a weighted regression, where the estimated residualized values $\hat{e}_i$ is regressed on an intercept, the treatment indicator T_i and covariates $\widetilde{\mathbf{X}}_i$, and using the sampling weights $\hat{w}_i$ as the weights. The coefficient of the treatment indicator is the post-residualized weighted least squares estimate for the PATE.

In a special case where no pre-treatment covariates are included, the post-residualized weighted least squares estimator is equivalent to the following post-residualized weighted estimator.

$$\hat{\tau}^{res}_{W} := \frac{\sum_{i \in \mathcal{S}} \hat{w}_i T_i \hat{e}_i}{\sum_{i \in \mathcal{S}} \hat{w}_i T_i} - \frac{\sum_{i \in \mathcal{S}} \hat{w}_i (1 - T_i) \hat{e}_i}{\sum_{i \in \mathcal{S}} \hat{w}_i (1 - T_i)}. \tag{3.10}$$

We summarize several key aspects of the post-residualized weighted least squares estimator here and formally discuss each point in the subsequent sections. First, the identification of the PATE is obtained under the same assumptions required for existing weighted estimators and the weighted least squares estimator, and we do not make any additional assumptions (Section 3.3). Most importantly, our proposed estimators are consistent for the PATE, regardless of the choice of the residualizing model. That is, we do not require the correct specification of the residualizing model $g(\mathbf{X}_i)$ to guarantee consistency of the proposed estimators. Therefore, akin with Rosenbaum et al. (2002) and Sales et al. (2018), the residualizing model $g(\mathbf{X}_i)$ can be seen as an "algorithmic model" in that the goal is to predict outcomes, rather than substantively explain an underlying probabilistic process.

Second, the proposed post-residualized weighted least squares estimator, $\hat{\tau}^{res}_{wLS}$, can achieve significant improvements in precision over the traditional weighted least squares estimator (equation (3.8)) when the residualizing model can predict outcomes in the experiment well (Section 3.3). We will show in Section 3.3 that while we maintain consistency regardless, how much efficiency gain we achieve depends on the predictive performance of the fitted residualizing model $\widehat{g}(\mathbf{X}_i)$. As such, researchers should, when possible, use not only simple models, such as ordinary least squares, but also more flexible machine learning models, such as random forests or other ensemble learning methods (Breiman, 2001; Polley and van der Laan, 2010) as the residualizing models to improve precision of the PATE estimation.

Finally, we derive a diagnostic measure that researchers can use to determine whether residualizing will likely lead to precision gains when estimating the PATE (Section 3.3). As emphasized in the second point above, when the residualizing model can predict outcomes in the experiment well, we can expect efficiency gains. However, when the residualizing

model fails to predict outcome measures in the experimental data, it is possible for post-residualizing to increase uncertainty of the PATE estimation. Our diagnostic measure helps researchers to estimate the expected efficiency gain, thereby deciding whether residualizing is beneficial in their applications.

Remark Our proposed post-residualized weighted least squares estimator is closely connected to the augmented inverse probability weighted estimators (AIPW) (Robins et al., 1994) developed for the PATE (Dahabreh et al., 2019) in that both estimators combine weighting and outcome-modeling. The process of estimating weights for both the post-residualized weighting estimators and AIPW is the same. However, the key difference between two approaches is that the AIPW estimates the outcome model using only the experimental data, thereby not exploiting the outcome variable available in the population data. In contrast, our post-residualized weighting estimator explicitly uses the outcome information available in the population data to estimate the residualizing model and improve precision. Furthermore, post-residualized weighting does not attempt to model both the treatment and control outcomes separately, and therefore, does not have the double robustness that the AIPW has. □

Remark Compared to the simpler post-residualized weighted estimator (equation (3.10)), there are two advantages to a more general, post-residualized weighted least squares estimator (equation (3.9)). First, it can leverage precision gains from pre-treatment covariates that are measured in the experimental data but not in the population data. That is, $\widetilde{\mathbf{X}}_i$ can include more covariates than $\mathbf{X}_i$. Second, $\hat{\tau}^{res}_{wLS}$ provides additional robustness over the post-residualized weighted estimator $\hat{\tau}^{res}_{W}$. More specifically, without further covariate adjustment, residualizing can be sensitive to differences between the population and experimental units in the covariate-outcome relationships. For example, considering JTPA, if earnings and employment depend heavily on the local economic condition, and thus the covariate relationships differ across sites, then residualizing may not provide efficiency gains. When this difference is large, residualizing can result in efficiency loss. However, by performing covariate adjustment on the residualized outcomes in the experimental data, we have an opportunity to correct for the difference in the covariate-outcome relationships between the experimental data and the population data. In other words, the post-residualized weighted least squares estimator, $\hat{\tau}^{res}_{wLS}$, gives researchers two opportunities to combat the precision loss of weighting: once from using the population data in the residualizing process, and a second from adjusting for covariates in the experimental data. □

Consistency

In this section, we show that the post-residualized weighted least squares estimator is a consistent estimator of the PATE regardless of the choice of the residualizing model $g(\mathbf{X}_i)$ and pre-treatment covariates $\widetilde{\mathbf{X}}_i$ that researchers adjust for in the weighted least squares

estimator. This emphasizes the point that $g(\mathbf{X}_i)$ need not be a correct specification of the underlying data generating process, but merely a function that predicts outcomes measured in the population.

Theorem 3.3.1 (Consistency of Post-residualized Weighted Least Squares Estimators) *Assume that sampling weights $\hat{w}_i$ are consistently estimated and Assumptions 3–5 hold with pre-treatment covariates $\mathbf{X}_i$. Then, the post-residualized weighted least squares estimator that adjusts for pre-treatment covariates $\widetilde{\mathbf{X}}_i$ (equation (3.9)) is a consistent estimator*

$$\hat{\tau}_{wLS}^{res} \xrightarrow{p} \tau,$$

with any residualizing model $g(\mathbf{X}_i)$ and any pre-treatment covariates $\widetilde{\mathbf{X}}_i$. The post-residualized weighted estimator (equation (3.10)) is also consistent as it is a special case when no covariate is included.

The proof of Theorem 3.3.1 can be found in Appendix B.1. This property allows for a large degree of flexibility in building the residualizing model, since consistency is guaranteed *regardless* of model specification or performance of $g(\mathbf{X}_i)$. We can obtain the consistency even for a misspecified residualizing model $g(\mathbf{X}_i)$ because the predicted experimental outcome $\hat{Y}_i = \hat{g}(\mathbf{X}_i)$ is only a function of the pre-treatment covariates $\mathbf{X}_i$, and thus, with randomized treatments (Assumption 3), its distribution is the same across treatment and control units on average for any sample size. As such, residualizing preserves the consistency of the original weighted estimator without requiring any additional assumptions.

A potential concern with covariate adjustment is that performing covariate adjustment within the experimental data can result in worsened asymptotic precision and invalid measures of uncertainty (Freedman, 2008). An alternative approach is to include interaction terms between the treatment indicator and covariates (Lin, 2013). Regardless, because the proposed post-residualized weighted least squares estimator is an extension of a weighted least squares estimator, we can compute valid standard errors with the standard Huber–White sandwich estimator.

While consistency is guaranteed, efficiency gains from residualizing *do* depend on the ability of the residualizing model to predict outcome measures in the experimental data. Theorem 3.3.1 allows for researchers to leverage complex, "black box" approaches (such as ensemble methods) to maximize the predictive accuracy, as interpretability of the residualizing model is secondary to being able to fit the data well. In the next section, we will formalize the criteria for variance reduction from residualizing.

Efficiency Gains

The post-residualized weighted estimator allows researchers to include information from the observational population data about the relationship between the pre-treatment covariates and the population outcomes into the estimation process. Whether or not we obtain precision gains, and the magnitude of these precision gains, will depend on the nature of the

residualizing model. In general, the better researchers are able to explain the outcomes measured in the experiment using the residualizing model, the greater the efficiency gains. For example, as shown in Section 3.6, we see greater gains from post-residualized weighting for earnings, where our predictive model performs better, than we do for employment, which is more difficult to predict with the auxiliary covariates.

To make these gains more explicit, we first define the *weighted variance* and *weighted covariance* as follows.

$$\text{var}_w(A_i) = \int \frac{1}{\pi(\mathbf{X}_i)^2} \cdot (A_i - \bar{A})^2 d\tilde{F}(\mathbf{X}_i, A_i), \tag{3.11}$$

$$\text{cov}_w(A_i, B_i) = \int \frac{1}{\pi(\mathbf{X}_i)^2} \cdot (A_i - \bar{A})(B_i - \bar{B}) d\tilde{F}(\mathbf{X}_i, A_i, B_i), \tag{3.12}$$

where $\bar{A} = \mathbb{E}_F(A_i)$ and $\bar{B} = \mathbb{E}_F(B_i)$.

To simplify the expression, we first describe the efficiency gain for the post-residualized weighted estimator (equation (3.10)) as follows.

Theorem 3.3.2 (Efficiency Gain for Post-residualized Weighted Estimators) *The difference between the asymptotic variance of $\hat{\tau}_W^{res}$ and that of $\hat{\tau}_W$ is:*

$$\begin{aligned} &asyvar_{\tilde{F}}(\hat{\tau}_W) - asyvar_{\tilde{F}}(\hat{\tau}_W^{res}) \\ &\quad = -\frac{1}{p(1-p)} var_w(\hat{Y}_i) + \frac{2}{p} cov_w(Y_i(1), \hat{Y}_i) + \frac{2}{1-p} cov_w(Y_i(0), \hat{Y}_i), \end{aligned} \tag{3.13}$$

where $asyvar_{\tilde{F}}(Z)$ denotes the scaled asymptotic variance of random variable Z over the sampling distribution $\tilde{F}$, i.e., $asyvar_{\tilde{F}}(Z) = \lim_{n\to\infty} var_{\tilde{F}}(\sqrt{n}Z)$. p is the probability of being treated within the experiment, i.e., $p = \Pr_{\tilde{F}}(T_i = 1)$.

The proof of Theorem 3.3.2 can be found in Appendix B.1. Theorem 3.3.2 decomposes the efficiency gain from post-residualized weighting into two components: (1) the variance of the predicted experimental outcomes $\text{var}_w(\hat{Y}_i)$, and (2) how related the predicted outcomes are to the actual outcomes in the experimental samples (represented by $\text{cov}_w(Y_i(1), \hat{Y}_i)$ and $\text{cov}_w(Y_i(0), \hat{Y}_i)$). If the covariance between the predicted outcomes and actual outcomes in the experimental sample is greater than the variance of the predicted outcomes, we expect precision gains. In other words, the gains to precision from residualizing depend on how well outcome measures in the experiment are explained by the residualizing model fitted to the population data.[4] As such, researchers should leverage the large amounts of data available at the population level to apply flexible modeling strategies in order to maximize the variation explained by the residualizing model.

More generally, we can formally write the efficiency gain for the post-residualized weighted least squared estimator (equation (3.9)) as follows.

[4]We note that the efficiency gain expression does not include uncertainty associated with estimating the residualizing model. This is because the chosen $\hat{g}(\mathbf{X}_i)$ is a dimension reducing function of the fixed pre-treatment covariates.

Theorem 3.3.3 (Efficiency Gain for Post-Residualized Weighted Least Squares Estimators) *The difference between the asymptotic variance of $\hat{\tau}_{wLS}$ and that of $\hat{\tau}^{res}_{wLS}$ is:*

$$
\begin{aligned}
&asyvar_{\tilde{F}}(\hat{\tau}_{wLS}) - asyvar_{\tilde{F}}(\hat{\tau}^{res}_{wLS}) \\
&= \frac{1}{p}\left\{var_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) - var_w(Y_i(1) - \hat{g}(\mathbf{X}_i))\right\} \\
&\qquad \underbrace{+ \frac{1}{1-p}\left\{var_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) - var_w(Y_i(0) - \hat{g}(\mathbf{X}_i))\right\}}
\end{aligned}
$$

(a) Explanatory power of residualizing model over linear regression

$$
\underbrace{+ \frac{2}{p} cov_w(\hat{e}_i(1), \tilde{\mathbf{X}}_i^\top \gamma^{res}_*) + \frac{2}{1-p} cov_w(\hat{e}_i(0), \tilde{\mathbf{X}}_i^\top \gamma^{res}_*) - \frac{1}{p(1-p)} var_w(\tilde{\mathbf{X}}_i^\top \gamma^{res}_*)}, \quad (3.14)
$$

(b) Remaining variation in residualized outcomes explained by linear regression on $\tilde{\mathbf{X}}_i$

where γ_* and γ^{res}_* are the true coefficients[5] *associated with the pre-treatment covariates, $\widetilde{\mathbf{X}}_i$ defined in the weighted least squares regression (equation (3.8)) and the post-residualized weighted least squares regression (equation (3.9)), respectively.*

When we include covariate adjustment to the experimental data, the gains to precision depend on two factors. The first factor, (a), compares the explanatory power of the residualizing model with the linear regression. More specifically, if $\hat{g}(\mathbf{X}_i)$ is able to explain more variation than the linear combination of $\tilde{\mathbf{X}}_i$, then we expect the first term to be positive. The second term, (b), represents the amount of variation in the residualized outcomes that can be explained by the pre-treatment covariates $\tilde{\mathbf{X}}_i$.

A natural question is why not directly adjust for covariates within the experimental sample instead of using a residualizing model? One advantage to using the post-residualized weighting over directly adjusting for covariates within the experimental sample arises from the fact that there is typically a larger amount of data available in the population data (i.e. $N \gg n$). While researchers could choose to use a flexible model within the experimental data to perform covariate adjustment, there is a greater restriction with respect to degrees-of-freedom to what type of model can be fit. The availability of large amounts of population data can be leveraged in the residualizing process to better estimate covariate-outcome relationships. Additionally, by using population data to build and tune the residualizing model, we protect the fidelity of inferences using the experimental data since it is only used for estimation of the PATE.

In the following subsection, we will describe a diagnostic measure that can help researchers determine whether or not they should expect precision gains from residualizing.

[5] We define the true coefficients as the coefficients that would be estimated as the experimental sample size $n \to \infty$. See Supplementary Materials for more information.

Diagnostics

As discussed above, while post-residualized weighting stands to greatly improve precision in estimation of the PATE, this is not guaranteed. To address this concern, we derive a diagnostic that evaluates when researchers should expect precision gains from residualizing.

Again to simplify the expression, we first start with the post-residualized weighted estimator (equation (3.10)). We can define a pseudo-R^2 measure as:

$$R_0^2 := 1 - \frac{\text{var}_w(\hat{e}_i(0))}{\text{var}_w(Y_i(0))}, \tag{3.15}$$

where we define $\hat{e}_i(t) = Y_i(t) - \hat{Y}_i$ for $t \in \{0, 1\}$.

R_0^2 can be interpreted as the weighted goodness-of-fit of the residualizing model for the potential outcomes under control for units in the experiment. Researchers can estimate R_0^2 using the estimated residuals across the control units in the experiment. When $R_0^2 > 0$, we expect an improvement in precision across the control units from residualizing.

More generally for the post-residualized weighted least squares estimator (equation (3.9)), we can define R_0^2 as:

$$R_0^2 = 1 - \frac{\text{var}_w(\hat{e}_i(0) - \widetilde{\mathbf{X}}_i^\top \gamma_*^{res})}{\text{var}_w(Y_i(0) - \widetilde{\mathbf{X}}_i^\top \gamma_*)}, \tag{3.16}$$

where we now include covariate adjustments from weighted least squares regression in our diagnostic. $\hat{e}_i(0) - \widetilde{\mathbf{X}}_i^\top \gamma_*^{res}$ are the residuals that arise from regressing the residualized outcomes under control on the pre-treatment covariates in the weighted regression. Similarly, the quantity $Y_i(0) - \widetilde{\mathbf{X}}_i^\top \gamma_*$ are the residuals from regressing the outcomes under control on the pre-treatment covariates. In this way, we are directly comparing the variance of the outcomes, following covariate adjustment, across the control units. The interpretation of this value is identical to that of the pseudo-R^2 value in the weighted estimator case. It is easy to see that R_0^2 in equation (3.15) is a special case of R_0^2 in equation (3.16) when $\widetilde{\mathbf{X}}_i = \emptyset$.

In line with Rubin's "locked box" approach (Rubin, 2008), we do not suggest estimating the analogous R_0^2 among treated units. However, if the variation in the control outcomes is greater than the overall treatment effect heterogeneity, then checking if R_0^2 is greater or less than zero is an effective diagnostic for whether or not we expect precision gains from residualizing. We formalize this in the following corollary, where we write the relative reduction from residualizing as a function of this proposed R_0^2 measure.

Corollary 3.3.1 (Relative Reduction from Residualizing)
With R_0^2 defined as in equation (3.15), define R_1^2 as the weighted goodness-of-fit of the residualizing model for the potential outcomes under treatment. Let $\xi = R_0^2 - R_1^2$, such that:

$$R_1^2 := 1 - \frac{var_w(\hat{e}_i(1) - \widetilde{\mathbf{X}}_i^\top \gamma_*^{res})}{var_w(Y_i(1) - \widetilde{\mathbf{X}}_i^\top \gamma_*)} = R_0^2 - \xi.$$

Furthermore, define the ratio $f = pvar_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*)/(1-p)var_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*)$. *Then the relative reduction in variance from residualizing is given by:*

$$\textit{Relative Reduction} := \frac{asyvar_{\tilde{F}}(\hat{\tau}_{wLS}) - asyvar_{\tilde{F}}(\hat{\tau}^{res}_{wLS})}{asyvar_{\tilde{F}}(\hat{\tau}_{wLS})} = R_0^2 - \frac{1}{1+f} \cdot \xi$$

Corollary 3.3.1, proof available in Appendix B.1. decomposes the overall relative reduction in variance of the weighted least squares estimator from residualizing into two components: (1) our proposed diagnostic measure R_0^2 and (2) a factor, represented by ξ, that measures the difference in prediction error between the experimental control and experimental treated potential outcomes. If the residualizing model explains similar amounts of variation across both the treated and control potential outcomes, then $R_1^2 \approx R_0^2$ and $\xi \approx 0$. In that scenario, R_0^2 will be roughly indicative of the expected relative reduction. When R_0^2 takes on a negative value, this is a strong indication that residualizing is unlikely to result in precision gains, since it is unlikely the prediction error will be significantly lower for treated units.

To summarize, R_0^2 can diagnose when one should expect improvements in precision from residualizing. When R_0^2 takes on negative values, researchers should not proceed with residualizing, as it is likely to result in precision loss.

3.4 Using the Predicted Outcomes as a Covariate

Thus far, we have discussed residualizing, or directly subtracting the predicted outcome values from the outcomes measured in the experimental sample. An alternative approach is to regress the outcomes measured in the experimental sample on the predicted outcomes $\hat{Y}_i$ from our residualizing model. In particular, we include $\hat{Y}_i$ as a covariate in a weighted linear regression:

$$\left(\hat{\tau}_W^{cov}, \hat{\beta}, \hat{\alpha}\right) = \underset{\tau,\beta,\alpha}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in \mathcal{S}} \hat{w}_i \left(Y_i - (\tau T_i + \beta \hat{Y}_i + \alpha)\right)^2.$$

We can extend this approach to also include pre-treatment covariates:

$$\left(\hat{\tau}_{wLS}^{cov}, \hat{\beta}, \hat{\gamma}, \hat{\alpha}\right) = \underset{\tau,\beta,\gamma,\alpha}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in \mathcal{S}} \hat{w}_i \left(Y_i - (\tau T_i + \beta \hat{Y}_i + \widetilde{\mathbf{X}}_i^\top \gamma + \alpha)\right)^2.$$

The residualizing methods we discussed in Section 3.3 can be seen as special cases of these methods where we set $\beta = 1$.

Residualizing by directly including $\hat{Y}_i$ as a covariate in the weighted least squares has many advantages. The primary advantage is that this approach allows researchers to flexibly use proxy outcomes measured in the target population. When the outcome of interest is not measured at the population level, or if the outcomes are measured in different ways across the experimental sample and the observed population data, researchers can estimate the residualizing model $g(\mathbf{X}_i)$ using alternative proxy outcomes $\tilde{Y}_i$ related to the outcome of

interest. However, use of these proxies can lead to scaling issues that limit the ability of the weighted and weighted least squares methods for post-residualizing to achieve efficiency gains. We show how including $\hat{Y}_i$ as a covariate addresses these concerns.

Additionally, as with our post-residualized estimators $\hat{\tau}_W^{res}$ and $\hat{\tau}_{wLS}^{res}$ discussed in Section 3.3, both $\hat{\tau}_W^{cov}$ and $\hat{\tau}_{wLS}^{cov}$ are consistent estimators of the PATE. Finally, including the predicted outcome $\hat{Y}_i$ as a covariate protects against efficiency loss, unlike $\hat{\tau}_W^{res}$ and $\hat{\tau}_{wLS}^{res}$ in the previous sections. This is true whether researchers rely on a proxy outcome $\tilde{Y}_i$, or if they build the residualizing model on Y_i.

Proxy Outcomes in the Population Data

There are many settings in which researchers may rely on a proxy outcome $\tilde{Y}_i$. First, an outcome measure used to estimate the residualizing model in the population data may differ from the outcome measure in the experiment. Second, even when the outcome measure used to estimate the residualizing model in the population data is in principle the same measure as the outcome of interest in the experimental data, there can be differences between $\tilde{Y}_i$ and Y_i that may arise due to differences in how the outcomes are measured or operationalized across the experimental sample and the population, or when the potential outcomes depend on context. For example, this might occur if the population is a mix of both treatment and control conditions with non-random treatment selection.

Example: JTPA Assume that we wish to generalize the impact of JTPA on employment in an experimental site to a new target site. However, in this target site, instead of current employment, we only have access to total weeks worked in the past year or whether an individual is collecting unemployment benefits, which differ from the employment indicator collected at the end-point in the experiment. These could serve as proxy measures for employment when using post-residualized weighting for generalizing the impact of JTPA to a target site. In Section 3.6 we use our two primary outcomes, earnings and an employment indicator, as proxies for one another.

Example: Get-Out-the-Vote (GOTV) Experiments Consider Get-Out-the-Vote experiments, again, where we are interested in the causal effect of a randomized GOTV message on voter turnout, which is measured by administrative voter files in the United States (e.g., Gerber and Green, 2000). Imagine, however, that we do not have administrative data available on our population, such as for all voters in the United States, but rather, we have a nationally representative survey. For many nationally representative surveys, it is infeasible to link administrative individual-level voting history data due to privacy issues and data constraints; as such, we do not have access to voter turnout. Instead, surveys often ask voters an "intent-to-vote" question, which can proxy for actual voter turnout. Our proposed method can use this "intent-to-vote" variable to build a residualizing model.

Example: Education Experiments Imagine that researchers are primarily interested in the causal effect of small class sizes not on standardized outcomes such as the SAT, but rather on a curriculum-based test score specific to a state collected during a given academic year. In this case, researchers may not have access to this curriculum-based measure in the state-level population data, but may have access to related standardized testing scores. These standardized test scores may be used as a proxy to the curriculum-based test score of interest that is measured in the experimental data when constructing the residualizing model.

When using proxy outcomes to estimate the residualizing model, the efficiency gain will be impacted by how similar the proxy outcomes are to the actual outcomes of interest. More formally, consider the following decomposition of the residuals $\hat{e}_i$:

$$\hat{e}_i = \underbrace{Y_i - \tilde{Y}_i}_{\substack{\text{(a)} \\ \text{Difference between} \\ \text{Outcomes in Experiment} \\ \text{and Proxy Outcome}}} + \underbrace{\tilde{Y}_i - \hat{Y}_i}_{\substack{\text{(b)} \\ \text{Prediction Error} \\ \text{for Proxy Outcome}}}, \tag{3.17}$$

where we define $\tilde{Y}_i$ as the proxy outcome. Conceptually, $\tilde{Y}_i$ represents the proxy outcome, had it been measured for the experimental data. For example, in the JTPA experiment, $\tilde{Y}_i$ could represent the variable for collecting unemployment, had it been measured for the experimental sample.

Equation (3.17) decomposes the residual term into two components. The second component (b) is the model prediction error. This is driven by how well the chosen residualizing model $g(\mathbf{X}_i)$ fits proxy outcomes measured in the population data. The first component (a) is how similar the proxy outcomes measured in the population data are to the outcome measures used in the experimental data. If the proxy outcomes differ substantially from the outcomes measured in the experimental data, while the post-residualized weighted estimators will still be consistent (see Theorem 3.3.1), there may be losses in efficiency from residualizing, regardless of how much we are able to minimize the prediction error in the second term (b).

Consistency

Like the previously proposed post-residualized weighted estimators $\hat{\tau}^{res}_W$ and $\hat{\tau}^{res}_{wLS}$, both $\hat{\tau}^{cov}_W$ and $\hat{\tau}^{cov}_{wLS}$ will be consistent estimators of the PATE. This follows from the fact that $\hat{Y}_i = \widehat{g}(\mathbf{X}_i)$ is just a function of pre-treatment covariates $\mathbf{X}_i$. In this sense, we can think of $\hat{\tau}^{cov}_W$ and $\hat{\tau}^{cov}_{wLS}$ as extensions of the weighted least squares estimator, where $\hat{Y}_i$ is an additional pre-treatment covariate included in the weighted linear regression. Thus, as shown in Section 3.3, both $\hat{\tau}^{cov}_W$ and $\hat{\tau}^{cov}_{wLS}$ are consistent estimators of the PATE.

Efficiency Gain and Diagnostics

There are two advantages to using $\hat{Y}_i$ as an additional covariate. First, because $\hat{Y}_i$ is treated as a covariate in a weighted regression, the estimated coefficient (i.e., $\hat{\beta}$) can capture any potential scaling differences between the proxy outcomes and the actual outcomes of interest. While the standard post-residualized weighted estimator can account for additive differences between the proxy outcome and actual outcome, including $\hat{Y}_i$ as a covariate in a weighted regression allows for our method to additionally account for scale differences between the proxy and actual outcomes. For example, returning to the Get-Out-the-Vote experiments, intent-to-vote is often measured on a Likert scale, while voter turnout is simply a binary variable of whether the individual voted or not. In such a scenario, residualizing directly on $\hat{Y}_i$ can lead to efficiency loss, despite the fact that intent-to-vote is correlated to voter turnout.

Second, treating $\hat{Y}_i$ as a covariate protects against precision loss when the proxy outcomes are significantly different from the outcomes of interest. At worst, $\hat{Y}_i$ is unrelated to Y_i, and we expect the coefficient in front of $\hat{Y}_i$ to be near zero. When this occurs, we expect the variance of the post-residualized weighted estimator when using $\hat{Y}_i$ as a covariate to be similar to the variance of a conventional estimator that does not include population-level outcome information. More formally:

Corollary 3.4.1 *The post-residualized weighted estimators using $\hat{Y}_i$ as a covariate will be at least as asymptotically efficient as the standard weighted estimators:*

$$asyvar(\hat{\tau}_W) - asyvar(\hat{\tau}_W^{cov}) \geq 0$$

$$asyvar(\hat{\tau}_{wLS}) - asyvar(\hat{\tau}_{wLS}^{cov}) \geq 0,$$

This result follows from Ding (2021), who shows that the variance of an estimator that accounts for pre-treatment covariates will be asymptotically less than or equal to the variance of an estimator that does not account for pre-treatment covariates.

To account for whether or not the re-scaled predicted outcomes sufficiently explain enough of the variation in the experimental sample, we extend our previously proposed diagnostic measures to the proxy outcome setting. To do so, we propose using sample splitting across the control units in the experimental sample. We regress $\hat{Y}_i$ on the control outcomes Y_i across one subset of the sample. This allows us to estimate $\hat{\beta}$. Then using $\hat{\beta}$, we can estimate residuals, accounting for the scaling factor (i.e., $Y_i - \hat{\beta}\hat{Y}_i$), across the held out sample, and calculate the $\hat{R}_0^2$ and $\hat{R}_{0,wLS}^2$ diagnostics from before. We finally conduct cross-fitting, i.e., repeating the same procedure by flipping the role of training and test data and then averaging diagnostics from both sample splits.

When to worry about external validity

When diagnostic measures indicate that post-residualized weighting is unsuitable for the data at hand, it is important to understand why. In particular, Equation (3.17) shows

that efficiency loss could occur from (1) the residualizing model's prediction error, and (2) the difference between the outcomes in the population and the outcomes measured in the experimental sample. Low diagnostic values indicate that post-residualizing methods may not provide efficiency gains, however it may also be indicative of contextual differences in the potential outcomes, which affect the validity of the PATE estimate.

The residualizing model's prediction error, from equation (3.17)-(b), can be estimated through cross validation using the population-level data. Researchers can hold out random subsets of the population-level data when estimating the residualizing model and calculate the prediction error across the held out sample. If the cross validated error is large, there will likely be little to no efficiency gains from using post-residualized weighting due to poor prediction, even if the true outcome Y_i is used to estimate $\hat{g}$. The difference between the outcomes Y_i and the proxy outcome $\tilde{Y}_i$, from equation (3.17)-(a), can be estimated when the proxy outcome is also measured in the experimental sample. For example, in the Get-Out-the-Vote experiments, researchers may have voters' intent-to-vote in the experimental sample. Alternatively, in the education experiments, researchers could measure both the curriculum-based test score and the standardized test score in the experimental sample. In JTPA, employment outcomes may be operationalized differently across sites.

In settings where $\tilde{Y}_i$ is not measured in the experimental data, researchers can still use the proposed diagnostic measures to determine if there are concerns about generalizability. For example, if the cross validated prediction error is low, but the diagnostics indicate that post-residualized weighting will not improve efficiency, then this indicates that the residualizing model predicts the population outcomes well, but does not predict outcomes measured in the experiment well. This could be due to two problems. First, if the population outcome is a proxy measure of the outcome measured in the experimental sample, then it could be that the measure used in the population data is not a good proxy for the experimental outcome. Alternatively, if researchers believe that the experimental and population outcomes are measured in the same way, then a low or negative R_0^2 measure, in conjunction with low cross validated prediction error, would indicate that the outcome-covariate relationships in the population are considerably different from the outcome-covariate relationships in the experimental sample. In this case, there may be limited external validity of the experiment due to a failure of the consistency of parallel studies assumption, since the potential outcomes may depend on context (see Egami and Hartman (2022) for more discussion).

3.5 Simulation

We now run a series of simulations to empirically examine the proposed post-residualizing method. In total, we consider four different data-generating scenarios, based on the following model for the potential outcomes under control:

$$\begin{aligned} Y_i(0) \;&= \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma_1 X_{1i}^2 + \gamma_2 \sqrt{|X_{2i}|} + \gamma_3 \big(X_{1i} \cdot X_{2i}\big) \\ &+\; \beta_S \cdot (1 - S_i) \cdot (\alpha + \beta_3 X_{1i} + \gamma_4 X_{1i} \cdot X_{2i}) + \varepsilon_i, \end{aligned}$$

Table 3.2: Summary of Different Simulation Scenarios

	Proxy and Experimental Sample Outcomes	DGP Type
Scenario 1	Identical DGP ($\beta_S = 0$)	Linear ($\gamma_\circ = 0$)
Scenario 2	Identical DGP ($\beta_S = 0$)	Nonlinear ($\gamma_\circ \neq 0$)
Scenario 3	Different DGP ($\beta_S \neq 0$)	Linear ($\gamma_\circ = 0$)
Scenario 4	Different DGP ($\beta_S \neq 0$)	Nonlinear ($\gamma_\circ \neq 0$)

where (X_{1i}, X_{2i}) are observed pre-treatment covariates, and $S_i \in \{0, 1\}$ is a binary indicator variable, taking the value of one when unit i is in the experimental data, and taking the value of zero when unit i is in the population data. β_S controls for differences between the experimental sample and population data outcomes, and the γ terms dictate the nonlinearity of the data generating processes.

We then define the treatment effect model as follows:

$$\tau_i = \alpha_\tau + X_{\tau,i},$$

where $X_{\tau,i}$ is an observed pre-treatment covariate that governs treatment effect heterogeneity. Therefore, the observed outcomes take on the following form: $Y_i = Y_i(0) + \tau_i \cdot T_i$. We provide additional details, including the sampling model and distributions of observed covariates in Appendix B.3.

The first two scenarios test the method when the outcome measures for both the experimental sample and the population data are drawn from the same underlying data generating process, to explore a setting where the outcome is measured identically across the experiment and target population (i.e., $\beta_S = 0$). The third and fourth scenarios use different data generating processes to simulate a context where the outcome measure differs between the experimental sample and the population (i.e., $\beta_S \neq 0$). This represents real-world settings in which the outcomes in the experimental sample and population are measured differently, or are situated in different contexts, which can result in differences in the outcome-covariate relationships. This setting also mimics the case in which researchers use a proxy outcome. For both of these settings, we consider a version of the data generating processes that is linear in the included covariates, and a second version that contains nonlinearities. Table 3.2 provides a summary of the different scenarios.

We compare conventional and post-residualized versions of two sets of estimators in each simulation. We perform post-residualizing in two different ways: the first directly residualizes the outcomes in the experimental sample by subtracting the predicted outcomes, and the second treats the predicted outcomes as a covariate in a weighted regression. Therefore, we compare a total of six different estimators: (1) the weighted estimators $\hat{\tau}_W$, $\hat{\tau}_W^{res}$, $\hat{\tau}_W^{cov}$, and (2) weighted least squares (wLS) $\hat{\tau}_{wLS}$, τ_{wLS}^{res}, and $\hat{\tau}_{wLS}^{cov}$. The difference-in-means estimator (DiM) is also provided as a baseline with no weighting adjustment.

The underlying sampling process is governed by a logit model. At each iteration of the simulation we draw both a biased experimental sample and a random sample of a larger target population as the population data. The population data is used to estimate the residualizing model and sampling weights. We use entropy balancing to estimate the sampling weights $\hat{w}_i$ for each simulation. Our residualizing model is a regression that contains all the pairwise interactions of the included covariates. The weighted least squares regression includes covariates additively without any interactions.[6]

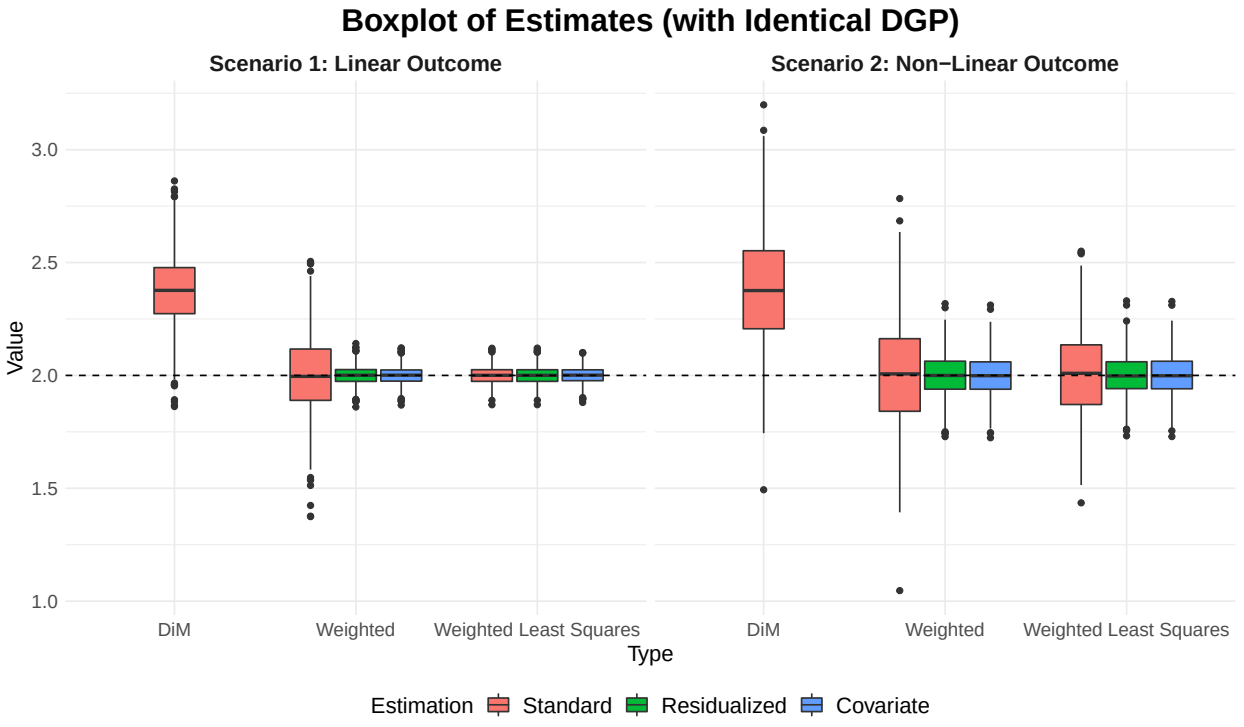

Figure 3.2: Summary of estimates across 1,000 simulations for Scenarios 1 and 2, in which the experimental sample and population outcomes are drawn from the same data generating process. The dashed line represents the super-population PATE.

Overall, we find that when the underlying outcome model is complex and contains nonlinear terms, our post-residualizing method exhibits large precision gains compared to conventional methods. When there is no difference between the population-level outcomes and the outcomes in the experimental sample, seen in Figure 3.2, direct residualizing and including $\hat{Y}_i$ as a covariate performs identically.

[6]It is possible in practice to include nonlinear transformation of pre-treatment covariates in the regression adjustment step. However, we have omitted it to illustrate the efficiency gains that can be obtained from accounting for nonlinearities through the residualizing step. This mimics how, in practice, we are able to fit more complex models to more data.

Scenario 1 When we consider a linear DGP, residualizing results in substantial precision gains for the weighted estimator. However, for the weighted least squares estimator, residualizing does not result in precision gains, because the covariate adjustment taking place in the weighted regression already includes the linear terms in the data generating process, and thus, the residualizing step does not model anything in the outcomes that is not already accounted for in the wLS regression.

Scenario 2 When we include nonlinear terms into the data generating process, residualizing results in precision gains for all of the estimators, because the residualizing model is able to account for some of the nonlinearities that the wLS regression does not account for. It is worth noting that the estimated residualizing model is not a correct specification of the underlying outcome model for the population data. However, because we have included the pairwise interactions between the covariates, the residualizing model is able to significantly reduce the variance for both estimators, even without accounting for all of the nonlinear terms in the underlying data generating process.

Scenarios 3 and 4 Next we consider a difference in the underlying data generating process between the experimental and population outcomes, presented in Figure 3.3. We operationalize this by including an interaction between treatment, the sampling indicator, and covariates. The degree to which the two processes differ is varied across different simulations using a single parameter, β_S. When the difference is relatively small (i.e. small $|\beta_S|$), the two methods used to residualize the experimental sample outcomes perform identically. This is evident by a lower RMSE when $|\beta_S| < 2$ for the post-residualized weighted estimators. When the difference in the DGP are large (i.e., $|\beta_S| > 2$), residualizing by directly subtracting the outcomes from the predicted outcomes results in precision loss, evident by a larger RMSE for the post-residualized weighted estimator $\hat{\tau}_W^{res}$, and for the post-residualized weighted least square estimator $\hat{\tau}_{wLS}^{res}$ when the true DGP is nonlinear. However, treating the predicted outcomes as a covariate in a weighted linear regression $\hat{\tau}_W^{cov}$ and $\hat{\tau}_{wLS}^{cov}$ allows for precision gain, even in these settings. We see that at worst, the covariate-based residualizing approach performs equivalently to the conventional estimators.

It is important to highlight that regardless of the degree of divergence between the population and experimental sample DGP's, post-residualized weighting is able to maintain nominal coverage. Furthermore, our proposed diagnostic measures adequately capture when we expect to gain or lose precision from residualizing. We provide coverage results and a summary of the diagnostic performance in the Appendix B.4.

3.6 Application: Job Training Partnership Act

To evaluate and benchmark how our proposed post-residualizing method may work in practice, we now turn to an empirical application. Recall that, while the original study evaluated the overall impact of JTPA, our focus is on generalizing the effect of each site individually

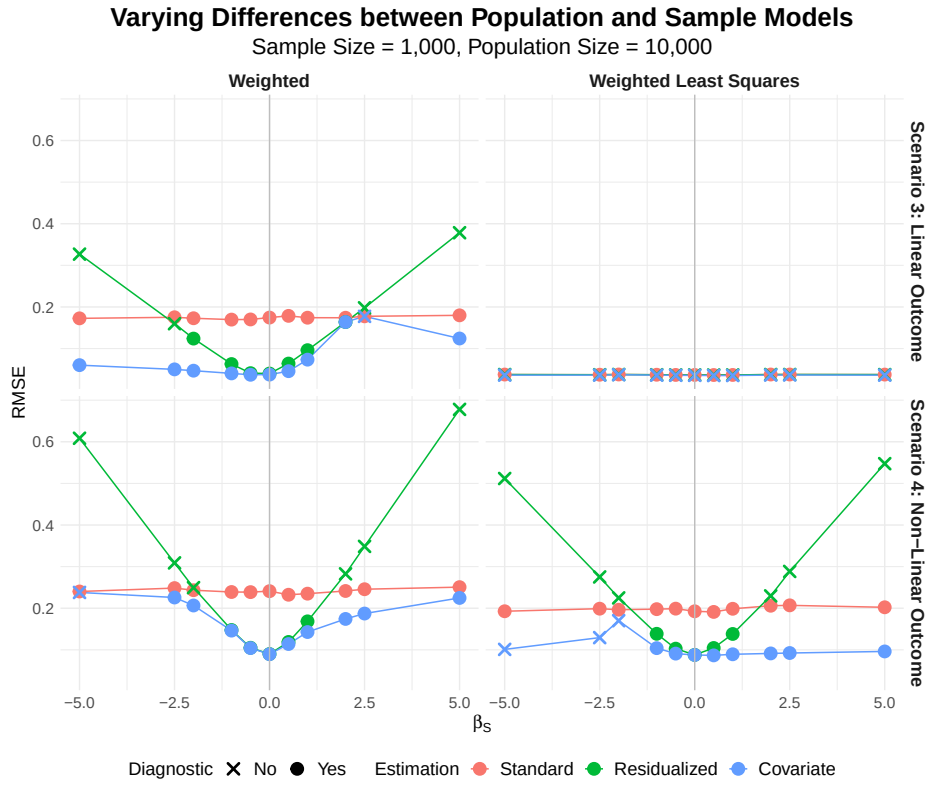

Figure 3.3: Plot of RMSE of the different estimators for Scenarios 3 and 4, in which the experimental sample and population outcomes are drawn from different data generating processes. β_S controls for how different the two processes are (i.e., the larger $|\beta_S|$ is, the larger the difference is between the two processes). The standard estimators are presented in black and the residualized estimators in gray and light gray. We label all the points for which the diagnostic measure estimates a loss in efficiency ($\times$) or gain ($\bullet$) from residualizing more than 50% of the time in the 1,000 iterations.

to the other 15 sites. More specifically, in our leave-one-out analysis for each site, we define the PATE as the average treatment effect among units in the remaining 15 sites. We then generalize the experimental results from one site to the population defined by the pooled remaining sites. This allows us to validate our method's performance by comparing our PATE estimators to the pooled experimental benchmark in the remaining sites. We evaluate generalizability for two outcomes: employment status (binary outcome) and total earnings (zero-inflated, continuous outcome).

Post-Residualized Weighting

Residualizing model

We include baseline covariates measured at the interview stage of the JTPA study. The covariates include measures of age, previous earnings, marital status, household composition, public assistance history, education and employment history, access to transportation, and ethnicity. More details about the pre-treatment covariates can be found in Appendix B.5.

We construct our residualizing model using an ensemble method, the *SuperLearner* (van der Laan et al., 2007). The ensemble model contains the Random Forest, with varying hyperparameters, and the LASSO, with hyperparameters chosen using cross validation. This allows us to capture nonlinearities in the data through the Random Forest, as well as linear relationships using the LASSO (van der Laan et al., 2007). We build separate models for the probability of employment and total earnings. We fit our residualizing model on the control units from the target population. Details can be found in the Appendix B.5.

Estimators

We estimate the PATE using two different estimators: the weighted estimator and the weighted least squares estimator (wLS). For each estimator, we consider the conventional estimators ($\widehat{\tau}_W$ and $\widehat{\tau}_{wLS}$), the post-residualized estimators directly subtracting the predicted outcomes from the outcomes in the experimental sample ($\widehat{\tau}^{res}_W$ and $\widehat{\tau}^{res}_{wLS}$), and the post-residualized estimators using the predicted outcomes as a covariate ($\widehat{\tau}^{cov}_W$ and $\widehat{\tau}^{cov}_{wLS}$). Sampling weights are estimated using entropy balancing in which we match main margins for age, education, previous earnings, race, and marital status (Hainmueller, 2012). Our weighted least squares (wLS) estimators include age, education level, and marital status as controls. Standard errors are estimated using heteroskedastic-consistent standard errors (HC2).

Diagnostics

For each site, we compute the pseudo-R^2 diagnostics. This can be done directly for the post-residualized weighted and weighted least squares estimators. When treating $\hat{Y}_i$ as a covariate, we use sample splitting to estimate the pseudo-R^2 values. Because some of the experimental sites comprise of relative few units (i.e., the experimental site of Montana contains only 38 units total), we perform repeated sample splitting, taking the average of the diagnostic across the repeated splits (Jacob, 2020; Chernozhukov et al., 2018).

Results

Bias

Because the conventional estimators and our proposed approach rely on the same identification assumptions, we first want to verify that the overall bias in the PATE estimation is not affected by the post-residualized weighting step. Across all 16 sites, the point estimates from post-residualized weighting do not change substantially from standard estimation approaches. Even in experimental sites in which it may not be advantageous to perform post-residualized weighting for efficiency gains, point estimates from post-residualized weighting methods are close to those from the conventional weighting estimators. We report the mean absolute error for all 16 sites in Appendix C.3.

Diagnostics

To evaluate whether the post-residualized weighting estimators provide efficiency gains over conventional approaches, we estimate our diagnostics. Supplementary Materials Table A9 summarizes the performance of the diagnostic measures across all 16 sites for both earnings and employment.

On average, we see that the proposed diagnostic measures are able to adequately capture when researchers should expect precision gains from residualizing. The $\hat{R}_0^2$ diagnostic has a high true positive rate for both directly residualizing and using $\hat{Y}_i$ as a covariate. As such, when the diagnostic measures indicate that researchers should residualize, residualizing results in precision gains. In the case when we are directly residualizing, the diagnostic measure also has a relatively high true negative rate, which implies that when $\hat{R}_0^2 < 0$, there is a loss in precision from directly residualizing. In the case of including $\hat{Y}_i$ as a covariate, there is a greater false negative rate, as the diagnostic tends to be more conservative in this setting. This is especially noticeable when employment is the outcome. Many of the false negatives here correspond to estimated $\hat{R}_0^2$ values that are negative, but very close to zero.

Efficiency Gain

Results on the efficiency gains to post-residualized weighting are summarized in Table 3.3, and graphically displayed in Figure 3.4. Restricting our attention to the sites for which the $\hat{R}_0^2$ values are greater than zero, there is a large reduction in variance overall from residualizing. When directly residualizing, for earnings, residualizing results in a 21% reduction in estimated variance for the weighted estimator and a 12% reduction for the weighted least squares estimator. For employment, directly residualizing leads to a 10% reduction in estimated variance for the weighted estimator and a 5% reduction for the weighted least squares estimator.

When using $\hat{Y}_i$ as a covariate, we see that including the predicted outcomes as a covariate results in a 25% reduction in variance for the weighted estimator and 16% reduction for weighted least squares when earnings is the outcome. For employment, adjusting for the

Summary of Standard Errors across Experimental Sites Subset by Diagnostic

	Earnings			Employment		
	Number of Sites	Conventional	Post-Resid. Weighting	Number of Sites	Conventional	Post-Resid. Weighting
Weighted						
Direct Residualizing	10	2.42	2.13	11	8.33	7.81
$\hat{Y}_i$ as Covariate	7	2.17	1.86	1	5.58	5.01
Weighted Least Squares						
Direct Residualizing	12	2.71	2.56	11	7.88	7.64
$\hat{Y}_i$ as Covariate	7	1.87	1.71	1	5.56	5.45

Table 3.3: Summary of gains to post-residualized weighting. Columns 1 and 4 give the number of sites for which the diagnostic measure indicates gains to post-residualized weighting. The average standard error among selected sites are presented for the conventional estimators (columns 2 and 5) and post-residualized estimators (columns 3 and 6).

predicted outcomes results in a 9% reduction in variance for the weighted estimator, and a 4% reduction for the weighted least squares.

There are several takeaways to highlight. First, we see that directly residualizing the outcomes can result in significant precision gain. In particular, the reduction in variance in the post-residualized weighted least squares demonstrates the advantage residualizing has over just using regression adjustment. Second, the larger reduction in variance from using $\hat{Y}_i$ as a covariate underscores the value of being able to capture the scaled relationships between the outcomes in the population data and in the experimental sample.

Figure 3.4 shows the relative variance of the PATE estimators to the unweighted SATE. It is well known that PATE estimators typically have higher variance than the SATE (Miratrix et al., 2018), however we see that with the post-residualized method, some of the precision loss incurred from the weighted PATE estimators can be offset. Table 3.3 provides a summary of the standard errors of the PATE estimators, relative to the difference-in-means estimators.

In the left panel of Figure 3.4, we also report the results when pooling all 16 sites together, which represents the setting in which researchers do not use the diagnostic and naively perform post-residualized weighting across all settings. We still generally see some improvements in precision from using post-residualized weighting. However, the improvements are much smaller than in the setting in which we subset to sites using the diagnostic measure. As such, we recommend that when possible, researchers should use the proposed diagnostic measures.

Figure 3.4: Reduction in Variance from using post-residualized weighting. We calculate the variance of the estimators, relative to the variance of the difference-in-means (DiM) estimator. We can interpret the y-axis as the amount of variance inflation that is incurred from generalization, and see that using the proposed method of incorporating population data can allow us to offset some of the precision loss incurred from re-weighting. We see that when using the proposed diagnostic measure, post-residualized weighting results in substantial precision gains across all four estimators.

3.7 Conclusion

In this paper we introduce post-residualized weighting as a method for mitigating the precision cost of generalizing experiments to larger populations. Existing estimators for population effects typically have high variance, especially if some sampling weights are extreme (Miratrix et al., 2018), making it difficult for policymakers and practitioners to draw conclusions about the impact of treatment in the target population. For example, in our stylized example, a single site from the JTPA might not be representative of the full experiment, so a generalized estimate based on it would potentially be too lacking in precision to inform any policy decision. Our precision gains come from leveraging a valuable type of data that has been typically unused in the generalizability literature so far: outcome data measured in the target population.

To assess the benefits of our approach in practice, we re-evaluate the impact of the Job Training Partnership Act (JTPA), using the multi-site nature of the experiment to benchmark the performance of our estimators relative to common methods using a within study comparison approach. We evaluate two outcomes, employment and earnings. We find that the post-residualized methods result in a 5-25% average reduction in variance, and that confidence intervals maintain nominal coverage. In particular, we achieve the most significant gains from including the predicted outcomes as a covariate, underscoring the value of this method when scaling issues may be present in the relationship between the outcomes in the

population data and in the experimental sample. Finally, our diagnostic measures accurately capture when the post-residualized estimators result in precision gains in estimation of the PATE.

In short, our proposed method first builds a flexible model using population outcome and covariate data, which is then used to residualize the experimental outcome data. We show that post-residualized weighting estimators, which rely on residualized outcomes, are consistent for the PATE under the same identifying assumptions as current methods. However, by utilizing residualized outcomes, the post-residualized weighting estimators can obtain large precision gains over conventional approaches. We propose three classes of post-residualized weighting estimators: a weighting estimator using the residualized experimental outcomes; a weighted least squares estimator based on the residualized experimental outcomes; and an extension of weighted least squares in which the predicted values of the residualizing model are included as a covariate.

Our proposed framework has many advantages. As discussed in Section 3.3, the residualizing model, $g(\mathbf{X}_i)$, is an "algorithmic model," which merely needs to adequately predict the outcomes measured in the experiment, but does not need to be correctly specified. This allows researchers a great deal of flexibility in constructing it. In Section 3.4 we discuss how researchers can leverage proxy outcomes that are correlated with, but different from, the outcome measured in the experimental setting. Finally, we provide diagnostic measures, based on the outcomes measured among experimental controls, that allow researchers to determine whether post-residualized weighting will likely improve precision in estimating the PATE.

We evaluate our three post-residualized estimators through simulation studies and an empirical application. Our simulations and JTPA application show significant precision gains from post-residualized weighting, and confirm the performance of the diagnostic measure to differentiate when researchers should expect precision gains from post-residualized weighting. We also find that including the predicted outcomes as a covariate ensures that post-residualized weighting does not hurt precision.

Chapter 4

Variance-based Sensitivity Analysis for Weighting Estimators

4.1 Introduction

In observational studies of causal effects, researchers must address possible confounding effects from non-random treatment assignment. Typically, one relies on pre-treatment covariates either to re-weight units based on propensity of treatment, or to model the outcome of interest. In practice, researchers have no way of knowing whether the observed covariates include all potential confounders. When confounders are omitted, the resulting estimators will be biased. Sensitivity analyses speak to this concern by allowing researchers to assess the robustness of their results to omitted confounders (Cornfield et al., 1959). In a sensitivity analysis, a researcher introduces a parameter describing the amount of unobserved confounding present and redoes the analysis under different values of this parameter, determining the set of values for which the results of the study will be reversed. The robustness of the study may then be evaluated by reasoning about the plausibility of these values.

In contrast to typical estimands, parameters in sensitivity analysis are inherently unindentifiable, because they are designed to describe an omitted variable. Thus, there exists a trade-off between how complex the sensitivity analysis is, and how informative the sensitivity analysis can be. For example, Dahabreh et al. (2019) proposed a sensitivity analysis in which researchers can obtain both an adjusted point estimate and the associated uncertainty from omitting a confounder. However, the sensitivity analysis requires researchers to directly model the bias that arises from omitting a confounder. In contrast, Zhao et al. (2019) introduced a sensitivity analysis that only requires one parameter and allows researchers to estimate confidence intervals that account for the unobserved confounder. However, the resulting intervals are often extremely wide, making it difficult to reason about whether or not there is sensitivity from omitting a confounder.

In the following paper, we introduce a new sensitivity model known as the *variance-based sensitivity model* that provide the flexibility and generality of existing sensitivity analyses,

while simultaneously allowing researchers to estimate narrower, more informative bounds for weighted estimators. The proposed sensitivity models constrain distributional differences in the weights that arise from omitting a confounder, and do not rely on additional assumptions on the outcome, confounder, or treatment assignment mechanism.

The paper provides three primary contributions. First, the proposed variance-based sensitivity model can be parameterized by a single sensitivity parameter, an R^2 measure. Unlike previously proposed sensitivity analyses, our parameter is both bounded and standardized on an interval of 0 to 1. We develop formal benchmarking approaches that allow researchers to use observed covariates to reason about plausible values for the R^2 value, providing much-needed interpretability.

Second, we introduce a method for estimation of valid confidence intervals under the variance-based sensitivity model. We give a closed-form solution for the maximum bias that can occur for a fixed set of sensitivity models, which we denote the *optimal bias bound*. We also provide extensions for incorporating additional substantive knowledge about the confounder into the optimal bias bound, further restricting the range of plausible bias.

Finally, we show that a variance-based sensitivity analysis can be formulated as a bias maximization problem, with a constraint on the weighted average error. We formalize the relationship between the variance-based sensitivity model and alternative approaches, which rely on constraining a worst-case error. By moving away from characterizing bias from the perspective of a worst-case error, variance-based sensitivity analysis is able to estimate more informative and stable bounds.

The paper is organized as follows. Section 4.2 gives set-up and notation. Section 4.3 introduces the variance-based sensitivity model. Section 4.4, compares the proposed sensitivity models to alternative sensitivity approaches. Proofs and extended discussion are provided in the Appendix.

4.2 Background

Set-Up and Notation

To begin, we consider an observational study with n individuals. Define Z_i as a binary treatment assignment variable, where $Z_i = 1$ when unit i is assigned to treatment, and 0 otherwise, $Y_i(1)$ and $Y_i(0)$ are potential outcomes, and $\mathcal{X}_i$ is a vector of pre-treatment covariates. Let the tuple $(Y_i(1), Y_i(0), X_i, Z_i)$ for all $i \in \{1, \ldots, n\}$ be independently and identically distributed from an arbitrary joint distribution, where $Y(1), Y(0) \in \mathbb{R}^n$, $Z_i \in \{0,1\}^n$ and $\mathcal{X} \in \mathbb{R}^{n \times p}$.

Throughout, we invoke the standard SUTVA assumption—i.e., no interference, with treatments identically administered across all units Rubin (1980). Thus observed outcomes Y_i can be written as $Y_i := Y_i(1) \cdot Z_i + Y_i(0) \cdot (1 - Z_i)$. Because treatments are not randomly assigned in an observational study, we must also invoke an additional assumption, known as the conditional ignorability of treatment assignment:

Assumption 6 (Conditional Ignorability of Treatment Assignment)

$$Y_i(1), Y_i(0) \perp\!\!\!\perp Z_i \mid \mathcal{X}_i$$

Assumption 6 states that conditional on a set of pre-treatment covariates $\mathcal{X}$, treatment assignment is independent of potential outcomes (i.e. no further confounding remains).

In addition to Assumption 6, we also assume overlap, such that conditional on some set of pre-treatment covariates, the probability of being assigned treatment is non-zero (Rosenbaum and Rubin, 1983):

Assumption 7 (Overlap) *For all* $x \in \mathcal{X}$, $0 < P(Z_i = 1 \mid \mathcal{X} = x) < 1$.

Our primary estimand is the average treatment effect for the treated (ATT):

$$\tau := \mathbb{E}(Y_i(1) - Y_i(0) \mid Z_i = 1).$$

However, the proposed method can easily be extended for estimating the average treatment effect (ATE). Furthermore, all proofs and derivations are done with respect to a general missingness indicator, such that the results can be applied to general, missing data settings, such as weighting for external validity or survey weighting (see Appendix C.1 for more discussion and details).

Weighted estimators, which are popular for estimating causal effects in observational setting, adjust for distributional differences in the pre-treatment covariates $\mathcal{X}$ across the treatment and control groups. For the ATT, a common weighted estimator is as follows:

$$\hat{\tau}_W = \frac{1}{\sum_{i=1}^n Z_i} \sum_{i=1}^n Z_i Y_i - \underbrace{\frac{\sum_{i=1}^n (1 - Z_i) Y_i w_i}{\sum_{i=1}^n (1 - Z_i) w_i}}_{\text{Weighted Control Mean}}.$$

A common choice of weights is inverse propensity weights, where $w_i = P(Z_i = 1 \mid \mathcal{X})/P(Z_i = 0 \mid \mathcal{X})$. These weights are often constructed using a logistic regression to predict the probability of treatment assignment. Recent balancing approaches provide a semi-parametric option for researchers to estimate weights by minimizing the distributional difference between the treatment and control groups, without modeling the underlying probabilities (see Ben-Michael et al. (2021) for a recent review on balancing weights).

Under the correct specification of the weights, if the correct set of covariates are included, the weighted estimator will be a consistent and unbiased estimate of the true treatment effect. However, in practice, there is no way of knowing whether Assumption 6 holds. We propose a set of sensitivity models that characterize the bias of a weighted estimator when researchers omit a variable from the set $\mathcal{X}$. More specifically, we will assume that $\mathcal{X} = \{\mathbf{X}, \mathbf{U}\}$, such that both $\mathbf{X}$ and $\mathbf{U}$ are necessary for Assumption 6 to hold. We assume that researchers

have otherwise correctly specified the weights.[1] We define w as the weights that include only $\mathbf{X}$, and w^* as the ideal weights that include both $\mathbf{X}$ and $\mathbf{U}$. We refer to the omitted variables, $\mathbf{U}$, as *confounders*. We note that the sensitivity framework will not explicitly account for settings in which (1) researchers estimate uniform weights (i.e., no adjustment for confounding), or (2) the true ideal weights are uniform. Finally, we will assume, without loss of generality, that both w and w^* are centered at mean 1.

Related Literature

A popular approach for assessing the robustness of weighted estimates to omitted confounders uses the marginal sensitivity model (Tan, 2006), in which researchers posit a bound, Λ, on the individual-level error in the weights that can arise under unobserved confounding:

$$\Lambda^{-1} \leq \frac{w_i^*}{w_i} \leq \Lambda, \qquad \text{for } i = 1, ..., n,$$

where $\Lambda \geq 1$. Λ represents the largest possible error that can arise from omitting a confounder. Researchers can bound the maximum and minimum bias that arises under a fixed Λ, and use a percentile bootstrap to estimate valid confidence intervals (Zhao et al., 2019).

In practice, the true Λ is unknown, so to conduct the sensitivity analysis, researchers posit increasing values of Λ until the estimated confidence intervals contain zero. The minimum Λ value for which the estimated intervals cross zero is denoted as Λ^*. If Λ^* is close to 1, even a small amount of error from omitting a confounder could result in an estimated effect becoming insignificant. On the other hand, if Λ^* is much larger than 1, estimated effects are only sensitive to very strong unmeasured confounders.

While the marginal sensitivity model guarantees valid intervals asymptotically, in practice, the intervals tend to be extremely conservative. This means that the estimated intervals often include the null estimate, even under low amounts of confounding. As such, when researchers' estimated bounds imply an estimated effect is no longer statistically significant, it is difficult to distinguish if this is a sign that there is sensitivity to an omitted confounder, or if the sensitivity model is overly pessimistic. Tightening these intervals often requires researchers to invoke additional constraints in the models, or parametrically model the outcomes in some way (Dorn and Guo, 2021; Nie et al., 2021). Furthermore, the underlying sensitivity parameter in the marginal sensitivity model is dependent on the worst-case error that arises from omitting a confounder. This is inherently difficult to reason about in practice, as the true value of the parameter will depend on outliers and, in asymptotic settings, can be infinitely large.

[1]For example, the framework does not explicitly account for cases in which researchers are using a probit model, when the true underlying data generating process is logistic. However, mis-specification concerns can also be addressed with the proposed framework, if researchers can write the mis-specification error as an omitted variable problem. A simple example of this is if a linear probability model is used, $\mathbf{U}$ can include non-linear functions of $\mathbf{X}$ that matter for modeling selection. We provide more discussion in Appendix C.1.

We now propose a new sensitivity model, the *variance-based sensitivity model*. The proposed framework can be viewed as a one parameter generalization of existing sensitivity frameworks in the literature that rely on bounding the error in the weights from omitting a confounder (e.g., Huang (2022), Hong et al. (2021), Shen et al. (2011), Ding and VanderWeele (2016)). We provide several key contributions. First, unlike the frameworks proposed by Hong et al. (2021) and Shen et al. (2011), the variance-based sensitivity model introduce a standardized and bounded parameterization of the confounding strength, which can help improve transparency and interpretability for applied researchers. Second, many of the aforementioned sensitivity analyses do not engage with how potential confounders may affect their inference, and are limited to discussions about movements in the point estimate. In contrast, the variance-based sensitivity model provides a method for researchers to estimate valid asymptotic confidence intervals for fixed level of confounding.

Furthermore, we formalize the connection between these variance-based approaches to alternative sensitivity approaches, which formulate sensitivity models as optimization problems. In particular, we demonstrate that the variance-based sensitivity model can be viewed as a constrained weighted L_2 norm problem, which provides a framework to compare the proposed sensitivity models with the marginal sensitivity model. Moving away from a worst-case error parameterization of the error allows researchers to obtain more informative and stable bounds under the variance-based sensitivity model. The benefits of constraining a weighted L_2 norm instead of a worst-case error is conceptually similar to the advantages highlighted in Zhang and Zhao (2022), in which authors consider a constraint on L_2 norms, and Kallus and Zhou (2018), which introduces an L_1 norm, with the added benefit of having an interpretable sensitivity parameter in the form of an R^2 value.

Running Example: NHANES

Throughout the paper, we perform a re-analysis of a study presented in Zhao et al. (2018) (as well as Zhao et al. (2019) and Soriano et al. (2021)), analyzing the effects of fish consumption on blood mercury levels. More specifically, we use data from the 2013-2014 National Health and Nutrition Examination Survey (NHANES).

Following the original study, we define the outcome of interest as the total blood mercury (in $\log_2$), measured in micrograms per liter. As such, an estimated treated-control outcome difference of 1 implies that a treated person's total blood mercury is twice that of an individual in control's total blood mercury. The treatment is defined by whether or not individuals consumed more than 12 servings of fish or shellfish in the preceding month. There are 234 total treated units and 873 control units. To account for the non-random treatment assignment, we use the available demographic data for the individuals in the survey, which include variables like gender, age, income, race, educational attainment, and smoking history to estimate entropy balancing weights (Hainmueller, 2012).

The unweighted estimate is 2.37; after accounting for pre-treatment covariates, we obtain a weighted ATT estimate of 2.15 (see Table 4.1 for a summary). Therefore, from our estimate,

we expect that on average, a treated individual who consumes more fish will have around 4 times as much total blood mercury than an individual in control.

	Unweighted (DiM)	IPW
Estimated Effect (ATT)	2.37 (0.10)	2.15 (0.11)

Table 4.1: Estimated effect of fish consumption on blood mercury levels. Standard errors are reported in parentheses.

4.3 The Variance-Based Sensitivity Model

We now introduce the *variance-based sensitivity model*. We begin by defining the sensitivity model and show that the model lends itself naturally to an R^2 parameterization. Then, we derive a closed-form solution for the maximum bias under the variance-based sensitivity model that can be directly estimated from the data, and introduce a method to estimate asymptotically valid confidence intervals under the proposed sensitivity model. Finally, we provide formal benchmarking tools to help researchers conduct their sensitivity analyses, and illustrate the proposed approach on the running example respectively. Appendix C.1 provides an extension for researchers to impose constraints on the strength of the relationship between the confounder and the outcome within the variance-based sensitivity model.

Defining a New Sensitivity Model

To begin, we define the following set as the "variance-based sensitivity model":

Definition 4.3.1 (Variance-Based Sensitivity Model)
Let R^2 be the residual variation in the true weights w^*, not explained by w:

$$R^2 := 1 - \underbrace{\frac{var(w_i \mid Z_i = 0)}{var(w_i^* \mid Z_i = 0)}}_{\substack{\text{Variation in } w^* \\ \text{explained by } w}}.$$

Then, for a fixed $R^2 \in [0, 1)$, we define the variance-based sensitivity model $\sigma(R^2)$:

$$\sigma(R^2) \equiv \left\{ w_i^* \in \mathbb{R}^n : 1 \leq \frac{var(w_i^* \mid Z_i = 0)}{var(w_i \mid Z_i = 0)} \leq \frac{1}{1 - R^2} \right\}.$$

In contrast to existing methods which constrain the worst-case, individual-level multiplicative error across the weights, the variance-based sensitivity model constrains the distributional

difference between the true weights w^* and the estimated weights w. This implicitly constrains the residual imbalance in the omitted variable. More formally, we decompose the true weight w^* into two components: (1) the weight w, and (2) the residual imbalance in $\mathbf{U}$:

$$w^* = \frac{P(Z = 1 \mid \mathbf{X}, \mathbf{U})}{1 - P(Z = 1 \mid \mathbf{X}, \mathbf{U})} = \underbrace{\frac{P(Z = 1 \mid \mathbf{X})}{1 - P(Z = 1 \mid \mathbf{X})}}_{\text{Weights } (w)} \cdot \underbrace{\frac{P(\mathbf{U} \mid \mathbf{X}, Z = 1)}{P(\mathbf{U} \mid \mathbf{X}, Z = 0)}}_{\text{Imbalance in } \mathbf{U}}, \tag{4.1}$$

where the imbalance term is a ratio of the conditional probability density function of the omitted variable across the treatment and control groups. The distributional difference between the weights w and the ideal weights w^* will be driven by the imbalance term. Intuitively, if the omitted variable $\mathbf{U}$ is very imbalanced, then accounting for the omitted variable results in very different values for w^* and w. Alternatively, if $\mathbf{U}$ is not very imbalanced, then including it will result in weights w^* that are very similar to w. Limiting the distributional difference between the true weights and estimated weights effectively restricts the amount of residual imbalance in the omitted confounder. In Section 4.4, we show that this is equivalent to constraining a weighted L_2 norm of the errors w_i^*/w_i, in contrast with the marginal sensitivity model, which constrains an L_∞ norm.

The distributional difference between the estimated weights and the true weights can be written as a function of an R^2 parameter. The R^2 parameter represents the residual variation in the true weights, not explained by the estimated weights, and is naturally bounded on an interval of $[0, 1]$.

Optimal Bias Bounds

Valid confidence intervals for a set of sensitivity models must account for two factors: (1) the bias that arises from omitting a confounder, and (2) the uncertainty associated with estimation. In the following subsection, we introduce optimal bias bounds that researchers can estimate for the variance-based sensitivity model, under a fixed R^2 value. Section 4.3 introduces a percentile bootstrap approach for researchers to simultaneously account for uncertainty in estimation.

In the following theorem, we show that for a fixed R^2, we can estimate the possible range of bias values. We refer to the minimum and maximum values of these potential bias values as the *optimal bias bounds*. However, unlike the marginal sensitivity models, in which researchers must solve a linear programming problem to identify the extrema, the variance-based sensitivity model admits a closed-form solution for the optimal bias. More specifically, the optimal bias bounds are a function of three different components: (1) a correlation bound, which represents the maximum correlation an omitted confounder can have with the outcome of interest; (2) the imbalance (represented by the R^2); and (3) a scaling factor.

Theorem 4.3.1 (Optimal Bias Bounds)
For a fixed $R^2 \in [0, 1)$, the maximum bias under $\sigma(R^2)$ (denoted as $\max_{\tilde{w} \in \sigma(R^2)} Bias(\hat{\tau}_W \mid \tilde{w})$)

can be written as a function of the following components:

$$\max_{\tilde{w}\in\sigma(R^2)} \; Bias(\hat{\tau}_W \mid \tilde{w})$$
$$= \underbrace{\sqrt{1 - cor(w_i, Y_i \mid Z_i = 0)^2}}_{\textit{(a) Correlation Bound}} \sqrt{\underbrace{\frac{R^2}{1-R^2}}_{\textit{(b) Imbalance}} \cdot \underbrace{var(Y_i|Z_i = 0) \cdot var(w_i|Z_i = 0)}_{\textit{(c) Scaling Factor}}}, \qquad (4.2)$$

with the minimum bias given as the negative of Equation (4.2)*. The optimal bias bounds are given by the minimum and maximum biases.*

Theorem 4.3.1 highlights the different components that affect the magnitude of the bias bounds. We provide more details about each component below.

Correlation Bound. The correlation bound, given by Equation (4.2)-(a), represents the maximum correlation between imbalance in an omitted confounder and outcome across the control group. Intuitively, this is similar to the marginal sensitivity model, in which Dorn and Guo (2021) demonstrated that the optimal bias bounds are obtained when the imbalance from the omitted confounder is maximally correlated with the outcome. From Equation (4.2)-(a), we see that the bound is a function of the correlation between the estimated weights and the outcome. If the estimated weights are highly correlated with the outcome, then the degree to which the residual imbalance in the omitted confounder can be correlated to the outcome is limited, and this bound is lower. However, if the correlation between the estimated weights and the outcome is relatively low, then the possible correlation between omitted-confounder imbalance and outcome has a much larger range. In the worst case, the estimated weights and the outcome are not correlated at all (i.e., $\text{cor}(w_i, Y_i \mid Z_i = 0) = 0$); then, the correlation bound will simply equal 1.

Residual Imbalance. The second component of the bias bound is the residual imbalance in an omitted confounder, and is a function of the R^2 parameter (Equation 4.2-(b)). Similar to a point made in Cinelli and Hazlett (2020), there exists an asymmetry in the drivers of bias. More specifically, as the correlation between the imbalance term and the outcome increases towards 1 (i.e., the correlation bound approaches 1), the overall impact on the bias bound is bounded at 1. In contrast, as the level of imbalance in the omitted confounder increases, the effect on the bias bounds is unbounded. In other words, as $R^2 \to 1$, the corresponding bias bounds will increase towards infinity.

Scaling Factor. The last factor in the bias bound is a scaling factor (represented by Equation (4.2)-(c)). The scaling factor comprises of the variance of the outcomes across the control units (i.e., $\text{var}(Y_i \mid Z_i = 0)$) and the variance of the estimated weights (i.e., $\text{var}(w_i \mid Z_i = 0)$). This represents the overall heterogeneity that is present in the data. More specifically, as the variance in the estimated weights increases, there is more imbalance

between the treatment and control groups that the weights are accounting for. Similarly, as the variance in the outcomes increases, there is more potential for heterogeneity to be related to the selection into treatment, making it more difficult to recover the true estimated effect.[2] The scaling factor is a function of the observed data, and is not related to the omitted confounder. However, the scaling factor serves as an amplification of any bias that would arise from omitting a confounder.

The key takeaway from Theorem 4.3.1 is that given a researcher-chosen R^2 value, the optimal bias bound in Theorem 4.3.1 is directly estimable from the data. As such, for a fixed R^2 value, researchers can directly calculate the range of possible bias values.

Constructing Confidence Intervals

We now introduce a method to construct valid asymptotic confidence intervals for the variance-based sensitivity models. Our method builds on the work of Zhao et al. (2019) and uses a percentile bootstrap to simultaneously accounts for the bias due to omitting a confounder and the uncertainty associated with estimation. Our approach is distinct from those in the partial identification literature that require known asymptotic distributions of the boundaries of the partially identified region (Imbens and Manski, 2004; Aronow and Lee, 2013); similar to the discussion provided in Zhao et al. (2019), it can difficult to characterize these distributions analytically in our sensitivity framework. Instead, the proposed bootstrap approach allows researchers to account for sampling uncertainty without explicitly characterizing the asymptotic distributions of the boundary estimates.

To begin, for a fixed R^2, we can define $\hat{\tau}(\tilde{w})$ as the weighted estimate, using weights $\tilde{w} \in \sigma(R^2)$. We can equivalently view $\hat{\tau}(\tilde{w})$ as an adjusted weighted estimate for some $\tilde{w} \in \sigma(R^2)$:

$$\hat{\tau}(\tilde{w}) := \hat{\tau}_W - \text{Bias}(\hat{\tau}_W \mid \tilde{w}), \tag{4.3}$$

where $\text{Bias}(\hat{\tau}_W \mid \tilde{w})$ is the bias of $\hat{\tau}_W$, assuming $\tilde{w}$ were the true weights (i.e., $\text{Bias}(\hat{\tau}_W \mid \tilde{w}) := \hat{\tau}_W - \hat{\tau}(\tilde{w})$), assuming the true weights are equal to $\tilde{w}$.

Applying the results from Zhao et al. (2019), for every $\tilde{w} \in \sigma(R^2)$, we can construct a confidence interval for $\tau(\tilde{w})$ using a percentile bootstrap:

$$[L(\tilde{w}), U(\tilde{w})] = \left[Q_{\alpha/2}\big(\hat{\tau}^{(b)}(\tilde{w})\big), Q_{1-\alpha/2}\big(\hat{\tau}^{(b)}(\tilde{w})\big)\right], \tag{4.4}$$

where $\hat{\tau}^{(b)}(\tilde{w})$ is the adjusted weighted estimator in bootstrap sample $b \in \{1, ..., B\}$, and $Q_\alpha(\cdot)$ denotes the α-th percentile in the bootstrap distribution. By the following theorem, $[L(\tilde{w}), U(\tilde{w})]$ will be an asymptotically valid (1-α) confidence interval for $\tau(\tilde{w})$:

Theorem 4.3.2 (Validity of Percentile Bootstrap)
Under mild regularity conditions (see Assumption 9 in the Appendix), for every $\tilde{w} \in \sigma(R^2)$:

$$\limsup_{n\to\infty} P(\tau(\tilde{w}) < L(\tilde{w})) \leq \frac{\alpha}{2} \text{ and } \limsup_{n\to\infty} P(\tau(\tilde{w}) > U(\tilde{w})) \leq \frac{\alpha}{2},$$

[2] We note that in settings when $\text{var}(Y_i \mid Z_i = 0) = 0$, there is *no* variation in the outcomes across the control group. As such, no amount of weighting will alter our estimate.

where $L(\tilde{w})$ and $U(\tilde{w})$ are defined as the $\alpha/2$ and $1-\alpha/2$-th quantiles of the bootstrapped estimates (i.e., Equation (4.4)).

Theorem 4.3.2 states that for any set of weights $\tilde{w}$, the percentile bootstrap can be applied to estimate valid confidence intervals for an adjusted, weighted estimate $\tau(\tilde{w})$. However, as discussed in the previous section, for a given R^2 value, there exists many different sets of weights $\tilde{w}$ that can be defined from $\sigma(R^2)$. As such, we apply the union method to estimate a conservative $(1-\alpha)\%$ confidence interval $\text{CI}(\alpha)$ for $\tau(\tilde{w})$:

$$\text{CI}(\alpha) = \left[Q_{\alpha/2}\left(\inf_{\tilde{w}\in\sigma(R^2)} \hat{\tau}^{(b)}(\tilde{w}) \right),\ Q_{1-\alpha/2}\left(\sup_{\tilde{w}\in\sigma(R^2)} \hat{\tau}^{(b)}(\tilde{w}) \right) \right]. \tag{4.5}$$

We can estimate $\text{CI}(\alpha)$ directly from the bootstrap samples by calculating the minimum and maximum adjusted weighted estimate for each bootstrap iteration, and then estimating the $\alpha/2$ and $1-\alpha/2$-th percentiles across the bootstrap distributions. The extrema of the adjusted weighted estimate follow directly from the results of Theorem 4.3.1:

$$\inf_{\tilde{w}\in\sigma(R^2)} \hat{\tau}(\tilde{w}) = \hat{\tau}_W - \max_{\tilde{w}\in\sigma(R^2)} \text{Bias}(\hat{\tau}_W \mid \tilde{w}) \qquad \sup_{\tilde{w}\in\sigma(R^2)} \hat{\tau}(\tilde{w}) = \hat{\tau}_W + \max_{\tilde{w}\in\sigma(R^2)} \text{Bias}(\hat{\tau}_W \mid \tilde{w})$$

As such, to estimate valid confidence intervals, researchers can use a percentile bootstrap, estimating the bias bound and calculate an adjusted point estimate in each bootstrap sample. We summarize the steps in Figure 4.1.

Conducting the Sensitivity Analysis

To conduct the sensitivity analysis, researchers estimate confidence intervals for increasing R^2 values until an estimated confidence interval just contains the null estimate; the corresponding R^2 is denoted as R^2_*. Because the R^2 is bounded on an interval $[0, 1]$, researchers are restricted by the range of values that they can posit for the different R^2 values. However, it can nonetheless be difficult to reason about the plausibility of a given R^2 value and the strength or weakness of confounders on the R^2 scale.

Previous papers have suggested the use of benchmarking to assess what may be plausible sensitivity parameters (Huang, 2022; Hartman and Huang, 2022; Cinelli and Hazlett, 2020; Hong et al., 2021; Carnegie et al., 2016; Hsu and Small, 2013). To perform benchmarking, researchers sequentially omit different observed covariates and re-estimate the weights. They can then directly calculate the error that arises from omitting each covariate and directly estimate the sensitivity parameters (or bounds for the sensitivity parameters). The estimated sensitivity parameters from omitting each covariate are then interpreted as the sensitivity parameters for omitted confounders with equivalent confounding strength to an observed covariate. If researchers have a strong substantive understanding of covariates that explain a lot of the variation in the treatment assignment mechanism or outcome, then benchmarking can be a useful tool for understanding the strength of hypothetical omitted variables.

Valid Confidence Intervals for the Variance-Based Sensitivity Model

Step 1. Fix $R^2 \in [0, 1)$ and generate B bootstrap samples of the data.

Step 2. For each bootstrap sample $b = 1, ..., B$:

1. Estimate weights $\hat{w}_i^{(b)}$ and the point estimate $\hat{\tau}_W^{(b)}$.
2. Calculate $\widehat{\text{var}}_b(Y_i)$, $\widehat{\text{cor}}_b(\hat{w}_i^{(b)}, Y_i)$, and $\widehat{\text{var}}_b(\hat{w}_i^{(b)})$, where the subscript b denotes the quantity calculated over the b-th bootstrap sample.
3. Use the optimal bias bounds (i.e., Equation (4.2)) to calculate the range of potential point estimates:

$$\hat{\tau}^{(b)}(\tilde{w}) \in \left[\hat{\tau}_W^{(b)} - \max_{\tilde{w} \in \sigma(R^2)} \text{Bias}(\hat{\tau}_W^{(b)} \mid \tilde{w}), \quad \hat{\tau}_W^{(b)} + \max_{\tilde{w} \in \sigma(R^2)} \text{Bias}(\hat{\tau}_W^{(b)} \mid \tilde{w})\right]$$

Step 3. From the B bootstrapped optimal bounds, estimate the $\alpha/2$ and $1-\alpha/2$-th percentiles of the minima and maxima values respectively to obtain valid confidence intervals (i.e., Equation (4.5)).

Figure 4.1: Summary of percentile bootstrap procedure for estimating confidence intervals.

We propose a formal benchmarking procedure for the variance-based sensitivity model. To begin, let there be p total observed covariates (i.e., $\mathbf{X} \in \mathbb{R}^{n \times p}$). Then for the j-th covariate, where $j \in \{1, ..., p\}$, we define the benchmarked weights $w^{-(j)}$ as the estimated weights, containing all covariates, except for the j-th covariate. Using $w^{-(j)}$, we can estimate the benchmarked R^2 value for an omitted confounder that is equivalently imbalanced as the j-th covariate:

$$\hat{R}^2_{(j)} = \frac{\hat{R}^2_{-(j)}}{1 + \hat{R}^2_{-(j)}}, \qquad \text{where } \hat{R}^2_{-(j)} := 1 - \frac{\text{var}(w_i^{-(j)})}{\text{var}(w_i)}. \tag{4.6}$$

$\hat{R}^2_{(j)}$ represents the R^2 value of an omitted variable that has the same amount of residual imbalance as the j-th covariate.[3] More specifically, $R^2_{(j)}$ corresponds to an omitted variable with the same amount of imbalance, after controlling for $\mathbf{X}$, as the j-th covariate, after controlling for $\mathbf{X}^{-(j)}$.

When interpreting the benchmarking results, it is important to consider that the magnitude of the benchmarked R^2 values is determined by the *residual* imbalance. More con-

[3]The reason we cannot directly use $\hat{R}^2_{-(j)}$ as an estimate for the benchmarked R^2 value comes from the fact that we must adjust for changes in the baseline variation of the weights between w and w^*. We refer readers to Cinelli and Hazlett (2020) and Huang (2022) for more discussion on this point.

cretely, we consider the variables *income* and *educational attainment* in the running example. We expect that both income and educational attainment will be predictive of individuals' propensity for fish consumption. However, omitting just income may not result in a very large R^2 value, because by balancing educational attainment, we have implicitly controlled for some of the imbalance in income. The benchmarked R^2 parameter thus represents the setting in which researchers have omitted a variable that, when controlling for all the other observed variables, has the same amount of residual imbalance as income after controlling for educational attainment. In cases when researchers wish to consider omitting a variable similar to a set of collinear variables, they can omit subsets of variables and perform the same benchmarking exercise.

Formal benchmarking can also be used to assess the plausibility of the event $R^2 \geq R^2_*$. More specifically, we can directly compare the benchmarked $\hat{R}^2_{(j)}$ values for $j \in \{1, ..., p\}$ with the estimated R^2_* to see how much more or less imbalanced an omitted confounder must be, relative to an observed covariate, in order to result in an R^2 value equal to R^2_*. We refer to this as the *minimum relative imbalance* (MRI):

$$\text{MRI}(j) = \frac{R^2_*}{\hat{R}^2_{(j)}}.$$

If the MRI is small (i.e., $\text{MRI}(j) < 1$), the omitted confounder need not be very imbalanced, relative to the j-th covariate, in order to make a null result plausible. In contrast, if the MRI is large (i.e., $\text{MRI}(j) > 1$), then the omitted confounder must be more imbalanced than the j-th observed covariate to make a null result plausible.

Formal benchmarking offers an opportunity for researchers to incorporate their substantive understanding into the sensitivity analysis and provides much-needed interpretability for the sensitivity framework. In particular, when researchers have strong priors about which underlying observed variables control the treatment assignment mechanism, formal benchmarking is very useful for reasoning about the plausibility of an omitted confounder strong enough to explain observed results in the absence of a true effect.

Illustration on NHANES

In our running example, we begin by varying the R^2 parameter across the range $[0, 1)$, and estimate the corresponding the 95% confidence intervals. We estimate $R^2_* = 0.57$, such that if $R^2 \geq 0.57$, the intervals contain the null estimate. This implies that if the omitted confounder explains 57% or more of the variation in the true weights, our estimated effect of fish consumption on blood mercury levels is no longer significantly different from the expected distribution under the null.

To assess the plausibility of an omitted confounder resulting in an R^2 value of 0.57, we perform formal benchmarking and estimate benchmarked $\hat{R}^2$ values for each covariate. Omitting a confounder like race, educational attainment, or income results in the largest R^2 values. More specifically, omitting a confounder with equivalent confounding strength

to race results in an R^2 of 0.19, while omitting a confounder with equivalent confounding strength to educational attainment or income results in an R^2 of 0.17 and 0.14, respectively. From these results, we see that an omitted confounder would have to explain around 3 times the variation in true weights as the strongest observed covariate, race, in order for the R^2 value to equal the cutoff value. We argue that while mathematically possible, the plausibility of a confounder resulting in the threshold $R^2_* = 0.57$ value is low.

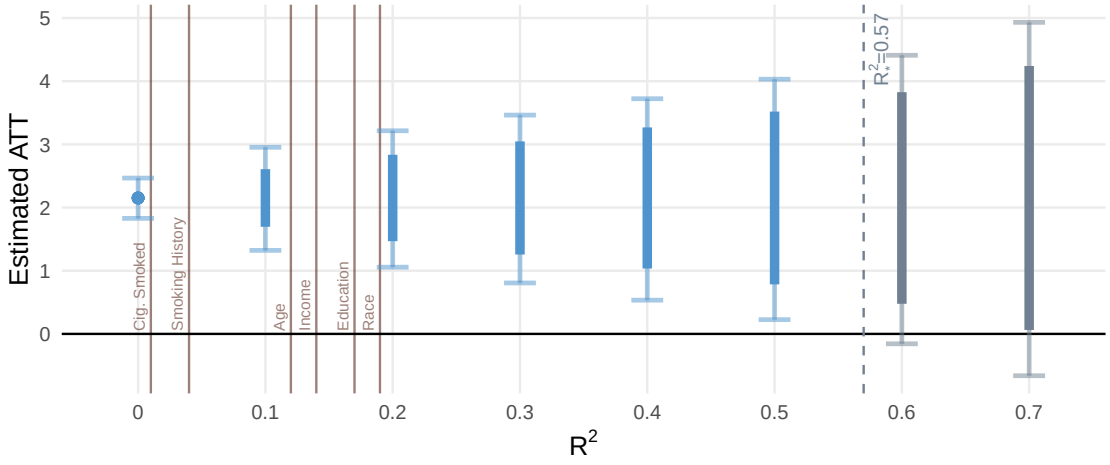

Figure 4.2: Results from the sensitivity analysis under the variance-based sensitivity model. We vary the R^2 measure across the x-axis and plot the range of estimated ATT values on the y-axis. The solid bar denotes the point estimate bounds for a specified R^2 value, estimated as the point estimate plus and minus the optimal bias bounds (Theorem 4.3.1). The lighter intervals represent the 95% confidence intervals. We also plot the benchmarking results for the observed covariates, where the lines represent the corresponding benchmarked R^2 values.

Bounding a Confounder's Relationship with the Outcome

Previous literature has highlighted two characteristics of the imbalance term in Equation (4.1) that affect the bias from omitting a variable: (1) the overall magnitude of the the imbalance term, and (2) the relationship between the imbalance term to the outcomes (e.g., Huang (2022); Hong et al. (2021); Cinelli and Hazlett (2020); Shen et al. (2011)). Like the marginal sensitivity model, the variance-based sensitivity model constrain the overall magnitude of the imbalance term, and implicitly assume that the imbalance is maximally correlated with the outcome. In settings when researchers wish to account for this additional characteristic of the imbalance term, the variance-based sensitivity model can be easily extended to allow researchers to bound the relationship between the imbalance and the outcome. In particular, unlike the marginal sensitivity model, in which researchers must solve a linear programming problem to identify the extrema, there exists a closed-form solution

for the optimal bias bounds under the variance-based sensitivity model. As such, researchers can choose to evaluate the optimal bias bounds and associated confidence intervals using less conservative values of the correlation bound.

While amplification approaches allowing researchers to examine the relationship between the outcomes and the confounder for a fixed level of imbalance exist for alternative sensitivity models, many of these methods require introducing additional complexities. (For example, Rosenbaum and Silber (2009) requires invoking parametric assumptions on the outcomes.) In contrast, the variance-based sensitivity model allow researchers to easily incorporate additional information about the confounder to directly bound the relationship between the outcome and the imbalance in an omitted confounder. We provide recommendations for alternative bounds that researchers can use in Appendix C.1, as well as benchmarking procedure that allows researchers to use observed covariate data to estimate plausible correlation bounds.

4.4 Relationship to the Marginal Sensitivity Model

We now examine the relationship between the variance-based sensitivity model and the marginal sensitivity model. We first show that both sets of sensitivity models can be written as norm-constrained optimization problems. The variance-based sensitivity model implicitly constrains a weighted L_2 norm, while the marginal sensitivity model constrains an L_∞ norm. We demonstrate that by moving away from a worst-case characterization of the error from an omitted variable, the variance-based sensitivity model can obtain narrower, more informative bounds. We illustrate the potential for narrower bounds using benchmarked results from the running example.

Sensitivity Models as an Optimization Problem

Re-formulating the variance-based sensitivity model as an optimization problem under a bounded norm enables comparison with the marginal sensitivity model, which can be written as bias maximization problems under a constrained L_∞ norm. We argue that constraining the weighted L_2 norm can produce less conservative estimated bounds.

To begin, we show that the variance-based sensitivity model is a bias maximization problem, given a fixed constraint on a weighted L_2 norm.

Theorem 4.4.1 (Weighted L_2 Norm Constraint)
Define the individual-level error in the weights as $\lambda_i := w_i^*/w_i$. Define the $L_{2,w}$ norm as follows:

$$||\lambda||_{2,w}^2 := \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 \cdot \nu(w_i) & \text{if } var(w_i) > 0, \\ \infty & \text{else} \end{cases},$$

where $\nu(w_i)$ is a function of the estimated weights. Then, the variance-based sensitivity model can equivalently be written as a norm-constrained optimization problem:

$$\max_{\tilde{w} \in \sigma(R^2)} Bias(\hat{\tau}_W \mid \tilde{w}) \iff \begin{cases} \max\limits_{\tilde{w}} Bias(\hat{\tau}_W \mid \tilde{w}) \\ s.t.\ ||\lambda||_{2,w} \leq \sqrt{\dfrac{k}{1-R^2}}, \end{cases}$$

where $k := 1 - R^2/\mathbb{E}(w_i^2)$. See Appendix C.2 for proof and details.

Theorem 4.4.1 is especially in concert with the following result from Zhao et al. (2019) showing that the marginal sensitivity model is constrained L_∞ problems:

$$\max_{\tilde{w} \in \varepsilon(\Lambda)} \text{Bias}(\hat{\tau}_W \mid \tilde{w}) \iff \begin{cases} \max\limits_{\tilde{w}} \quad \text{Bias}(\hat{\tau}_W \mid \tilde{w}) \\ \text{s.t. } \Lambda^{-1} \leq ||\lambda||_\infty \leq \Lambda. \end{cases}$$

These constrained-norm representations provide insight into the benefits expected from the variance-based sensitivity model. Because the marginal sensitivity model optimizes over the set of weights defined by a worst-case error, the estimated bounds on the bias always correspond to cases in which *all* units are exposed to this worst-case error. However, in settings when one or two subjects are subject to much larger levels of confounding than others, this can result in an overly pessimistic view of the potential bias (Fogarty and Hasegawa, 2019; Zhao et al., 2019). In contrast, the variance-based sensitivity model is optimizing over a set of weights defined by average weighted error, and thus allow a small number of weights to be exposed to large amounts of error, even at moderate levels of overall confounding.

Comparison of Estimated Bounds

While constrained-norm representations provide intuition for why the variance-based sensitivity model may obtain narrower bounds than the marginal sensitivity model, in practice it is difficult to directly compare the bounds estimated under the two families of models. This is because the two approaches are using two different parameters and are fundamentally characterizing the error from omitting a confounder in a different manner. In the following subsection, we consider a setting in which researchers can estimate bounds using the true sensitivity parameter and compare the size of the associated confidence intervals. While in practice, researchers do not have access to the true sensitivity parameters, this approach provides intuition for the relative performances of the two sensitivity models.

First, we formalize a condition under which the variance-based sensitivity model will result in narrower bounds than the marginal sensitivity model. In general, we expect the variance-based sensitivity model to result in narrower bounds if the worst-case error (Λ) is much larger than the true average weighted error (proxied by R^2). If the difference in the worst-case error and average weighted error is not very large, then there will not be much improvement in the estimated bounds from using the variance-based sensitivity

model. Theorem 4.4.2 provides a maximum threshold for the size of R^2 relative to the Λ value sufficient for strictly narrower bounds under the variance-based sensitivity model.

Theorem 4.4.2 (Narrower Bounds under the Variance-Based Sensitivity Model) *Let $\psi(\Lambda)$ represent the difference in the estimated point estimate bounds under the marginal sensitivity model $\varepsilon(\Lambda)$ for a given $\Lambda \geq 1$:*

$$\psi(\Lambda) := \max_{\tilde{w} \in \varepsilon(\Lambda)} \frac{\sum_{i:Z_i=0} Y_i Z_i \tilde{w}_i}{\sum_{i:Z_i=0} Z_i \tilde{w}_i} - \min_{\tilde{w} \in \varepsilon(\Lambda)} \frac{\sum_{i:Z_i=0} Y_i Z_i \tilde{w}_i}{\sum_{i:Z_i=0} Z_i \tilde{w}_i}.$$

Then if the true R^2 parameter is lower than the following threshold,

$$R^2 \leq \frac{\psi(\Lambda)^2}{4 \underbrace{(1 - cor(w_i, Y_i \mid Z_i = 0)^2)}_{Correlation\ Bound} \cdot \underbrace{var(w_i \mid Z_i = 0) var(Y_i \mid Z_i = 0)}_{Scaling\ Factor} + \psi(\Lambda)^2}, \tag{4.7}$$

the bounds under the variance-based sensitivity model will be narrower than the bounds for the marginal sensitivity model.

Besides the worst-case error Λ, the R^2 threshold is determined by the correlation between the estimated weights and the outcome, $\text{cor}(w_i, Y_i \mid Z_i = 0)$, and the scaling factor, $\text{var}(w_i \mid Z_i = 0) \cdot \text{var}(Y_i \mid Z_i = 0)$. These components affect the estimated bounds under both sets of sensitivity models. Both $\text{cor}(w_i, Y_i \mid Z_i = 0)$ and the scaling factor are direct inputs into the optimal bias bounds under the variance-based sensitivity model. In addition, increases in these quantities either lead to larger outcome values, or more extreme weights; because the optimal bounds under the marginal sensitivity model are estimated by scaling the weights and outcomes by Λ (or Λ^{-1}), $\psi(\Lambda)$ will also increase.

Theorem 4.4.2 does not guarantee that the variance-based sensitivity model will always result in narrower bounds than the marginal sensitivity model, but we consider two specific scenarios that highlight the practical advantages from using variance-based sensitivity model. First, we consider an asymptotic setting. We show that in many cases, the worst-case error Λ will diverge to infinity, regardless of the omitted variable's confounding strength. In contrast, the R^2 parameter is a direct function of the confounding strength and retains a more stable interpretation across different data scales. Second, we consider a finite-sample setting, in which the outcomes and probability of treatment are highly correlated (which we refer to as *limited outcome overlap*). In this setting, the marginal sensitivity model may produce narrower intervals, but these intervals can be misleadingly narrow and fail to provide nominal coverage. In contrast, while the intervals under the variance-based sensitivity model will be wider in this setting, the intervals adequately account for the limited outcome overlap and continue to provide nominal coverage.

Remark. Several recent extensions of the marginal sensitivity model allow researchers to mitigate some the conservative nature of the method by adding in additional constraints beyond bounding the worst-case error (Kallus and Zhou, 2018; Dorn and Guo, 2021; Dorn et al.,

2021). However, these methods usually require adding additional sensitivity parameters or performing some form of outcome modeling. The variance-based sensitivity model offers stable and informative bounds via a one-parameter sensitivity analysis without additional assumptions, constraints, or complexities.

Infinite Worst-Case Error in Asymptotic Settings

We begin by considering the asymptotic setting. We show that when the omitted confounder results in an error that can be arbitrarily small or large and the outcomes Y_i are unbounded, the asymptotic confidence intervals for the variance-based sensitivity model will necessarily be narrower than the confidence intervals estimated under the marginal sensitivity model. Consider the following instructive example.

Example 4.4.1 (Behavior of Λ for a Logit Model)
Assume researchers use a logit model to estimate the weights using $\mathbf{X}_i$, but omit a confounder U_i. The estimated and ideal weights take on the following forms:

$$\hat{w}_i = \exp(\hat{\gamma}^\top \mathbf{X}_i) \qquad \hat{w}_i^* = \exp(\hat{\gamma}^{*\top} \mathbf{X}_i + \hat{\beta} U_i)$$

Then let $\hat{\Lambda}$ be the maximum error across our observed sample:

$$\hat{\Lambda} := \max_{1 \leq i \leq n} \{\hat{w}_i^*/\hat{w}_i, \hat{w}_i/\hat{w}_i^*\}.$$

Assume $\left[\mathbf{X}_i, U_i\right] \overset{iid}{\sim} MVN(0, I)$. Then $\mathbb{E}(\hat{\Lambda}) \to \infty$ as $n \to \infty$:

$$\lim_{n \to \infty} \frac{\mathbb{E}(\hat{\Lambda})}{\exp(\sqrt{2\nu^2 \log(n)})} \geq 1,$$

where $\nu^2 = (\gamma^* - \gamma)^\top (\gamma^* - \gamma) + \beta^2$ (where γ^*, γ, and β represent the population counterparts to the estimated coefficients), and the results follow immediately from Wainwright (2019)'s tail bound on normally distributed random variables. See Appendix C.2 for more details.

Example 4.4.1 highlights that the worst-case error will increase towards infinity as the sample size grows, *regardless* of the confounding strength of the omitted variable (represented by $\hat{\beta}$). The omitted confounder could explain little of the treatment assignment process, but the worst-case bound on the error would still be infinitely large. This means that researchers would have to specify an infinitely large Λ value for the marginal sensitivity model to be valid. In other words, while we can find a $\hat{\Lambda}$ in an observed sample that provides an upper bound on the difference between the realized and the ideal weights, the marginal sensitivity model will not hold for *any* value of Λ in the population (Jin et al., 2022). Intuitively, this occurs because $\hat{\Lambda}$ is a function of the largest U_i value in the sample: as the sample size increases, the probability of observing an extreme U_i value increases.

A secondary issue arises from the decoupling of the magnitude of the sensitivity parameter and the underlying confounding strength of the omitted variable. In particular, reasoning about whether or not Λ values is large or small is can be challenging. Even with the aid of benchmarking, researchers can at best estimate the worst-case error that arises from omitting different covariates, but reasoning about whether or not it is plausible for such an error to arise from an omitted variable amounts to reasoning about whether or not it is plausible for potential outliers to occur. Furthermore, as the sample size increases, researchers must take into account the potential for more outliers to occur.

In contrast, we can calculate the R^2 value for the variance-based sensitivity model under the same setting as Example 4.4.1. It is a function of $\hat{\gamma}$, $\hat{\gamma}^*$, and $\hat{\beta}$ that does not depend on the sample size.

Example 4.4.2 (Behavior of R^2 for a Logit Model)
Consider the same setting as Example 4.4.1. Then, the R^2 value can be written as follows:

$$R^2 = 1 - \frac{\exp(\hat{\gamma}^\top \hat{\gamma}) - 1}{\exp(\hat{\gamma}^{*\top} \hat{\gamma}^* + \hat{\beta}^2) - 1} \cdot \frac{\exp(\hat{\gamma}^\top \hat{\gamma})}{\exp(\hat{\gamma}^{*\top} \hat{\gamma}^* + \hat{\beta}^2)}.$$

Example 4.4.1 gives one setting in which Λ will be infinitely large, regardless of the confounding strength of the omitted variable. More generally, the following corollary to Theorem 4.4.2 shows that under any setting when the error from omitting a confounder can take on values that are arbitrarily small or large, the variance-based sensitivity model necessarily produces narrower bounds in sufficiently large samples.

Corollary 4.4.1 (Narrower Bounds under the Variance-Based Sensitivity Model)
Consider the set of confounders, in which for all $\delta > 0$, $P(w_i^*/w_i < \delta) > 0$, or $P(w_i^*/w_i > \delta) > 0$. Then, if the outcomes are unbounded, $\psi(\Lambda)$ will diverge in probability to infinity, and the threshold from Theorem 4.4.2 will converge in probability to 1:

$$\frac{\psi(\Lambda)^2}{4(1 - cor(w_i, Y_i)^2) \cdot var(w_i) var(Y_i) + \psi(\Lambda)^2} \xrightarrow{p} 1$$

Because $R^2 < 1$ by definition, for sufficiently large n, the variance-based sensitivity model will produce narrower bounds.

Corollary 4.4.1 highlights that in certain asymptotic settings, if the outcomes are unbounded, the marginal sensitivity models will result in infinitely large bias bounds. This will occur, regardless of whether the omitted confounder is strong or weak. In contrast, because the variance-based sensitivity model is not using a worst-case characterization of error, the resulting bias bounds will be less susceptible to extreme values and be narrower by construction.

Limited Overlap in Finite-Samples

We now consider a finite-sample setting. We show that paradoxically, in certain finite-samples, the marginal sensitivity model can result in narrower bounds than the variance-based sensitivity model even when Λ is very large. However, the narrower bounds come with a risk of substantial undercoverage, especially when the sample size is small or there is limited outcome overlap. In these settings, the variance-based sensitivity model will tend to return wider intervals, but maintain nominal coverage.

The key to this phenomenon is a property of the marginal sensitivity model, referred to as *sample boundedness*. Sample boundedness implies that even at infinitely large Λ values, the worst-case bounds under the marginal sensitivity model approach but cannot exceed the range of the observed control outcomes (i.e., $\lim_{\Lambda\to\infty}\psi(\Lambda) \leq \max_{i:Z_i=0} Y_i - \min_{i:Z_i=0} Y_i$).

In contrast, the variance-based sensitivity model is not inherently sample bounded. In settings with relatively large amounts of confounding, the marginal sensitivity model will have narrower intervals than the variance-based sensitivity model, since as R^2 increases towards 1, the estimated bounds under the variance-based sensitivity model will be adequately wide. However, sample boundedness may prohibit the construction of valid confidence intervals unless in the absence of a key implicit assumption on the distribution of the unobserved potential outcomes. Consider the following toy example:

Example 4.4.3 (Misleading Optimism from Sample Boundedness)
Consider the following population of 4 units, with the following potential outcomes, treatment assignment, and the estimated probability of treatments for each unit:

i	$Y_i(0)$	$Y_i(1)$	$\hat{P}(Z_i = 1)$	Z_i
1	-10	-10	0.1	0
2	5	5	0.2	0
3	10	10	0.9	1
4	20	20	0.95	1

The true ATT is zero, but the estimated ATT is equal to 14.6, so substantial confounding is present. However, since the sample bounds for the ATT are the interval $[10, 25]$*, no value of* Λ *can produce an estimated interval (under the marginal sensitivity model) containing zero, erroneously suggesting the presence of a true effect highly robust to substantial confounding.*

While this example is somewhat contrived, it highlights the problems with sample boundedness if the potential outcome ranges in the two groups have limited overlap, which may occur when potential outcomes are strongly correlated with the probability of treatment. For a formal characterization of this outcome overlap condition, see Appendix C.1. When there exists limited outcome overlap, estimated intervals from the marginal sensitivity model may be misleadingly optimistic, especially for dramatic levels of potential confounding, but intervals constructed under the variance-based sensitivity model, which are not sample bounded,

are not affected. Figure 4.3 illustrates the behavior and coverage rates of both sets of sensitivity models under varying amounts of outcome overlap and sample sizes in an empirical example, described in greater detail in Appendix C.1.

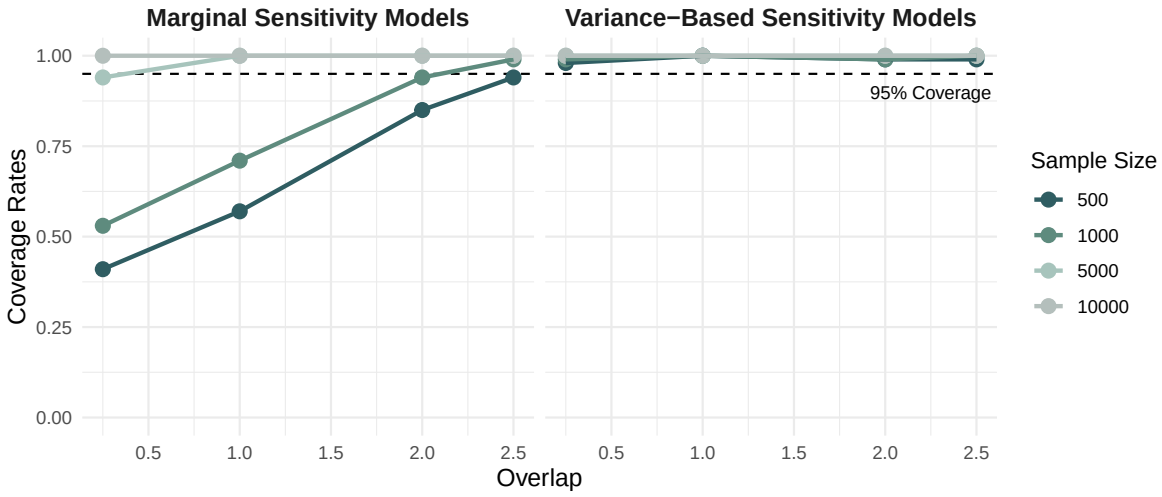

Figure 4.3: Coverage rates for the marginal sensitivity model and the variance-based sensitivity model, assuming an oracle bias setting when researchers have full knowledge of the true underlying sensitivity parameter. The x-axis represents a parameter σ_v^2, which represents how much outcome overlap there is between the treatment and control groups (i.e., as σ_v^2 increases, the amount of outcome overlap increases). See Appendix C.1 for more details on the data generating process.

Remark. We note that sample boundedness is not necessarily a negative feature in the context of *estimation*. The bias-variance tradeoff of using a stabilized weighted estimator has been extensively studied (e.g., Robins et al. (2007)). However, in the context of a sensitivity analysis, in which we are explicitly interested in examining the potential bias that can arise under varying levels of confounding, requiring sample boundedness can lead to misleading conclusions, and potential issues with outcome overlap should be considered carefully when interpreting results.

Illustration on NHANES

We now conduct formal benchmarking for the variance-based models and the marginal sensitivity model in our running example. We then estimate the corresponding bounds and intervals under both approaches. See Figure 4.4 for a visualization. We see that for each of the covariates, omitting a confounder like any of the observed covariates would result in wider bounds under the marginal sensitivity model than the variance-based models.

Notably, under the marginal sensitivity model, omitting a confounder like *educational attainment* would be sufficient to explain the entire observed effect under the null hypothesis of no effect. In contrast, under the variance-based sensitivity model, omitting a confounder with equivalent confounding strength to any of the observed covariates would not be sufficient to explain the observed data under the null. While the average error from omitting a variable like *educational attainment* is relatively low, the maximum error that occurs is relatively large. The marginal sensitivity model, which assumes such a maximal error could occur in the unobserved confounder for most or all data points, thus produces much wider intervals than the variance-based model, which is much less responsive to individual outliers.

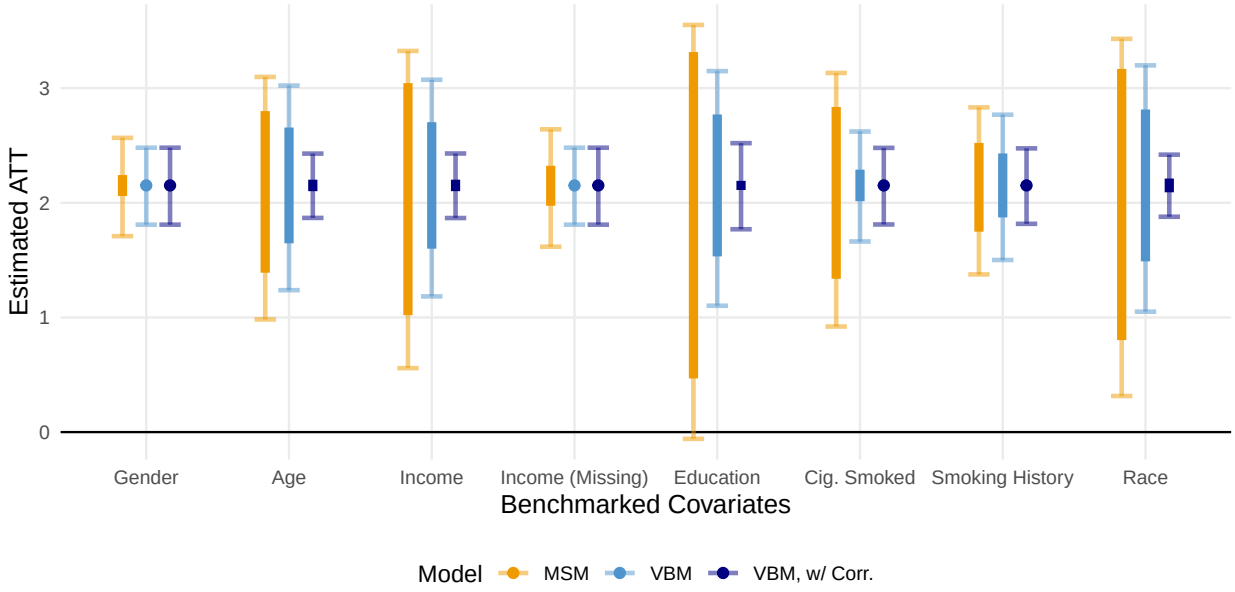

Figure 4.4: Estimated intervals for both the marginal sensitivity model (in yellow) and the variance-based sensitivity model (in light blue), and the variance-based sensitivity model, using a less conservative correlation bound (in dark blue). The darker intervals represent the point estimate bounds, while the lighter intervals represent the 95% confidence intervals. The intervals are estimated using the benchmarked Λ and R^2 values for each covariate, and are interpreted as the resulting intervals for an omitted confounder with equivalent confounding strength to the observed covariate.

We also estimate intervals (and bounds) under the variance-based sensitivity model using a relaxed correlation bound. In particular, we choose the correlation bound by benchmarking an optional correlation parameter, giving the correlation between the outcome and the imbalance in an omitted confounder, to the observed correlation between the outcome and each observed covariate. (See Appendix C.1 for more details.) By accounting for the relationship between the confounder and the outcome, we are able to obtain much narrower

intervals. In particular, we see that even in cases where a potential omitted confounder is highly imbalanced (e.g., omitting a confounder like *age* results in an R^2 value of 0.12, and Λ value of 2.1), the overall bias that occurs from omitting it may be relatively low if the imbalance is largely unrelated to the outcome. By considering this additional dimension of the bias–which can be easily done using the variance-based sensitivity model– researchers are able to better characterize the types of confounders that may lead to large amounts of bias and obtain a more holistic understanding of the sensitivity in their estimated effects.

4.5 Conclusion

We have introduced a novel sensitivity model, the variance-based sensitivity model, which characterizes the error from omitting a confounder by using the distributional differences between the estimated weights and true weights. We show that the variance-based sensitivity model can be parameterized using an R^2 measure that represents the degree of residual imbalance in an omitted confounder, and provide methods for benchmarking the R^2 value of an omitted confounder against residual imbalances for observed covariates. We also derive a closed-form solution for the maximum possible bias and introduce a method for estimation of asymptotically valid confidence intervals under the sensitivity model.

We also formalize the connection between the proposed sensitivity model and existing sensitivity analyses. We highlight that the variance-based sensitivity model has several notable advantages over the existing approaches that rely on worst-case bounds on the confounding strength. First, by characterizing bias in terms of distributional differences instead of a worst-case error bound, the variance-based sensitivity model can estimate narrower, less conservative bounds. Second, we show empirically that the variance-based sensitivity model obtains nominal coverage even in finite sample settings where the standard marginal sensitivity model dramatically undercovers due to issues with outcome overlap. Finally, because the variance based sensitivity model admits a closed-form solution for the optimal bias, we can introduce a natural two-parameter extension that uses constraints on the relationship between the omitted confounder and the outcome to produce narrower bounds.

We suggest several directions for future work. First, we demonstrated that variance-based sensitivity model, like the marginal sensitivity model, can be written as a norm-constrained optimization problem. Exploring other possible norms under which to constrain unobserved confounding could lead to a broad unified sensitivity framework, helping contextualize a wider variety of different sensitivity methods with their own strengths and weaknesses. Second, it is natural to ask what factors under a researcher's control at the design stage may influence the degree of robustness to unmeasured bias exhibited under the variance-based sensitivity analysis. While the closed form for the optimal bias bound already provides insights in this direction, developing a metric akin to design sensitivity for matched studies (Rosenbaum, 2004, 2010b) would provide valuable further guidance about how to design weighting estimators for maximum robustness. Finally, while we focused on a choice between bounding a weighted average error and bounding a worst-case error, future work could in-

corporate both constraints in the same study. We anticipate that this would result in further narrowing of sensitivity bounds.

Bibliography

Aronow, P. M. and D. K. Lee (2013). Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika 100*(1), 235–240.

Athey, S., R. Chetty, and G. Imbens (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research.

Athey, S., J. Tibshirani, S. Wager, et al. (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Baldassarri, D. and M. Abascal (2017). Field experiments across the social sciences. *Annual Review of Sociology 43*, 41–73.

Banerjee, A. V. and E. Duflo (2009). The experimental approach to development economics. *Annu. Rev. Econ. 1*(1), 151–178.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Bareinboim, E. and J. Pearl (2016). Causal Inference and the Data-Fusion Problem. *Proceedings of the National Academy of Sciences 113*(27), 7345–7352.

Ben-Michael, E., A. Feller, D. A. Hirshberg, and J. R. Zubizarreta (2021). The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*.

Bloom, H. S. et al. (1993). The national jtpa study. title ii-a impacts on earnings and employment at 18 months.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Buchanan, A. L., M. G. Hudgens, S. R. Cole, K. R. Mollan, P. E. Sax, E. S. Daar, A. A. Adimora, J. J. Eron, and M. J. Mugavero (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 181*(4), 1193–1209.

Carnegie, N. B., M. Harada, and J. L. Hill (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness 9*(3), 395–420.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research.

Cinelli, C. and C. Hazlett (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(1), 39–67.

Cole, S. R. and E. A. Stuart (2010). Generalizing evidence from randomized clinical trials to target populations: The actg 320 trial. *American journal of epidemiology 172*(1), 107–115.

Colnet, B., I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang (2020). Causal Inference Methods for Combining Randomized Trials and Observational Studies: A Review. *arXiv preprint arXiv:2011.08047*.

Cornfield, J., W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute 22*(1), 173–203.

Dahabreh, I. J., S. E. Robertson, E. J. Tchetgen, E. A. Stuart, and M. A. Hernán (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics 75*(2), 685–694.

Dahabreh, I. J., J. M. Robins, S. J. Haneuse, I. Saeed, S. E. Robertson, E. A. Stuart, and M. A. Hernán (2019). Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. *arXiv preprint arXiv:1905.10684*.

Deaton, A. and N. Cartwright (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine 210*, 2–21.

Deville, J.-C. and C.-E. Särndal (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association 87*(418), 376–382.

Ding, P. (2021). Two seemingly paradoxical results in linear models: the variance inflation factor and the analysis of covariance. *Journal of Causal Inference 9*(1), 1–8.

Ding, P., A. Feller, and L. Miratrix (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association 114*(525), 304–317.

Ding, P. and T. J. VanderWeele (2016). Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.) 27*(3), 368.

Djebbari, H. and J. Smith (2008). Heterogeneous impacts in progresa. *Journal of Econometrics 145*(1-2), 64–80.

Dorn, J. and K. Guo (2021). Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *arXiv preprint arXiv:2102.04543*.

Dorn, J., K. Guo, and N. Kallus (2021). Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*.

Egami, N. and E. Hartman (2019). Covariate selection for generalizing experimental results. *arXiv preprint arXiv:1909.02669*.

Egami, N. and E. Hartman (2022). Elements of external validity: Framework, design, and analysis. *APSR*.

Falk, A. and J. J. Heckman (2009). Lab experiments are a major source of knowledge in the social sciences. *science 326*(5952), 535–538.

Fogarty, C. B. and R. B. Hasegawa (2019). Extended sensitivity analysis for heterogeneous unmeasured confounding with an application to sibling studies of returns to education. *The Annals of Applied Statistics 13*(2), 767–796.

Ford, I. and J. Norrie (2016). Pragmatic trials. *New England Journal of Medicine 375*(5), 454–463. PMID: 27518663.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon, 3ˆ e serie, Sciences, Sect. A 14*, 53–77.

Freedman, D. A. (2008, 03). On regression adjustments in experiments with several treatments. *Ann. Appl. Stat. 2*(1), 176–196.

Gerber, A. S. and D. P. Green (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *The American Political Science Review 94*(3), 653–663.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis 20*(1), 25–46.

Hartman, E., R. Grieve, R. Ramsahai, and J. S. Sekhon (2015). From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 178*(3), 757–778.

Hartman, E., C. Hazlett, and C. Sterbenz (2021). Kpop: A kernel balancing approach for reducing specification assumptions in survey weighting. *arXiv preprint arXiv:2107.08075*.

Hartman, E. and M. Huang (2022). Sensitivity analysis for survey weights. *arXiv preprint arXiv:2206.07119*.

Hoeffding, W. (1941). Masstabinvariante korrelationsmasse für diskontinuierliche verteilungen. *Archiv für mathematische Wirtschafts-und Sozialforschung 7*, 49–70.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association 81*(396), 945–960.

Hong, G., F. Yang, and X. Qin (2021). Did you conduct a sensitivity analysis? a new weighting-based approach for evaluations of the average treatment effect for the treated. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 184*(1), 227–254.

Hsu, J. Y. and D. S. Small (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics 69*(4), 803–811.

Huang, M. (2022). Sensitivity analysis in the generalization of experimental results. *arXiv preprint arXiv:2202.03408*.

Huang, M., N. Egami, E. Hartman, and L. Miratrix (2021). Leveraging population outcomes to improve the generalization of experimental results. *arXiv preprint arXiv:2111.01357*.

Huang, M. and S. Pimentel (2022). Variance-based sensitivity models for weighted estimators result in more informative bounds. *arXiv preprint arXiv:2208.01691*.

Huang, M. Y., B. G. Vegetabile, L. F. Burgette, C. Setodji, and B. A. Griffin (2022). Higher moments matter for optimal balance weighting in causal estimation. *Epidemiology (Cambridge, Mass.)*.

Huber, J. (2013). Is theory getting lost in the "identification revolution"?

Imai, K., G. King, and E. A. Stuart (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society) 171*(2), 481–502.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Jacob, D. (2020). Cross-fitting and averaging for machine learning estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2007.02852*.

Jin, Y., Z. Ren, and Z. Zhou (2022). Sensitivity analysis under the f-sensitivity models: Definition, estimation and inference. *arXiv preprint arXiv:2203.04373*.

Josey, K. P., S. A. Berkowitz, D. Ghosh, and S. Raghavan (2021). Transporting experimental results with entropy balancing. *Statistics in Medicine*.

Kallus, N. and X. Mao (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*.

Kallus, N. and A. Zhou (2018). Confounding-robust policy improvement. *Advances in neural information processing systems 31*.

Kang, J. D., J. L. Schafer, et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science 22*(4), 523–539.

Kern, H. L., E. A. Stuart, J. Hill, and D. P. Green (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness 9*(1), 103–127.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604–620.

Lin, W. (2013, 03). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Ann. Appl. Stat. 7*(1), 295–318.

Lu, B., E. Ben-Michael, A. Feller, and L. Miratrix (2021). Is it who you are or where you are? accounting for compositional differences in cross-site treatment variation. *arXiv preprint arXiv:2103.14765*.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine 23*(19), 2937–2960.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics 12*(2), 685–726.

Miratrix, L. W., J. S. Sekhon, A. G. Theodoridis, and L. F. Campos (2018). Worth weighting? how to think about and use weights in survey experiments. *Political Analysis 26*(3), 275–291.

Miratrix, L. W., J. S. Sekhon, and B. Yu (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(2), 369–396.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles (with discussion). Section 9 (translated). *Statistical Science 5*(4), 465–472.

Nguyen, T. Q., C. Ebnesajjad, S. R. Cole, E. A. Stuart, et al. (2017). Sensitivity analysis for an unobserved moderator in rct-to-target-population generalization of treatment effects. *The Annals of Applied Statistics 11*(1), 225–247.

Nie, X., G. Imbens, and S. Wager (2021). Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*.

Olkin, I. (1981). Range restrictions for product-moment correlation matrices. *Psychometrika 46*(4), 469–472.

Olsen, R. B., L. L. Orr, S. H. Bell, and E. A. Stuart (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management 32*(1), 107–121.

O'Muircheartaigh, C. and L. V. Hedges (2014). Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 63*(2), 195–210.

Pearl, J. and E. Bareinboim (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science 29*(4), 579–595.

Polley, E. C. and M. J. van der Laan (2010). Super Learner in Prediction.

Raudenbush, S. W. and H. S. Bloom (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation 36*(4), 475–499.

Resnick, S. I. (2008). *Extreme values, regular variation, and point processes*, Volume 4. Springer Science & Business Media.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science 22*(4), 544–559.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association 89*(427), 846–866.

Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika 91*(1), 153–164.

Rosenbaum, P. R. (2010a). *Design of observational studies*, Volume 10. Springer.

Rosenbaum, P. R. (2010b). Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association 105*(490), 692–702.

Rosenbaum, P. R. et al. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science 17*(3), 286–327.

Rosenbaum, P. R. and D. B. Rubin (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological) 45*(2), 212–218.

Rosenbaum, P. R. and J. H. Silber (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association 104*(488), 1398–1405.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology 66*(5), 688.

Rubin, D. B. (1980). Discussion of 'Randomization analysis of experimental data: The Fisher randomization test comment' by Basu. *Journal of the American Statistical Association 75*(371), 591–593.

Rubin, D. B. (2008). For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics 2*(3), 808–840.

Sales, A. C., B. B. Hansen, and B. Rowan (2018). Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. *Journal of Educational and Behavioral Statistics 43*(1), 3–31.

Särndal, C.-E., B. Swensson, and J. Wretman (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Shen, C., X. Li, L. Li, and M. C. Were (2011). Sensitivity analysis for causal inference using inverse probability weighting. *Biometrical Journal 53*(5), 822–837.

Soriano, D., E. Ben-Michael, P. J. Bickel, A. Feller, and S. D. Pimentel (2021). Interpretable sensitivity analysis for balancing weights. *arXiv preprint arXiv:2102.13218*.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics 25*(1), 1.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174*(2), 369–386.

Stuart, E. A. and A. Rhodes (2017). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Evaluation review 41*(4), 357–388.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association 101*(476), 1619–1637.

Tan, Z. (2007). Comment: Understanding or, ps and dr. *Statistical Science 22*(4), 560–568.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics 38*(3), 239–266.

Tipton, E. (2014). How generalizable is your experiment? an index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics 39*(6), 478–501.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology 6*(1).

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge University Press.

Wang, Y. and J. R. Zubizarreta (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika 107*(1), 93–105.

Westreich, D. and S. R. Cole (2010). Invited commentary: positivity in practice. *American journal of epidemiology 171*(6), 674–677.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica: Journal of the econometric society*, 1–25.

Word, E. R. et al. (1990). The state of tennessee's student/teacher achievement ratio (star) project: Technical report (1985-1990).

Zhang, Y. and Q. Zhao (2022). Bounds and semiparametric inference in l_∞ and l_2-sensitivity analysis for observational studies. *arXiv preprint arXiv:2211.04697*.

Zhao, Q. and D. Percival (2016). Entropy balancing is doubly robust. *Journal of Causal Inference 5*(1).

Zhao, Q., D. S. Small, and B. B. Bhattacharya (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*(4), 735–761.

Zhao, Q., D. S. Small, and P. R. Rosenbaum (2018). Cross-screening in observational studies that test many hypotheses. *Journal of the American Statistical Association 113*(523), 1070–1084.

Zheng, J., A. D'Amour, and A. Franks (2021). Copula-based sensitivity analysis for multi-treatment causal inference with unobserved confounding. *arXiv preprint arXiv:2102.09412*.

Appendix A

Sensitivity Analysis for Generalizing Experimental Results

A.1 Extensions and Additional Discussion

Extension of Sensitivity Framework for Balancing Weights

The proposed sensitivity framework can be extended for balancing weights. Balancing weights directly optimize for covariate balance (i.e., Hainmueller (2012); Ben-Michael et al. (2021); Wang and Zubizarreta (2020), to name a few). There is a connection between balancing weights and propensity scores; for example, Wang and Zubizarreta (2020) show that balancing weights are a more general formulation of regularized propensity scores.

We argue that for the class of balancing weights that meet the following conditions, the sensitivity framework can be directly applied:
Condition 1. $\mathbb{E}_{\mathcal{S}}(w_i)/\mathbb{E}_{\mathcal{S}}(w_i^*) = 1$
Condition 2. $\mathbb{E}_{\mathcal{S}}(w_i^* \mid \mathbf{X}_i) = w_i$

When Condition 1 is met, this implies that the bias decomposition introduced in Theorem 2.3.1 will hold. When Condition 2 is met, this implies that the bounds derived for R_ε^2 and $\rho_{\varepsilon,\tau}$ will apply. Condition 1 states that the estimated weights and the ideal weights must be centered at the same value. This is not a very stringent condition, as most weights (by definition) will be centered at mean 1. Condition 2 states that by conditioning on the observed covariates $\mathbf{X}_i$, the ideal weights must be centered at the estimated weights w_i. The sensitivity analysis can still be applied to balancing weights that meet Condition 1, but not 2; however, the estimated bounds on the parameters may not necessarily hold.

Extended Details on Bounding σ_τ^2

To begin, decompose σ_τ^2 as:

$$\sigma_\tau^2 = \text{var}_{\mathcal{S}}(Y_i(1)) + \text{var}_{\mathcal{S}}(Y_i(0)) - 2\text{cov}_{\mathcal{S}}(Y_i(1), Y_i(0))$$

The decomposition illustrates that the magnitude of treatment effect heterogeneity will be driven by two factors: (1) the total variation in the outcomes (i.e., $\text{var}_{\mathcal{S}}(Y_i(1)) + \text{var}_{\mathcal{S}}(Y_i(0))$), and (2) how correlated the potential outcomes are. Because we cannot estimate the covariance between the potential outcomes, σ^2_τ can never be identified. However, Ding et al. (2019) showed that sharp bounds for σ^2_τ can be obtained by applying Fréchet-Hoeffding bounds Hoeffding (1941); Fréchet (1951):

$$\int_0^1 \Big\{F^{-1}_{Y_1}(u) - F^{-1}_{Y_0}(u)\Big\}du \leq \sigma^2_\tau \leq \int_0^1 \Big\{F^{-1}_{Y_1}(u) - F^{-1}_{Y_0}(1-u)\Big\}du, \tag{A.1}$$

where F_{Y_1} and F_{Y_0} represent the empirical cumulative distribution functions of the treatment and control potential outcomes, respectively. Intuitively, the lower bound of σ^2_τ is reached when the potential outcomes are perfectly correlated (i.e., $\text{cor}_{\mathcal{S}}(Y_i(1), Y_i(0)) = 1$). The upper bound of σ^2_τ is reached when the potential outcomes are perfectly anti-correlated (i.e., $\text{cor}_{\mathcal{S}}(Y_i(1), Y_i(0)) = -1$). As such, researchers may use the upper bound detailed in Equation A.1 as a conservative estimate for σ^2_τ. The bound in Equation (A.1) will *always* hold. However, it can span a large range of values. If researchers are willing to impose additional assumptions, a tighter bound on σ^2_τ can be obtained.

We will discuss two examples of assumptions that researchers may wish to impose. Figure A.1 provides a summary.

Directional sign of the correlation between τ_i and $Y_i(0)$: Ding et al. (2019) and Raudenbush and Bloom (2015) show that information about the correlation between the individual-level treatment effect and the control potential outcomes could be inferred from m (i.e., the ratio of variance between the treatment and control outcomes). More specifically, when $m < 1$ (i.e., the variance of the control outcomes is greater than the variance of the treatment outcomes), then the correlation between the individual treatment effect and the control potential outcome is negative, and the lower bound on σ^2_τ can be tightened:[1]

$$\text{var}_{\mathcal{S}}(Y_i(1)) \leq \sigma^2_\tau \tag{A.2}$$

Unfortunately, the converse cannot be shown to be true (i.e., $m > 1$ does not necessarily imply a positive relationship). However, researchers may have substantive knowledge to justify a positive relationship. For example, in the original JTPA study, researchers compared the estimated impact of jobs training programs across women by previous earnings and

[1]This follows simply from the fact that we may rewrite the decomposed treatment effect heterogeneity as:

$$\begin{aligned}\sigma^2_\tau &= \text{var}_{\mathcal{S}}(Y_i(1)) + \text{var}_{\mathcal{S}}(Y_i(0)) - 2\text{cov}(Y_i(1), Y_i(0)) \\ &= \text{var}_{\mathcal{S}}(Y_i(1)) - \text{var}_{\mathcal{S}}(Y_i(0)) - 2\text{cov}_{\mathcal{S}}(\tau_i, Y_i(0))\end{aligned}$$

Because $\sigma^2_\tau \geq 0$, then $2\text{cov}_{\mathcal{S}}(\tau_i, Y_i(0)) \leq \underbrace{\text{var}_{\mathcal{S}}(Y_i(1)) - \text{var}_{\mathcal{S}}(Y_i(0))}_{(*)}$. When $m < 1$, this implies that the term in $(*)$ is going to be negative, which in turn, implies $\text{cov}_{\mathcal{S}}(\tau_i, Y_i(0)) \leq 0$.

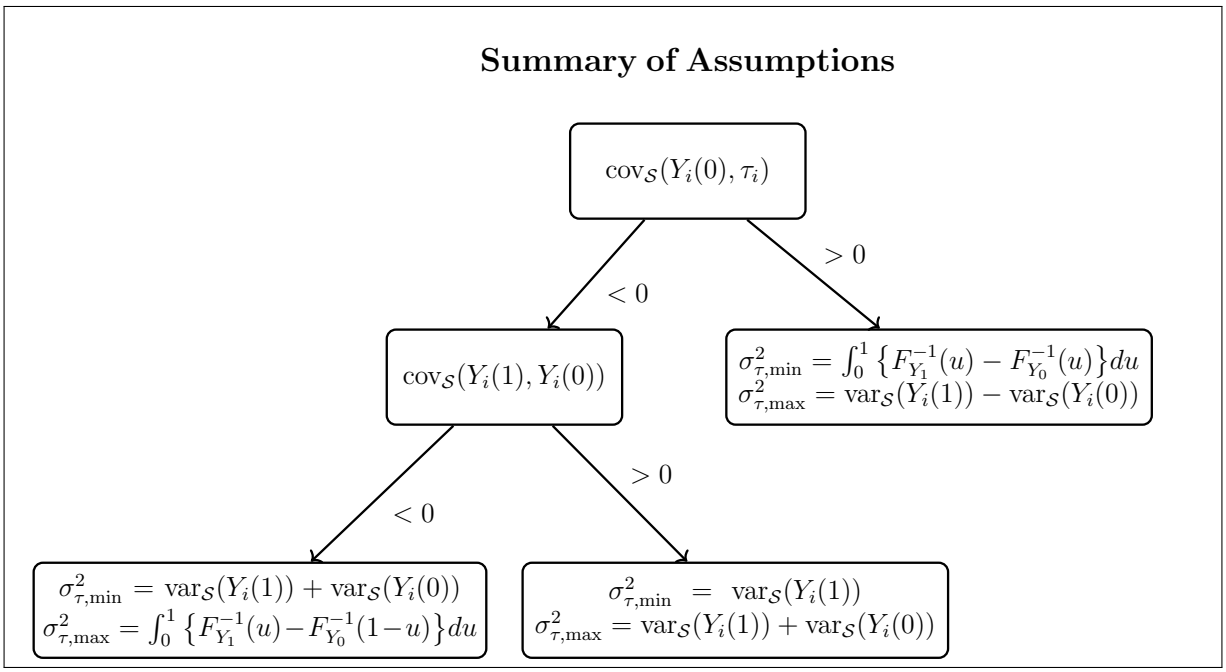

Figure A.1: Summary of assumptions that researchers may invoke to help tighten the bound on σ^2_τ. The above diagram provides the tightened minimum and maximum values for σ^2_τ, depending on the assumption researchers wish to invoke. Researchers can estimate m to check if $m < 1$. If $m < 1$, then it is guaranteed that $\text{cov}_{\mathcal{S}}(Y_i(0), \tau_i) < 0$. Substantive knowledge can be used to help justify the different assumptions.

employment history (Bloom et al. (1993)). They found that women who had a higher hourly wage in their work history had a higher estimated impact from accessibility to jobs training programs. Similarly, women who came from families with greater household income also saw a greater impact from jobs training programs. As such, we assume there exists a non-negative association between the individual-level treatment effect and the outcomes under control (i.e., $\text{cov}(Y_i(0), \tau_i) > 0$)

In cases where researchers are willing to assume that $\text{cov}_{\mathcal{S}}(\tau_i, Y_i(0)) \geq 0$, the upper bound of σ^2_τ becomes:

$$\sigma^2_\tau \leq \text{var}_{\mathcal{S}}(Y_i(1)) - \text{var}_{\mathcal{S}}(Y_i(0)) \tag{A.3}$$

Directional sign of the correlation between potential outcomes: Alternatively, researchers may assume information about the relationship between $Y_i(1)$ and $Y_i(0)$. In particular, if researchers believe that the correlation between $Y_i(1)$ and $Y_i(0)$ is non-negative, then a tighter upper bound may be obtained on σ^2_τ:

$$\sigma^2_\tau \leq \text{var}_{\mathcal{S}}(Y_i(1)) + \text{var}_{\mathcal{S}}(Y_i(0)) \tag{A.4}$$

Alternatively, if researchers assume the correlation between $Y_i(1)$ and $Y_i(0)$ is negative, then a tighter lower bound is obtained:

$$\text{var}_{\mathcal{S}}(Y_i(1)) + \text{var}_{\mathcal{S}}(Y_i(0)) \leq \sigma_\tau^2 \tag{A.5}$$

We note that in order for the correlation between $Y_i(1)$ and $Y_i(0)$ to be negative, $\text{cov}(Y_i(0), \tau_i)$ must be negative.[2] (The converse is not true–i.e., $\text{cov}(\tau_i, Y_i(0))$ may be negative, without $\text{cov}(Y_i(1), Y_i(0)) < 0$.)

These two assumptions can be combined in conjunction to help tighten the bound on σ_τ^2. We recommend researchers first estimate m to determine whether or not $\text{cov}(Y_i(0), \tau_i)$ must be negative, which can help narrow down the plausible assumptions that can be used. A summary is provided in Figure A.1.

Another approach to tighten the bound on plausible σ_τ^2 values is to directly model the individual-level treatment effect (e.g., see Athey et al. (2019)) Many existing approaches leverage flexible, machine learning methods to estimate $\hat{\tau}_i$ without relying heavily on parametric assumptions, such as linearity. Therefore, researchers can model $\hat{\tau}_i$, and then directly estimate $\text{var}_{\mathcal{S}}(\hat{\tau}_i)$ to understand what may be plausible values for σ_τ^2. Because we are only concerned about treatment effect heterogeneity across the experimental sample, this can be especially advantageous in settings where researchers have a richer set of covariates within the experimental sample that may not be measured across the population.[3]

We highlight two ways researchers can leverage parametrically modeling τ_i to help bound σ_τ^2. First, similar to Ding et al. (2019), researchers may use the estimated $\text{var}_{\mathcal{S}}(\hat{\tau}_i)$ and posit how many times larger the actual variation in individual-level treatment effect is. In general, we caution researchers from directly using $\text{var}_{\mathcal{S}}(\hat{\tau}_i)$ as the estimate for σ_τ^2. Even in the scenario that the true conditional expectation of the individual-level treatment effect is used to estimate τ_i, $\text{var}_{\mathcal{S}}(\hat{\tau}_i)$ will be an underestimation of the true variation in the individual-level treatment effect.[4] The second way researchers can benefit from parametrically modeling τ_i is

[2]This follows from the following:

$$\begin{aligned} \text{cov}(Y_i(0), Y_i(1)) < 0 &\implies \text{cov}(Y_i(0), Y_i(0)) + \text{cov}(Y_i(0), \tau_i) < 0 \\ &\implies \text{cov}(Y_i(0), \tau_i) < -\text{var}(Y_i(0)) < 0 \end{aligned}$$

[3]We note that in cases when researchers have access to a rich set of covariates across both the population and the experimental sample and strongly believe that they can accurately parametrically model τ_i, it may be advantageous to use a doubly robust estimator, instead of just the weighted estimator. (See Section A.1 for discussion.)

[4]This can be formalized in the following. Assume $g(\mathbf{X}_i) = \mathbb{E}(\tau_i \mid \mathbf{X}_i)$. Thus, we may decompose τ_i into the component that can be explained by $g(\mathbf{X}_i)$, and the component that cannot:

$$\tau_i = \mathbb{E}(\tau_i \mid \mathbf{X}_i) + u_i$$

Because the covariance between u_i and the estimated $\hat{\tau}_i$ values must be 0: $\sigma_\tau^2 = \text{var}_{\mathcal{S}}(\hat{\tau}_i) + \text{var}_{\mathcal{S}}(u_i)$, and $\sigma_\tau^2 \geq \text{var}_{\mathcal{S}}(\hat{\tau}_i)$.

by using $\hat{\tau}_i$ to help aid the substantive justification of one of the assumptions used to tighten the bounds on σ^2_τ. For example, if researchers wish to assume that $\text{cov}_\mathcal{S}(Y_i(0), \tau_i) > 0$, they can use the estimated $\hat{\tau}_i$ to check if $\text{cov}_\mathcal{S}(Y_i(0), \hat{\tau}_i)$ is positive.

Extreme Scenario Analysis

We propose an extreme scenario analysis for researchers to evaluate the bias when the error term ε_i is maximally correlated to the individual-level treatment effect. Under this scenario, the maximum values that $\rho_{\varepsilon,\tau}$ and R^2_ε (referred to as ρ_{max} and R^2_{max}, respectively) can take on will be a function of $1 - \text{cor}_\mathcal{S}(w_i, \tau_i)^2$:

$$\rho^2_{max} = R^2_{max} = 1 - \text{cor}_\mathcal{S}(w_i, \tau_i)^2$$

As such, evaluating the bias at $\left(\rho_{max}, R^2_{max}\right)$ will result in an upper bound on the bias.

In practice, this can be an extremely conservative estimate of the bias. Researchers can choose to evaluate less conservative estimates of $\left(\rho_{max}, R^2_{max}\right)$ by relaxing how much variation in the treatment effect they believe the true weights w^*_i can explain. More detail is provided in Appendix A.1. More specifically, Miratrix et al. (2018) demonstrated that the survey weights from a survey experiment were weakly correlated with the treatment effect heterogeneity. For example, in the JTPA application, calculating the extreme scenario bound using our conservative estimate of $\text{cor}_\mathcal{S}(w_i, \tau_i)$ results in a bound of 0.99 for the maximum value $\rho_{\varepsilon,\tau}$ and R^2_ε. The extreme scenario would arise if the error term explained 99% of the variation in the individual-level treatment effect and the ideal weights.

The extreme scenario bound allows researchers to evaluate the bias when the error term ε_i explains all residual variation in the individual-level treatment effect. When $\rho_{\varepsilon,\tau} = \rho_{max}$, the maximum value of R^2_ε is equal to $1 - \text{cor}_\mathcal{S}(w_i, \tau_i)^2$. However, we can use Lemma 2.3.2 to show that when $\rho_{\varepsilon,\tau}$ is equal to the upper bound of $1 - \text{cor}_\mathcal{S}(w_i, \tau_i)^2$, then R^2_ε can actually take on a range of values, defined by the following:

$$R^2_{max} = 1 - \left(\text{cor}_\mathcal{S}(w_i, \tau_i) \cdot \text{cor}_\mathcal{S}(w^*_i, \tau_i) \pm \sqrt{(1 - \text{cor}_\mathcal{S}(w^*_i, \tau_i)^2) \cdot (1 - \text{cor}_\mathcal{S}(w_i, \tau_i)^2)}\right)^2 \quad \text{(A.6)}$$

Equation (A.6) represents the degree of imbalance that must be present in order for $\rho_{\varepsilon,\tau}$ to equal to the upper bound of ρ_{max}. However, Equation (A.6) depends on $\text{cor}_\mathcal{S}(w^*_i, \tau_i)$ (i.e., the relationship between the true selection weights w^*_i and the individual-level treatment effect), which cannot be estimated. Evaluating Equation (A.6) for the extreme case that $|\text{cor}(w^*_i, \tau_i) = 1|$ removes the dependency on $\text{cor}(w^*_i, \tau_i)$ and results in the upper bound proposed earlier:

$$R^2_{max} \leq 1 - \text{cor}_\mathcal{S}(w_i, \tau_i)^2$$

Researchers may not wish to assume that $\text{cor}(w^*_i, \tau_i)$ is at -1 or 1. As such, evaluating Equation (A.6) at lower values of $\text{cor}(w^*_i, \tau_i)$ will result in a less conservative bound on R^2_ε. One approach researchers can take to posit plausible values for $\text{cor}(w^*_i, \tau_i)$ is by using the

estimated $\text{cor}(w_i, \tau_i)$ and specifying how much *additional* variation in τ_i they believe w_i^* is able to explain. For example, if $\widehat{\text{cor}}(w_i, \tau_i)$ is very low (i.e., ≈ 0.1), it may be unlikely that the true weights w_i^* would be $10\times$ more correlated with the individual-level treatment effect, such that $\text{cor}(w_i^*, \tau_i) \approx 1$. This allows researchers to obtain less conservative estimates of an extreme scenario bound, depending on what they deem is a "reasonable" choice for $\text{cor}_{\mathcal{S}}(w_i^*, \tau_i)$.

Accounting for Uncertainty in the Sensitivity Tools

To account for potential changes in inference from omitting a variable, we extend the percentile bootstrap framework proposed in Huang and Pimentel (2022) and Zhao et al. (2019). More specifically, we can define the adjusted weighted estimator as:

$$\hat{\tau}_W^*(R_\varepsilon^2, \rho_{\varepsilon,\tau}) := \hat{\tau}_W + \text{Bias}(\hat{\tau}_W \mid R_\varepsilon^2, \rho_{\varepsilon,\tau}, \sigma_\tau^2),$$

and estimate confidence intervals around the adjusted weighted estimator for a grid of $\{R_\varepsilon^2, \rho_{\varepsilon,\tau}, \sigma_\tau^2\}$ values. They can then identify the maximum bias that can occur before the intervals contain the null estimate, which can be then be used to define the killer confounder region. We provide the technical details of the procedure for a fixed set of values $(R_\varepsilon^2, \rho_{\varepsilon,\tau}, \sigma_\tau^2)$ below.

Furthermore, we note that researchers may calculate the sensitivity tools over repeated bootstrap iterations to account for estimation uncertainty associated with the robustness value and the formal benchmarking results.

Details on Augmented Weighted Estimators

Interpreting the Parameters

Correlation between ε_i and ξ_i (i.e., $\rho_{\varepsilon,\xi}$) The correlation term between ε_i and ξ_i represents the relationship between the error in the weight estimation, and the error in the treatment effect modeling. In other words, $\rho_{\varepsilon,\xi}$ is a measure for how related the residual imbalance in the omitted confounder $\mathbf{U}_i$ is to the residuals in the individual-level treatment effect model. In general, we expect $|\rho_{\varepsilon,\xi}|$ to be less than $|\rho_{\varepsilon,\tau}|$ because the residual imbalance in the omitted confounder is likely to be less correlated to the residuals ξ_i than the overall individual-level treatment effect τ_i.

We can extend Lemma 2.3.2 to bound $\rho_{\varepsilon,\xi}$ on the following range:

$$\left[-\sqrt{1 - \text{cor}_{\mathcal{S}}(w_i, \xi_i)}, \sqrt{1 - \text{cor}_{\mathcal{S}}(w_i, \xi_i)}\right].$$

If the estimated weights w_i are highly correlated with the residuals ξ_i, then the range of values that $\rho_{\varepsilon,\xi}$ may take on will be more restricted.

Valid Confidence Intervals

Step 1. Fix some $(R^2_\varepsilon, \rho_{\varepsilon,\tau}, \sigma^2_\tau)$. Generate B bootstrap samples of the data.

Step 2. For each bootstrap sample $b = 1, ..., B$:

1. Estimate weights $\hat{w}_i^{(b)}$ and the point estimate $\hat{\tau}_W^{(b)}$.
2. Using the fixed σ^2_τ value, calculate $\widehat{\text{cor}}_b(\hat{w}_i^{(b)}, \tau_i)$, and $\widehat{\text{var}}_b(\hat{w}_i^{(b)})$, where the subscript b denotes the quantity calculated over the b-th bootstrap sample.
3. Using the bootstrapped quantities, calculate the adjusted weighted estimator for the b-th bootstrap:

$$\hat{\tau}_W^{*(b)}(R^2_\varepsilon, \rho_{\varepsilon,\tau}, \sigma^2_\tau) := \hat{\tau}_W^{(b)} + \text{Bias}(\hat{\tau}_W^{(b)} \mid R^2_\varepsilon, \rho_{\varepsilon,\tau}, \sigma^2_\tau)$$

Step 3. From the B bootstrapped optimal bounds, estimate the $\alpha/2$ and $1-\alpha/2$-th percentiles of the minima and maxima values respectively to obtain valid confidence intervals:

$$CI(\alpha) = \left[Q_{\alpha/2}(\hat{\tau}_W^{*(b)}(R^2_\varepsilon, \rho_{\varepsilon,\tau}, \sigma^2_\tau), Q_{1-\alpha/2}(\hat{\tau}_W^{*(b)}(R^2_\varepsilon, \rho_{\varepsilon,\tau}, \sigma^2_\tau)\right],$$

where $\liminf_{n\to\infty} P(\tau^*_W \subseteq CI(\alpha)) \geq 1 - \alpha$.

Table A.1: Procedure for estimating valid confidence intervals.

Variation in ξ_i (i.e., σ^2_ξ) σ^2_ξ is the total variation leftover in the treatment effect heterogeneity that is not explained by the estimated treatment effect model. σ^2_ξ is often referred to in the literature as the *idiosyncratic treatment effect variation* (Ding et al. (2019); Djebbari and Smith (2008)). σ^2_ξ can be written as a function of σ^2_τ:

$$\sigma^2_\xi = \sigma^2_\tau - \text{var}_\mathcal{S}(\hat{\tau}_i) - 2\text{cov}_\mathcal{S}(\hat{\tau}_i, \xi_i),$$

where both $\text{var}(\hat{\tau}_i)$ and $\text{cov}(\hat{\tau}_i, \xi_i)$ can be estimated from observed data. Thus, researchers can use the same bounds derived in Section 2.3 to estimate an upper bound for σ^2_τ (denoted as $\sigma^2_{\tau,max}$, and bound σ^2_ξ in the following manner:

$$\sigma^2_\xi \leq \sigma^2_{\tau,max} - \text{var}_\mathcal{S}(\hat{\tau}_i) - 2\text{cov}_\mathcal{S}(\hat{\tau}_i, \xi_i) \tag{A.7}$$

Alternatively, researchers can choose to bound σ^2_ξ directly. For example, the bound from Equation (A.1) can be extended for the residuals across the potential outcomes (Ding et al., 2019). The derived bounds can be sharpened by invoking additional assumptions on the residuals between the potential outcomes.

Summary of Sensitivity Framework

We summarize the sensitivity analysis framework for augmented weighted estimators below.

Summary of Sensitivity Framework for Augmented Weighted Estimators

Step 1. Estimate a conservative upper bound for σ_ξ^2 (i.e., $\sigma_{\xi,\max}^2$).

Step 2. Using $\sigma_{\xi,\max}^2$, estimate $\widehat{\text{cor}_{\mathcal{S}}}^2(w_i, \xi_i)$ (as a conservative bound for $\text{cor}_{\mathcal{S}}^2(w_i, \xi_i)$).

Step 3. Vary $\rho_{\varepsilon,\xi}$ from $-\sqrt{1 - \widehat{\text{cor}_{\mathcal{S}}}^2(w_i, \xi_i)}$ to $\sqrt{1 - \widehat{\text{cor}_{\mathcal{S}}}^2(w_i, \xi_i)}$.

Step 4. Vary R_ε^2 from the range of $[0, 1)$.

Step 5. Evaluate the bias.

Relationship with Sensitivity Analysis from Nguyen et al. (2017)

Consider the case in which only the treatment effect heterogeneity is modeled, using $\hat{\tau}(\mathbf{X}_i)$. Denote this estimator as $\hat{\tau}_{model}$. The bias formula for failing to account for $\mathbf{U}_i$ in the individual-level treatment model is:

$$\text{Bias}(\hat{\tau}_{model}) = \rho_{w^*,\xi} \cdot \sqrt{\text{var}_{\mathcal{S}}(w_i^*) \cdot \sigma_\xi^2}, \tag{A.8}$$

where $\rho_{w^*,\xi} := \text{cor}_{\mathcal{S}}(w_i^*, \xi_i)$. If we assume the following linear model:

$$\mathbb{E}(\tau_i) = \tau + \beta_X \mathbf{X}_i + \beta_U \mathbf{U}_i,$$

where $\mathbf{X}_i \perp\!\!\!\perp \mathbf{U}_i \mid S_i$, then Equation (A.8) is equivalent to the bias formula from Nguyen et al. (2017):

$$\beta_U \cdot \big(\mathbb{E}(\mathbf{U}_i | S_i = 0) - \mathbb{E}(\mathbf{U}_i | S_i = 1)\big)$$

Proof: Assume we estimate the following $\hat{\tau}(\mathbf{X}_i)$ model:

$$\hat{\tau}(\mathbf{X}_i) := \beta_X \mathbf{X}_i$$

This is equivalent to fitting two linear regressions to the control and treatment potential outcomes, using only $\mathbf{X}_i$. As such,

$$\xi_i = \tau_i - \hat{\tau}(\mathbf{X}_i) = \beta_U \mathbf{U}_i$$

Therefore, using the bias formula:

$$cor_{\mathcal{S}}(w_i^*, \xi_i) \cdot \sqrt{\text{var}_{\mathcal{S}}(w_i^*) \cdot \sigma_\xi^2}$$

$$
\begin{aligned}
&\equiv \text{cov}_{\mathcal{S}}(\xi_i, w_i^*) \\
&= \mathbb{E}\left(\beta_U \mathbf{U}_i \cdot w_i^* | S_i = 1\right) - \mathbb{E}(\beta_U \mathbf{U}_i | S_i = 1) \cdot \mathbb{E}(w_i^* | S_i = 1)
\end{aligned}
$$

Using the decomposition of $w_i^* = w_i \cdot P(\mathbf{U}_i|S_i = 0)/P(\mathbf{U}_i|S_i = 1)$ from Lemma A.2.1:

$$
\begin{aligned}
&= \mathbb{E}\left(\beta_U \mathbf{U}_i \cdot w_i \cdot \frac{P(\mathbf{U}_i|S_i = 0)}{P(\mathbf{U}_i|S_i = 1)} \middle| S_i = 1 \right) - \mathbb{E}(\beta_U \mathbf{U}_i | S_i = 1) \cdot \mathbb{E}(w_i^* | S_i = 1) \\
&= \mathbb{E}(w_i | S_i = 1) \cdot \mathbb{E}\left(\beta_U \mathbf{U}_i \cdot \frac{P(\mathbf{U}_i|S_i = 0)}{P(\mathbf{U}_i|S_i = 1)} \middle| S_i = 1 \right) - \\
&\quad \mathbb{E}(\beta_U \mathbf{U}_i | S_i = 1) \cdot \mathbb{E}(w_i^* | S_i = 1) \\
&= \mathbb{E}(w_i | S_i = 1) \cdot \beta_U \cdot \underbrace{\mathbb{E}\left(\mathbf{U}_i \cdot \frac{P(\mathbf{U}_i|S_i = 0)}{P(\mathbf{U}_i|S_i = 1)} \middle| S_i = 1 \right)}_{:= \mathbb{E}(\mathbf{U}_i | S_i = 0)} - \\
&\quad \beta_U \cdot \mathbb{E}(\mathbf{U}_i | S_i = 1) \cdot \mathbb{E}(w_i^* | S_i = 1)
\end{aligned}
$$

By definition of balancing weights:

$$
\begin{aligned}
&= \mathbb{E}(w_i | S_i = 1) \cdot \beta_U \cdot \mathbb{E}\left(\mathbf{U}_i \mid S_i = 0\right) - \beta_U \cdot \mathbb{E}(\mathbf{U}_i \mid S_i = 1) \cdot \mathbb{E}(w_i^* | S_i = 1) \\
&= \beta_U \cdot \Big(\mathbb{E}\left(\mathbf{U}_i \mid S_i = 0\right) - \mathbb{E}(\mathbf{U}_i \mid S_i = 1) \Big),
\end{aligned}
$$

which is equivalent to the expression from Nguyen et al. (2017). □

Tools for Sensitivity Analysis for the Augmented Weighted Estimator

Robustness Value An analogous robustness value to the one introduced in Section 2.3 can be derived for the augmented weighted estimator. In particular:

$$
RV_q^{Aug} = \frac{1}{2}\left(\sqrt{b_q^2 + 4b_q} - b_q \right), \quad \text{where} \quad b_q = \frac{q^2 \cdot \left(\hat{\tau}_W^{Aug}\right)^2}{\sigma_\xi^2 \cdot \text{var}(w_i)}
$$

The primary difference between RV_q^{Aug} and the previously proposed RV_q is that the robustness value for the augmented weighted estimator is a function of σ_ξ^2, instead of σ_τ^2. This highlights the fact that the relative robustness of the augmented weighted estimator, compared to the weighted estimator, depends directly on how much variation is explained by the individual-level treatment effect model $\hat{\tau}_i$.

Extreme Scenario Analysis In the augmented weighted estimator setting, the extreme scenario analysis represents the case in which the error term ε_i is able to explain all residual variation in the idiosyncratic treatment effect (i.e., $\rho_{\varepsilon,\xi}^2 = 1 - \text{cor}(w_i, \xi_i)^2$). Thus, the maximum parameter values may be evaluated at $\rho_{\varepsilon,\xi}^2 = R_\varepsilon^2 = 1 - \text{cor}(w_i, \xi_i)^2$. In practice,

the correlation between the estimated weights and the residual component of the treatment effect heterogeneity, unexplained by the observed covariates, is likely to be relatively low. As such, we expect the extreme scenario analysis to be conservative in nature. Researchers can employ similar methods to the weighted estimator case to evaluate less conservative scenarios (see Section A.1).

Formal Benchmarking To formally benchmark the sensitivity parameters in the augmented weighted estimator framework, we must also account for the error from misspecifying the treatment effect heterogeneity model. More specifically, let $\hat{\tau}(\mathbf{X}_i^{-(j)})$ be the estimated individual-level treatment effect, omitting covariates $\mathbf{X}_i^{-(j)}$. Then, define the following error term:

$$\xi_i^{-(j)} := \hat{\tau}(\mathbf{X}_i) - \hat{\tau}(\mathbf{X}_i^{-(j)})$$

$\xi_i^{-(j)}$ represents the error incurred from omitting $\mathbf{X}_i^{(j)}$ from estimating τ_i.

We define the following:

$$k_\rho^\xi := \frac{\text{cor}_{\mathcal{S}}(\varepsilon_i, \xi_i)}{\text{cor}_{\mathcal{S}}(\varepsilon_i^{-(j)}, \xi_i^{-(j)})},$$

where k_ρ^ξ compares the amount of variation that ε_i can explain in ξ_i, relative to the amount of variation that $\varepsilon_i^{-(j)}$ can explain in $\xi_i^{-(j)}$. To calibrate $\rho_{\varepsilon,\xi}$, researchers can estimate $\text{cor}_{\mathcal{S}}(\varepsilon_i^{-(j)}, \xi_i^{-(j)})$ and scale by the inputted k_ρ^ξ value. At $k_\rho^\xi = 1$, this implies that the correlation between ε_i and ξ_i is equivalent to the correlation between $\varepsilon_i^{-(j)}$ and $\xi_i^{-(j)}$. It is worth noting that researchers can choose to additionally benchmark σ_ξ^2. However, if researchers are bounding σ_ξ^2 using Equation (A.7), there is no need to calibrate σ_ξ^2 because we will have bounded it using $\sigma_{\tau,max}^2$ (which is not dependent on any covariates) and two estimable quantities.

A.2 Proofs for Theorems and Lemmas

We begin with a lemma that shows the error term can be decomposed into two different components.

Lemma A.2.1 (Error Decomposition)
When using inverse propensity weights, the estimated weights and the ideal weights are written as:

$$w_i = \frac{P(S_i = 1)}{P(S_i = 0)} \cdot \frac{1 - P(S_i = 1|\mathbf{X}_i)}{P(S_i = 1|\mathbf{X}_i)} \qquad w_i^* = \frac{P(S_i = 1)}{P(S_i = 0)} \cdot \frac{1 - P(S_i = 1|\mathbf{X}_i, \mathbf{U}_i)}{P(S_i = 1|\mathbf{X}_i, \mathbf{U}_i)}$$

Then, the error in weight estimation from omitting $\mathbf{U}_i$ can be decomposed in the following manner:

$$\varepsilon_i = w_i - w_i^*$$

$$= \underbrace{\frac{P(S_i=1)}{P(S_i=0)} \cdot \frac{P(S_i=0|\mathbf{X}_i)}{P(S_i=1|\mathbf{X}_i)}}_{\textit{Estimated Weights } (w_i)} \cdot \underbrace{\left(\frac{P(\mathbf{U}_i|\mathbf{X}_i, S_i=1) - P(\mathbf{U}_i|\mathbf{X}_i, S_i=0)}{P(\mathbf{U}_i|\mathbf{X}_i, S_i=1)}\right)}_{\textit{Residual Imbalance in } \mathbf{U}_i},$$

where $P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 1) - P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 1)$ *represents the difference in the underlying probability density function of the omitted confounder* $\mathbf{U}_i$*, conditioned on* $\mathbf{X}_i$*, across the target population (*$S_i = 0$*) and the experimental sample (*$S_i = 1$*).*

Proof: We will substitute in the IPW forms for both w_i and w_i^* and then apply Baye's rule:

$$\begin{aligned}
\varepsilon_i = w_i - w_i^* =& \frac{P(S_i=1)}{P(S_i=0)} \cdot \frac{P(S_i=0|\mathbf{X}_i)}{P(S_i=1|\mathbf{X}_i)} - \frac{P(S_i=1)}{P(S_i=0)} \cdot \frac{P(S_i=0|\mathbf{X}_i, U_i)}{P(S_i=1|\mathbf{X}_i, \mathbf{U}_i)} \\
=& \frac{P(S_i=1)}{P(S_i=0)} \cdot \left(\frac{1}{P(S_i=1|\mathbf{X}_i)} - 1 - \frac{1}{P(S_i=1|\mathbf{X}_i, \mathbf{U}_i)} + 1\right) \\
=& \frac{P(S_i=1)}{P(S_i=0)} \cdot \underbrace{\left(\frac{1}{P(S_i=1|\mathbf{X}_i)} - \frac{1}{P(S_i=1|\mathbf{X}_i, \mathbf{U}_i)}\right)}_{(*)}
\end{aligned}$$

Using Baye's Rule, we can show that ε_i is proportional to the imbalance in the omitted confounder $\mathbf{U}_i$, conditional on $\mathbf{X}_i$. This is done by re-writing the term $(*)$:

$$\begin{aligned}
&\frac{1}{P(S_i=1|\mathbf{X}_i)} - \frac{1}{P(S_i=1|\mathbf{X}_i, \mathbf{U}_i)} \\
&= \frac{1}{P(S_i=1|\mathbf{X}_i)} - \frac{P(\mathbf{U}_i|\mathbf{X}_i)}{P(\mathbf{U}_i|S_i=1, \mathbf{X}_i) \cdot P(S_i=1|\mathbf{X}_i)} \\
&= \frac{1}{P(S_i=1|\mathbf{X}_i)} \cdot \left(1 - \frac{P(\mathbf{U}_i|\mathbf{X}_i)}{P(\mathbf{U}_i|S_i=1, \mathbf{X}_i)}\right) \\
&= \frac{1}{P(S_i=1|\mathbf{X}_i)} \cdot \left(\frac{P(\mathbf{U}_i|\mathbf{X}_i, S_i=1)(1 - P(S_i=1|\mathbf{X}_i)) - P(\mathbf{U}_i|\mathbf{X}_i, S_i=0)P(S_i=0|\mathbf{X}_i)}{P(\mathbf{U}_i|S_i=1, \mathbf{X}_i)}\right) \\
&= \frac{P(S_i=0|\mathbf{X}_i)}{P(S_i=1|\mathbf{X}_i)} \left(\frac{P(\mathbf{U}_i|S_i=1, \mathbf{X}_i) - P(\mathbf{U}_i|S_i=0, \mathbf{X}_i)}{P(\mathbf{U}_i|S_i=1, \mathbf{X}_i)}\right)
\end{aligned}$$

□

Proof of Lemma 2.3.1 (Variance Decomposition of w_i^*)

For inverse propensity score weights, the variance of the true weights w_i^* can be decomposed linearly into two components:

$$\text{var}_{\mathcal{S}}(w_i^*) = \text{var}_{\mathcal{S}}(w_i) + \text{var}_{\mathcal{S}}(\varepsilon_i)$$

Therefore, R^2_ε is bounded between 0 and 1.

Proof: The proof of Lemma 2.3.1 will proceed in two parts. To begin, we will first show that for inverse propensity score weights, $\mathbb{E}_{\mathcal{S}}(w_i^*|\mathbf{X}_i) = w_i$. Then, we will show that $\text{var}(w_i^*)$ can be written as the sum of the variance of the estimated weights w_i and the error term ε_i.

Recall from Lemma A.2.1, we showed that w_i^* could be decomposed in the following terms:

$$w_i^* = \frac{P(S_i = 0)}{P(S_i = 1)} \frac{P(S_i = 0 \mid \mathbf{X}_i)}{P(S_i = 1 \mid \mathbf{X}_i)} \frac{P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 0)}{P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 1)} \equiv w_i \cdot \frac{P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 0)}{P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 1)}$$

We can then show that the expectation of w_i^*, conditioned on $\mathbf{X}_i$, will be equal to w_i:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}(w_i^* \mid \mathbf{X}_i) &= \mathbb{E}_{\mathcal{S}} \left(w_i \cdot \frac{P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 0)}{P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 1)} \middle| \mathbf{X}_i \right) \\
&= w_i \cdot \mathbb{E}_{\mathcal{S}} \left(\frac{P\mathbf{U}_i \mid \mathbf{X}_i, S_i = 0)}{P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 1)} \middle| \mathbf{X}_i \right) \\
&= w_i \cdot \mathbb{E} \left(\frac{P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 0)}{P(\mathbf{U}_i \mid \mathbf{X}_i, S_i = 1)} \middle| \mathbf{X}_i, S_i = 1 \right) \\
&= w_i \cdot \left(\sum_{u \in \mathcal{U}} \frac{P(\mathbf{U}_i = u \mid \mathbf{X}_i, S_i = 0)}{P(\mathbf{U}_i = u \mid \mathbf{X}_i, S_i = 1)} P(\mathbf{U}_i = u \mid \mathbf{X}_i, S_i = 1) \right) \\
&= w_i \cdot \underbrace{\left(\sum_{u \in \mathcal{U}} P(\mathbf{U}_i = u \mid \mathbf{X}_i, S_i = 0) \right)}_{=1} \\
&= w_i
\end{aligned}$$

Now we will show that the variance of ε_i can be written as the difference between the variance of w_i and the variance of w_i^*:

$$\begin{aligned}
\text{var}_{\mathcal{S}}(\varepsilon_i) &= \text{var}_{\mathcal{S}}(w_i - w_i^*) \\
&= \text{var}_{\mathcal{S}}(w_i) + \text{var}_{\mathcal{S}}(w_i^*) - 2\text{cov}_{\mathcal{S}}(w_i, w_i^*) \\
&= \text{var}_{\mathcal{S}}(w_i) + \text{var}_{\mathcal{S}}(w_i^*) - 2\left(\mathbb{E}_{\mathcal{S}}(w_i \cdot w_i^*) - \mathbb{E}_{\mathcal{S}}(w_i)\mathbb{E}_{\mathcal{S}}(w_i^*)\right)
\end{aligned}$$

Making use of the fact that $\mathbb{E}_{\mathcal{S}}(w_i) = \mathbb{E}_{\mathcal{S}}(w_i^*)$ and by Law of Iterated Expectation:

$$= \text{var}_{\mathcal{S}}(w_i) + \text{var}_{\mathcal{S}}(w_i^*) - 2\left(\mathbb{E}_{\mathcal{S}}(\mathbb{E}_{\mathcal{S}}(w_i \cdot w_i^* | \mathbf{X}_i = x)) - \mathbb{E}_{\mathcal{S}}(w_i)^2\right)$$

From above, we have shown that $\mathbb{E}_{\mathcal{S}}(w_i^*|\mathbf{X}_i = x) = w_i$:

$$\begin{aligned}
&= \text{var}_{\mathcal{S}}(w_i) + \text{var}_{\mathcal{S}}(w_i^*) - 2\left(\mathbb{E}_{\mathcal{S}}(w_i^2) - \mathbb{E}_{\mathcal{S}}(w_i)^2\right) \\
&= \text{var}_{\mathcal{S}}(w_i) + \text{var}_{\mathcal{S}}(w_i^*) - 2\text{var}_{\mathcal{S}}(w_i) \\
&= \text{var}_{\mathcal{S}}(w_i^*) - \text{var}_{\mathcal{S}}(w_i)
\end{aligned}$$

Thus, we have shown that $\text{var}_{\mathcal{S}}(w_i^*)$ can be decomposed into the sum of $\text{var}_{\mathcal{S}}(w_i)$ and $\text{var}_{\mathcal{S}}(\varepsilon_i)$. It naturally follows that $R_\varepsilon^2 := \text{var}_{\mathcal{S}}(\varepsilon_i)/\text{var}_{\mathcal{S}}(w_i^*)$ is bounded on the interval $[0, 1]$.

Remark. The extension for balancing weights in Section A.1 states that for Lemma A.2.1 to hold for a set of balancing weights, the condition of $\mathbb{E}_{\mathcal{S}}(w_i^* \mid \mathbf{X}_i) = w_i$ must hold. □

Proof of Lemma 2.3.2 (Correlation Decomposition)

The correlation between ε_i and the individual-level treatment effects can be decomposed in the following manner:

$$\rho_{\varepsilon,\tau} = \begin{cases} \text{cor}_{\mathcal{S}}(w_i, \tau_i)\sqrt{\frac{1-R_\varepsilon^2}{R_\varepsilon^2}} - \text{cor}_{\mathcal{S}}(w_i^*, \tau_i) \cdot \frac{1}{\sqrt{R_\varepsilon^2}} R_\varepsilon^2 > 0 \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{when } R_\varepsilon^2 = 0 \end{cases}$$

Furthermore, $\rho_{\varepsilon,\tau}$ is bounded by the following range:

$$-\sqrt{1 - \text{cor}^2(w_i, \tau_i)} \leq \rho_{\varepsilon,\tau} \leq \sqrt{1 - \text{cor}^2(w_i, \tau_i)}$$

Proof: To begin, we can rewrite $\rho_{\varepsilon,\tau}$ as follows:

$$\begin{aligned}
\rho_{\varepsilon,\tau} &= \frac{\text{cov}_{\mathcal{S}}(w_i, \tau_i) - \text{cov}(w_i^*, \tau_i)}{\sqrt{\text{var}_{\mathcal{S}}(\varepsilon_i) \cdot \text{var}_{\mathcal{S}}(\tau_i)}} \\
&= \frac{\text{cor}_{\mathcal{S}}(w_i, \tau_i) \cdot \sqrt{\text{var}_{\mathcal{S}}(w_i) \cdot \text{var}_{\mathcal{S}}(\tau_i)} - \text{cor}_{\mathcal{S}}(w_i^*, \tau_i) \cdot \sqrt{\text{var}_{\mathcal{S}}(w_i^*) \cdot \text{var}_{\mathcal{S}}(\tau_i)}}{\sqrt{\text{var}_{\mathcal{S}}(\varepsilon_i) \cdot \text{var}_{\mathcal{S}}(\tau_i)}} \\
&= \frac{\text{cor}_{\mathcal{S}}(w_i, \tau_i) \cdot \sqrt{\text{var}_{\mathcal{S}}(w_i)} - \text{cor}_{\mathcal{S}}(w_i^*, \tau_i) \cdot \sqrt{\text{var}_{\mathcal{S}}(w_i^*)}}{\sqrt{\text{var}_{\mathcal{S}}(w_i^*) \cdot R_\varepsilon^2}} \\
&= \text{cor}_{\mathcal{S}}(w_i, \tau_i)\sqrt{\frac{\text{var}_{\mathcal{S}}(w_i)}{\text{var}_{\mathcal{S}}(w_i^*)}} \cdot \frac{1}{\sqrt{R_\varepsilon^2}} - \text{cor}_{\mathcal{S}}(w_i^*, \tau_i) \cdot \frac{1}{\sqrt{R_\varepsilon^2}} \\
&= \text{cor}_{\mathcal{S}}(w_i, \tau_i)\sqrt{\frac{1-R_\varepsilon^2}{R_\varepsilon^2}} - \text{cor}_{\mathcal{S}}(w_i^*, \tau_i) \cdot \frac{1}{\sqrt{R_\varepsilon^2}} \qquad \text{(A.9)}
\end{aligned}$$

Now, note that $\text{cor}(w_i^*, \tau_i)$ can be bounded using the recursive formula of partial correlation:[5]

$$\text{cor}_{\mathcal{S}}(w_i^*, \tau_i) \in \text{cor}_{\mathcal{S}}(w_i, \tau_i) \cdot \text{cor}_{\mathcal{S}}(w_i, w_i^*) \pm \sqrt{1 - \text{cor}_{\mathcal{S}}^2(w_i, \tau_i)}\sqrt{1 - \text{cor}_{\mathcal{S}}^2(w_i, w_i^*)}$$

Because $\text{cor}_{\mathcal{S}}(w_i, w_i^*) = \sqrt{\frac{\text{var}_{\mathcal{S}}(w_i)}{\text{var}_{\mathcal{S}}(w_i^*)}}$, the above simplifies to the following:

$$\text{cor}_{\mathcal{S}}(w_i^*, \tau_i) \in \text{cor}_{\mathcal{S}}(w_i, \tau_i)\sqrt{1 - R_\varepsilon^2} \pm \sqrt{1 - \text{cor}^2(w_i, \tau_i)}\sqrt{R_\varepsilon^2}.$$

[5]This follows from applying the recursive formula of partial correlation for a single variable, and applying the fact that the partial correlation must be bounded by 1 and -1.

Thus, substituting in the bounds for $\text{cor}(w_i^*, \tau_i)$ into Equation (A.9), we obtain the bound:

$$-\sqrt{1 - \text{cor}_{\mathcal{S}}^2(w_i, \tau_i)} \leq \rho_{\varepsilon,\tau} \leq \sqrt{1 - \text{cor}_{\mathcal{S}}^2(w_i, \tau_i)}$$

□

Proof of Theorem 2.3.1 (Bias of Weighted Estimator)

Assume $Y_i(1) - Y_i(0) \perp\!\!\!\perp S_i \mid \{\mathbf{X}_i, \mathbf{U}_i\}$. Let w_i be the weights estimated using only $\mathbf{X}_i$, and let w_i^* be the (correct) weights, obtained using $\{\mathbf{X}_i, \mathbf{U}_i\}$. The bias of a weighted estimator from using w_i instead of w_i^* is given as:

$$\text{Bias}(\hat{\tau}_W) = \rho_{\varepsilon,\tau} \cdot \sqrt{\text{var}_{\mathcal{S}}(w_i) \cdot \frac{R_\varepsilon^2}{1 - R_\varepsilon^2} \cdot \sigma_\tau^2},$$

where ε_i is defined as the difference between the estimated weights and the correct weights (i.e., $\varepsilon_i = w_i - w_i^*$), and τ_i is the individual-level treatment effect.

Proof: I will first show the proof for a Horvitz-Thompson style weighted estimator. The proof for a Hajek style weighted estimator (with stabilized weights) follows similarly, but with the addition of a finite-sample bias term. A Horvitz-Thompson style weighted estimator is defined as:

$$\hat{\tau}_W = \frac{1}{n_1} \sum_{i \in \mathcal{S}} w_i T_i Y_i - \frac{1}{n_0} \sum_{i \in \mathcal{S}} w_i (1 - T_i) Y_i$$

We begin by showing that if we were to have estimated weights with the full separating set $\mathcal{X}_i$, the weighted estimator will be an unbiased estimator for PATE. We will denote the weighted estimator using w_i^* as $\hat{\tau}_W^*$. We will denote expectations with a subscript $\mathcal{S}$ as the expectation over the experimental sample (i.e., $\mathbb{E}_{\mathcal{S}}(\cdot) = \mathbb{E}(\cdot \mid S_i = 1)$), and expectations with a subscript $\mathcal{P}$ as the expectation over the target population. Expectations with no subscripts will represent the expectation over both the experimental sample and the target population. We define $\mathcal{D}$ as the set of all indices corresponding to units in the experimental sample and the target population.

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_W^*) =& \mathbb{E}\left(\frac{1}{n_1} \sum_{i \in \mathcal{S}} w_i^* T_i Y_i - \frac{1}{n_0} \sum_{i \in \mathcal{S}} w_i^* T_i Y_i\right) \\
=& \mathbb{E}\left(\frac{1}{n_1} \sum_{i \in \mathcal{D}} w_i^* T_i Y_i S_i - \frac{1}{n_0} \sum_{i \in \mathcal{D}} w_i^* (1 - T_i) Y_i S_i\right) \\
=& \frac{1}{n_1} \mathbb{E}\left(\sum_{i \in \mathcal{D}} w_i^* S_i T_i Y_i(1)\right) - \frac{1}{n_0} \mathbb{E}\left(\sum_{i \in \mathcal{D}} w_i^* (1 - T_i) Y_i(0)\right)
\end{aligned}$$

By Linearity of Expectation:

$$=\frac{1}{n_1}\sum_{i\in\mathcal{D}}\mathbb{E}\left(w_i^* S_i T_i Y_i(1)\right) - \frac{1}{n_0}\sum_{i\in\mathcal{D}}\mathbb{E}\left(w_i^*(1-T_i)Y_i(0)\right)$$

By Law of Total Expectation:

$$\begin{aligned}
=&\frac{1}{n_1}\sum_{i\in\mathcal{D}}\mathbb{E}\left(w_i^* S_i T_i Y_i(1)|S_i=1,T_i=1\right)P(S_i=1 \text{ and } T_i=1)+\\
&\frac{1}{n_0}\sum_{i\in\mathcal{D}}\mathbb{E}\left(w_i^* S_i (1-T_i) Y_i(0)|S_i=1,T_i=0\right)\\
=&\frac{1}{n_1}\sum_{i\in\mathcal{D}}\frac{n}{n+N}\cdot\frac{n_1}{n}\mathbb{E}\left(w_i^* S_i T_i Y_i(1)|S_i=1,T_i=1\right)+\\
&\frac{1}{n_0}\sum_{i\in\mathcal{D}}\frac{n}{n+N}\cdot\frac{n_0}{n}\mathbb{E}\left(w_i^* S_i (1-T_i) Y_i(0)|S_i=1,T_i=0\right)\\
=&\mathbb{E}(w_i^* S_i Y_i(1)|S_i=1,T_i=1) - \mathbb{E}(w_i^* S_i Y_i(0)|S_i=1,T_i=0)
\end{aligned}$$

From random treatment assignment:

$$\begin{aligned}
=&\mathbb{E}(w_i^* S_i Y_i(1)|S_i=1) - \mathbb{E}(w_i^* S_i Y_i(0)|S_i=1)\\
=&\mathbb{E}(w_i^* S_i (Y_i(1)-Y_i(0))|S_i=1)\\
\equiv&\mathbb{E}_{\mathcal{S}}(w_i^*\tau_i)
\end{aligned} \tag{A.10}$$

To show that $\mathbb{E}_{\mathcal{S}}(w_i^*\tau_i)=\tau$, we first apply Baye's Rule:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}(w_i^*\tau_i) &= \sum_{\tau_i,\mathcal{X}_i} w_i^*\tau_i\cdot P(\mathcal{X}_i,\tau_i|S_i=1)\\
&= \sum_{\tau,\mathcal{X}_i} w_i^*\tau_i\cdot\frac{P(S_i=1|\mathcal{X}_i,\tau_i)\cdot P(\mathcal{X}_i,\tau_i)}{P(S_i=1)}
\end{aligned}$$

By the conditional ignorability assumption that $\tau_i \perp\!\!\!\perp S_i \mid \mathcal{X}_i$:

$$\begin{aligned}
&= \sum_{\tau,\mathcal{X}_i} w_i^*\tau_i\cdot\frac{P(S_i=1|\mathcal{X}_i)\cdot P(\mathcal{X}_i,\tau_i)}{P(S_i=1)}\\
&= \sum_{\tau,\mathcal{X}_i} \frac{P(S_i=1)}{P(S_i=0)}\frac{P(S_i=0|\mathcal{X}_i)}{P(S_i=1|\mathcal{X}_i)}\tau_i\cdot\frac{P(S_i=1|\mathcal{X}_i)\cdot P(\mathcal{X}_i,\tau_i)}{P(S_i=1)}\\
&= \sum_{\tau,\mathcal{X}_i} \tau_i\cdot\frac{P(S_i=0|\mathcal{X}_i)\cdot P(\mathcal{X}_i,\tau_i)}{P(S_i=0)}\\
&= \sum_{\tau,\mathcal{X}_i} \tau_i\cdot P(\mathcal{X}_i,\tau_i|S_i=0)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathcal{P}}(\tau_i) \\
&= \mathbb{E}_{\mathcal{P}}(\tau_i) \equiv \tau
\end{aligned}
$$

As such, the bias of a weighted estimator when omitting a confounder is:

$$\text{Bias}(\hat{\tau}_W) = \mathbb{E}_{\mathcal{S}}(\hat{\tau}_W) - \tau$$

Using the result from Equation (A.10) with w_i and the fact that $\mathbb{E}_{\mathcal{S}}(w_i^* \tau_i) = \tau$:

$$
\begin{aligned}
&= \mathbb{E}_{\mathcal{S}}(w_i \tau_i) - \mathbb{E}_{\mathcal{S}}(w_i^* \tau_i) \\
&= \mathbb{E}_{\mathcal{S}}(\underbrace{(w_i - w_i^*)}_{:=\varepsilon_i} \tau_i) \\
&= \mathbb{E}_{\mathcal{S}}(\varepsilon_i \tau_i)
\end{aligned}
$$

By construction, $\mathbb{E}_{\mathcal{S}}(w_i) = \mathbb{E}_{\mathcal{S}}(w_i^*) = 1$, which implies that $\mathbb{E}_{\mathcal{S}}(\varepsilon_i) = 0$:

$$
\begin{aligned}
&= \mathbb{E}_{\mathcal{S}}(\varepsilon_i \tau_i) - \mathbb{E}_{\mathcal{S}}(\varepsilon_i) \cdot \mathbb{E}_{\mathcal{S}}(\tau_i) \\
&= \text{cov}_{\mathcal{S}}(\varepsilon_i, \tau_i) \\
&= \text{cor}_{\mathcal{S}}(\varepsilon_i, \tau_i) \cdot \sqrt{\text{var}_{\mathcal{S}}(\varepsilon_i) \cdot \text{var}_{\mathcal{S}}(\tau_i)}
\end{aligned}
$$

Define $R^2_\varepsilon := \text{var}_{\mathcal{S}}(\varepsilon_i)/\text{var}_{\mathcal{S}}(w_i^*)$ and making use of Lemma 2.3.1:

$$
\begin{aligned}
&= \text{cor}_{\mathcal{S}}(\varepsilon_i, \tau_i) \cdot \sqrt{\text{var}_{\mathcal{S}}(w_i) \cdot \frac{R^2_\varepsilon}{1 - R^2_\varepsilon} \cdot \text{var}_{\mathcal{S}}(\tau_i)} \\
&\equiv \rho_{\varepsilon,\tau} \cdot \sqrt{\text{var}_{\mathcal{S}}(w_i) \cdot \frac{R^2_\varepsilon}{1 - R^2_\varepsilon} \cdot \sigma^2_\tau}
\end{aligned}
$$

□

Proof of Theorem 2.4.1

Let k_σ and k_ρ be defined as in Equation (2.10). Let $R^{2-(j)}_\varepsilon := \text{var}_{\mathcal{S}}(\varepsilon_i^{-(j)})/\text{var}_{\mathcal{S}}(w_i)$, and $\rho^{-(j)}_{\varepsilon,\tau} := \text{cor}_{\mathcal{S}}(\varepsilon_i^{-(j)}, \tau_i)$. The sensitivity parameters R^2_ε and $\rho_{\varepsilon,\tau}$ can be written as a function of k_σ and k_ρ:

$$R^2_\varepsilon = \frac{k_\sigma \cdot R^{2-(j)}_\varepsilon}{1 + k_\sigma \cdot R^{2-(j)}_\varepsilon}, \qquad \rho_{\varepsilon,\tau} = k_\rho \cdot \rho^{-(j)}_{\varepsilon,\tau}$$

Proof: It follows immediately from Equation (2.10) that $\rho_{\varepsilon,\tau} = k_\rho \cdot \rho_{\varepsilon,\tau}^{-(j)}$. Therefore, we just need to show that R_ε^2 can be written as a function of $R_\varepsilon^{2-(j)}$.

$$R_\varepsilon^2 = \frac{\text{var}_\mathcal{S}(\varepsilon_i)}{\text{var}_\mathcal{S}(w_i^*)}$$

By Equation (2.10):

$$\begin{aligned}
&= k_\sigma \cdot \frac{\text{var}_\mathcal{S}(\varepsilon_i^{-(j)})}{\text{var}_\mathcal{S}(w_i^*)} \\
&= k_\sigma \cdot \frac{\text{var}_\mathcal{S}(\varepsilon_i^{-(j)})}{\text{var}_\mathcal{S}(w_i) + \text{var}_\mathcal{S}(\varepsilon_i)} \\
&= \frac{k_\sigma \cdot \text{var}_\mathcal{S}(\varepsilon_i^{-(j)})/\text{var}_\mathcal{S}(w_i)}{1 + k_\sigma \text{var}_\mathcal{S}(\varepsilon_i^{-(j)})/\text{var}_\mathcal{S}(w_i)} \\
&= \frac{k_\sigma \cdot R_\varepsilon^{2-(j)}}{1 + k_\sigma \cdot R_\varepsilon^{2-(j)}}
\end{aligned}$$

□

Proof of Theorem 2.5.1

The bias of an augmented weighted estimator when both the weight and outcome model are mis-specified is given as:

$$\text{Bias}(\hat{\tau}_W^{Aug}) = \rho_{\varepsilon,\xi} \cdot \sqrt{\text{var}_\mathcal{S}(w_i) \cdot \frac{R_\varepsilon^2}{1 - R_\varepsilon^2} \cdot \text{var}_\mathcal{S}(\xi_i)}$$

where ε_i is defined consistent with before, and ξ_i represents the difference between the true individual-level treatment effect and estimated treatment effect (i.e., $\xi_i = \tau_i - \hat{\tau}_i$).

Proof:

$$\hat{\tau}_W^{Aug} = \hat{\tau}_W - \underbrace{\frac{1}{n}\sum_{i\in\mathcal{S}} w_i \hat{\tau}(\mathbf{X}_i) + \frac{1}{N}\sum_{i\in\mathcal{P}} \hat{\tau}(\mathbf{X}_i)}_{\text{Augmented Component}}$$

From Theorem 2.3.1, we showed that $\mathbb{E}(\hat{\tau}_W) = \mathbb{E}_\mathcal{S}(w_i\tau_i)$. We will now derive the expectation of the augmented component. To begin, we take the expectation of the $1/n\sum_{i\in\mathcal{S}} w_i\hat{\tau}(\mathbf{X}_i)$ component:

$$\mathbb{E}\left(\frac{1}{n}\sum_{i\in\mathcal{S}} w_i\hat{\tau}(\mathbf{X}_i)\right) = \frac{1}{n}\mathbb{E}_\mathcal{S}\left(\sum_{i\in\mathcal{S}} w_i\hat{\tau}(\mathbf{X}_i)\right)$$

$$
\begin{aligned}
&= \frac{1}{n}\mathbb{E}_{\mathcal{S}}\left(\sum_{i\in\mathcal{D}} S_i w_i \hat{\tau}(\mathbf{X}_i)\right) \\
&= \frac{1}{n}\sum_{i\in\mathcal{D}} \mathbb{E}(S_i w_i \hat{\tau}(\mathbf{X}_i)) \\
&= \frac{1}{n}\sum_{i\in\mathcal{D}} \mathbb{E}(S_i w_i \hat{\tau}(\mathbf{X}_i) \mid S_i = 1) P(S_i = 1) \\
&= \mathbb{E}(w_i \hat{\tau}(\mathbf{X}_i) \mid S_i = 1) \\
&= \mathbb{E}_{\mathcal{S}}(w_i \hat{\tau}(\mathbf{X}_i)
\end{aligned}
$$

For the $1/N \sum_{i\in\mathcal{P}} \hat{\tau}(\mathbf{X}_i)$ component:

$$
\begin{aligned}
\mathbb{E}\left(\frac{1}{N}\sum_{i\in\mathcal{P}} \hat{\tau}(\mathbf{X}_i)\right) &= \mathbb{E}\left(\frac{1}{N}\sum_{i\in\mathcal{D}} (1 - S_i)\cdot \hat{\tau}(\mathbf{X}_i)\right) \\
&= \frac{1}{N}\sum_{i\in\mathcal{D}} \mathbb{E}((1 - S_i)\hat{\tau}(\mathbf{X}_i)) \\
&= \frac{1}{N}\sum_{i\in\mathcal{D}} \mathbb{E}((1 - S_i)\hat{\tau}(\mathbf{X}_i) \mid S_i = 0) P(S_i = 0) \\
&= \mathbb{E}(\hat{\tau}(\mathbf{X}_i) \mid S_i = 0) \\
&= \mathbb{E}_{\mathcal{P}}(\hat{\tau}(\mathbf{X}_i))
\end{aligned}
$$

As such, the bias of the augmented weighted estimator can be written as follows:

$$
\begin{aligned}
\text{Bias}(\hat{\tau}_W^{Aug}) &= \mathbb{E}(\hat{\tau}_W^{Aug}) - \tau \\
&= \mathbb{E}_{\mathcal{S}}(w_i(\tau_i - \hat{\tau}(\mathbf{X}_i)) + \mathbb{E}_{\mathcal{P}}(\hat{\tau}(\mathbf{X}_i)) - \mathbb{E}_{\mathcal{P}}(\tau_i) \\
&= \mathbb{E}_{\mathcal{S}}(w_i(\tau_i - \hat{\tau}(\mathbf{X}_i)) - \mathbb{E}_{\mathcal{P}}(\tau_i - \hat{\tau}_i)
\end{aligned}
$$

By definition, $\varepsilon_i = w_i - w_i^*$:

$$
\begin{aligned}
&= \mathbb{E}_{\mathcal{S}}(\varepsilon_i(\tau_i - \hat{\tau}_i)) + \mathbb{E}_{\mathcal{S}}(w_i^*(\tau_i - \hat{\tau}_i)) - \mathbb{E}_{\mathcal{P}}(\tau_i - \hat{\tau}_i) \\
&= \mathbb{E}_{\mathcal{S}}(\varepsilon_i(\tau_i - \hat{\tau}_i)) + \mathbb{E}_{\mathcal{P}}(\tau_i - \hat{\tau}_i) - \mathbb{E}_{\mathcal{P}}(\tau_i - \hat{\tau}_i) \\
&= \mathbb{E}_{\mathcal{S}}(\varepsilon_i(\tau_i - \hat{\tau}_i))
\end{aligned}
$$

Defining $\xi_i := \tau_i - \hat{\tau}_i$:

$$
\begin{aligned}
&= \mathbb{E}_{\mathcal{S}}(\varepsilon_i \cdot \xi_i) \\
&= \text{cov}_{\mathcal{S}}(\varepsilon_i, \xi_i) \\
&= \text{cor}_{\mathcal{S}}(\varepsilon_i, \xi_i) \cdot \sqrt{\text{var}_{\mathcal{S}}(\varepsilon_i) \cdot \text{var}_{\mathcal{S}}(\xi_i)} \\
&= \rho_{\varepsilon,\xi} \cdot \sqrt{\text{var}_{\mathcal{S}}(w_i) \cdot \frac{R_\varepsilon^2}{1 - R_\varepsilon^2} \cdot \text{var}_{\mathcal{S}}(\xi_i)}
\end{aligned}
$$

□

A.3 Additional Derivations

Robustness Value

To derive the robustness value, recall that we are interested in the bias that arises from a confounder with equal impact on the overall imbalance and the individual-level treatment effect. In particular, we define the robustness value such that $RV = \rho^2_{\varepsilon,\tau} = R^2_\varepsilon$. Thus, for the bias to equal $q \times 100\%$ of a given point estimate:

$$\text{Bias}(\hat{\tau}_W) = q \cdot \hat{\tau}_W$$

From Theorem 2.3.1:

$$\begin{aligned}
\implies \rho_{\varepsilon,\tau}\sqrt{\text{var}_{\mathcal{S}}(w_i) \cdot \frac{R^2_\varepsilon}{1-R^2_\varepsilon} \cdot \sigma^2_\tau} &= q \cdot \hat{\tau}_W \\
\rho^2_{\varepsilon,\tau} \cdot \text{var}_{\mathcal{S}}(w_i) \cdot \frac{R^2_\varepsilon}{1-R^2_\varepsilon} \cdot \sigma^2_\tau &= q^2 \cdot \hat{\tau}^2_W \\
\rho^2_{\varepsilon,\tau} \cdot \frac{R^2_\varepsilon}{1-R^2_\varepsilon} &= \underbrace{\frac{q^2 \cdot \hat{\tau}^2_W}{\text{var}_{\mathcal{S}}(w_i) \cdot \sigma^2_\tau}}_{:=a_q}
\end{aligned}$$

Defining $RV = \rho^2_{\varepsilon,\tau} = R^2_\varepsilon$:

$$RV \cdot \frac{RV}{1-RV} = a_q \tag{A.11}$$

Let RV_q be the value of RV for a given q. Thus, solving Equation (A.11) for RV_q:

$$RV_q = \frac{1}{2}\left(\sqrt{a_q^2 + 4a_q} - a_q\right)$$

A similar derivation can be applied for the augmented weighted estimator, with the primary difference being that instead of a_q, the robustness value is a function of b_q, where:

$$b_q = \frac{q^2 \cdot \left(\hat{\tau}^{Aug}_W\right)^2}{\sigma^2_\xi \cdot \text{var}(w_i)}$$

Extreme Bounds

To derive R^2_{max}, we set $\rho_{\varepsilon,\tau}$ to be at the extreme bounds of $\pm\sqrt{1-\text{cor}^2_{\mathcal{S}}(w_i,\tau_i)}$, and solve for R^2_ε using the correlation decomposition from Lemma 2.3.2.

$$\text{cor}_{\mathcal{S}}(w_i,\tau_i) \cdot \sqrt{\frac{1-R^2_\varepsilon}{R^2_\varepsilon}} - \text{cor}_{\mathcal{S}}(w^*_i,\tau_i) \cdot \frac{1}{\sqrt{R^2_\varepsilon}} \leq \sqrt{1-\text{cor}^2_{\mathcal{S}}(w_i,\tau_i)}$$

$$\text{cor}_{\mathcal{S}}(w_i, \tau_i)\sqrt{1 - R^2_\varepsilon} - \text{cor}_{\mathcal{S}}(w_i^*, \tau_i) \leq \sqrt{1 - \text{cor}^2_{\mathcal{S}}(w_i, \tau_i)} \cdot \sqrt{1 - \text{cor}^2_{\mathcal{S}}(w_i^*, \tau_i)}$$

Using the quadratic formula, we solve for $\sqrt{R^2_\varepsilon}$:

$$\sqrt{R^2_\varepsilon} = 1 - \text{cor}_{\mathcal{S}}(w_i, \tau_i) \cdot \text{cor}_{\mathcal{S}}(w_i^*, \tau_i) \pm \sqrt{(1 - \text{cor}^2_{\mathcal{S}}(w_i^*, \tau_i)} \cdot \sqrt{(1 - \text{cor}^2_{\mathcal{S}}(w_i, \tau_i)}$$

Setting $\text{cor}_{\mathcal{S}}(w_i^*, \tau_i) = \pm 1$, we find that

$$\underbrace{R^2_\varepsilon}_{:=R^2_{max}} = 1 - \text{cor}_{\mathcal{S}}(w_i, \tau_i)^2$$

A.4 Extended Results for Empirical Application

Bounding σ^2_τ

To estimate a bound for σ^2_τ, we follow Figure A.1 and use the upper bound estimated in Equation (A.3) (i.e., for cases when we assume $\text{cov}_{\mathcal{S}}(\tau_i, Y_i(0)) \geq 0$). This is in line with the substantive findings from the original JTPA study. To reiterate the example from Section A.1, researchers found that women with the greatest estimated impact from JTPA services also had higher hourly wages in their work history, and or came from families with greater household income. (See Bloom et al. (1993) for more discussion.)

Therefore, we bound σ^2_τ by taking the difference between the estimated variance in the treated outcomes and the estimated variance in the control outcomes and obtain a bound of 29.01.

Estimation Uncertainty with Benchmarking

To account for potential estimation uncertainty associated with benchmarking results, we use a similar approach from Hong et al. (2021) and perform benchmarking across 1,000 bootstrap iterations. We then check the number of bootstrap iterations in which the benchmarked covariates are strong enough to be a killer confounder (i.e., either strong enough to reduce the estimate to zero, or strong enough to alter the statistical significance of the estimate). We report this under "% Con." in Table A.2.

We see that the benchmarking results are relatively stable across bootstrap iterations. More specifically, in assessing if omitting a confounder with equivalent confounding strength to one of the observed covariates will result in enough bias for the estimated effect to be reduced to zero, we see that only 1-2% of the bootstrap iterations resulted in killer confounders. This implies that even accounting for estimation uncertainty, we can generally expect that omitting a confounder with equivalent confounding strength as one of the observed covariates will not result in a killer confounder that reduces the estimate to zero. In contrast, in assessing changes to statistical significance, we see that omitting a confounder with equivalent confounding strength as whether or not an individual is Black would have resulted in a statistically insignificant effect in 7% of the bootstrap iterations.

				Estimated Effect = 0				Changes in Signif.			
Variable	$\hat{R}^2_{\varepsilon}$	$\hat{\rho}_{\varepsilon,\tau}$	$\widehat{\text{Bias}}$	MRCS	k^{min}_{σ}	k^{min}_{ρ}	% Con.	MRCS	k^{min}_{σ}	k^{min}_{ρ}	% Con.
Prev. Earnings	0.01	-0.41	-0.12	-23.4	72.9	-1.8	2%	-2.4	10.7	-0.7	4%
Age	0.00	0.04	0.00	—	—	18.2	1%	—	—	7.0	1%
Married	0.05	-0.19	-0.14	-20.7	12.3	-4.0	1%	-2.1	1.8	-1.5	1%
Hourly Wage	0.03	-0.24	-0.14	-20.6	20.2	-3.1	1%	-2.1	3.0	-1.2	1%
Black	0.17	-0.11	-0.16	-17.7	3.4	-7.1	2%	-1.8	0.5	-2.7	7%
Hispanic	0.24	-0.14	-0.26	-10.8	2.3	-5.5	0%	-1.1	0.3	-2.1	4%
HS/GED	0.07	-0.04	-0.04	-76.1	8.1	-18.5	1%	-7.7	1.2	-7.1	3%
Education	0.07	-0.10	-0.09	-30.0	7.5	-7.6	1%	-3.0	1.1	-2.9	4%

Point Estimate ($\hat{\tau}_W$): 2.81; $\hat{\sigma}^2_{\tau,\max} = 29.01$; $RV_1 = 0.56$; $RV_{\alpha=0.05} = 0.08$

Table A.2: Formal benchmarking results for Coosa Valley, Georgia. The estimated bias is reported in thousands of USD.

Applying the Augmented Weighted Sensitivity Analysis

To illustrate the sensitivity analysis for augmented weighted estimators, we return to our JTPA application. To estimate the individual-level treatment effect model, we use a causal random forest, estimated on the same set of covariates included in the weights (Athey et al. (2019)). We then estimate the individual-level treatment effect for all units across both the experimental sample and the target population. Using the bound from Equation (A.7), we estimate an upper bound for σ^2_{ξ} to be 28.5. After obtaining the upper bound for σ^2_{ξ}, we proceed with the sensitivity analysis.

Summarizing Sensitivity. To begin, we visualize the bias contour plot, as well as estimate the robustness value and the extreme scenario bound. See Figure A.2 for the bias contour plot.

	Unweighted	Aug-Weighted	$RV^{Aug}_{q=1}$
Impact of JTPA access on earnings	1.63	2.84	0.56

$\hat{\sigma}^2_{\xi,\max} = 28.50$; $\widehat{\text{cor}}_{\mathcal{S}}(w_i, \xi_i) = 0.24$

Table A.3: Summary of point estimates and sensitivity statistics.

The robustness value for the augmented weighted estimator is 0.56, which implies that if the error from omitting the confounder can explain 56% of the variation in the idiosyncratic treatment effect (i.e., ξ_i), as well as 56% of the variation in the ideal weights, then the bias will be large enough to reduce the point estimate to 0. We see that the robustness value for the augmented weighted estimator is slightly higher than the robustness value for the

weighted estimator. This is likely due to the fact that we have modeled some of the variation in τ_i with our estimated treatment effect heterogeneity model.

Formal Benchmarking Results. We now perform formal benchmarking across the observed covariates for the augmented weighted estimator. The formally benchmarked parameter values for the R^2_ε parameter will be identical to the formally benchmarked R^2_ε values in the weighted estimator setting. In general, we that the estimated bias values from formal benchmarking in the augmented weighted estimator case is lower than the estimated bias values for the weighted estimator case; this is likely due to the fact that the bound on the idiosyncratic treatment effect variation is lower than the bound on the overall treatment effect heterogeneity (i.e., $\hat{\sigma}^2_{\xi,\max} \leq \hat{\sigma}^2_{\tau,\max}$).

Covariate	R^2_ε	$\rho_{\varepsilon,\xi}$	Est. Bias	MRCS	k^{min}_σ	k^{min}_ρ
Prev. Earnings	0.01	0.02	0.00	576.67	73.72	44.59
Age	0.00	0.00	0.00	—	—	439.56
Married	0.05	0.05	0.04	78.05	12.40	15.01
Hourly Wage	0.03	0.20	0.11	25.76	20.46	3.82
Black	0.17	0.03	0.05	57.90	3.39	22.78
Hisp.	0.24	0.11	0.22	13.14	2.31	6.57
HS/GED	0.07	-0.09	-0.08	-34.39	8.14	-8.26
Years of Educ.	0.07	-0.13	-0.13	-22.32	7.56	-5.58

Table A.4: Formal benchmarking for Coosa Valley, Georgia, for an augmented weighted estimator. We see a greater degree of robustness in omitting a confounder with equivalent confounding strength to the observed covariates for the augmented weighted estimator, relative to the weighted estimator. This is reflected in the larger MRCS, k^{min}_σ and k^{min}_ρ values.

Extreme Scenario Analysis

For the extreme scenario analysis, we examine the potential bias that may occur if the correlation term is equal to the maximum possible value of $\sqrt{1 - \text{cor}_{\mathcal{S}}(w_i, \tau_i)}$. Then, we evaluate the R^2_ε value that corresponds to this maximum correlation term, when $|\text{cor}_{\mathcal{S}}(w^*_i, \tau_i)| = 1$. In general, we expect this to be an extremely conservative estimate for the maximum amount of bias incurred by an omitted confounder. We provide the results in Table A.5.

The general plausibility of an omitted confounder with the degree of explanatory power and imbalance seems relatively low. In particular, comparing ρ_{max} and R^2_{max} with the benchmarked parameters shows that the omitted confounder would have to be significantly stronger than any of the observed covariates for the extreme scenario to occur.

In cases when researchers do not feel that the benchmarked parameters are representative of the potential confounders, it can be difficult to justify the plausibility or implausibility

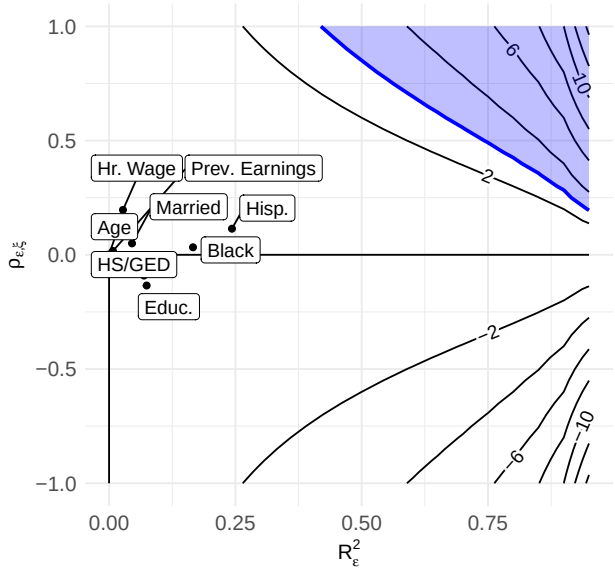

Figure A.2: Bias Contour Plot for Coosa Valley, Georgia, using an Augmented Weighted Estimator. Akin to Figure 2.1, the shaded blue region represents the killer confounder region, for which a confounder will result in a directional change of the point estimate. We also plot the formal benchmarking results. We see that the points are even further away from the killer confounder region than in the weighted estimator setting.

	Estimate	ρ_{max}	R^2_{max}	Est. Bias
Weighted	2.81	0.93	0.86	7.84
Augmented	2.84	0.93	0.87	8.21

Table A.5: Extreme Scenario Analysis. We note that $\rho_{\max}$ and $R^2_{\max}$ are far larger than any of the benchmarked parameters.

of such an extreme $\rho_{\varepsilon,\tau}$ (or $\rho_{\varepsilon,\xi}$) term. An alternative approach is for researchers to vary different $\text{cor}_{\mathcal{S}}(w_i^*, \tau_i)$ (or $\text{cor}_{\mathcal{S}}(w_i^*, \xi_i)$) values, which can be easier to assess the plausibility of, because they can directly compare the posited $\text{cor}_{\mathcal{S}}(w_i^*, \tau_i)$ (or $\text{cor}_{\mathcal{S}}(w_i^*, \xi_i)$) with the observed correlation values calculated using the estimated weights. $\text{cor}_{\mathcal{S}}(w_i^*, \tau_i)$ represents the maximum amount of variation that the selection weights can explain in the treatment effect heterogeneity. For example, if researchers assume that the (true) selection weights are highly correlated with the treatment effect heterogeneity, then $\text{cor}_{\mathcal{S}}(w_i^*, \tau_i)$ should be close to 1.

To visually represent this, we generate plots where the x-axis represents the R^2 value, and the y-axis represents the adjusted point estimate. We fix $\text{cor}_{\mathcal{S}}(w_i^*, \tau_i)$ and $\text{cor}_{\mathcal{S}}(w_i^*, \xi_i)$

to a set of values: {-0.5, 0.25, 0.25, 0.5, 0.9}. The estimated correlation value between the estimated weights and the individual-level treatment effect is 0.07, while the estimated correlation value between the estimated weights and the idiosyncratic treatment effect is 0.11. Thus, even for the case that $|\text{cor}_{\mathcal{S}}(w_i^*, \tau_i)|$ or $|\rho(w_i^*, \xi_i)|$ to equal 0.25 would imply that additionally balancing on an omitted confounder would result in a significantly higher amount of variation explained. We see that for both the weighted and augmented weighted estimators, it is only when the correlation term switches signs that the point estimate is at risk of being zero, or negative. In other words, additionally balancing on the omitted confounder would have to alter the direction of the correlation between the weights and τ_i (or ξ_i) for the point estimate to become negative.

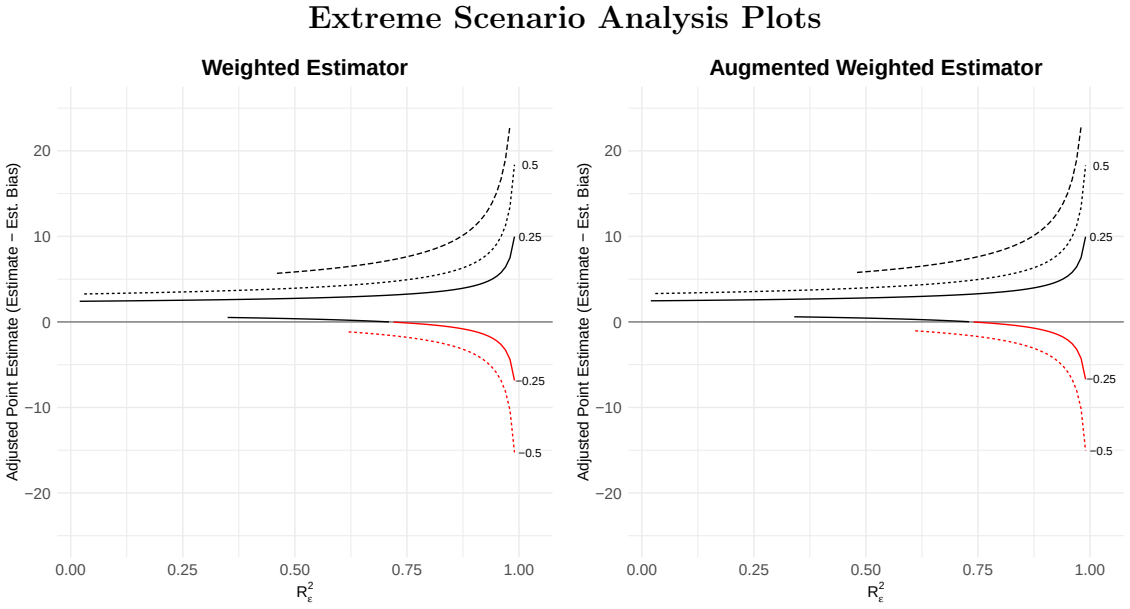

Figure A.3: We vary different values of $\text{cor}_{\mathcal{S}}(w_i^*, \tau_i)$ (and $\text{cor}(w_i^*, \xi_i)$). We set the x-axis to be different R^2_ε values, and the y-axis to be the adjusted point estimate (i.e., the point estimate minus the estimated bias). The lines marked by red represent results that would alter the interpretation of the point estimate.

Conducting the Sensitivity Analysis Across the Other Experimental Sites

In the main text, we conducted the sensitivity analysis across the experimental site of Coosa Valley, Georgia to illustrate the different sensitivity tools proposed in the paper. We now illustrate the sensitivity analysis across all 16 experimental sites. We provide a summary of the sensitivity statistics, as well as the benchmark PATE in Table B.5.

Site	n	N	Target PATE	Within-Site Estimate	Weighted Estimate	SE	$\hat{\sigma}^2_{\tau,\max}$	$RV_{q=1}$	$\widehat{\text{cor}}_{\mathcal{S}}(w_i, \tau_i)$
NE	636	5466	1.25	1.11	1.37	1.37	8.40	0.45	0.10
LC	485	5617	1.21	1.61	0.36	1.62	53.76	0.06	-0.23
HF	234	5868	1.28	0.95	1.81	1.99	27.87	0.43	0.25
IN	1392	4710	1.10	1.73	1.63	0.89	10.45	0.54	-0.06
CV	788	5314	1.18	1.63	2.81	1.21	29.01	0.56	0.37
CC	524	5578	1.37	-0.21	0.54	2.36	5.02	0.15	0.24
JK	353	5749	1.19	2.16	2.95	1.77	9.41	0.53	0.21
MT	38	6064	1.27	-5.21	-11.88	4.83	248.52	0.52	-0.49
PR	463	5639	1.12	3.03	3.28	2.23	54.81	0.38	0.04
MN	179	5923	1.32	-1.43	-1.69	3.06	391.43	0.09	-0.01
MD	177	5925	1.23	1.24	0.66	2.55	21.07	0.21	-0.19
SM	401	5701	1.29	0.60	0.92	1.63	299.61	0.06	0.02
OH	74	6028	1.30	-2.99	-4.35	2.69	216.54	0.37	-0.14
CI	190	5912	1.24	1.35	0.34	3.25	57.99	0.05	-0.18
OK	87	6015	1.24	1.83	5.09	5.04	83.45	0.34	0.28
JC	81	6021	1.27	-0.53	-7.45	5.94	333.44	0.26	-0.26

Table A.6: Sensitivity Statistics Across JTPA Experimental Sites

We see that several factors drive the sensitivity. First, the estimated effect size affects how much robustness is reflected. Experimental sites like Jackson, Missouri (JK), or Marion, Ohio (OH) have relatively large estimated effects, which mean that the amount of bias necessary to reduce the estimated effect to zero or change sign must be larger. This is reflected in the larger robustness values. Second, we see that the bound on the variation in the individual-level treatment effect also affects the overall robustness. In sites like Northwest, Minnesota (MN), the maximum σ^2_τ value estimated (i.e., $\hat{\sigma}^2_{\tau,\max} = 391.43$) is much larger than that of the maximum σ^2_τ value in Fortwayne, Indiana (IN) (i.e., $\hat{\sigma}^2_{\tau,\max} = 10.45$). Thus, despite the estimated effects being roughly the same in magnitude, there is much more robustness in the site of IN than in MN, as there is less potential for the confounding in selection to be correlated to the treatment effect heterogeneity.

We also examine the benchmarking results across the different experimental sites. We compare the actual error in estimation with the benchmarked parameter values for confounders like whether or not an individual is black or Hispanic, as well as previous earnings.[6] In general, more imbalance from omitting a confounder like these covariates corresponds to more greater error in recovering the target PATE. We note that there is a large degree of variation across the different experimental sites from the resulting parameter values that would occur from omitting a variable like the observed covariates. In particular, the benchmarked parameter values correspond to how much inherent imbalance there is between the experimental sample and the target population there is. As such, in certain sites, the experimental sample is much more representative of the target population than others, and as such, the benchmarked parameter values will be lower.

While we see that there is an association between the benchmarking results and robustness values with the error in recovering the target PATE, this by no means implies that a small robustness value (or large MRCS value) is *indicative* of an omitted variable. The plausibility of omitting a confounder that would explain the minimum variation determined by the robustness value, or a confounder with specified MRCS value, still depends on substantive justification. We once again re-iterate that we should not, and cannot, use naive cutoff values for the sensitivity statistics and the benchmarking results to determine whether or not an estimated effect is robust or not.

Finally, we note that many of the estimated effects across the experimental site were statistically insignificant. This is consistent with the generalizability literature, in which re-weighting data often results in a loss in precision (Miratrix et al., 2018). Researchers can, when possible, utilize model-assisted approaches like post-residualized weighting, to obtain more precise estimates (Huang et al., 2021). An advantage to the proposed sensitivity framework is that it can be easily extended for these alternative estimation approaches.

[6]These variables are chosen because substantively, we expect race and previous earnings to explain much of the variation in both the individual-level treatment effect, as well as the selection process.

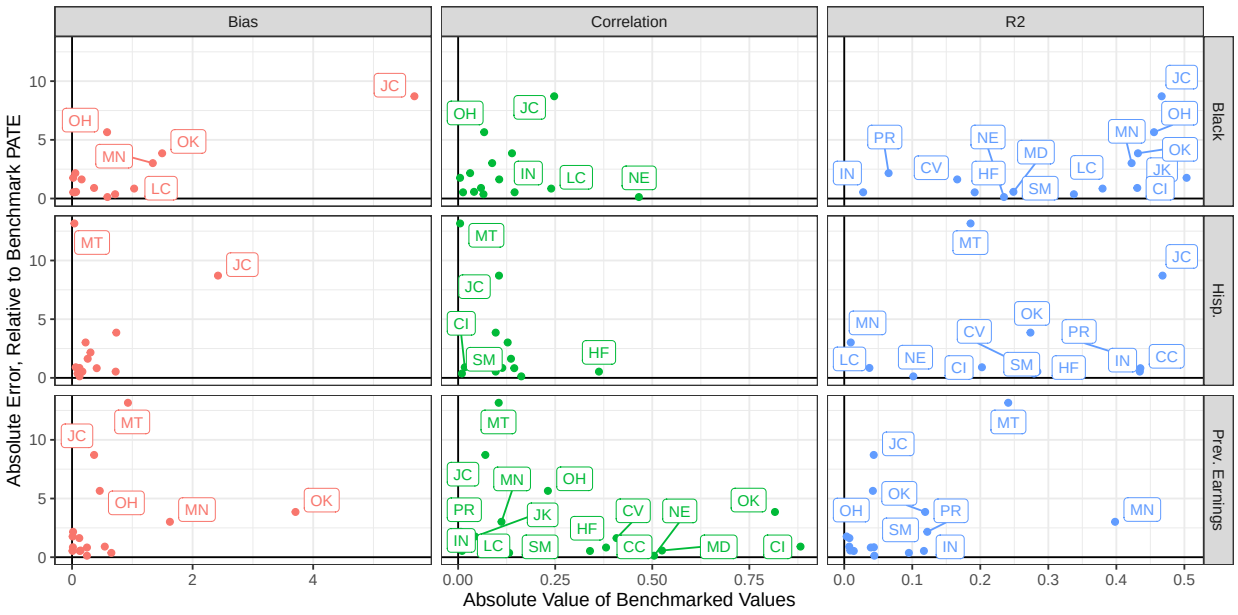

Figure A.4: We compare the absolute error in recovering the target PATE with three key quantities from benchmarking–(1) the estimated bias, (2) $\hat{\rho}_{\varepsilon,\tau}$, and (3) $\hat{R}^2_{\varepsilon}$. We see that in general, greater sensitivity that was reflected from the benchmarking results was consistent with larger amounts of error in recovering the target PATE.

Appendix B

Leveraging Population Outcomes to Improve the Generalization of Experimental Results

B.1 Proofs and Derivations

Derivation of Variance Terms

Consider a countably infinite population of $(\mathbf{X}_i, Y_i(t)) \sim F$, where $t \in \{0, 1\}$, with density $dF(\mathbf{X}_i, Y_i(t))$. This is our target population. We define the sampling distribution for the experimental data to be $(\mathbf{X}_i, Y_i(t)) \sim \tilde{F}$ with density $d\tilde{F}(\mathbf{X}_i, Y_i(t))$. Because we consider settings where the selection into the experiment from the target population is biased, $F \neq \tilde{F}$. Let $\mathcal{S}$ be the set of all indices for all units sampled in the experimental sample. As we can consider the treatment and control groups to be independent samples from an infinite population, we will focus below on one potential outcome $Y_i(t)$.

We defined a relative density in equation (6) as follows.

$$\pi(\mathbf{X}_i) = \frac{d\tilde{F}(\mathbf{X}_i)}{dF(\mathbf{X}_i)}.$$

over the support of F, where $dF(\mathbf{X}_i) > 0$. The $\pi(\mathbf{X}_i)$ is our infinite analog to the sampling propensity score. It scales our distribution. We further assume that $\pi(\mathbf{X}_i) > 0$ (this is an overlap assumption, saying our realized sampling distribution is not missing parts of the underlying distribution). $\pi(\mathbf{X}_i)$ captures the relative density of our realized distribution to the real population. Smaller $\pi(\mathbf{X}_i)$ correspond to areas where there is a lot more in the target population than in our sample. Larger $\pi(\mathbf{X}_i)$ are where we are over-sampling.

We assume known weights for any unit, dependent on $\mathbf{X}_i$, with $w_i = \kappa/\pi(\mathbf{X}_i)$ (the κ is a fixed constant allowing our weights to be normalized on some arbitrary scale).

For the remainder of the Supplementary Materials, the distribution over which a quantity is computed will be denoted by subscript. For example, the expectation over the realized sampling distribution will be written as $\mathbb{E}_{\tilde{F}}(\cdot)$, while the expectation over the target population will be written as $\mathbb{E}_F(\cdot)$.

Lemma B.1.1 (Variance of a Hájek estimator) *Define $\hat{\mu}$ as a Hájek estimator:*

$$\hat{\mu}_t = \frac{\sum_{i \in \mathcal{S}} w_i Y_i(t)}{\sum_{i \in \mathcal{S}} w_i},$$

where consistent with before, $w_i = \kappa/\pi(\mathbf{X}_i)$, and $(\mathbf{X}_i, Y_i(t)) \sim \tilde{F}$. The approximate asymptotic variance of a Hájek estimator is:

$$asyvar_{\tilde{F}}(\hat{\mu}_t) \approx \int \frac{1}{\pi(\mathbf{X}_i)^2}(Y_i(t) - \mu_t)^2 d\tilde{F}(\mathbf{X}_i, Y_i(t)),$$

where the asymptotic variance is being taken with respect to the realized sampling distribution, and $\mu_t = \mathbb{E}_F(Y_i(t))$ (i.e., the expected value of $Y_i(t)$ over the target population).

Proof: To begin, we write the Hájek estimator as a ratio estimator of the following form:

$$\begin{aligned}\hat{\mu}_t &= \frac{\sum_{i \in \mathcal{S}} w_i Y_i(t)}{\sum_{i \in \mathcal{S}} w_i} \\ &= \frac{\frac{1}{n}\sum_{i \in \mathcal{S}} w_i Y_i(t)}{\frac{1}{n}\sum_{i \in \mathcal{S}} w_i}\end{aligned}$$

where we define n to be the sample size, i.e., $n = |\mathcal{S}|$.

We then define $\hat{A} = \frac{1}{n}\sum_{i \in \mathcal{S}} w_i Y_i(t)$ and $\hat{B} = \frac{1}{n}\sum_{i \in \mathcal{S}} w_i$ for notational simplicity. If we define $A = \mathbb{E}_{\tilde{F}}(\hat{A})$, $A = \kappa\mu_t$. Similarly, if we define $B = \mathbb{E}_{\tilde{F}}(\hat{B})$, $B = \kappa$.
To derive the variance expression, we will use the delta method below for a ratio, i.e., a function $h(a, b) = a/b$. For this ratio, we have

$$\frac{d}{da}h(a,b) = \frac{1}{b} \qquad \frac{d}{db}h(a,b) = -\frac{a}{b^2}.$$

Therefore, using the Delta Method for a ratio,

$$\begin{aligned}
\hat{\mu}_t &= \frac{\frac{1}{n}\sum_{i\in\mathcal{S}} w_i Y_i(t)}{\frac{1}{n}\sum_{i\in\mathcal{S}} w_i} \\
&= \frac{\hat{A}}{\hat{B}} \\
&\approx \frac{A}{B} + \frac{1}{B}(\hat{A} - A) - \frac{A}{B^2}(\hat{B} - B) \\
&= \frac{A}{B} - \frac{A}{B} + \frac{A}{B} + \frac{1}{B}\hat{A} - \frac{A}{B^2}\hat{B} \\
&= \mu_t + \frac{1}{\kappa}\frac{1}{n}\sum_{i\in\mathcal{S}} w_i Y_i(t) - \frac{\mu_t}{\kappa}\frac{1}{n}\sum_{i\in\mathcal{S}} w_i \\
&= \mu_t + \frac{1}{n\kappa}\sum_{i\in\mathcal{S}} w_i (Y_i(t) - \mu_t)
\end{aligned}$$

where the first and second equalities follow from the definition of $\hat{\mu}_t$ and $(\hat{A}, \hat{B})$, the third from the delta method, the fourth from simple algebra, the fifth from the definition of (A, B), and the sixth from re-arrangement of the terms.
Finally,

$$\begin{aligned}
\mathrm{var}_{\tilde{F}}\left(\hat{\mu}_t\right) &= \mathrm{var}_{\tilde{F}}\left(\hat{\mu}_t - \mu_t\right) && \text{(B.1)} \\
&\approx \frac{1}{n^2\kappa^2} \cdot \mathrm{var}_{\tilde{F}}\left(\sum_{i\in\mathcal{S}} w_i (Y_i(t) - \mu_t)\right) \\
&= \frac{1}{n^2\kappa^2} n \int \frac{\kappa^2}{\pi(\mathbf{X}_i)^2}(Y_i(t) - \mu_t)^2 d\tilde{F}(\mathbf{X}_i, Y_i(t)) \\
&= \frac{1}{n}\int \frac{1}{\pi(\mathbf{X}_i)^2}(Y_i(t) - \mu_t)^2 d\tilde{F}(\mathbf{X}_i, Y_i(t)) && \text{(B.2)}
\end{aligned}$$

As such, $\mathrm{asyvar}_{\tilde{F}}\left(\hat{\mu}_t\right) = \lim_{n\to\infty} \mathrm{var}(\sqrt{n}\hat{\mu}_t) = \int \frac{1}{\pi(\mathbf{X}_i)^2}(Y_i(t) - \mu_t)^2 d\tilde{F}(\mathbf{X}_i, Y_i(t))$. □

Lemma B.1.2 (Weighted Variance) *Define the weighted variance and the weighted covariance as:*

$$var_w(A_i) = \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i - \bar{A})^2 d\tilde{F}(\mathbf{X}_i, A_i)$$

$$cov_w(A_i, B_i) = \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i - \bar{A})(B_i - \bar{B}) d\tilde{F}(\mathbf{X}_i, A_i, B_i)$$

Under this definition, common variance and covariance properties apply:

$$var_w(A_i + B_i) = var_w(A_i) + var_w(B_i) + 2cov_w(A_i, B_i)$$

$$cov_w(A_i + B_i, C_i) = cov_w(A_i, C_i) + cov_w(B_i, C_i)$$

Proof:

$$\begin{aligned}
\mathrm{var}_w(A_i+B_i) &= \int \frac{1}{\pi(\mathbf{X}_i)^2}\left(A_i+B_i-(\bar{A}+\bar{B})\right)^2 d\tilde{F}(\mathbf{X}_i,A_i,B_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}\left((A_i-\bar{A})^2+(B_i-\bar{B})^2+2(A_i-\bar{A})(B_i-\bar{B})\right) d\tilde{F}(\mathbf{X}_i,A_i,B_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i-\bar{A})^2 d\tilde{F}(\mathbf{X}_i,A_i,B_i) + \int \frac{1}{\pi(\mathbf{X}_i)^2}(B_i-\bar{B})^2 d\tilde{F}(\mathbf{X}_i,A_i,B_i) + \\
&\quad 2\int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i-\bar{A})(B_i-\bar{B}) d\tilde{F}(\mathbf{X}_i,A_i,B_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i-\bar{A})^2 d\tilde{F}(\mathbf{X}_i,A_i) + \int \frac{1}{\pi(\mathbf{X}_i)^2}(B_i-\bar{B})^2 d\tilde{F}(\mathbf{X}_i,B_i) + \\
&\quad 2\int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i-\bar{A})(B_i-\bar{B}) d\tilde{F}(\mathbf{X}_i,A_i,B_i) \\
&= \mathrm{var}_w(A_i)+\mathrm{var}_w(B_i)+2\mathrm{cov}_w(A_i,B_i)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{cov}_w(A_i+B_i,C_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}\left(A_i+B_i-(\bar{A}+\bar{B})\right)\left(C_i-\bar{C}\right) d\tilde{F}(\mathbf{X}_i,A_i,B_i,C_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}\left((A_i-\bar{A})(B_i-\bar{B})\right)\left(C_i-\bar{C}\right) d\tilde{F}(\mathbf{X}_i,A_i,B_i,C_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}\left((A_i-\bar{A})(C_i-\bar{C})+(B_i-\bar{B})(C_i-\bar{C})\right) d\tilde{F}(\mathbf{X}_i,A_i,B_i,C_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i-\bar{A})(C_i-\bar{C}) d\tilde{F}(\mathbf{X}_i,A_i,B_i,C_i) + \\
&\quad \int \frac{1}{\pi(\mathbf{X}_i)^2}(B_i-\bar{B})(C_i-\bar{C}) d\tilde{F}(\mathbf{X}_i,A_i,B_i,C_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i-\bar{A})(C_i-\bar{C}) d\tilde{F}(\mathbf{X}_i,A_i,C_i) + \int \frac{1}{\pi(\mathbf{X}_i)^2}(B_i-\bar{B})(C_i-\bar{C}) d\tilde{F}(\mathbf{X}_i,B_i,C_i) \\
&= \mathrm{cov}_w(A_i,C_i)+\mathrm{cov}_w(B_i,C_i)
\end{aligned}$$

□

Lemma B.1.3 (Asymptotic Variance of a Weighted Estimator)
The asymptotic variance of a Hájek-style weighted estimator is:

$$\begin{aligned}
&asyvar_{\tilde{F}}(\hat{\tau}_W) \\
&\quad = asyvar_{\tilde{F}}(\hat{\mu}_1) + asyvar_{\tilde{F}}(\hat{\mu}_0) \\
&\quad \approx \frac{1}{p}\int \frac{1}{\pi(\mathbf{X}_i)^2}(Y_i(1)-\mu_1)^2 d\tilde{F}(\mathbf{X}_i, Y_i(1)) + \frac{1}{1-p}\int \frac{1}{\pi(\mathbf{X}_i)^2}(Y_i(0)-\mu_0)^2 d\tilde{F}(\mathbf{X}_i, Y_i(0)) \\
&\quad = \frac{1}{p} var_w(Y_i(1)) + \frac{1}{1-p} var_w(Y_i(0)),
\end{aligned}$$

where $var_w(\cdot)$ is defined in equation (9). p is the probability of treatment assignment, i.e., $p = \Pr_{\tilde{F}}(T_i = 1)$. $\mu_1 = \mathbb{E}_F(Y_i(1))$ and $\mu_0 = \mathbb{E}_F(Y_i(0))$.

Proof: Because we are sampling from an infinite super-population, the treatment and control groups can be treated as two separate samples from the infinite super-population. We directly apply Lemma B.1.1 to arrive at the final result.

□

Lemma B.1.4 (Asymptotic Variance of Weighted Least Squares Estimator)
The asymptotic variance of a weighted least squares estimator is:

$$asyvar(\hat{\tau}_{wLS}) = \frac{1}{p} var_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) + \frac{1}{1-p} var_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*),$$

where γ_* is the vector of true coefficients associated with the pretreatment covariates $\tilde{\mathbf{X}}_i$ defined as:

$$(\tau_{wLS}, \alpha_*, \gamma_*) = \underset{\tau,\alpha,\gamma}{\operatorname{argmin}} \mathbb{E}_{\tilde{F}} \left\{ \hat{w}_i \left(Y_i - (\tau T_i + \alpha + \widetilde{\mathbf{X}}_i^\top \gamma) \right)^2 \right\} \tag{B.3}$$

Proof: To begin, analogous with Lin (2013) (Lemma 6), the weighted least squares estimator can be written as:

$$\hat{\tau}_{wLS} = \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} w_i T_i (Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) - \frac{1}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \sum_{i\in\mathcal{S}} w_i (1-T_i)(Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) \tag{B.4}$$

Akin with Ding (2021), we define δ_X as:

$$\delta_X = \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} w_i T_i \tilde{\mathbf{X}}_i^\top - \frac{1}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \sum_{i\in\mathcal{S}} w_i (1-T_i) \tilde{\mathbf{X}}_i^\top$$

δ_X represents any residual imbalance between the treatment and control groups in the weighted pre-treatment covariates. We can re-write Equation (B.4) as:

$$
\begin{aligned}
\hat{\tau}_{wLS} =& \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} w_i T_i (Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) - \frac{1}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \sum_{i\in\mathcal{S}} w_i (1-T_i)(Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) \\
=& \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} w_i T_i (Y_i(1) - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) - \frac{1}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \sum_{i\in\mathcal{S}} w_i (1-T_i)(Y_i(0) - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) \\
=& \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} w_i T_i (Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_* + \tilde{\mathbf{X}}_i^\top \gamma_* - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) - \\
& \frac{1}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \sum_{i\in\mathcal{S}} w_i (1-T_i)(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_* + \tilde{\mathbf{X}}_i^\top \gamma_* - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) \\
=& \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} \Big(w_i T_i (Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) + w_i T_i \tilde{\mathbf{X}}_i^\top (\gamma_* - \hat{\gamma}) \Big) - \\
& \frac{1}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \sum_{i\in\mathcal{S}} \Big(w_i (1-T_i)(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) + w_i (1-T_i) \tilde{\mathbf{X}}_i^\top (\gamma_* - \hat{\gamma}) \Big) \\
=& \underbrace{\frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} w_i T_i (Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \frac{1}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \sum_{i\in\mathcal{S}} w_i (1-T_i)(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*)}_{:=\hat{\tau}^*_{wLS}} + \\
& \underbrace{\frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} w_i T_i \tilde{\mathbf{X}}_i^\top (\gamma_* - \hat{\gamma}) - \frac{1}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \sum_{i\in\mathcal{S}} w_i (1-T_i) \tilde{\mathbf{X}}_i^\top (\gamma_* - \hat{\gamma})}_{=\delta_X(\gamma_* - \hat{\gamma})} \\
=& \hat{\tau}^*_{wLS} + \delta_X (\gamma_* - \hat{\gamma}),
\end{aligned}
$$

where $\hat{\tau}^*_{wLS}$ represents the potential outcomes, adjusted for the pre-treatment covariates using the *true* coefficients γ_*.

Under standard regularity conditions for least squares, $\gamma_* - \hat{\gamma} = o_p(1)$ (White, 1982). Furthermore, $\sqrt{n}\delta_X = O_p(1)$:

$$
\begin{aligned}
\lim_{n\to\infty} \mathrm{var}_{\bar{F}}(\delta_X) &= \lim_{n\to\infty} \left(\frac{1}{n_1} \mathrm{var}_w(\tilde{\mathbf{X}}_i) + \frac{1}{n_0} \mathrm{var}_w(\tilde{\mathbf{X}}_i) \right) \\
&= \lim_{n\to\infty} \frac{1}{n} \cdot \left(\frac{1}{p} + \frac{1}{1-p} \right) \mathrm{var}_w(\tilde{\mathbf{X}}_i) \\
&= \lim_{n\to\infty} \frac{1}{n} \cdot \frac{1}{p(1-p)} \mathrm{var}_w(\tilde{\mathbf{X}}_i)
\end{aligned}
$$

Assuming $\mathrm{var}_w(\tilde{\mathbf{X}}_i)$ is finite, $\delta_X = O_p(\sqrt{n}^{-1}) \implies \sqrt{n}\delta_X = O_p(1)$.

Therefore, as $n \to \infty$:

$$\sqrt{n}(\hat{\tau}_{wLS} - \tau) = \sqrt{n}(\hat{\tau}^*_{wLS} - \tau) + \underbrace{\sqrt{n}\delta_X(\gamma_* - \hat{\gamma})}_{\xrightarrow{p} 0}$$
$$\xrightarrow{d} N(0, \text{var}(\hat{\tau}^*_{wLS})),$$

where $\text{var}_{\tilde{F}}(\hat{\tau}^*_{wLS}) \approx \frac{1}{p}\text{var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) + \frac{1}{1-p}\text{var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*)$ (this result follows from applying Lemma 1 on the adjusted potential outcomes).

□

Proof of Theorem 1

Suppose Assumption 2 holds with $\mathbf{X}_i$, the Post-Residualized Weighted Least Squares Estimator is a consistent estimator for the PATE:

$$\hat{\tau}^{res}_{wLS} \xrightarrow{p} \tau$$

Proof: To begin, we can write $\hat{\tau}^{res}_{wLS}$ as the above estimator on the residuals of the initial population regression:

$$\begin{aligned}\hat{\tau}^{res}_{wLS} =& \frac{1}{(\sum_{i\in\mathcal{S}} w_i T_i)}\left(\sum_{i\in\mathcal{S}} w_i T_i(\hat{e}_i - \mathbf{X}_i\hat{\gamma}^{res})\right) - \\ & \left(\frac{1}{(\sum_{i\in\mathcal{S}} w_i(1-T_i))}\sum_{i\in\mathcal{S}} w_i(1-T_i)(\hat{e}_i - \mathbf{X}_i\hat{\gamma}^{res})\right) \\ =& \underbrace{\frac{\sum_{i\in\mathcal{S}} w_i T_i \hat{e}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)\hat{e}_i}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}}_{=\hat{\tau}^{res}_W} - \underbrace{\left(\frac{\sum_{i\in\mathcal{S}} w_i T_i \mathbf{X}_i\hat{\gamma}^{res}}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)\mathbf{X}_i\hat{\gamma}^{res}}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}\right)}_{(*)},\end{aligned}$$

where $\hat{\gamma}^{res}$ represents the estimated coefficients for the covariates $\mathbf{X}_i$ in the weighted regression run on the residualized outcomes $\hat{e}_i$. Note that the above represents two distinct regression steps: $\hat{e}_i$ is the result of the first population regression. $\hat{\gamma}^{res}$ is estimated for the covariates $\mathbf{X}_i$ from the second regression using the residualized sample outcomes, $\hat{e}_i$.

We begin by showing that $\hat{\tau}^{res}_W \xrightarrow{p} \tau$. We will begin by the proof by showing that $\hat{\tau}^{res}_W$ can be written as the difference between $\hat{\tau}_W$, and a weighted estimator computed over the fitted values $\hat{Y}_i$, which we will define as $\hat{\tau}_{\hat{Y}}$. Following the generalization literature, we treat the

weights as known, as well as the observed sampled population:

$$\begin{aligned}
\hat{\tau}_W^{res} =& \frac{\sum_{i\in\mathcal{S}} w_i T_i \cdot \hat{e}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \cdot \hat{e}_i}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \\
=& \frac{\sum_{i\in\mathcal{S}} w_i T_i \cdot (Y_i - \hat{Y}_i)}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \cdot (Y_i - \hat{Y}_i)}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \\
=& \underbrace{\frac{\sum_{i\in\mathcal{S}} w_i T_i \cdot Y_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \cdot Y_i}{\sum_{i\in\mathcal{S}} w_i (1-T_i)}}_{=\hat{\tau}_W} - \\
& \underbrace{\left(\frac{\sum_{i\in\mathcal{S}} w_i T_i \cdot \hat{Y}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \cdot \hat{Y}_i}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \right)}_{=\hat{\tau}_{\hat{Y}}} \\
=& \hat{\tau}_W - \hat{\tau}_{\hat{Y}}
\end{aligned}$$

We will begin by showing that $\hat{\tau}_W \xrightarrow{p} \tau$. To begin:

$$\hat{\tau}_W = \frac{\sum_{i\in\mathcal{S}} w_i T_i \cdot Y_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \cdot Y_i}{\sum_{i\in\mathcal{S}} w_i (1-T_i)}$$

By Law of Large Numbers and the Continuous Mapping Theorem:

$$\hat{\tau}_W \xrightarrow{p} \underbrace{\frac{\mathbb{E}_{\tilde{F}}(w_i T_i Y_i)}{\mathbb{E}_{\tilde{F}}(w_i T_i)}}_{(1)} - \underbrace{\frac{\mathbb{E}_{\tilde{F}}(w_i (1-T_i) Y_i)}{\mathbb{E}_{\tilde{F}}(w_i (1-T_i))}}_{(2)}$$

We will now show that the first term (i.e., (1)) is equal to $\mathbb{E}_F(Y_i(1))$. We first evaluate the expectation in the denominator.

$$\begin{aligned}
\mathbb{E}_{\tilde{F}}(w_i T_i) &= \frac{n_1}{n} \mathbb{E}_{\tilde{F}}(w_i) \\
&= \frac{n_1}{n} \mathbb{E}_{\tilde{F}} \left(\frac{\kappa}{\pi(\mathbf{X}_i)} \right) \\
&= \frac{n_1}{n} \cdot \kappa \int \frac{1}{\pi(\mathbf{X}_i)} d\tilde{F}(\mathbf{X}_i) \\
&= \frac{n_1}{n} \cdot \kappa \underbrace{\int \frac{1}{\pi(\mathbf{X}_i)} \pi(\mathbf{X}_i) dF(\mathbf{X}_i)}_{=1} \\
&= \frac{n_1}{n} \cdot \kappa
\end{aligned}$$

For the numerator:

$$\begin{aligned}
\mathbb{E}_{\tilde{F}}(w_i T_i Y_i) &= \mathbb{E}_{\tilde{F}}(w_i T_i Y_i(1)) \\
&= \frac{n_1}{n}\mathbb{E}_{\tilde{F}}(w_i Y_i(1)) \\
&= \frac{n_1}{n}\mathbb{E}_{\tilde{F}}\left(\frac{\kappa}{\pi(\mathbf{X}_i)}Y_i(1)\right) \\
&= \frac{n_1}{n}\cdot\kappa\mathbb{E}_{\tilde{F}}\left(\frac{1}{\pi(\mathbf{X}_i)}Y_i(1)\right) \\
&= \frac{n_1}{n}\cdot\kappa\int\frac{Y_i(1)}{\pi(\mathbf{X}_i)}d\tilde{F}(\mathbf{X}_i, Y_i(1)) \\
&= \frac{n_1}{n}\cdot\kappa\int\frac{Y_i(1)}{\pi(\mathbf{X}_i)}\cdot\pi(\mathbf{X}_i)dF(\mathbf{X}_i, Y_i(1)) \\
&= \frac{n_1}{n}\cdot\kappa\int Y_i(1)dF(\mathbf{X}_i, Y_i(1)) \\
&= \frac{n_1}{n}\kappa\cdot\mathbb{E}_F(Y_i(1))
\end{aligned}$$

Therefore, re-writing (1):

$$\begin{aligned}
\frac{\mathbb{E}_{\tilde{F}}(w_i T_i Y_i)}{\mathbb{E}_{\tilde{F}}(w_i T_i)} &= \frac{p\kappa\cdot\mathbb{E}_F(Y_i(1))}{p\cdot\kappa} \\
&= \mathbb{E}_F(Y_i(1))
\end{aligned}$$

Similarly, we can show that the second term, $\mathbb{E}_{\tilde{F}}(w_i(1-T_i)Y_i)/\mathbb{E}_{\tilde{F}}(w_i(1-T_i))$, is equal to $\mathbb{E}_F(Y_i(0))$. Therefore:

$$\begin{aligned}
\mathbb{E}_{\tilde{F}}(\hat{\tau}_W) &\xrightarrow{p} \mathbb{E}_F(Y_i(1)) - \mathbb{E}_F(Y_i(0)) \\
&= \tau
\end{aligned}$$

Now we will show that $\hat{\tau}_{\hat{Y}} \xrightarrow{p} 0$. Once again, applying Law of Large Numbers and the Continuous Mapping Theorem:

$$\begin{aligned}
\hat{\tau}_{\hat{Y}} &= \frac{\sum_{i\in\mathcal{S}} w_i T_i \hat{Y}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i)\hat{Y}_i}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \\
&\xrightarrow{p} \frac{\mathbb{E}_{\tilde{F}}(w_i T_i \hat{Y}_i)}{\mathbb{E}_{\tilde{F}}(w_i T_i)} - \frac{\mathbb{E}_{\tilde{F}}(w_i (1-T_i)\hat{Y}_i)}{\mathbb{E}_{\tilde{F}}(w_i (1-T_i))} \\
&= \frac{p\cdot\mathbb{E}_{\tilde{F}}(w_i\hat{Y}_i)}{p\mathbb{E}_{\tilde{F}}(w_i)} - \frac{(1-p)\cdot\mathbb{E}_{\tilde{F}}(w_i\hat{Y}_i)}{(1-p)\mathbb{E}_{\tilde{F}}(w_i)} \\
&= \frac{\mathbb{E}_{\tilde{F}}(w_i\hat{Y}_i)}{\mathbb{E}_{\tilde{F}}(w_i)} - \frac{\mathbb{E}_{\tilde{F}}(w_i\hat{Y}_i)}{\mathbb{E}_{\tilde{F}}(w_i)} \\
&= 0
\end{aligned}$$

where the third line follows from the fact that treatment assignment is randomized and independent of weights. Therefore, by the Continuous Mapping Theorem, $\hat{\tau}_W^{res} \xrightarrow{p} \tau$.

Now looking just at the $(*)$ term:

$$\frac{\sum_{i\in\mathcal{S}} w_i T_i \mathbf{X}_i \hat{\gamma}^{res}}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \mathbf{X}_i \hat{\gamma}^{res}}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} = \left(\frac{\sum_{i\in\mathcal{S}} w_i T_i \mathbf{X}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \mathbf{X}_i}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \right) \hat{\gamma}^{res}$$

Under standard regularity conditions for least squares, $\hat{\gamma}^{res}$ converges to γ_*^{res}. Furthermore, using Law of Large Numbers and the Continuous Mapping Theorem:

$$\begin{aligned}
\frac{\sum_{i\in\mathcal{S}} w_i T_i \mathbf{X}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \mathbf{X}_i}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} &\xrightarrow{p} \frac{\mathbb{E}_{\tilde{F}}(w_i T_i \mathbf{X}_i)}{\mathbb{E}_{\tilde{F}}(w_i T_i)} - \frac{\mathbb{E}_{\tilde{F}}(w_i (1-T_i) \mathbf{X}_i)}{\mathbb{E}_{\tilde{F}}(w_i (1-T_i))} \\
&= \frac{\mathbb{E}_{\tilde{F}}(w_i \mathbf{X}_i)}{\mathbb{E}_{\tilde{F}}(w_i)} - \frac{\mathbb{E}_{\tilde{F}}(w_i \mathbf{X}_i)}{\mathbb{E}_{\tilde{F}}(w_i)} \\
&= 0
\end{aligned}$$

As such, we see that the term in $(*)$ will converge in probability to zero. Therefore, $\hat{\tau}_{wLS}^{res} \xrightarrow{p} \tau$. □

Proof of Theorem 2

The difference between the asymptotic variance of $\hat{\tau}_W^{res}$ and the asymptotic variance of $\hat{\tau}_W$ is:

$$\begin{aligned}
&\text{asyvar}_{\tilde{F}}(\hat{\tau}_W) - \text{asyvar}_{\tilde{F}}(\hat{\tau}_W^{res}) \\
&\qquad = -\frac{1}{p(1-p)} \text{var}_w(\hat{Y}_i) + \frac{2}{p} \text{cov}_w(Y_i(1), \hat{Y}_i) + \frac{2}{1-p} \text{cov}_w(Y_i(0), \hat{Y}_i),
\end{aligned}$$

Proof: From Lemma B.1, the asymptotic variance of a weighted estimator is:

$$\text{asyvar}_{\tilde{F}}(\hat{\tau}_W) = \frac{1}{p} \text{var}_w(Y_i(1)) + \frac{1}{1-p} \text{var}_w(Y_i(0))$$

Using the residualized potential outcomes $\hat{e}_i(1)$ and $\hat{e}_i(0)$, the asymptotic variance of a weighted residualized estimator is:

$$\text{asyvar}_{\tilde{F}}(\hat{\tau}_W^{res}) = \frac{1}{p} \text{var}_w(\hat{e}_i(1)) + \frac{1}{1-p} \text{var}_w(\hat{e}_i(0)).$$

From the definition of potential residuals, we can write the potential residuals as a function of the original outcome values and the fitted values:

$$\begin{aligned}\text{var}_w(\hat{e}_i(0)) &= \text{var}_w(Y_i(0) - \hat{Y}_i) \\ &= \text{var}_w(Y_i(0)) + \text{var}_w(\hat{Y}_i) - 2\text{cov}_w(Y_i(0), \hat{Y}_i) \end{aligned} \tag{B.5}$$

$$\begin{aligned}\text{var}_w(\hat{e}_i(1)) &= \text{var}_w(Y_i(1) - \hat{Y}_i) \\ &= \text{var}_w(Y_i(1)) + \text{var}_w(\hat{Y}_i) - 2\text{cov}_w(Y_i(1), \hat{Y}_i) \end{aligned} \tag{B.6}$$

Therefore, the difference in variances of our two estimators is

$$\begin{aligned}&\text{asyvar}_{\tilde{F}}(\hat{\tau}_W) - \text{asyvar}_{\tilde{F}}(\hat{\tau}_W^{res}) \\ &\quad= \left\{\frac{1}{p}\text{var}_w(Y_i(1)) + \frac{1}{1-p}\text{var}_w(Y_i(0)))\right\} - \left\{\frac{1}{p}\text{var}_w(\hat{e}_i(1)) + \frac{1}{1-p}\frac{1}{n_0}\text{var}_w(\hat{e}_i(0))\right\} \\ &\quad= \frac{1}{p} \cdot \big(\text{var}_w(Y_i(1)) - \text{var}_w(\hat{e}_i(1))\big) + \frac{1}{1-p} \cdot \big(\text{var}_w(Y_i(0)) - \text{var}_w(\hat{e}_i(0)))\big)\end{aligned}$$

Plugging in (B.5) and (B.6):

$$\begin{aligned}&= -\frac{1}{p} \cdot \Big\{\text{var}_w(Y_i(1)) + \text{var}_w(\hat{Y}_i) - 2\text{cov}_w(Y_i(1), \hat{Y}_i) - \text{var}_w(Y_i(1)\Big\} \\ &\quad - \frac{1}{1-p} \cdot \Big\{\text{var}_w(Y_i(0)) + \text{var}_w(\hat{Y}_i) - 2\text{cov}_w(Y_i(0), \hat{Y}_i) - \text{var}_w(Y_i(0)\Big\} \\ &= -\frac{1}{p(1-p)} \cdot \text{var}_w(\hat{Y}_i) + \frac{2}{p} \cdot \text{cov}_w(Y_i(1), \hat{Y}_i) + \frac{2}{1-p} \cdot \text{cov}_w(Y_i(0), \hat{Y}_i)\end{aligned}$$

□

Proof of Theorem 3

The difference between the asymptotic variance of $\hat{\tau}_{wLS}$ and the asymptotic variance of $\hat{\tau}^{res}_{wLS}$ is:

$$\begin{aligned}
&\text{asyvar}_{\tilde{F}}(\hat{\tau}_{wLS}) - \text{asyvar}_{\tilde{F}}(\hat{\tau}^{res}_{wLS}) \\
&= \frac{1}{p}\left\{\text{var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \text{var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i))\right\} \\
&\quad + \frac{1}{1-p}\left\{\text{var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \text{var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i))\right\} \\
&\quad + \frac{2}{p}\text{cov}_w(\hat{e}_i(1), \tilde{\mathbf{X}}_i^\top \gamma^{res}_*) + \frac{2}{1-p}\text{cov}_w(\hat{e}_i(0), \tilde{\mathbf{X}}_i^\top \gamma^{res}_*) - \frac{1}{p(1-p)}\text{var}_w(\tilde{\mathbf{X}}_i^\top \gamma^{res}_*),
\end{aligned}$$

where γ_* and γ^{res}_* are the true coefficients associated with the pre-treatment covariates, $\widetilde{\mathbf{X}}_i$ defined in the weighted least squares regression (equation (13)) and the post-residualized weighted least squares regression (equation (14)), respectively. Formally, γ_* and γ^{res}_* are formally defined as the solution to the following optimization problems.

$$(\tau_{wLS}, \alpha_*, \gamma_*) = \underset{\tau,\alpha,\gamma}{\text{argmin}}\, \mathbb{E}_{\tilde{F}}\left\{\hat{w}_i\left(Y_i - (\tau T_i + \alpha + \widetilde{\mathbf{X}}_i^\top \gamma)\right)^2\right\} \tag{B.7}$$

$$(\tau^{res}_{wLS}, \alpha^{res}_*, \gamma^{res}_*) = \underset{\tau,\alpha,\gamma}{\text{argmin}}\, \mathbb{E}_{\tilde{F}}\left\{\hat{w}_i\left(\widehat{e}_i - (\tau T_i + \alpha + \widetilde{\mathbf{X}}_i^\top \gamma)\right)^2\right\} \tag{B.8}$$

Proof:

$$\begin{aligned}
&\text{asyvar}_{\tilde{F}}(\hat{\tau}_{wLS}) - \text{asyvar}_{\tilde{F}}(\hat{\tau}^{res}_{wLS}) &\text{(B.9)}\\
&= \left\{\frac{1}{p}\text{var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) + \frac{1}{1-p}\text{var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*)\right\} \\
&\quad - \left\{\frac{1}{p}\text{var}_w(\hat{e}_i(1) - \tilde{\mathbf{X}}_i^\top \gamma^{res}_*) + \frac{1}{1-p}\text{var}_w(\hat{e}_i(0) - \tilde{\mathbf{X}}_i^\top \gamma^{res}_*)\right\} &\text{(B.10)}
\end{aligned}$$

The adjusted residualized outcomes can be re-written as a function of the residualized outcomes and the fitted values from the regression. First, for the treatment outcomes:

$$\begin{aligned}
\text{var}_w(\hat{e}_i(1) - \tilde{\mathbf{X}}_i^\top \gamma^{res}_*) &= \text{var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i) - \tilde{\mathbf{X}}_i^\top \gamma^{res}_*) \\
&= \text{var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i)) + \text{var}_w(\tilde{\mathbf{X}}_i^\top \gamma^{res}_*) - 2\text{cov}_w(Y_i(1) - \hat{g}(\mathbf{X}_i), \tilde{\mathbf{X}}_i^\top \gamma^{res}_*)
\end{aligned}$$

Similarly,

$$\begin{aligned}\text{var}_w(\hat{e}_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) &= \text{var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) \\ &= \text{var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i)) + \text{var}_w(\tilde{\mathbf{X}}_i^\top \gamma_*^{res}) - 2\text{cov}_w(Y_i(0) - \hat{g}(\mathbf{X}_i), \tilde{\mathbf{X}}_i^\top \gamma_*^{res})\end{aligned}$$

Plugging into Equation (B.10):

$$\begin{aligned}&\text{asyvar}_{\tilde{F}}(\hat{\tau}_{wLS}) - \text{asyvar}_{\tilde{F}}(\hat{\tau}_{wLS}^{res}) \\ &= \frac{1}{p}\left\{\text{var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \text{var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i))\right\} \\ &\quad + \frac{1}{1-p}\left\{\text{var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \text{var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i))\right\} \\ &\quad - \left\{\frac{1}{p(1-p)}\text{var}_w(\tilde{\mathbf{X}}_i^\top \gamma_*^{res}) - \frac{2}{p}\text{cov}_w(Y_i(1) - \hat{g}(\mathbf{X}_i), \tilde{\mathbf{X}}_i^\top \gamma_*^{res})\right. \\ &\qquad \left. - \frac{2}{1-p}\text{cov}_w(Y_i(0) - \hat{g}(\mathbf{X}_i), \tilde{\mathbf{X}}_i^\top \gamma_*^{res})\right\} \\ &= \frac{1}{p}\left\{\text{var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \text{var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i))\right\} \\ &\quad + \frac{1}{1-p}\left\{\text{var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \text{var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i))\right\} \\ &\quad + \left\{-\frac{1}{p(1-p)}\text{var}_w(\tilde{\mathbf{X}}_i^\top \gamma_*^{res}) + \frac{2}{p}\text{cov}_w(\hat{e}_i(1), \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) + \frac{2}{1-p}\text{cov}_w(\hat{e}_i(0), \tilde{\mathbf{X}}_i^\top \gamma_*^{res})\right\}\end{aligned}$$

□

Proof of Corollary 1

The relative reduction in variance from residualizing is given by:

$$\text{Relative Reduction} := \frac{\text{asyvar}_{\tilde{F}}(\hat{\tau}_{wLS}) - \text{asyvar}_{\tilde{F}}(\hat{\tau}_{wLS}^{res})}{\text{asyvar}_{\tilde{F}}(\hat{\tau}_{wLS})} = R_0^2 - \frac{1}{1+f} \cdot \xi$$

Proof: Let $C_1 = 1/p$ and $C_0 = 1/1 - p$. Furthermore, let $\epsilon_i(1) := Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*$, $\epsilon_i(0) := Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*$, $\epsilon_i^{res}(1) := \hat{e}_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}$, $\epsilon_i^{res}(0) := \hat{e}_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}$. Then, we can write the variance of the weighted least squares estimator (i.e., Lemma B.1.4) as:

$$\text{var}_{\tilde{F}}(\hat{\tau}_{wLS}) = \frac{1}{p}\text{var}_w(\epsilon_i(1)) + \frac{1}{1-p}\text{var}_w(\epsilon_i(0)),$$

and similarly, the variance of the residualized weighted least squares estimator as:

$$\text{var}_{\tilde{F}}(\hat{\tau}_{wLS}^{res}) = \frac{1}{p}\text{var}_w(\epsilon_i^{res}(1)) + \frac{1}{1-p}\text{var}_w(\epsilon_i^{res}(0)),$$

Then, we may re-write the relative reduction as follows:

$$\begin{aligned}
&\frac{\text{asyvar}_{\tilde{F}}(\hat{\tau}_{wLS}) - \text{asyvar}_{\tilde{F}}(\hat{\tau}^{res}_{wLS})}{\text{asyvar}_{\tilde{F}}(\hat{\tau}_{wLS})} \\
&= \frac{C_1\text{var}_w(\epsilon_i(1)) + C_0\text{var}_w(\epsilon_i(0)) - (C_1\text{var}_w(\epsilon^{res}_i(1)) + C_0\text{var}_w(\epsilon^{res}_i(0)))}{C_1\text{var}_w(\epsilon_i(1)) + C_0\text{var}_w(\epsilon_i(0))} \\
&= \frac{C_1\text{var}_w(\epsilon_i(1)) - C_1\text{var}_w(\epsilon^{res}_i(1)) + C_0\text{var}_w(\epsilon_i(0)) - C_0\text{var}_w(\epsilon^{res}_i(0)))}{C_1\text{var}_w(\epsilon_i(1)) + C_0\text{var}_w(\epsilon_i(0))}
\end{aligned}$$

Dividing the numerator and denominator by $C_1 \cdot \text{var}(\epsilon_i(1))$, and defining f to be equal to $C_0\text{var}_w(\epsilon_i(0))/C_1\text{var}_w(\epsilon_i(1))$:

$$\begin{aligned}
&= \frac{1 - \text{var}_w(\epsilon^{res}_i(1))/\text{var}_w(\epsilon_i(1)) + f - f \cdot \text{var}_w(\epsilon^{res}_i(0))/\text{var}_w(\epsilon_i(0))}{1+f} \\
&= \frac{1}{1+f}\left(R^2_1 + fR^2_0\right)
\end{aligned}$$

Using the definition of $\xi = R^2_0 - R^2_1$:

$$\begin{aligned}
&= \frac{1}{1+f}\left(R^2_0 - \xi + fR^2_0\right) \\
&= R^2_0 - \frac{1}{1+f} \cdot \xi
\end{aligned}$$

□

B.2 Diagnostic Measure

We detail how to estimate the diagnostic measures in this section. To estimate the diagnostic for the post-residualized weighted estimator, we compute the estimated weighted variance of both the residuals and the outcomes for the units assigned to control:

$$\begin{aligned}
\hat{R}^2_0 &= 1 - \frac{\widehat{\text{var}}_{w,0}(\hat{e}_i)}{\widehat{\text{var}}_{w,0}(Y_i)} \\
&= 1 - \frac{\sum_{i\in\mathcal{S}} w^2_i(1-T_i)(\hat{e}_i - \hat{\mu}^{res}_0)^2}{\sum_{i\in\mathcal{S}} w^2_i(1-T_i)(Y_i - \hat{\mu}_0)^2}
\end{aligned} \tag{B.11}$$

where $\hat{\mu}_0$ and $\hat{\mu}^{res}_0$ are defined as:

$$\hat{\mu}_0 = \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)Y_i}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}, \qquad \hat{\mu}^{res}_0 = \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)\hat{e}_i}{\sum_{i\in\mathcal{S}} w_i(1-T_i)} \tag{B.12}$$

For the post-residualized weighted least squares estimator, estimating the diagnostic follows similarly, but we now have to account for the covariate adjustment taking place:

$$\begin{aligned}\hat{R}^2_{0,wLS} &= 1 - \frac{\widehat{\text{var}}_{w,0}(\hat{e}_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}_0^{res})}{\widehat{\text{var}}_{w,0}(Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}_0)} \\ &= 1 - \frac{\sum_{i\in\mathcal{S}} w_i^2(1-T_i)(\hat{\epsilon}_i^{res} - \hat{\epsilon}_0^{res})^2}{\sum_{i\in\mathcal{S}} w_i^2(1-T_i)(\hat{\epsilon}_i - \hat{\epsilon}_0)^2},\end{aligned} \tag{B.13}$$

where $\hat{\epsilon}_i$ represents the residuals estimated from regressing the outcomes Y_i on the pre-treatment covariates $\tilde{\mathbf{X}}_i$, across the subset of units assigned to control (i.e., $Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}_0$, where $\hat{\gamma}_0$ is estimated by running the regression $Y_i \sim \tilde{\mathbf{X}}_i$ across units assigned to control). $\hat{\epsilon}_i^{res}$ is analogously defined for the residualized outcomes $\hat{e}_i$. $\hat{\epsilon}_0$ and $\hat{\epsilon}_0^{res}$ are the weighted average of both $\hat{\epsilon}_i$ and $\hat{\epsilon}_i^{res}$, respectively.

When treating $\hat{Y}_i$ as a covariate, the diagnostic can be estimated in an analogous way, but by first performing sample splitting. More specifically, the procedure for including $\hat{Y}_i$ as a covariate for the weighted estimator is as follows:

1. Across the subset of units assigned to control, randomly partition the units into two subsets: S_1 and S_2. Without loss of generality, we will use S_1 as our training sample, and S_2 as our testing sample.
2. Regress $\hat{Y}_i$ on the outcomes across S_1 to obtain a $\hat{\beta}$ value.
3. Using $\hat{\beta}$, estimate the out-of-sample residuals $\hat{e}_i^{oos}$ across S_2, where $\hat{e}_i^{oos} := Y_i - \hat{\beta}\hat{Y}_i$.
4. Estimate the diagnostic using $\hat{e}_i^{oos}$ and the outcomes Y_i across S_2 using Equation (B.11).
5. Cross-fit: repeat steps 1-3, but flipping S_1 and S_2 (i.e., regress $\hat{Y}_i$ on the outcomes across S_2 to obtain a $\hat{\beta}$ value, and estimate the diagnostic across S_1).
6. Average the two diagnostic values together.

When including $\hat{Y}_i$ as a covariate for the weighted least squares estimator, researchers can repeat the procedure above; however, when estimating the diagnostic using $\hat{e}_i^{oos}$, researchers must account for $\tilde{\mathbf{X}}_i$. More specifically:

1. Follow Steps 1-3 above to obtain $\hat{e}_i^{oos}$ across S_1.
2. Regress $\hat{e}_i^{oos}$ on $\tilde{\mathbf{X}}_i$, and regress Y_i on $\tilde{\mathbf{X}}_i$ across S_2. Use Equation (B.13) to estimate the diagnostic value.
3. Cross fit, and average the two diagnostic values together.

When researchers have relatively small sample sizes, it can be advantageous to perform repeated sample splitting, and take the average of the diagnostic across all the repeated splits (see Jacob (2020) for more details).

B.3 Simulations

This section provides details associated with the simulations described in Section 6 of the main manuscript.

Simulation Set-Up

To begin, we randomly generate a set of covariates $\begin{bmatrix} X_1 & X_2 & X_S & X_\tau \end{bmatrix} \sim MVN(\mathbf{0}, \Sigma)$ with the following covariance structure:

$$\Sigma = \begin{bmatrix} 1 & 0 & 0.45 & 0.5 \\ 0 & 1 & 0 & 0 \\ 0.45 & 0 & 1 & 0.9 \\ 0.5 & 0 & 0.9 & 1 \end{bmatrix}$$

where, recall, (X_{1i}, X_{2i}) are observed pre-treatment covariates, X_{Si} controls the probability of inclusion in the experimental sample, and $X_{\tau i}$ determines the treatment effect.

Unit i's propensity for being included in the experimental sample (recorded as $S_i = 1$) is governed by a logit model on the covariate X_{Si}:

$$P(S_i = 1) \propto \frac{\exp(X_{Si})}{1 + \exp(X_{Si})}.$$

At each iteration of the simulation, an experimental sample is drawn using the propensity score, as well as a random sample of the population. The sampled population is used to estimate the residualizing model and sampling weights.

Each specific data generating process for the potential outcome under control is determined by the values of the βs and γs and α. Below, we provide the parameter values and simplified DGP for $Y_i(0)$.

- Scenario 1: Linear Data Generating Process, identical population/sample DGP
  $\beta_1 = 2, \beta_2 = 1, \beta_3 = 0, \beta_S = 0, \gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0, \gamma_4 = 0, \alpha = 0$, yielding:

$$Y_i(0) = 2X_{1i} + X_{2i} + \varepsilon_i$$

- Scenario 2: Nonlinear Data Generating Process, identical population/sample DGP
  $\beta_1 = 2, \beta_2 = 1, \beta_3 = 0, \beta_S = 2.5, \gamma_1 = 0.5, \gamma_2 = 3, \gamma_3 = 2.5, \gamma_4 = 0, \alpha = 0$, yielding:

$$Y_i(0) = 2X_{1i} + X_{2i} + 0.5X_{1i}^2 + 3\sqrt{|X_{2i}|} + 2.5\big(X_{1i} \cdot X_{2i}\big) + \varepsilon_i$$

- Scenario 3: Linear Data Generating Process, different population/sample DGP
  $\beta_1 = 2, \beta_2 = 1, \beta_3 = -1, \beta_S = \beta_S, \gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0, \gamma_4 = 0, \alpha = 0.5$, yielding:

$$Y_i(0) = 2X_{1i} + X_{2i} + \beta_S \cdot (1 - S_i) \cdot (0.5 - X_{1i}) + \varepsilon_i,$$

- Scenario 4: Nonlinear Data Generating Process, different population/sample DGP
  $\beta_1 = 2, \beta_2 = 1, \beta_3 = -1, \beta_S = \beta_S, \gamma_1 = 0.5, \gamma_2 = 3, \gamma_3 = 2.5, \gamma_4 = 1.5, \alpha = 0.5$, yielding:

$$\begin{aligned} Y_i(0) =& 2X_{1i} + X_{2i} + 0.5X_{1i}^2 + 3\sqrt{|X_{2i}|} + 2.5\big(X_{1i} \cdot X_{2i}\big) \\ & \beta_S \cdot (1 - S_i) \cdot (0.5 - X_{1i} + 1.5X_{1i} \cdot X_{2i}) + \varepsilon_i, \end{aligned}$$

For Scenarios 3 and 4, β_S takes on values $\{-5, -2, -1, 0, 1, 2, 5\}$.

B.4 Supplementary Tables

Table B.1 presents summary results for estimator performance under Scenarios 1 and 2, including MSE, Bias, and SE. Column 1 presents the baseline results for the difference-in-means (DiM). Columns 2-4 present the results for the weighted estimators and columns 5-7 present results for the weighted least squares estimator. For the weighted and weighted least squares estimators we present the standard estimator without residualizing, the directly residualized estimator and inclusion of $\hat{Y}$ as a covariate.

Table B.2 presents summary results for estimator performance under Scenarios 3 and 4, including MSE and Bias. In these scenarios we vary the value of β_S, presented in column 1, which controls the degree of alignment between the experimental sample outcomes and the population outcomes. We fix the experimental sample size at $n = 1,000$. Columns 2-3 presents the baseline results for the difference-in-means (DiM). Columns 4-9 present the results for the weighted estimators and columns 10-15 present results for the weighted least squares estimator. For the weighted and weighted least squares estimators we present the standard estimator without residualizing, the directly residualized estimator and inclusion of $\hat{Y}$ as a covariate.

In Table B.3 we summarize the true positive and true negative rates for the diagnostic measures for the post-residualized estimators.[1] Column 1 presents the value of β_S. Columns 2-9 present the post-residualized weighted, post-residualized weighted least squares, the post-residualized weighted estimator with $\hat{Y}$ as a covariate, and the post-residualized weighted least squares estimator with $\hat{Y}$ as a covariate, respectively. We see that in general, the diagnostic measures are able to adequately capture when residualizing results in precision gain. We see that using sample splitting to estimate the pseudo-R^2 measure for the case in which we include $\hat{Y}_i$ as a covariate can sometimes be conservative, which results in a low true positive rate in cases when the divergence between the experimental sample and population are rather large. In cases where residualizing always leads to losses or gains in precision, the total number of true positive or true negative rates is zero (respectively).

Finally, in Table B.4 we evaluate the 95% coverage rates for the proposed post-residualized estimators. We see that in all scenarios, we achieve at least nominal coverage. When the

[1] True positive rates were calculated by taking the total number of true positives (i.e., cases where the diagnostic correctly indicated there would be efficiency gain from residualizing) and dividing by the total number of cases in which residualizing led to efficiency gain. True negatives are similarly defined.

Summary of Estimator Performance (N=10,000)

			Weighted			Weighted Least Squares		
		DiM	$\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Scenario 1: Linear Outcome Model								
n=100	MSE	36.44	30.05	1.48	1.34	1.34	1.34	1.30
	Bias	3.60	-0.13	0.05	0.12	0.19	0.19	0.27
	SE	4.85	5.48	1.22	1.15	1.14	1.14	1.11
n=1000	MSE	16.41	2.98	0.17	0.15	0.14	0.14	0.13
	Bias	3.74	0.00	-0.01	0.00	0.00	0.00	0.01
	SE	1.56	1.73	0.41	0.38	0.38	0.38	0.36
n=5000	MSE	14.39	0.64	0.04	0.03	0.03	0.03	0.03
	Bias	3.72	0.01	0.00	0.00	0.01	0.01	0.01
	SE	0.72	0.80	0.19	0.19	0.18	0.18	0.18
Scenario 2: Nonlinear Outcome Model								
n=100	MSE	70.71	58.80	8.25	8.20	36.59	8.16	8.04
	Bias	3.44	-0.30	0.09	0.14	0.04	0.23	0.26
	SE	7.68	7.67	2.87	2.86	6.05	2.85	2.83
n=1000	MSE	20.37	5.58	0.82	0.80	3.53	0.79	0.78
	Bias	3.78	0.05	-0.00	-0.00	0.05	0.00	0.01
	SE	2.46	2.36	0.91	0.90	1.88	0.89	0.89
n=5000	MSE	14.80	1.17	0.18	0.18	0.83	0.17	0.17
	Bias	3.68	-0.02	-0.01	-0.01	-0.03	-0.01	-0.00
	SE	1.12	1.08	0.42	0.42	0.91	0.42	0.42

Table B.1: Summary of estimator performance for Scenarios 1 and 2. The population is fixed at $N = 10,000$, and $1,000$ iterations were run for each sample size. MSE is scaled by 100, and the bias and standard error are scaled by 10.

population and sample data generating processes diverge significantly, we showed in the previous sections that there could be a loss in efficiency from using post residualized weighting. However, coverage rates are not affected by residualizing.

B.5 Additional Information for Empirical Application

As discussed in Section 7, we construct our target population using a leave-one-out procedure. Table B.5 provides a summary of the site specific and target population average treatment effects. More specifically, the difference-in-means (DiM) columns denote the experimental estimate in the specific site. The target PATE is defined as the average difference-in-means estimate across the other 15 sites. Standard errors are presented in parentheses. Certain

Summary of Estimator Performance - Scenario 3 and 4 (N = 10,000)

					Weighted						Weighted Least Squares			
	DiM		$\widehat{\tau}_W$		$\widehat{\tau}_W^{res}$		$\widehat{\tau}_W^{cov}$		$\widehat{\tau}_{wLS}$		$\widehat{\tau}_{wLS}^{res}$		$\widehat{\tau}_{wLS}^{cov}$	
β_S	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
Scenario 3: Linear Outcome														
-5	16.41	3.74	2.98	0.00	10.69	-0.11	0.36	-0.03	0.14	0.00	0.14	0.00	0.13	0.01
-2.5	15.83	3.67	3.07	-0.06	2.55	0.06	0.25	0.01	0.14	0.02	0.14	0.02	0.13	0.04
-2	16.05	3.72	2.99	0.01	1.54	0.02	0.22	0.02	0.14	0.02	0.14	0.02	0.14	0.04
-1	16.11	3.73	2.88	0.05	0.39	-0.02	0.16	0.00	0.14	-0.00	0.14	-0.00	0.13	0.02
-0.5	16.37	3.75	2.89	0.07	0.17	-0.02	0.14	-0.00	0.13	0.00	0.13	0.00	0.13	0.02
0	16.50	3.75	3.04	0.06	0.16	0.00	0.14	0.00	0.13	0.00	0.13	0.00	0.13	0.02
0.5	16.38	3.74	3.19	0.04	0.41	0.01	0.21	0.01	0.13	0.01	0.13	0.01	0.12	0.02
1	16.11	3.72	3.03	0.00	0.92	0.01	0.54	0.02	0.13	-0.01	0.13	-0.01	0.12	0.01
2	16.23	3.74	3.03	0.01	2.68	0.04	2.68	0.05	0.14	-0.00	0.14	-0.00	0.13	0.01
2.5	16.09	3.71	3.15	-0.01	3.92	0.01	3.15	-0.00	0.14	-0.01	0.14	-0.01	0.13	0.01
5	16.33	3.71	3.23	0.00	14.32	-0.01	1.54	0.02	0.14	-0.00	0.14	-0.00	0.13	0.01
Scenario 4: Nonlinear Outcome														
-5	20.31	3.74	5.77	0.04	37.03	-0.01	5.66	0.04	3.72	0.05	26.19	0.10	1.02	0.03
-2.5	20.31	3.74	6.17	-0.01	9.55	0.10	5.10	0.04	3.96	0.05	7.57	0.06	1.67	0.04
-2	19.50	3.65	5.92	-0.08	6.22	-0.00	4.27	-0.04	3.86	-0.04	5.05	-0.05	2.89	-0.00
-1	19.77	3.73	5.71	-0.02	2.18	-0.08	2.14	-0.07	3.91	-0.09	1.92	-0.05	1.08	0.02
-0.5	19.75	3.68	5.70	-0.06	1.10	-0.03	1.09	-0.03	3.96	-0.12	1.06	-0.01	0.83	0.05
0	19.74	3.69	5.81	-0.04	0.81	0.01	0.80	0.01	3.71	-0.05	0.77	0.02	0.77	0.02
0.5	20.49	3.83	5.40	0.09	1.42	0.03	1.30	0.03	3.65	0.04	1.09	0.01	0.75	0.02
1	20.24	3.80	5.52	0.08	2.84	-0.05	2.04	-0.01	3.95	0.06	1.91	-0.07	0.80	-0.01
2	20.03	3.72	5.83	0.05	7.99	-0.02	3.04	0.02	4.24	0.03	5.27	-0.06	0.84	-0.00
2.5	20.45	3.74	6.04	0.06	12.15	-0.11	3.51	-0.01	4.28	0.05	8.32	-0.09	0.85	-0.00
5	20.80	3.75	6.29	0.08	45.95	-0.25	5.05	0.02	4.09	0.06	29.97	-0.27	0.92	-0.02

Table B.2: Summary of estimator performance for Scenarios 3 and 4, where $n = 1,000$ and $N = 10,000$. 1,000 iterations were run for each β_S value. The bias is scaled by 10, and the MSE is scaled by 100.

Diagnostic Performance across Simulations

	$\widehat{\tau}_W^{res}$		$\widehat{\tau}_W^{cov}$		$\widehat{\tau}_{wLS}^{res}$		$\widehat{\tau}_{wLS}^{cov}$	
β_S	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
Scenario 3: Linear Outcomes								
-5	0/0	1000/1000	1000/1000	0/0	207/472	329/528	338/705	166/295
-2.5	1/942	58/58	1000/1000	0/0	203/499	304/501	308/694	177/306
-2	999/1000	0/0	1000/1000	0/0	216/514	288/486	310/689	175/311
-1	1000/1000	0/0	1000/1000	0/0	219/525	287/475	293/689	188/311
-0.5	1000/1000	0/0	1000/1000	0/0	214/519	282/481	293/689	183/311
0	1000/1000	0/0	1000/1000	0/0	223/523	275/477	283/683	177/317
0.5	1000/1000	0/0	1000/1000	0/0	222/536	268/464	260/666	194/334
1	1000/1000	0/0	1000/1000	0/0	233/519	283/481	254/669	199/331
2	999/1000	0/0	998/1000	0/0	228/490	321/510	297/695	175/305
2.5	0/0	999/1000	188/490	346/510	209/466	336/534	341/705	149/295
5	0/0	1000/1000	1000/1000	0/0	214/486	303/514	322/699	155/301
Scenario 4: Nonlinear Outcomes								
-5	0/0	1000/1000	360/718	224/282	0/0	1000/1000	58/1000	0/0
-2.5	0/0	998/1000	881/985	10/15	0/0	1000/1000	0/1000	0/0
-2	87/217	738/783	950/996	2/4	0/0	998/1000	0/994	5/6
-1	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0
-0.5	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0
0	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0
0.5	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0
1	1000/1000	0/0	1000/1000	0/0	999/1000	0/0	1000/1000	0/0
2	13/28	907/972	1000/1000	0/0	22/28	906/972	1000/1000	0/0
2.5	0/0	1000/1000	1000/1000	0/0	0/0	1000/1000	1000/1000	0/0
5	0/0	1000/1000	999/1000	0/0	0/0	1000/1000	1000/1000	0/0

Table B.3: True positive rates (TPR) and true negative rates (TNR) for the diagnostic measures.

sites, such as MT (Butte, MT) contain only 38 experimental units, and the point estimate of the experimental site DiM is vastly different from the target PATE. Thus, we expect the task of generalizing to be more difficult for these sites.

Estimating the Residualizing Model

Pre-treatment covariates were taken from the baseline survey conducted at the beginning of the original JTPA experiment, to assess whether or not individuals were eligible for JTPA services. A full list of the covariates included in the residualizing model is provided in Table B.6. In addition to the pre-treatment covariates, we also include normalized measures of previous earnings. Specifically, we include the z-score of an individual's previous earnings,

Coverage Rates

	Weighted			Weighted Least Squares		
β_S	$\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Scenario 3: Linear Outcome						
-5	0.95	0.95	0.97	0.99	0.99	0.99
-2.5	0.95	0.96	0.97	0.98	0.98	0.98
-2	0.95	0.97	0.98	0.97	0.97	0.98
-1	0.95	0.97	0.98	0.98	0.97	0.98
-0.5	0.95	0.98	0.99	0.98	0.98	0.98
0	0.95	0.99	0.98	0.99	0.99	0.99
0.5	0.95	0.97	0.98	0.99	0.99	0.99
1	0.96	0.95	0.95	0.98	0.98	0.98
2	0.95	0.94	0.94	0.98	0.98	0.98
2.5	0.94	0.94	0.94	0.98	0.98	0.98
5	0.94	0.94	0.95	0.98	0.98	0.99
Scenario 4: Nonlinear Outcome						
-5	0.95	0.96	0.95	0.96	0.96	0.96
-2.5	0.94	0.96	0.95	0.96	0.96	0.95
-2	0.96	0.96	0.96	0.96	0.96	0.96
-1	0.96	0.97	0.97	0.95	0.96	0.95
-0.5	0.95	0.95	0.96	0.95	0.95	0.96
0	0.94	0.96	0.96	0.95	0.96	0.96
0.5	0.95	0.96	0.96	0.96	0.96	0.97
1	0.95	0.94	0.96	0.95	0.96	0.96
2	0.95	0.95	0.96	0.94	0.95	0.96
2.5	0.96	0.95	0.96	0.94	0.95	0.96
5	0.96	0.94	0.94	0.94	0.95	0.96

Table B.4: 95% coverage rates of Normal approximation confidence intervals across 1000 simulations.

Summary of Experimental Sites and Target Population

					Earnings (in $1000)		Employment (Percentage)	
Site	Location	Expt. Size (n)	Target Pop Size (N)	Prob. of Treatment	DiM	Target PATE	DiM	Target PATE
CC	Corpus Christi, TX	524	5578	0.65	-0.21 (1.16)	1.37 (1.16)	-0.28 (3.2)	1.8 (3.2)
CI	Cedar Rapids, IA	190	5912	0.63	1.35 (1.89)	1.24 (1.89)	-0.77 (5.07)	1.71 (5.07)
CV	Coosa Valley, GA	788	5314	0.66	1.63 (0.95)	1.18 (0.95)	5.95 (2.63)	0.98 (2.63)
HF	Heartland, FL	234	5868	0.73	0.95 (1.38)	1.28 (1.38)	6.8 (5.07)	1.42 (5.07)
IN	Fort Wayne, IN	1392	4710	0.67	1.73 (0.83)	1.1 (0.83)	-0.4 (1.58)	2.23 (1.58)
JC	Jersey City, NJ	81	6021	0.64	-0.53 (3.01)	1.27 (3.01)	-2.39 (9.66)	1.67 (9.66)
JK	Jackson, MO	353	5749	0.67	2.16 (1.22)	1.19 (1.22)	5.66 (4.16)	1.38 (4.16)
LC	Larimer County, CO	485	5617	0.69	1.61 (1.32)	1.21 (1.32)	-1.97 (3.24)	1.93 (3.24)
MD	Decatur, IL	177	5925	0.70	1.24 (2.5)	1.23 (2.5)	0.03 (5.24)	1.67 (5.24)
MN	Northwest MN	179	5923	0.67	-1.43 (2.3)	1.32 (2.3)	-0.52 (6.26)	1.69 (6.26)
MT	Butte, MT	38	6064	0.71	-5.21 (4.1)	1.27 (4.1)	-7.41 (5.14)	1.67 (5.14)
NE	Omaha, NE	636	5466	0.66	1.11 (0.98)	1.25 (0.98)	-1.15 (2.56)	1.98 (2.56)
OH	Marion, OH	74	6028	0.70	-2.99 (2.71)	1.3 (2.71)	-6.82 (10.37)	1.74 (10.37)
OK	Oakland, CA	87	6015	0.64	1.83 (3.48)	1.24 (3.48)	3.34 (10.77)	1.57 (10.77)
PR	Providence, RI	463	5639	0.69	3.03 (1.34)	1.12 (1.34)	6.78 (4.58)	1.34 (4.58)
SM	Springfield, MO	401	5701	0.67	0.6 (1.31)	1.29 (1.31)	5.44 (3.34)	1.36 (3.34)

Table B.5: Summary of the JTPA study.

relative to the experimental site, as well as the z-score of an individual's previous earnings, relative to the entire population.

Baseline Covariates included in Residualizing Models

Ethnicity		**Weeks Worked**†		**Public Assistance History**		**Family Income**†	
White		Zero		Food Stamps		Less than $3,000	
Black		1-26 weeks		Cash Welfare, other than AFDC		$3,000-$6,000	
Hispanic		27-52 weeks		Unemployment Benefits		More than $6,000	
AAPI							
		Earnings		**AFDC Histories**		**Accessibility**	
Education		Previous Earnings‡	*	Ever AFDC case head		Driver's License	
ABE/ESL		Weekly Pay	*	Case head anytime†		Car available for regular use	
High school diploma		Quantile within Site	*	Received AFDC†		Telephone at home	
GED certificate		$< 25\%$	*	Years as AFDC case head:			
Some college		$> 50\%$	*	Less than 2 years		**Household Composition**	
Occupational Training		$> 90\%$	*	2-5 years		Marital Status	
Technical certificate		Quantile across Experiment	*	More than 5 years		Spouse present	
Job search assistance		$< 25\%$	*			Household Size	
Years of Education‡	*	$> 50\%$	*	**Age**		Number of children present	
		$> 90\%$	*	Age‡	*	Child under 6 present	
Work History		Non-Zero Previous Earnings	*	Age Buckets			
Ever employed		UI Reported Earnings		20-21		**Geographic Region**	
Employed upon application				22-29		West	*
Total earnings†		**Living in Public Housing**		30-44		Midwest	*
Hourly earnings		Yes		45-54		South	*
Hours worked	*			55 or older		North	*

Table B.6: We provide a list of all of the covariates included in the Super Learner. Many of these variables were included in the original JTPA study's regression model. Any variable denoted with an asterisk ($*$) was not included in the original JTPA study's regression model. † indicates that the measure is from the past 12 months prior to the baseline survey, ‡ indicates higher order terms included of that variable.

	Weighted Estimator			Weighted Least Squares		
	$\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Earnings	2.37	2.07	2.19	2.46	2.28	2.24
Employment ($\times$ 100)	8.53	8.06	8.21	7.95	7.65	8.01

Table B.7: Mean absolute error across sites.

Numerical Results for Empirical Application

Table B.7 provides numerical results for the mean absolute error across all 16 experimental sites for the six different estimators. We note that the mean absolute error of the point estimates do not vary substantially from using post-residualized weighting. This supports the results in Section 7.2.1.

Table B.8 reports the estimated standard errors (columns 3-5 for weighted estimators and columns 8-10 for weighted least squares estimators) for each site, along with the estimated diagnostics (columns 6-7 for weighted estimators and columns 11-12 for weighted least squares estimators). In general, the diagnostics are able to adequately determine whether or not we expect there to be improvements in standard error for accounting for the population outcome information, as discussed in Section 7.2.2.

Finally, Table B.9 presents the true positive rate and false positive rate for our diagnostics across the sites where the diagnostic indicated residualizing would increase precision (or not). We present these counts for both outcomes, separately.

Using Proxy Outcomes

To illustrate use of a proxy outcome, we run the same analysis as in Section 7, except we use employment as a proxy for earnings, and vice versa when building the residualizing model. This mimics a situation in which we have access to a related, but different outcome measure in our target population. Because employment is binary while earnings are continuous, we expect that direct residualizing may not result in substantial efficiency gains, and thus that our diagnostic measures would indicate not to residualize. However, treating $\hat{Y}_i$ as a covariate should still result in efficiency gain, as earnings and employment are correlated and the model can adjust for the scaling differences.

Bias

Table B.10 presents the mean absolute error of the different estimation methods. When earnings is the outcome, both directly residualizing and using $\hat{Y}_i$ as a covariate result in relatively stable performance. However, when employment is the outcome, the scaling differences between earnings (in $1000) and the binary employment measure lead to large residuals. We see a loss to precision from direct residualizing, and exacerbated finite sample bias. However,

when including $\hat{Y}_i$ as a covariate, we are able to account for the scaling differences, and the mean absolute error is lower.

Diagnostics

We estimate the same diagnostics as in Section 7.2.1 to determine when to expect precision gains from performing post-residualized weighting. We summarize the true positive and true negative rates of the diagnostic in Table B.11. We see that the performance of the diagnostic is good for direct residualization. However, we see that the diagnostic for including $\hat{Y}_i$ as a covariate is relatively conservative, and fails to identify all the cases in which it is beneficial to residualize. However, the true negative rate of the diagnostic for including $\hat{Y}_i$ as a covariate is very high (almost 100%), which indicates that the diagnostic is very effective at identifying when residualizing fails to lead to precision gain.

Table B.12 provides the standard errors and diagnostic measures for each site and estimator. Within the "Weighted" and "Weighted Least Squares" sections, the left three columns present the standard error for the corresponding estimator for each site, and the right two columns present the diagnostic measure. One key takeaway is that, when employment is the outcome, using earnings as a proxy outcome results in large scaling differences between our residualizing model, captured by $\hat{Y}_i$, and the true outcome measure. This is unsurprising since earnings is continuous and employment is binary. As a result, the $\hat{R}_0^2$ measures for the estimators that use direct residualizing (i.e., $\hat{\tau}_W^{res}$ and $\hat{\tau}_{wLS}^{res}$) are all negative, indicating that we should not use direct residualizing in that setting. However, even in this scenario, the diagnostic for using $\hat{Y}_i$ as a covariate does not indicate significant gains. When using employment as a proxy for earnings, the diagnostics indicate small gains to direct residualizing across most sites, and gains from including $\hat{Y}_i$ as a covariate across about half of sites.

Efficiency Gain

Table B.12 presents the standard errors of each weighting method, with and without post-residualizing, for each site. Table B.13 presents the average standard error across sites for post-residualized weighting using proxy outcomes, where we restrict our attention to the sites identified by the diagnostic measures for when we expect precision gains. When using employment as a proxy for earnings, direct residualizing indicates small gains in 13/16 sites, and including $\hat{Y}_i$ as a covariate indicates gains in just under half of sites. The relative improvement in variance is small due to the differences in magnitude between $\hat{Y}_i$ and Y_i. In particular, we see around a 0.3-0.4% reduction in variance from performing direct residualizing. However, when including $\hat{Y}_i$ as a covariate, which accounts for the scaling difference, the improvements are more substantial. In particular, when using $\hat{Y}_i$ as a covariate in the weighted estimator, there is a 14% reduction in variance. Using weighted least squares, there is a 9% reduction in variance from including $\hat{Y}_i$ as a covariate. The primary takeaway to highlight is that using $\hat{Y}_i$ as a covariate to perform post-residualized weighting can allow us

to leverage proxy outcomes that exist on different scales than the outcome of interest, where we expect greater gains the more closely related the outcome and proxy outcome are.

For employment, we do not consider direct residualizing because the diagnostic measure did not identify any experimental sites in which directly residualizing would lead to precision gains. When including $\hat{Y}_i$ as a covariate the diagnostic indicated 5 sites that indicate gains from post-residualized weighting; among these we see a 5% reduction in variance when using $\hat{Y}_i$ as a covariate in the weighted estimator, and a 1% reduction in variance in the weighted least squares estimator. Finally, we emphasize that estimating the PATE results in variance inflation relative to the within-sample difference-in-means, as expected. However, we see that post-residualized weighting can offset some of this loss in precision.

This exercise shows how a proxy outcome can be used for building the residualizing model. When the two variables are on very different scales, we expect that direct residualizing would not be beneficial, as evidenced here and captured by our diagnostic measures. Including $\hat{Y}_i$ as a covariate addresses scaling concerns, although as we see when using earnings as a proxy for employment, does not always allow for gains. We see that even using proxy outcomes, our diagnostic measures can accurately capture when there is potential for precision gains, and our post-residualized weighting method can lead to precision gains in estimation of the target PATE.

Standard Errors and Diagnostics for Residualizing Models

		Weighted					Weighted Least Squares				
Site	n	$\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\hat{R}_0^2$	$\hat{R}_{0,cov}^2$	$\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$	$\hat{R}_{0,wLS}^2$	$\hat{R}_{0,wLS,cov}^2$
Outcome: Earnings											
NE	636	1.70	1.53	1.53	0.23	0.22	1.58	1.53	1.53	0.08	0.06
LC	485	2.46	2.02	2.11	0.42	0.32	2.40	2.08	2.14	0.38	0.26
HF	234	1.88	1.63	1.66	0.36	0.19	1.87	1.66	1.69	0.42	0.18
IN	1392	1.03	0.93	0.92	0.25	0.26	1.00	0.91	0.91	0.22	0.21
CV	788	1.40	1.25	1.22	0.04	0.01	1.36	1.23	1.20	0.08	0.07
CC	524	2.51	2.52	2.48	-0.06	-0.18	2.42	2.42	2.39	-0.13	-0.23
JK	353	2.29	2.28	2.25	0.19	-0.25	2.19	2.18	2.16	0.30	0.10
MT	38	6.44	7.04	8.40	-0.36	-9.31	4.64	4.83	6.09	0.40	-3.90
PR	463	2.69	2.61	2.60	0.08	-0.16	2.82	2.75	2.71	0.03	-0.17
MN	179	4.79	4.70	4.80	-0.03	-0.35	3.72	4.26	4.20	-0.31	-0.56
MD	177	2.87	2.46	2.48	0.33	0.24	2.67	2.30	2.32	0.30	0.13
SM	401	2.07	2.28	2.12	-0.30	-0.13	2.13	2.23	2.11	-0.14	-0.09
OH	74	3.97	3.27	3.42	0.33	-0.22	3.94	3.75	3.77	0.29	-0.37
CI	190	3.84	3.33	3.07	0.41	0.31	3.47	3.15	2.94	0.28	-0.18
OK	87	4.69	5.07	4.64	-0.05	-0.43	4.61	4.39	4.22	0.14	-0.19
JC	81	7.24	8.81	8.50	-0.75	-1.15	6.14	7.51	6.56	-0.19	-0.49
Outcome: Employment											
NE	636	0.04	0.04	0.04	0.03	-0.01	0.04	0.04	0.04	0.02	-0.00
LC	485	0.06	0.06	0.05	0.19	0.20	0.06	0.06	0.05	0.11	0.09
HF	234	0.06	0.06	0.06	0.04	-0.03	0.06	0.06	0.06	0.02	-0.03
IN	1392	0.02	0.02	0.02	-0.15	-0.04	0.02	0.02	0.02	-0.21	-0.08
CV	788	0.03	0.03	0.03	-0.01	-0.01	0.03	0.03	0.03	-0.03	-0.01
CC	524	0.06	0.06	0.06	-0.09	-0.10	0.06	0.06	0.06	-0.10	-0.08
JK	353	0.10	0.09	0.09	0.13	-1.53	0.09	0.09	0.09	0.08	-0.85
MT	38	0.13	0.13	0.13	—	—	0.13	0.15	0.14	—	—
PR	463	0.06	0.06	0.06	0.03	-0.05	0.07	0.06	0.07	0.03	-0.03
MN	179	0.13	0.12	0.11	0.23	-3.09	0.12	0.11	0.11	0.21	-0.89
MD	177	0.09	0.08	0.08	0.19	-0.06	0.08	0.08	0.08	0.20	-4.0e28
SM	401	0.06	0.06	0.06	0.07	-0.15	0.06	0.06	0.06	0.05	-0.03
OH	74	0.07	0.06	0.07	0.09	-1.78	0.08	0.07	0.08	0.08	-1.8e28
CI	190	0.04	0.04	0.05	0.19	-0.07	0.05	0.05	0.05	0.27	-4.4e28
OK	87	0.21	0.19	0.19	0.12	-2.03	0.17	0.17	0.18	0.21	-1.35
JC	81	0.16	0.14	0.14	-0.88	-3.88	0.12	0.13	0.13	-0.92	-0.58

Table B.8: Standard error and diagnostic values for post-residualized weighting across the 16 experimental sites for two primary outcomes–earnings and employment. The diagnostic values for the site of Butte, Montana (MT) are null when outcome is employment, because all units in the control group were unemployed.

	Weighted Estimator		Weighted Least Squares	
	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Earnings				
True Positive Rate	10/11	7/12	11/11	7/13
True Negative Rate	5/5	4/4	4/5	3/3
Employment				
True Positive Rate	11/13	1/12	10/10	1/7
True Negative Rate	3/3	4/4	5/6	8/9

Table B.9: Performance of proposed diagnostic measures, as measured through the true positive rate and false positive rate.

Estimator Performance Summary with Proxy Outcomes

	Weighted			Weighted Least Squares		
	$\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Earnings	2.37	2.35	2.14	2.46	2.44	2.21
Employment ($\times$ 100)	8.53	66.15	7.85	7.95	65.33	7.45

Table B.10: Mean absolute errors for each estimator, across all experimental sites when using proxy outcomes.

	Weighted Estimator		Weighted Least Squares	
	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Earnings				
True Positive Rate	13/14	7/12	12/13	6/13
True Negative Rate	2/2	4/4	2/3	3/3
Employment				
True Positive Rate	–	2/12	–	3/10
True Negative Rate	16/16	4/4	16/16	5/6

Table B.11: Performance of proposed diagnostic measures using proxy outcomes, as measured through the true positive rate and false positive rate.

Standard Errors and Diagnostics Using Proxy Outcomes

		Weighted					Weighted Least Squares				
Site	n	$\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\hat{R}_0^2$	$\hat{R}_{0,cov}^2$	$\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$	$\hat{R}_{0,wLS}^2$	$\hat{R}_{0,wLS,cov}^2$
Outcome: Earnings											
NE	636	1.70	1.70	1.62	0.00	0.09	1.58	1.58	1.53	0.00	0.03
LC	485	2.46	2.45	2.39	0.00	0.15	2.40	2.40	2.37	0.00	0.05
HF	234	1.88	1.88	1.76	0.01	0.15	1.87	1.86	1.78	0.01	0.08
IN	1392	1.03	1.03	0.96	0.01	0.27	1.00	0.99	0.95	0.01	0.21
CV	788	1.40	1.40	1.39	0.00	-0.06	1.36	1.36	1.37	0.00	-0.03
CC	524	2.51	2.51	2.46	0.01	-0.03	2.42	2.41	2.40	0.00	-0.10
JK	353	2.29	2.28	2.10	0.01	0.07	2.19	2.18	2.07	0.01	0.04
MT	38	6.44	6.44	9.60	-0.00	-9.23	4.64	4.65	7.32	0.01	-5.86
PR	463	2.69	2.69	2.70	0.00	-0.16	2.82	2.82	2.82	-0.00	-0.15
MN	179	4.79	4.78	4.13	0.00	0.13	3.72	3.71	3.71	0.00	-0.14
MD	177	2.87	2.87	2.61	0.01	0.14	2.67	2.66	2.43	0.01	0.16
SM	401	2.07	2.07	2.04	0.00	-0.07	2.13	2.12	2.06	0.00	-0.02
OH	74	3.97	3.97	4.00	0.00	-0.44	3.94	3.93	3.75	0.00	-0.50
CI	190	3.84	3.84	3.40	0.00	-0.03	3.47	3.47	3.07	0.00	-0.22
OK	87	4.69	4.71	4.51	-0.01	-0.88	4.61	4.61	4.06	-0.01	-0.67
JC	81	7.24	7.26	8.52	-0.01	-0.82	6.14	6.17	6.75	-0.01	-0.83
Outcome: Employment											
NE	636	0.04	0.56	0.04	-352.40	-0.00	0.04	0.49	0.04	-248.43	-0.01
LC	485	0.06	0.70	0.05	-220.79	0.13	0.06	0.56	0.05	-193.43	0.02
HF	234	0.06	0.94	0.06	-260.76	-0.02	0.06	0.90	0.06	-282.80	0.06
IN	1392	0.02	0.34	0.02	-391.67	-0.04	0.02	0.32	0.02	-354.59	-0.05
CV	788	0.03	0.42	0.03	-151.95	0.02	0.03	0.38	0.03	-129.68	0.03
CC	524	0.06	1.00	0.06	-284.99	-0.21	0.06	0.89	0.06	-236.05	-0.18
JK	353	0.10	1.10	0.08	-104.99	-3.17	0.09	0.92	0.08	-88.67	-2.13
MT	38	0.13	2.42	0.12	—	—	0.13	2.49	0.13	—	—
PR	463	0.06	0.93	0.06	-228.05	-0.05	0.07	0.80	0.07	-200.67	-0.06
MN	179	0.13	1.57	0.13	-207.13	-14.37	0.12	1.61	0.12	-189.16	-3.35
MD	177	0.09	0.93	0.08	-66.00	-0.14	0.08	0.81	0.08	-77.66	-4.7e28
SM	401	0.06	0.76	0.06	-95.56	-0.12	0.06	0.68	0.06	-84.67	-0.10
OH	74	0.07	1.77	0.07	-1202.77	-0.56	0.08	1.60	0.08	-985.02	-2.4e28
CI	190	0.04	1.40	0.05	-1312.75	-0.47	0.05	1.41	0.05	-1241.70	-1.1e28
OK	87	0.21	3.24	0.16	-249.20	-1.29	0.17	2.44	0.16	-65.65	-0.28
JC	81	0.16	3.20	0.17	-6487.60	-4.51	0.12	1.70	0.14	-300.75	-0.24

Table B.12: Standard error and diagnostic values for post-residualized weighting using proxy outcomes across the 16 experimental sites for two primary outcomes–earnings and employment. Once again, the diagnostics for MT are null when employment is the outcome, because all the units in the control group are unemployed.

Summary of Standard Errors Subset by Diagnostic, using Proxy Outcomes

	Earnings				Employment			
	Number of Sites	DiM	Standard	Post Resid. Weighting	Number of Sites	DiM	Standard	Post Resid. Weighting
Weighted								
Direct Residualizing	13	1.53	2.58	2.57	0	–	–	–
$\hat{Y}_i$ as Covariate	7	1.50	2.43	2.23	2	2.93	4.01	4.00
Weighted Least Squares								
Direct Residualizing	13	1.74	2.57	2.57	0	–	–	–
$\hat{Y}_i$ as Covariate	6	1.37	1.95	1.85	3	3.65	4.68	4.63

Table B.13: Summary of standard errors across the 16 experimental sites identified by the diagnostic measures.

Appendix C

Variance-Based Sensitivity Analysis for Weighting Estimators

C.1 Additional Discussion

Missingness

In the man manuscript, the estimand of interest is the average treatment effect, across the treated. However, we note that the sensitivity framework introduced can be applied to more general settings, in which we consider missingness conditionally at random:

$$Y_i \perp\!\!\!\perp A_i \mid \mathcal{X}$$

This provides a very flexible framework to consider many settings of interest. Table C.1 summarizes several settings of interest, along with the associated conditional ignorability assumption to be relaxed by sensitivity analysis.

Setting	Missingness Indicator	Ignorability Statement
Survey Response	R_i (Response)	$Y_i \perp\!\!\!\perp R_i \mid \mathcal{X}$
Internal Validity	Z_i (Treatment Assignment)	$Y_i(1), Y_i(0) \perp\!\!\!\perp Z_i \mid \mathcal{X}$
External Validity	S_i (Inclusion in Experimental Sample)	$Y_i(1) - Y_i(0) \perp\!\!\!\perp S_i \mid \mathcal{X}$

Table C.1: Summary of different common missingness settings.

Relationship with Extensions for Sharper Bounds

Recently, several papers have demonstrated that the worst-case bounds derived under the marginal sensitivity model result in w^* that fail to recover the causal estimand. Thus,

these worst-case bounds tend to be unnecessarily conservative, and may not necessarily be sharp. Both Dorn and Guo (2021) and Nie et al. (2021) introduce additional optimization constraints for slightly tighter (and sometimes sharp) bounds. However, these additional optimization constraints come at a cost. In particular, the approach in Dorn and Guo (2021) require imposing parametric assumptions of the conditional quantiles of the outcomes; furthermore, the method does not accommodate discrete outcome variables. Nie et al. (2021) include additional balancing constraints when solving for the bounds; however, Dorn and Guo (2021) show that doing so can result in unstable performance in finite-sample settings. For this paper, we restrict our discussion to the marginal sensitivity model, but similar comparisons could be made to extensions of these models. Because these methods are all extensions of the marginal sensitivity model, the shortcomings and drawbacks that are discussed about the marginal sensitivity model similarly apply to these approaches as well.

Parametric Assumption of Conditional Ignorability

In practice, when researchers estimate weights, they are implicitly assuming a parametric version of Assumption 6. Following Hartman et al. (2021), we formalize the parametric version of Assumption 6:

Assumption 8 (Linear ignorability in $\boldsymbol{\phi(\mathbf{X})}$)
Let $\phi(\cdot)$ be a feature mapping of $\mathbf{X}_i$. Then, write the outcome Y_i as follows:

$$Y_i = \phi(\mathbf{X}_i)^\top \beta + \delta_i.$$

Similarly, write $P(Z_i = 1 \mid \mathbf{X}_i)$ as follows:

$$Pr(Z_i = 1 \mid \mathbf{X}_i) = g(\phi(\mathbf{X}_i)^\top \theta) + \eta_i,$$

where $g(\cdot) : \mathcal{R} \mapsto [0, 1]$. Then, linear ignorability holds when $\delta_i \perp\!\!\!\perp \eta_i$.

Linear ignorability in $\phi(\mathbf{X}_i)$ implies that the part of the outcome that is orthogonal to $\phi(\mathbf{X}_i)$ is independent to the part of the treatment assignment process that is orthogonal to $\phi(\mathbf{X}_i)$.

The distinction between the non-parametric version of conditional ignorability (i.e., Assumption 6) and the parametric version (i.e., Assumption 8) arises from the types of violations that matter for omitted variable bias. Under the non-parametric version of conditional ignorability, only variables that are fully unobserved (or omitted) will result in bias. However, under Assumption 8, in addition to including all of the correct variables, the choice of feature mapping also matters. For example, if researchers only include first-order moments in their weights estimation, then $\phi(\mathbf{X}_i) = \mathbf{X}_i$. However, if the true feature map necessary for linear ignorability to hold also includes higher-order terms or non-linear interactions between covariates, then using only the first-order moments will result in bias (Huang et al., 2022). As such, omitted variables in such a setting would also include any transformations of existing covariates that have not been explicitly accounted for in the estimated weights. We refer

readers to Hartman and Huang (2022) for more discussion about the two assumptions in the context of sensitivity analysis. We note that the proposed sensitivity framework is valid, regardless of which version of conditional ignorability researchers are interested in using.

Moving Away from Worst-Case Correlation Bounds

Theorem 4.3.1 allows researchers to calculate the maximum bias that can occur for a fixed R^2. This is done by assuming the correlation between the imbalance in the omitted confounder is maximally correlated with the outcome. This can be conservative in practice. We provide several recommendations for researchers who may wish to relax this bound. Doing so can result in narrower bounds, at the cost of having to reason about an additional parameter. Throughout this section, we will refer to the correlation bound as ρ^*, such that the maximum bias is written as:

$$\rho^* \cdot \sqrt{\frac{R^2}{1-R^2} \cdot \text{var}(Y_i \mid A_i = 1) \cdot \text{var}(w_i \mid A_i = 1)}$$

We suggest several different approaches for researchers to estimate less conservative bounds.

Estimating Bounds using Relative Correlation Applying the results from Huang (2022), we can decompose the correlation between the imbalance and the outcome into a function of the R^2 value, the correlation between the estimated weights and the outcomes, and the correlation between the true weights and the outcomes:

$$\text{cor}(w_i, Y_i)\sqrt{\frac{1-R^2}{R^2}} - \text{cor}(w_i^*, Y_i) \cdot \sqrt{\frac{1}{R^2}} \tag{C.1}$$

As such, an intuitive way to evaluate bounds for the correlation term is to posit a bound for the correlation between the true weights and the outcomes by a relative scaling constant k:

$$k := \frac{\text{cor}(w_i^*, Y_i)}{\text{cor}(w_i, Y_i)},$$

where k represents how many more times correlated the true weights are to the outcomes, relative to the estimated weights. k will be naturally upper-bounded at $1/\text{cor}(w_i, Y_i)$. Using Equation (C.1), researchers can then obtain a new upper bound for ρ^*:

$$\rho^* \leq \frac{\text{cor}(w_i, Y_i)}{\sqrt{R^2}} \left(\sqrt{1-R^2} - k\right)$$

It is worth noting that the correlation bound will change, depending on the R^2 parameter.

Benchmarking the Correlation Term In practice, researchers may also perform formal benchmarking to estimate what may be plausible correlation values. More specifically, researchers can calculate the error from omitting the j-th covariate and evaluate the correlation between the residual imbalance in the j-th covariate and the outcome, using this as the upper bound for ρ^*:

$$\rho^*_{(j)} \leq \widehat{\text{cor}}(w_i - w_i^{-(j)}, Y_i \mid A_i = 1)$$

Evaluating the bias at $\rho^*_{(j)}$ and $\hat{R}^2_{(j)}$ provides researchers with an estimate of the bias if they omitted a confounder with residual imbalance that is (1) equivalent in magnitude as the residual imbalance of the j-th covariate, and (2) equivalently as correlated with the outcome as the residual imbalance of the j-th covariate. Researchers can then estimate the associated confidence intervals by fixing both the correlation term and R^2.

Extended discussion for sample boundedness

Proposition C.1.1 (Necessary Condition for Validity of Sample Bounds) *Define $\mathcal{A}$ as the set of all observed Y_i values across the sample $A_i = 1$. For sample boundedness to be true (i.e., $\mathbb{E}(Y_i \mid A_i = 0) \in [\min_{i:A_i=1} Y_i, \max_{i:A_i=1} Y_i]$), the expectation of the outcomes not contained in the sample range must be constrained by the following:*

$$\mathbb{E}(Y_i \mid A_i = 0, Y_i \notin \mathcal{A}) \in \left[\frac{1}{1-p_{\mathcal{A}}} \min_{i:A_i=1} Y_i - \frac{p_{\mathcal{A}}}{1-p_{\mathcal{A}}} \max_{i:A_i=1} Y_i,\ \frac{1}{1-p_{\mathcal{A}}} \max_{i:A_i=1} Y_i - \frac{p_{\mathcal{A}}}{1-p_{\mathcal{A}}} \min_{i:A_i=1} Y_i\right],$$

where $p_{\mathcal{A}} := P(Y_i \in \mathcal{A} \mid A_i = 0)$ represents the proportion of unobserved outcomes that fall within the observed sample range.

The bound specified above represents how much overlap there must exist in the observed and unobserved potential outcomes. The bound is a function of (1) the proportion of unobserved units with outcomes that are outside the range of outcomes across the observed sample units (i.e., $1 - p_{\mathcal{A}} = P(Y_i \notin \mathcal{A} \mid A_i = 0)$), and (2) the sample bounds. If a small proportion of the outcomes in the unobserved population fall outside the sample bounds, then the bound will be relatively wide. However, if a large proportion of outcomes in the unobserved population fall outside the sample bounds, then the bound will be more narrow.

We also simulate the behavior of both sensitivity models under varying amounts of overlap.

Example C.1.1 (Coverage Rates in Limited Outcome Overlap Settings) *Define the treatment assignment mechanism as a logit model, and the outcome model as a linear model:*

$$P(Z_i = 1 \mid \mathcal{X}) \propto \frac{\exp(\gamma_1 X_{i,1} + \gamma_2 X_{i,2} + \beta U_i)}{1 + \exp(\gamma_1 X_{i,1} + \gamma_2 X_{i,2} + \beta U_i)} \qquad Y_i = \gamma_1 X_{i,1} + \gamma_2 X_{i,2} + \beta U_i + v_i,$$

where $X_{i,1}, X_{i,2}$ and U_i are standard normal random variables, and $v_i \sim N(0, \sigma_v^2)$. v_i represents a noise parameter that controls for how much outcome overlap there is. When σ_v^2 is large, then there is increased overlap between the treatment and control groups, as the treatment probability is less correlated with the outcome.

We vary $\sigma_v^2 \in \{0, 0.1, 0.25, 1, 2, 2.5\}$, and set $\gamma_1 = 2.5$, $\gamma_2 = 5$, and $\beta = 1$. For each iteration of the simulation, we assume that researchers omit U_i, and estimate confidence intervals using both the marginal sensitivity model and the variance-based sensitivity model, using the true sensitivity parameters. We visualize the coverage rates across simulations in Figure 4.3. We see that even in low overlap scenarios and small sample sizes, the variance-based sensitivity model have nominal coverage. However, the marginal sensitivity model struggles to achieve nominal coverage in limited overlap settings.

Example C.1.1 highlights that in small sample settings and limited overlap, the marginal sensitivity model fails to obtain nominal coverage, *even with the true Λ value*. In contrast, the variance-based sensitivity model consistently has nominal coverage.

We see that within finite-sample settings, the marginal sensitivity model may obtain narrower bounds than the variance-based sensitivity model, due to their inherent sample boundedness. However, these narrower bounds risk not being valid in settings with smaller sample size and limited outcome overlap, and can risk large amounts of under-coverage. Thus, the estimated confidence intervals under the variance-based sensitivity model are technically wider, but appropriately so, providing at least nominal coverage, even in cases with severely limited outcome overlap.

C.2 Proofs and Derivations

Theorem 4.3.1

For a fixed $R^2 \in [0, 1)$, then the maximum bias under $\sigma(R^2)$ can be written as a function of the following components:

$$\max_{\tilde{w} \in \sigma(R^2)} \text{Bias}(\hat{\tau}_W \mid \tilde{w}) = \underbrace{\sqrt{1 - \text{cor}(w_i, Y_i \mid A_i = 1)^2}}_{\text{(a) Correlation Bound}} \sqrt{\underbrace{\frac{R^2}{1 - R^2}}_{\text{(b) Imbalance}} \cdot \underbrace{\text{var}(Y_i | A_i = 1)\text{var}(w_i | A_i = 1)}_{\text{(c) Scaling Factor}}},$$

with the minimum bias given as the negative of Equation (4.2). The optimal bias bounds are thus given by the minimum and maximum biases.

Proof: We will start by deriving the optimal bounds. To begin, we can decompose the bias of a weighted estimator as follows:

$$\begin{aligned}
\text{Bias}(\hat{\tau}_W) &= \mathbb{E}\left(\hat{\tau}_W\right) - \tau \\
&\quad \text{By conditional ignorability:} \\
&= \mathbb{E}\left(\sum_{i\in\mathcal{A}} w_i Y_i\right) - \mathbb{E}\left(\sum_{i\in\mathcal{A}} w_i^* Y_i\right) \\
&= \mathbb{E}(w_i Y_i \mid A_i = 1) - \mathbb{E}(w_i^* Y_i \mid A_i = 1) \\
&= \mathbb{E}((w_i - w_i^*) \cdot Y_i \mid A_i = 1) \\
&\quad \text{By construction, } \mathbb{E}(w_i \mid A_i = 1) = \mathbb{E}(w_i^* \mid A_i = 1): \\
&= \mathbb{E}((w_i - w_i^*) \cdot Y_i \mid A_i = 1) - \mathbb{E}(w_i - w_i^* \mid A_i = 1) \cdot \mathbb{E}(Y_i \mid A_i = 1) \\
&= \text{cov}(w_i - w_i^*, Y_i \mid A_i = 1) \\
&= \text{cor}(w_i - w_i^*, Y_i \mid A_i = 1) \cdot \sqrt{\text{var}(w_i - w_i^* \mid A_i = 1) \cdot \text{var}(Y_i \mid A_i = 1)} \qquad \text{(C.2)}
\end{aligned}$$

This is similar to the derivation provided in Shen et al. (2011) and Hong et al. (2021). However, we will go a step further to amplify the term, $\text{var}(w_i - w_i^* \mid A_i = 1)$, into an R^2 value and the variance of the estimated weights. To do so, we extend the results from Huang (2022), which examined the bias in the context of an external validity setting, and thus, focused on re-weighting an individual-level treatment effect τ_i. We instead apply the results to a general missingness setting, in which we are re-weighting outcomes Y_i. We re-write the variance of the error in the weights in Equation (C.2) as a function of the R^2 parameter and the variance of the estimated weights, providing the following bias decomposition:

$$\text{Bias}(\hat{\tau}_W) = \text{cor}(w_i - w_i^*, Y_i \mid A_i = 1) \cdot \sqrt{\frac{R^2}{1 - R^2} \cdot \text{var}(Y_i \mid A_i = 1) \cdot \text{var}(w_i \mid A_i = 1)},$$

where R^2 is defined in Definition 4.3.1. Because we are fixing $R^2 \in [0, 1)$,[1] and $\text{var}(Y_i \mid A_i = 1) \cdot \text{var}(w_i \mid A_i = 1)$ are directly estimable from the data, to maximize the bias, we must maximize the correlation term.

Applying results from Huang (2022), we note that the error in the weights (i.e., $w_i - w_i^*$) is orthogonal to the estimated weights w_i (i.e., $\text{cov}(w_i - w_i^*, w_i \mid A_i = 1) = 0$). Then, applying the recursive formula of partial correlation, we obtain the following bounds for the correlation:[2]

$$-\sqrt{1 - \text{cor}(w_i, Y_i \mid A_i = 1)^2} \leq \text{cor}(w_i - w_i^*, Y_i \mid A_i = 1) \leq \sqrt{1 - \text{cor}(w_i, Y_i \mid A_i = 1)^2}$$

Thus, Equation 4.2 in Theorem 4.3.1 directly follows. . □

[1] In settings when $R^2 = 1$, this implies that researchers have effectively explained *none* of the variation in the true weights–i.e., in settings when researchers use uniform weights. However, if researchers have at least included one covariate that is at least correlated with a variable in the separating set $\mathcal{X}$, then $R^2 < 1$.

[2] This follows from results in Olkin (1981).

Theorem 4.3.2

For every $\tilde{w} \in \sigma(R^2)$:

$$\limsup_{n\to\infty} P(\tau(\tilde{w}) < L(\tilde{w})) \leq \frac{\alpha}{2} \text{ and } \limsup_{n\to\infty} P(\tau(\tilde{w}) > U(\tilde{w})) \leq \frac{\alpha}{2},$$

where $L(\tilde{w})$ and $U(\tilde{w})$ are defined as the $\alpha/2$ and $1-\alpha/2$-th quantiles of the bootstrapped estimates (i.e., Equation (4.4)).

Proof: We may re-write our bootstrapped estimate $\hat{\tau}^{(b)}(\tilde{w})$ as:

$$\begin{aligned}\hat{\tau}^{(b)}(\tilde{w}) &= \hat{\tau}_W^{(b)} - \text{Bias}(\hat{\tau}_W \mid \tilde{w}) \\ &= \hat{\tau}_W^{(b)} - \rho \cdot \sqrt{\text{var}(\hat{w}_i^{(b)}) \frac{R^2}{1-R^2} \cdot \text{var}(Y_i^{(b)})}\end{aligned}$$

Because ρ and R^2 are fixed (across bootstrap samples), the components that drive variation across bootstrap samples are: $\hat{\tau}_W^{(b)}$, $\text{var}(\hat{w}_i^{(b)})$, and $\text{var}(Y_i^{(b)})$.

An overview of the proof is as follows. Similar to Zhao et al. (2019), we will use a Z-estimation framework. In particular, we will add in three additional parameters: $\hat{\mu}_w^2$, $\hat{\mu}_Y$, $\hat{\mu}_Y^2$, which represent the second order moment of the weights, the average of the outcomes, and the second order moment of the outcomes, respectively. Then, we will invoke the asymptotic normality of bootstrapped Z-estimators. In the following proof, we will show the validity of the percentile bootstrap in the case that researchers are using inverse propensity score weights; however, we note that researchers can invoke the results in Soriano et al. (2021) to show validity of the results for balancing weights.

To begin, define μ_w as the expectation of the weights:

$$\mu_w = \mathbb{E}(Aw) \equiv \mathbb{E}(A \cdot (1 + \exp(-\beta X))).$$

Then, we define μ as:

$$\mu = \frac{\mathbb{E}(AY(1 + \exp(-\beta^\top X)))}{\mu_w}.$$

Define $\mu_w^2 = \mathbb{E}(Aw^2)$ and $\sigma_Y^2 = \mathbb{E}(AY^2)$ as the second moment of the weights and the outcomes, respectively. Then, we define the vector $\theta = (\mu, \mu_w, \beta, \mu_w^2, \mu_Y, \mu_Y^2)^\top \in \Theta$. Define the function $Q : 0, 1 \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d+5}$, where for $t = (a, x^\top, y) \in \{0, 1\} \times \mathbb{R}^d \times \mathbb{R}$:

$$Q(t \mid \theta) = \begin{pmatrix} Q_1(t|\theta) \\ Q_2(t|\theta) \\ Q_3(t|\theta) \\ Q_4(t|\theta) \\ Q_5(t|\theta) \\ Q_6(t|\theta) \end{pmatrix} := \begin{pmatrix} \left(a - \frac{\exp(\beta^\top x)}{1+\exp(\beta^\top x)}\right) x \\ \mu_w - a\left(1 + \exp(-\beta^\top x)\right) \\ \mu_w \mu - ay\left(1 + \exp(-\beta^\top x)\right) \\ \mu_w^2 - a\left(1 + \exp(-\beta^\top x)\right)^2 \\ \mu_y - ay \\ \mu_y^2 - ay^2 \end{pmatrix} \tag{C.3}$$

Finally, we define $\Phi(\theta)$ as:

$$\Phi(\theta) = \int Q(t|\theta) d\mathbb{P}(t),$$

where $T = (A, X^\top, AY)^\top \sim \mathbb{P}$, where $\mathbb{P}$ represents the true distribution generating the data. It is simple to see that $\Phi(\theta^*) = 0$, when θ^* is equal to the true parameter values. Then, the Z-estimates $\hat{\theta}$:

$$\Phi_n(\hat{\theta}) := \frac{1}{n}\sum_{i=1}^{n} Q(T_i|\hat{\theta}) \tag{C.4}$$

$$= \begin{pmatrix} \left(\frac{1}{n}\sum_{i=1}^n A_i - \frac{\exp(\hat{\beta}^\top \mathbf{X}_i)}{1+\exp(\hat{\beta}^\top \mathbf{X}_i)}\right)\mathbf{X}_i \\ \hat{\mu}_w - \frac{1}{n}\sum_{i=1}^n A_i\left(1+\exp(-\hat{\beta}^\top \mathbf{X}_i)\right) \\ \hat{\mu}_w\mu - \frac{1}{n}\sum_{i=1}^n A_i Y_i\left(1+\exp(-\hat{\beta}^\top \mathbf{X}_i)\right) \\ \hat{\mu}_w^2 - \frac{1}{n}\sum_{i=1}^n A_i\left(1+\exp(-\hat{\beta}^\top \mathbf{X}_i)\right)^2 \\ \hat{\mu}_y - \frac{1}{n}\sum_{i=1}^n A_i Y_i \\ \hat{\mu}_y^2 - \frac{1}{n}\sum_{i=1}^n (A_i Y_i^2) \end{pmatrix} = 0 \tag{C.5}$$

We define $\Sigma := \mathbb{E}(Q(t \mid \theta)Q(t \mid \theta)^\top)$. We will invoke the following regularity conditions, consistent with Zhao et al. (2019).

Assumption 9 (Regularity Conditions)
Assume that the parameter space Θ is compact, and that θ is in the interior of Θ. Furthermore, $(Y, \mathbf{X})$ satisfies the following:

1. $\mathbb{E}(Y^4) < \infty$
2. $\det\left(\mathbb{E}\left(\frac{\exp(\beta^\top \mathbf{X})}{(1+\exp(\beta^\top \mathbf{X}))^2}\mathbf{X}\mathbf{X}^\top\right)\right) > 0$
3. $\forall$ *compact subsets* $S \subset \mathbb{R}^d$, $\mathbb{E}(\sup_{\beta \in S} \exp(\beta^\top \mathbf{X})) < \infty$

To show asymptotic normality of bootstrapped Z-estimators, we must first verify that $\dot{\Phi}_0$ and Σ are well-defined (Kosorok, 2008).

$$\dot{\Phi}_0 = \mathbb{E}\left(\nabla_{\theta=\theta_0} Q(T|\theta)\right)$$

$$= \begin{pmatrix} 0 & 0 & -\mathbb{E}\left(\frac{\exp(\beta_0^\top \mathbf{X})}{1+\exp(\beta_0^\top \mathbf{X})^2}\mathbf{X}\mathbf{X}^\top\right) & 0 & 0 & 0 \\ 0 & 1 & \mathbb{E}(A\mathbf{X}^\top \exp(-\beta_0^\top \mathbf{X}) & 0 & 0 & 0 \\ \mu_w & \mu & \mathbb{E}(AY\mathbf{X}^\top(\exp(\beta_0^\top \mathbf{X}) & 0 & 0 & 0 \\ 0 & 0 & \mathbb{E}(A\mathbf{X}^\top(\exp(\beta_0^\top \mathbf{X}) + \exp(-2\beta_0^\top \mathbf{X}))) & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

By Leibniz Formula for determinants:

$$\begin{aligned}\left|\det(\dot{\Phi}_0)\right| &= \left|\det\begin{pmatrix} 0 & 0 & -\mathbb{E}\left(\frac{\exp(\beta_0^\top \mathbf{X})}{1+\exp(\beta_0^\top \mathbf{X})^2}\mathbf{X}\mathbf{X}^\top\right) \\ 0 & 1 & \mathbb{E}(A\mathbf{X}^\top \exp(-\beta_0^\top \mathbf{X}) \\ \mu_w & \mu & \mathbb{E}(AY\mathbf{X}^\top(\exp(\beta_0^\top \mathbf{X}) \end{pmatrix} \det\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right| \\ &= \mu_w \left|\det \mathbb{E}\left(\frac{\exp(\beta_0^\top \mathbf{X})}{(1+\exp(\beta_0^\top \mathbf{X}))^2}\mathbf{X}\mathbf{X}^\top\right)\right| > 0,\end{aligned}$$

which follows by regularity condition (2). As such, $\dot{\Phi}_0$ is invertible. Furthermore, by regularity condition (1), $\Sigma < \infty$.

As such, we simply need to verify the three conditions for asymptotic normality of bootstrapped Z-estimators:

1. The class of functions $t \to Q(t|\theta) : \theta \in \Theta$ is $\mathbb{P}$-Glivenko-Cantelli.
2. $||\Phi(\theta)||_1$ is strictly positive outside every open neighborhood of θ_0.
3. The class of functions is $\mathbb{P}$-Donsker, and $\mathbb{E}((Q(T|\theta_n) - Q(T|\theta_0))^2) \to 0$ whenever $||\theta_n - \theta_0||_1 \to 0$.

It is worth noting that the first three parameters (μ, μ_w, β) are special cases from Zhao et al. (2019), in which we do not perform any shifting in the weights (i.e., $h(x, y) = 0$). We will then show that the three conditions still hold after additionally accounting for the last three parameters. The proof for each condition is provided below.

Condition 1: The class of functions $t \to Q(t|\theta) : \theta \in \Theta$ is $\mathbb{P}$-Glivenko-Cantelli.

$$||Q(t|\theta) \leq ||Q_1(t|\theta)||_1 + \sum_{b=2}^{5} |Q_b(t|\theta)|$$

Zhao et al. (2019) show that $||Q_1(t|\theta)||_1 + |Q_2(t|\theta)| + |Q_3(t|v)|$ is bounded as a function of x, y, and some absolute constant M_1:

$$||Q_1(t|\theta)||_1 + |Q_2(t|\theta)| + |Q_3(t|v)| \leq ||x||_1 + |y| + \exp(-\beta^\top x)(1 + |y|) + M_1.$$

As such, all that is left to show is to show that $|Q_4(t|\theta)| + |Q_5(t|\theta)| + |Q_6(t|\theta)|$ is finite. To begin:

$$\begin{aligned}|Q_4(t|\theta)| &= |\mu_w^2 - (a(1+\exp(-\beta^\top x)^2| \\ &\leq \mu_w^2 + (1+\exp(-\beta^\top x))^2 \\ |Q_5(t|\theta)| &= |\mu_y - ay| \\ &\leq |\mu_y| + |y| \\ |Q_6(t|\theta)| &= |\mu_y^2 - ay^2| \\ &\leq \mu_y^2 + |y^2|\end{aligned}$$

As such:

$$|Q_4(t|\theta)| + |Q_5(t|\theta)| + |Q_6(t|\theta)| \leq M_2 + (1 + \exp(-\beta^\top x))^2 + |y| + |y^2|,$$

where M_2 is some absolute constant. As such, where M is an absolute constant:

$$||Q(t|\theta||_1 \leq ||x||_1 + 2|y| + |y^2| + \exp(-\beta^\top x)(1 + |y|) + (1 + \exp(-\beta^\top x))^2 + M,$$

where $M < \infty$ by regularity condition (1). Therefore, $\mathbb{E}(\sup_{\theta\in\Theta} ||Q(t|\theta)||_1) < \infty$, and $\{t \to Q(t|\theta) : \theta \in \Theta\}$ is $\mathbb{P}$-Glivenko-Cantelli.

Condition 2: $||\Phi(\theta)||_1$ is strictly positive outside every open neighborhood of θ_0.
Following Zhao et al. (2019), we fix some $\varepsilon > 0$. If $||\beta - \beta_0||_1 > \varepsilon/M$, then it is trivial to show that $||\Phi(\theta)||_1 > 0$. Zhao et al. (2019) show that when $||\beta - \beta_0||_1 \leq \varepsilon/M$, if $|\mu_w - \mu_{w,0}| > \varepsilon/4K$, where $K = \sup_{\theta\in\Theta} |\mu| \in (0, \infty)$, then $||\Phi(\theta)||_1 > 0$. Furthermore, when $||\beta - \beta_0||_1 \leq \varepsilon/M$ and $|\mu_w - \mu_{w,0}| \leq \varepsilon/4K$ and $|\mu - \mu_0| > \varepsilon/2\mu_w$, then $||\Phi(\theta)||_1 > 0$.

Thus, we must show for the remaining 3 parameters that when $||\mu_w^2 - \mu_{w,0}^2||$, $||\mu_y - \mu_{y,0}||$, or $||\mu_y^2 - \mu_{y,0}^2||_1$ are greater than some ε, $||\Phi(\theta)||_1 > 0$. Assume $|||\beta - \beta_0||_1 \leq \varepsilon/M$. Then:

$$\begin{aligned}
\left|\mathbb{E}\left[A\exp(-\beta^\top \mathbf{X})^2 + A\exp(-\beta_0^\top \mathbf{X})^2\right]\right| &= \left|\mathbb{E}\left[A\exp(-2\beta^\top \mathbf{X}) + A\exp(-2\beta_0^\top \mathbf{X})\right]\right| \\
&\leq \left|\mathbb{E}\left(\exp(-2\beta^\top \mathbf{X} - 2\beta_0^\top \mathbf{X})\right)\right| \\
&\leq 2||\beta - \beta_0||_\infty \mathbb{E}\left(||\mathbf{X}||_1 \exp(-t^*)\right) \text{ for } t^* \in [\beta_0^\top \mathbf{X}, \beta^\top \mathbf{X}] \\
&\leq 2 \cdot \frac{\varepsilon}{64K} = \frac{\varepsilon}{32K}
\end{aligned}$$

As such, if $||\mu_w^2 - \mu_{w,0}^2|| > \varepsilon/32K$:

$$||\Phi(\theta)||_1 \geq \left|\mu_w^2 - \mu_{w,0}^2 + \mathbb{E}\left[A\exp(-\beta^\top \mathbf{X})^2 + A\exp(-\beta_0^\top \mathbf{X})^2\right]\right| > 0 \tag{C.6}$$

For the final two parameters, it is worth noting that there is no dependency on the other parameter estimates. As such, regardless of whether the other parameters are smaller than some ϵ, if $||\mu_y - \mu_{y,0}||_1 > \varepsilon$:

$$\begin{aligned}
||\Phi(\theta)||_1 &\geq |\mu_y - \mathbb{E}(AY) - (\mu_{y,0} - \mathbb{E}(AY))| \\
&= |\mu_y - \mu_{y,0}| > 0
\end{aligned} \tag{C.7}$$

Similarly, if $||\mu_y^2 - \mu_{y,0}^2|| > \varepsilon$

$$\begin{aligned}
||\Phi(\theta)||_1 &\geq \left|\mu_y^2 - \mathbb{E}(AY^2) - (\mu_{y,0}^2 - \mathbb{E}(AY^2))\right| \\
&= \left|\mu_{y,0}^2 - \mu_y\right| > 0
\end{aligned} \tag{C.8}$$

As such, combining Equation (C.6), (C.7), (C.8), as well as the results from Zhao et al. (2019), we have shown that for all $\delta > 0$, $\inf\{||\Phi(\theta)||^2 : ||\theta - \theta_0||_1 > \delta\} > 0$.

Condition 3: The class of functions is $\mathbb{P}$-Donsker, and $\mathbb{E}((Q(T|\theta_n) - Q(T|\theta_0))^2) \to 0$ whenever $||\theta_n - \theta_0||_1 \to 0$.
From Zhao et al. (2019), we obtain a bound for the first three terms (i.e., $Q_1(t|\theta)$, $Q_2(t|\theta)$, and $Q_3(t|\theta)$). Then, for the 4th term:[3]

$$
\begin{aligned}
&|Q_4(t|\theta_2) - Q_4(t|\theta_1)| \\
&\quad = |\mu_{w,2}^2 - (a(1+\exp(-\beta_2^\top x))^2 - (\mu_{w,1}^2 - (a(1+\exp(-\beta_1^\top x))^2| \\
&\quad \leq |\mu_{w,2}^2 - \mu_{w,1}^2| + |(1+\exp(-\beta_2^\top x))^2 - (1+\exp(-\beta_1^\top x))^2| \\
&\quad = |\mu_{w,2}^2 - \mu_{w,1}^2| + \big|2\big(\exp(-\beta_2^\top x) - \exp(-\beta_1^\top x)\big) + \exp(-2\beta_2^\top x) - \exp(-2\beta_1^\top x)\big| \\
&\quad \leq |\mu_{w,2}^2 - \mu_{w,1}^2| + \big|2\big(\exp(-\beta_2^\top x) - \exp(-\beta_1^\top x)\big)\big| + \big|\exp(-2\beta_2^\top x) - \exp(-2\beta_1^\top x)\big|
\end{aligned}
$$

Applying the Mean Value Theorem (equivalently, results from Zhao et al. (2019)):

$$
\begin{aligned}
&\lesssim |\mu_{w,2}^2 - \mu_{w,1}^2| + 2||\beta_2 - \beta_1||_2||x||_2 \sup_{\beta\in\Theta} \exp(-\beta^\top x) + ||2\beta_2 - 2\beta_1||_2||x||_2 \sup_{\beta\in\Theta} \exp(-2\beta^\top x) \\
&= |\mu_{w,2}^2 - \mu_{w,1}^2| + 2||\beta_2 - \beta_1||_2||x||_2 \sup_{\beta\in\Theta} \exp(-\beta^\top x) + 2||\beta_2 - \beta_1||_2||x||_2 \sup_{\beta\in\Theta} \exp(-2\beta^\top x) \\
&= |\mu_{w,2}^2 - \mu_{w,1}^2| + 2||\beta_2 - \beta_1||_2||x||_2 \sup_{\beta\in\Theta} \exp(-\beta^\top x) \left(1 + \sup_{\beta\in\Theta} \exp(-\beta^\top x)\right) \\
&\lesssim M_4(x) \left(|\mu_{w,2}^2 - \mu_{w,1}^2| + ||\beta_2 - \beta_1||_1\right)
\end{aligned}
$$

Finally, for the 5th and 6th terms:

$$
\begin{aligned}
&|Q_5(t|\theta_2) - Q_5(t|\theta_1)| \\
&\quad = |\mu_{y,2} - ay - (\mu_{y,1} - ay)| \\
&\quad = |\mu_{y,2} - \mu_{y,1}| \\
&|Q_6(t|\theta_2) - Q_6(t|\theta_1)| \\
&\quad = |\mu_{y,2}^2 - ay^2 - (\mu_{y,1}^2 - ay^2)| \\
&\quad \leq |\mu_{y,2}^2 - \mu_{y,1}^2|
\end{aligned}
$$

Combining results with Zhao et al. (2019), we see that:

$$
||Q(t|\theta_2) - Q(t|\theta_1)||_1 = \sum_{b=1}^{6} ||Q_b(t|\theta_2) - Q_b(t|\theta_1)|| \lesssim M(x,y)||\theta_2 - \theta_1||_1
$$

Since $\mathbb{E}(M(X,Y)^2) < \infty$, we have shown that the class of functions is $\mathbb{P}$-Donsker, and furthermore, that whenever $||\theta_n - \theta_0||_1 \to 0$, $\mathbb{E}\big[(Q(t|\theta_n) - Q(t|\theta_0))^2\big] \to 0$.

[3]Consistent with Zhao et al. (2019), for some $a, b \in \mathbb{R}$, and some constant $C > 0$, if $a \leq C \cdot b$, then we write $a \lesssim b$.

Then, by invoking Kosorok (2008), Theorem 10.16:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \dot{\Phi}_0^{-1}\Sigma\dot{\Phi}_0\right), \quad \text{and} \quad \sqrt{n}(\hat{\theta}^{(b)} - \theta) \xrightarrow{d} N\left(0, \dot{\Phi}_0^{-1}\Sigma\dot{\Phi}_0\right), \tag{C.9}$$

As such, applying Delta Method and results from Appendix C3 in Zhao et al. (2019) concludes the proof. □

Theorem 4.4.1 (Weighted L_2 Analog)

Define the individual-level error in the weights as $\lambda_i := w_i^*/w_i$. Define the $L_{2,w}$ norm as follows:

$$||\lambda||_{2,w}^2 := \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 \cdot \nu(w_i) & \text{if } \text{var}(w_i) > 0, \\ \infty & \text{else} \end{cases},$$

where $\nu(w_i)$ is a function of the estimated weights. Then, the variance-based sensitivity model can equivalently be written as a norm-constrained optimization problem:

$$\max_{\tilde{w} \in \sigma(R^2)} \text{Bias}(\hat{\tau}_W \mid \tilde{w}) \iff \begin{cases} \max_{\tilde{w}} \text{Bias}(\hat{\tau}_W \mid \tilde{w}) \\ \text{s.t. } ||\lambda||_{2,w} \leq \sqrt{\frac{k}{1 - R^2}}, \end{cases}$$

where $k := 1 - R^2/\mathbb{E}(w_i^2)$.

Proof: Define $\lambda_i := w_i^*/w_i$. Then, following results from Huang (2022):

$$\text{var}(w_i - w_i^*) = \text{var}(w_i^*) - \text{var}(w_i)$$

We can then substitute in λ_i:

$$\begin{aligned} &= \text{var}(\lambda_i \cdot w_i) - \text{var}(w_i) \\ &= \mathbb{E}(\lambda_i^2 \cdot w_i^2) - \underbrace{\mathbb{E}(\lambda_i \cdot w_i)^2}_{\equiv \mathbb{E}(w_i^*)^2 = 1} - \text{var}(w_i) \\ &= \text{cov}(\lambda_i^2, w_i^2) + \mathbb{E}(\lambda_i^2)\mathbb{E}(w_i^2) - 1 - \text{var}(w_i) \\ &= \text{cov}(\lambda_i^2, w_i^2) + \mathbb{E}(\lambda_i^2)\text{var}(w_i) + \mathbb{E}(\lambda_i^2) - 1 - \text{var}(w_i) \\ &= \text{cov}(\lambda_i^2, w_i^2) + (\mathbb{E}(\lambda_i^2) - 1) \cdot (\text{var}(w_i) + 1) \\ &= \text{cov}(\lambda_i^2, w_i^2) + (\mathbb{E}(\lambda_i^2) - 1) \cdot \mathbb{E}(w_i^2) \\ \implies \frac{\text{var}(w_i - w_i^*)}{\mathbb{E}(w_i^2)} &= \frac{\text{cov}(\lambda_i^2, w_i^2)}{\mathbb{E}(w_i^2)} + (\mathbb{E}(\lambda_i^2) - 1) \end{aligned}$$

Re-arranging the terms:

$$\begin{aligned}
\frac{\text{cov}(\lambda_i^2, w_i^2)}{\mathbb{E}(w_i^2)} + \mathbb{E}(\lambda_i^2) &= 1 + \frac{\text{var}(w_i - w_i^*)}{\mathbb{E}(w_i^2)} \\
&= 1 + \frac{\mathbb{E}(w_i^2) - \mathbb{E}(w_i)^2}{\mathbb{E}(w_i^2)} \cdot \frac{R^2}{1-R^2} \\
&= 1 + \frac{R^2}{1-R^2} - \frac{\mathbb{E}(w_i)^2}{\mathbb{E}(w_i^2)} \cdot \frac{R^2}{1-R^2} \\
&= \frac{1}{1-R^2} - \underbrace{\frac{\mathbb{E}(w_i)^2}{\mathbb{E}(w_i^2)}}_{1/\mathbb{E}(w_i^2)} \cdot \frac{R^2}{1-R^2} \\
&= \frac{1}{1-R^2} \underbrace{\left(1 - \frac{R^2}{\mathbb{E}(w_i^2)}\right)}_{:=k}
\end{aligned}$$

By setting R^2, we are also setting the value for $\frac{\text{cov}(\lambda_i^2, w_i^2)}{\mathbb{E}(w_i^2)} + \mathbb{E}(\lambda_i^2)$.

We now re-write $\frac{\text{cov}(\lambda_i^2, w_i^2)}{\mathbb{E}(w_i^2)} + \mathbb{E}(\lambda_i^2)$ as a weighted sum:

$$\begin{aligned}
\mathbb{E}(\lambda_i^2) + \frac{\text{cov}(\lambda_i^2, w_i^2)}{\mathbb{E}(w_i^2)} &= \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 + \frac{1}{\mathbb{E}(w_i^2)} \cdot \frac{1}{n}\sum_{i=1}^{n} (\lambda_i^2 - \mathbb{E}(\lambda_i^2))(w_i^2 - \mathbb{E}(w_i^2)) \\
&= \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 + \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 \cdot \frac{w_i^2 - \mathbb{E}(w_i^2)}{\mathbb{E}(w_i^2)} - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\lambda_i^2) \cdot \frac{w_i^2 - \mathbb{E}(w_i^2)}{\mathbb{E}(w_i^2)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 \cdot \underbrace{\left(1 + \frac{w_i^2 - \mathbb{E}(w_i^2)}{\mathbb{E}(w_i^2)}\right)}_{:=\nu(w_i)} + \mathbb{E}(\lambda_i^2) \underbrace{\frac{1}{n}\sum_{i=1}^{n} \frac{w_i^2 - \mathbb{E}(w_i^2)}{\mathbb{E}(w_i^2)}}_{:=0} \\
&= \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 \cdot \nu(w_i)
\end{aligned}$$

As such, we can define the $L_{2,w}$ norm as follows:

$$||\lambda||_{2,w}^2 := \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 \cdot \nu(w_i) & \text{if } \text{var}(w_i) > 0 \\ \infty & \text{else} \end{cases}$$

We will show that $L_{2,w}$ meets the criteria for being a semi-norm.

1. Triangle Inequality: $||\lambda_1 + \lambda_2||_{2,w} \leq ||\lambda_1||_{2,w} + ||\lambda_2||_{2,w}$

$$\begin{aligned}||\lambda_1 + \lambda_2||_{2,w}^2 &= \sum_{i=1}^{n}(\lambda_{i,1} + \lambda_{i,2})^2 \cdot \nu(w_i) \\ &= \sum_{i=1}^{n}\lambda_{i,1}^2\nu(w_i) + \sum_{i=1}^{n}\lambda_{i,2}^2\nu(w_i) + 2\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,2}\nu(w_i)\end{aligned}$$

Applying Cauchy-Schwarz, and noting the following:
$\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,2}\nu(w_i)\right)^2 \leq \left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,1}^2\nu(w_i)\right)^2 \left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,2}^2\nu(w_i)\right)^2$. Then:

$$\begin{aligned}&\leq \sum_{i=1}^{n}\lambda_{i,1}^2\nu(w_i) + \sum_{i=1}^{n}\lambda_{i,2}^2\nu(w_i) + \\ &\quad 2\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,1}^2\nu(w_i)\right)\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,2}^2\nu(w_i)\right) \\ &= \left(||\lambda_1||_{2,w} + ||\lambda_2||_{2,w}\right)^2\end{aligned}$$

2. Absolute homogeneity:

$$\begin{aligned}||k \cdot \lambda||_{2,w} &= \sqrt{\sum_{i=1}^{n}(k \cdot \lambda_i)^2 \cdot \nu(w_i)} \\ &= k\sqrt{\sum_{i=1}^{n}\lambda_i^2 \cdot \nu(w_i)} \\ &= k \cdot ||\lambda||_{2,w}\end{aligned}$$

□

Theorem 4.4.2

Let $\psi(\Lambda)$ represent the difference in the estimated point estimate bounds under the marginal sensitivity model $\varepsilon(\Lambda)$ for a given $\Lambda \geq 1$:

$$\psi(\Lambda) := \max_{\tilde{w} \in \varepsilon(\Lambda)} \frac{\sum_{i:Z_i=0} Y_i Z_i \tilde{w}_i}{\sum_{i:Z_i=0} Z_i \tilde{w}_i} - \min_{\tilde{w} \in \varepsilon(\Lambda)} \frac{\sum_{i:Z_i=0} Y_i Z_i \tilde{w}_i}{\sum_{i:Z_i=0} Z_i \tilde{w}_i}.$$

Then if the true R^2 parameter is lower than the following threshold,

$$R^2 \leq \frac{\psi(\Lambda)^2}{4 \underbrace{(1 - \text{cor}(w_i, Y_i)^2)}_{\text{Correlation Bound}} \cdot \underbrace{\text{var}(w_i)\text{var}(Y_i)}_{\text{Scaling Factor}} + \psi(\Lambda)^2},$$

the bounds under the variance-based sensitivity model will be narrower than the bounds for the marginal sensitivity model.

Proof: The length of the point estimate bounds under the variance-based sensitivity model $\sigma(R^2)$ is equal to two times the maximum bias bound:

$$\begin{aligned} &\max_{\tilde{w} \in \sigma(R^2)} \tau(\tilde{w}) - \min_{\tilde{w} \in \sigma(R^2)} \tau(\tilde{w}) \\ &\qquad = 2 \cdot \sqrt{1 - \text{cor}(w_i, Y_i \mid A_i = 1)^2} \cdot \sqrt{\frac{R^2}{1 - R^2} \cdot \text{var}(w_i) \cdot \text{var}(Y_i \mid A_i = 1)} \end{aligned}$$

By definition, the length of the estimated point estimate bounds under the marginal sensitivity model is represented by $\psi(\Lambda)$. Thus, we want to solve for the R^2 value such that the following inequality holds:

$$2 \cdot \sqrt{1 - \text{cor}(w_i, Y_i \mid A_i = 1)^2} \cdot \sqrt{\frac{R^2}{1 - R^2} \cdot \text{var}(w_i \mid A_i = 1) \cdot \text{var}(Y_i \mid A_i = 1)} \leq \psi(\Lambda)$$

Solving for the R^2 value:

$$\begin{aligned} \frac{R^2}{1 - R^2} &\leq \frac{\psi(\Lambda)^2/4}{(1 - \text{cor}(w_i, Y_i \mid A_i = 1)^2) \cdot \text{var}(w_i \mid A_i = 1) \cdot \text{var}(Y_i \mid A_i = 1)} \\ R^2 &\leq \frac{\psi(\Lambda)^2/4(1 - \text{cor}(w_i, Y_i \mid A_i = 1)^2)\text{var}(w_i \mid A_i = 1)\text{var}(Y_i \mid A_i = 1)}{1 + \psi(\Lambda)^2/4(1 - \text{cor}(w_i, Y_i \mid A_i = 1)^2)\text{var}(w_i \mid A_i = 1)\text{var}(Y_i \mid A_i = 1)} \\ &= \frac{\psi(\Lambda)^2}{4(1 - \text{cor}(w_i, Y_i \mid A_i = 1)^2)\text{var}(w_i \mid A_i = 1)\text{var}(Y_i \mid A_i = 1) + \psi(\Lambda)^2}, \end{aligned}$$

The results from the corollary directly follow. □

Example 4.4.1

Assume researchers use a logit model to estimate the weights using $\mathbf{X}_i$, but omit a confounder U_i. The estimated and ideal weights take on the following forms:

$$\hat{w}_i = \exp(\hat{\gamma}^\top \mathbf{X}_i) \qquad \hat{w}_i^* = \exp(\hat{\gamma}^{*\top} \mathbf{X}_i + \hat{\beta} U_i)$$

Assume $\left[\mathbf{X}_i, U_i\right] \overset{iid}{\sim} MVN(0, I)$. Then $\mathbb{E}(\hat{\Lambda}) \to \infty$ as $n \to \infty$:

$$\lim_{n\to\infty} \frac{\mathbb{E}(\hat{\Lambda})}{\exp(\sqrt{2\nu^2 \log(n)})} \geq 1,$$

where $\nu^2 = (\gamma^* - \gamma)^\top (\gamma^* - \gamma) + \beta^2$, and the results follow immediately from Wainwright (2019).

Proof: The multiplicative error between w_i^* and w_i is written as:

$$\frac{\hat{w}_i^*}{\hat{w}_i} = \frac{\exp(\hat{\gamma}^{*\top} \mathbf{X}_i + \hat{\beta} U_i)}{\exp(\hat{\gamma}^\top \mathbf{X}_i)} = \exp((\hat{\gamma}^* - \hat{\gamma})^\top \mathbf{X}_i + \hat{\beta} U_i),$$

and $\hat{\Lambda}$ is defined as the maximum:

$$\hat{\Lambda} = \max_{1 \leq i \leq n} \exp(|(\hat{\gamma}^* - \hat{\gamma})^\top \mathbf{X}_i + \hat{\beta} U_i|)$$

We will show that $\mathbb{E}(\hat{\Lambda}) \to \infty$, as $n \to \infty$. To begin, define V_i as:

$$V_i := (\hat{\gamma}^* - \hat{\gamma})^\top \mathbf{X}_i + \hat{\beta} U_i$$

Because $\mathbf{X}_i$ and U_i are normally distributed, V_i will be normally distributed, with mean 0, and variance $\nu^2 := (\hat{\gamma}^* - \hat{\gamma})^\top + \hat{\beta}^2$. Let $V^{(1)}, ..., V^{(n)}$ be the ordered set of V such that $V^{(1)} \leq ... \leq V^{(n)}$. Without loss of generality, assume $|V^{(n)}| \geq |V^{(1)}|$. Then, $\mathbb{E}(\hat{\Lambda}) = \mathbb{E}(\exp(|V^{(n)}|)$. Using Jensen's inequality, the expectation of $\hat{\Lambda}$ may be lower bounded:

$$\mathbb{E}(\hat{\Lambda}) = \mathbb{E}(\exp(|V^{(n)}|) \geq \exp(\mathbb{E}(|V^{(n)}|))$$

Then, we may invoke a well-studied result that for any set of n normally distributed random variables (Wainwright (2019)):

$$\lim_{n\to\infty} \frac{\mathbb{E}(V^{(n)})}{\sqrt{2\nu^2 \log(n)}} = 1$$

Because $\mathbb{E}(|V^{(n)}|) \geq \mathbb{E}(V^{(n)})$

$$\lim_{n\to\infty} \frac{\mathbb{E}(|V^{(n)}|)}{\sqrt{2\nu^2 \log(n)}} \geq 1$$

As such, as $n \to \infty$, $\mathbb{E}(\hat{\Lambda}) \to \infty$. □

Example 4.4.2

Consider the same setting as Example 4.4.1. Then, the R^2 value can be written as follows:

$$R^2 = 1 - \frac{\exp(\hat{\gamma}^\top \hat{\gamma}) - 1}{\exp(\hat{\gamma}^{*\top} \hat{\gamma}^* + \hat{\beta}^2) - 1} \cdot \frac{\exp(\hat{\gamma}^\top \hat{\gamma})}{\exp(\hat{\gamma}^{*\top} \hat{\gamma}^* + \hat{\beta}^2)}.$$

Proof: Because $[\mathbf{X}_i, U_i] \overset{iid}{\sim} MVN(0, I)$, $\hat{w}_i$ and $\hat{w}_i^*$ both are lognormal random variables, by definition, the variance of $\hat{w}_i^*$ is: $(\exp(\hat{\gamma}^{*\top}\gamma^* + \hat{\beta}^2) - 1) \cdot \exp(\hat{\gamma}^{*\top}\hat{\gamma} + \hat{\beta}^2)$, and similarly, the variance of $\hat{w}_i$ is: $(\exp(\hat{\gamma}^\top \hat{\gamma}) - 1) \cdot \exp(\hat{\gamma}^\top \hat{\gamma})$. Then, the result of the example immediately follows, using $R^2 := 1 - \text{var}(w_i \mid A_i = 1)/\text{var}(w_i^* \mid A_i = 1)$.

□

Corollary 4.4.1

Consider the set of confounders, in which for all $\delta > 0$, $P(w_i^*/w_i < \delta) > 0$, or $P(w_i^*/w_i > \delta) > 0$. Then, if the outcomes are unbounded, the threshold from Theorem 4.4.2 will converge in probability to 1:

$$\frac{\psi(\Lambda)^2}{4(1 - \text{cor}(w_i, Y_i)^2) \cdot \text{var}(w_i)\text{var}(Y_i) + \psi(\Lambda)^2} \xrightarrow{p} 1$$

Proof: To begin, for simplicity of notation, we define $g(\psi(\Lambda); Y_i, w_i)$ as the threshold from Theorem 4.4.2:

$$g(\psi(\Lambda); Y_i, w_i) := \frac{\psi(\Lambda)^2}{4(1 - \text{cor}(w_i, Y_i)^2) \cdot \text{var}(w_i)\text{var}(Y_i) + \psi(\Lambda)^2}$$

We will use results from Resnick (2008), who show that for a sequence of random variables drawn i.i.d., the maximum of the sequence will converge in probability towards the upper bound of the support. We provide the derivation for completeness. Let $\{W_i\}_{i=1}^n$ be drawn i.i.d. from a distribution F. Then by i.i.d.:

$$P(W^{(n)} \leq w) = P\left(\bigcap_{i=1}^{n} \{W_i \leq w\}\right) = F_Y^n(w),$$

where $W^{(1)} \leq ... \leq W^{(n)}$. Then define $w_0 = \sup\{w : F_W(w) < 1\}$. Then for some $w' < w_0$, $P(W^{(n)} \leq w') = F_W^n(w') = 0$, since $F_W(w') < 1$. As such, $W^{(n)}$ converges almost surely, and by extension, in probability, to w_0. We note that the same result can be applied for the minima of $\{W_i\}_{i=1}^n$ by using $-\{W_i\}_{i=1}^n$.

Now, define $\lambda_i := w_i^*/w_i$. We have restricted the set of plausible λ_i such that $\liminf\{\lambda : 1 - F_\lambda(\lambda) < 1\} = 0$, or $\limsup\{\lambda : F_\lambda(\lambda) < 1\} \to \infty$. First consider the setting for $\liminf\{\lambda : 1 - F_\lambda(\lambda) < 1\} = 0$. We can apply the results from above to show that for a sequence of random $\lambda_1, ..., \lambda_n$, the minimum of the sequence will converge in probability towards zero Because $\Lambda = \max_{1 \leq i \leq n} \{\lambda_i, 1/\lambda_i\}$, this implies that Λ will diverge in probability towards infinity. Similarly, for $\limsup\{\lambda : F_\lambda(\lambda) < 1\} \to \infty$, the maximum of the sequence will diverge in probability towards infinity, which implies that Λ will diverge in probability towards infinity.

Applying Continuous Mapping Theorem and sample boundedness, the length of the point estimate bounds under the marginal sensitivity model ($\psi(\Lambda)$) will be equal to the range of the observed control outcomes. Thus, if the outcomes Y_i are unbounded as well (i.e., $F_Y(y) < 1$ for all $y \in \mathbb{R}$.), $\psi(\Lambda)$ will diverge in probability to infinity.

Thus, we have shown that $\psi(\Lambda)$ will diverge in probability to infinity. Applying Continuous Mapping Theorem again, $g(\psi(\Lambda); Y_i, w_i) \xrightarrow{p} 1$, which concludes the proof. □

C.3 Extended Tables

Benchmarking Results

Covariate	$\hat{\Lambda}$	MSM	$\hat{R}^2$	$\hat{\rho}$	VBM	VBM, w/ Corr.
Gender	1.1	[1.71, 2.57]	0.00	0.01	[1.81, 2.48]	[1.81, 2.48]
Age	2.1	[0.98, 3.10]	0.12	-0.01	[1.24, 3.02]	[1.87, 2.43]
Income	2.9	[0.56, 3.32]	0.14	-0.09	[1.18, 3.07]	[1.87, 2.43]
Income (Missing)	1.2	[1.62, 2.64]	0.00	-0.05	[1.81, 2.48]	[1.81, 2.48]
Education	4.7	[-0.06, 3.55]	0.17	0.06	[1.10, 3.15]	[1.77, 2.52]
Cig. Smoked	2.2	[0.92, 3.13]	0.01	-0.01	[1.66, 2.62]	[1.81, 2.48]
Smoking History	1.5	[1.38, 2.83]	0.04	-0.02	[1.50, 2.77]	[1.82, 2.47]
Race	3.5	[0.31, 3.43]	0.19	-0.09	[1.05, 3.20]	[1.88, 2.42]

Table C.2: Benchmarking results for both the marginal sensitivity model and the variance-based sensitivity model. We include both the benchmarked parameter values, as well as the estimated 95% confidence intervals.