Assessing Sensitivity to Unconfoundedness: Estimation and Inference

Matthew A. Masten[*] Alexandre Poirier[†] Linqi Zhang[‡]

July 29, 2024

Abstract

This paper provides a set of methods for quantifying the robustness of treatment effects estimated using the unconfoundedness assumption. Specifically, we estimate and do inference on bounds on various treatment effect parameters, like the average treatment effect (ATE) and the average effect of treatment on the treated (ATT), under nonparametric relaxations of the unconfoundedness assumption indexed by a scalar sensitivity parameter c. These relaxations allow for limited selection on unobservables, depending on the value of c. For large enough c, these bounds equal the no assumptions bounds. Using a non-standard bootstrap method, we show how to construct confidence bands for these bound functions which are uniform over all values of c. We illustrate these methods with an empirical application to the National Supported Work Demonstration program. We implement these methods in the companion Stata module `tesensitivity` for easy use in practice.

Keywords: Causal Inference, Treatment Effects, Sensitivity Analysis, Partial Identification

[*] Department of Economics, Duke University, `matt.masten@duke.edu`

[†] Department of Economics, Georgetown University, `alexandre.poirier@georgetown.edu`

[‡] Department of Economics, Boston College, `linqi.zhang@bc.edu`

1 Introduction

A core goal of causal inference is to identify and estimate effects of a treatment variable on an outcome variable. A common assumption used to identify such effects is unconfoundedness, which says that potential outcomes are independent of treatment conditional on covariates. This assumption is also known as conditional independence, selection on observables, ignorability, or exogenous selection; see Imbens (2004) for a survey. This assumption is not refutable, meaning that the observational data alone cannot tell us whether it is true. Nonetheless, empirical researchers may wonder: How important is this assumption in their analyses? Put differently: How sensitive are their results to unconfoundedness failures?

A large literature on sensitivity analysis has developed to answer this question. Moreover, researchers widely acknowledge that answering this question is an important step in empirical research. For example, in their figure 1, Caliendo and Kopeinig (2008) describe the workflow of a standard analysis using selection on observables. Their fifth and final step in this workflow is to perform sensitivity analysis to the unconfoundedness assumption. Imbens and Wooldridge (2009, section 6.2), Imbens and Rubin (2015, chapter 22), and Athey and Imbens (2017) all recommend that researchers conduct sensitivity analyses to assess the importance of non-refutable identifying assumptions. In particular, Athey and Imbens (2017) describe these methods as "a systematic way of doing the sensitivity analyses that are routinely done in empirical work, but often in an unsystematic way."

Most of the existing approaches to assessing unconfoundedness rely on strong auxiliary assumptions, however. For example, they often assume that all unobserved confounding arises due to a single unobserved variable whose distribution is parametrically specified, like a binary or normal distribution (e.g., Rosenbaum and Rubin 1983, Imbens 2003, Ichino, Mealli, and Nannicini 2008). They also often rely on linear models or parametric functional forms (e.g., Robins, Rotnitzky, and Scharfstein 2000, Altonji, Elder, and Taber 2005, 2008, Hosman, Hansen, and Holland 2010, Krauth 2016, Oster 2019, and Cinelli and Hazlett 2020). These assumptions—which are not needed for identification of the baseline model when unconfoundedness holds—raise a new question: Are the findings of these sensitivity analyses *themselves* sensitive to these extra auxiliary assumptions?

In this paper, we provide a set of tools for assessing the sensitivity of the unconfoundedness assumption which do not rely on strong auxiliary assumptions. We do this by studying nonparametric relaxations of the unconfoundedness assumption with binary treatments. Specifically, we build on the identification results of Masten and Poirier (2018), who consider a class of assumptions called *conditional c-dependence*. This class measures relaxations of conditional independence by a single scalar parameter $c \in [0, 1]$. This parameter c is the largest difference between the propensity score and the probability of treatment conditional on covariates and an unobserved potential outcome. Hence it has a straightforward interpretation as a deviation from conditional independence, as measured in probability units. For any positive c, conditional independence only partially holds, and so we cannot learn the exact value of our treatment effect parameters, like the average treatment effect (ATE) or the average effect of treatment on the treated (ATT). Instead, we only get bounds. Masten and Poirier (2018) derive closed-form expressions for these bounds as a function of c. Setting $c = 0$ yields the baseline model where unconfoundedness holds. Setting $c = 1$ yields the other extreme where no assumptions on selection are made, and hence gives the no assumption bounds as in Manski (1990). The bounds are monotonic in c, so that small values of c give narrow bounds while larger values of c give wider bounds. Just how wide these bounds are—and hence how sensitive one's results are—depends on the data.

While Masten and Poirier (2018) studied identification of treatment effects under nonparametric relaxations of unconfoundedness, they did not study estimation or inference. We do that in this paper. First we propose sample analog estimators of the bounds on the conditional quantile treatment effect (CQTE), the conditional average treatment effect

(CATE), the ATE, and the ATT. We do this using flexible parametric first step estimators of the propensity score and the conditional quantile function of the observed outcomes given treatment and covariates. Although such parametric restrictions are not required for our identification theory, the analysis of inference is complicated and non-standard even with these parametric first step estimators. Note that our approach of using nonparametric identification results paired with flexible parametric estimators is analogous to what is commonly done in the baseline model which imposes unconfoundedness: Identification is shown nonparametrically but many commonly used estimators are based on flexible parametric first step estimators. For example, see chapter 13 in Imbens and Rubin (2015). In appendix H we use simulations to show that these parametric estimators can perform well even under misspecification. Finally, note that the parametric assumptions we impose are testable.

Our bound estimators are non-smooth functionals. This implies that usual delta method arguments do not apply and the nonparametric bootstrap is inconsistent. We therefore derive their asymptotic distributions and show consistency of a non-standard bootstrap by using more recent methods as in Fang and Santos (2019). We show how to construct confidence bands for the bound functions which are uniform over all values of $c \in [0, 1]$. We also provide a sufficient condition on the propensity score and the distribution of the covariates under which we can do inference using the standard nonparametric bootstrap. Finally, in section 6, we use our results to study the effects of a well known supported work program: the National Supported Work (NSW) demonstration project. Our findings suggest that the program does not pass a cost-benefit analysis, even if we allow for a large amount of selection on unobservables. Using the techniques developed in this paper, and implemented in accompanying Stata module `tesensitivity`, researchers can quantify the robustness of treatment effects estimated using the unconfoundedness assumption.

Related Literature

We conclude this section with a brief literature review. As mentioned earlier, there is a large existing literature that studies how to relax unconfoundedness. Here we discuss the most closely related work and several recent papers.

The sensitivity analysis in Ichino et al. (2008) is nonparametric, but requires that all variables are discretely distributed. In contrast, we allow for continuous outcomes, covariates, and unobservables. Moreover, unlike us, they do not provide any formal results for doing estimation or inference. Another important and early approach was developed by Rosenbaum starting in the 1980's (Rosenbaum 1984, 1987, 1988, 1991, 1995, among many others). He proposed a sensitivity analysis for unconfoundedness which does not rely on a parametric model for outcomes or treatment assignment probabilities. He does design based inference for a finite set of units (also called randomization inference; see chapter 5 of Imbens and Rubin 2015 for an overview). This approach to inference is conceptually distinct from the approach we use based on repeated sampling from a large super-population. For this reason, we view these different approaches to inference in sensitivity analyses as complementary. Recent work on sensitivity analysis following Rosenbaum's tradition includes Fogarty and Small (2016), Fogarty, Shi, Mikkelsen, and Small (2017), Rosenbaum (2018), and Fogarty (2020). Kallus, Mao, and Zhou (2019) study bounds on CATE under the same nonparametric relaxations defined by Rosenbaum, but instead using a super-population approach. They propose sample analog kernel estimators based on an implicit characterization of the identified set using extrema. They show consistency of these estimators, but they do not provide any inference results. As we discuss later, this is a key distinction because inference in this setting is non-standard. Masten and Poirier (2020) also used c-dependence to relax unconfoundedness. Their focus was substantially different from the current paper since they do not do inference on parameter bound functions, as we do in this paper. In particular, they do not provide any formal results for estimation and inference on the ATE or ATT bound functions. However, in their appendix they sketched some related inference results, but only under two very strong assumptions: First, they require all covariates to be discrete. Second, they assume that the sensitivity parameter is smaller than the propensity score for *all* values of the covariates. This assumption rarely holds in practice; in fact, it can *never* hold if some covariate values drive the propensity score to zero or one. In contrast, our results are valid for all values of the sensitivity parameters as well as for continuous covariates and hence can be used to study the impact of any magnitude of selection on unobservables on one's baseline findings. Finally, a few recent papers provide methods for assessing omitted variable bias in linear models. This includes Oster (2019) and Cinelli and Hazlett (2020). Their results are restricted to estimands which can be written as OLS coefficients while our analysis is instead based on the nonparametric selection on observables framework.

2 Population Bounds on Treatment Effects

In this section we briefly review the population model from the previous literature that our estimation and inference results in sections 3–5 use. We first summarize standard results on point identification of treatment effects under unconfoundedness. We then describe how we relax unconfoundedness. Finally, we review the bounds on treatment effects derived by Masten and Poirier (2018) when unconfoundedness is relaxed. We provide additional discussion and intuition for these bounds in appendix A.

Model and Baseline Point Identification Results

We use the standard potential outcomes model. Let $X \in \{0,1\}$ be an observed binary treatment. Let Y_1 and Y_0 denote the unobserved potential outcomes. The observed outcome is $Y = XY_1 + (1-X)Y_0$. Let $W \in \mathbb{R}^{d_W}$ denote a vector of observed covariates, which may be discrete, continuous, or mixed. Let $\mathcal{W} = \text{supp}(W)$ denote the support of W. Let

$p_{1|w} = \mathbb{P}(X = 1 \mid W = w)$ denote the propensity score and let $p_{0|w} = \mathbb{P}(X = 0 \mid W = w)$.

It is well known that the conditional distributions of potential outcomes $Y_1 \mid W$ and $Y_0 \mid W$ are point identified under the following two assumptions:

Unconfoundedness: $X \perp\!\!\!\perp Y_1 \mid W$ and $X \perp\!\!\!\perp Y_0 \mid W$.

Overlap: $p_{1|w} \in (0, 1)$ for all $w \in \mathcal{W}$.

Consequently, any functional of the distributions of $Y_1 \mid W$ and $Y_0 \mid W$ is also point identified. We focus on two leading examples: The average treatment effect, $\text{ATE} = \mathbb{E}(Y_1 - Y_0)$ and the average treatment effect for the treated, $\text{ATT} = \mathbb{E}(Y_1 - Y_0 \mid X = 1)$. We also consider the conditional quantile treatment effects $\text{CQTE}(\tau \mid w) = Q_{Y_1|W}(\tau \mid w) - Q_{Y_0|W}(\tau \mid w)$ and the conditional average treatment effect $\text{CATE}(w) = \mathbb{E}(Y_1 - Y_0 \mid W = w)$.

Sensitivity Analysis: Relaxing Unconfoundedness

The overlap assumption is refutable and hence can be directly verified from the data. The unconfoundedness assumption, however, is not refutable. Consequently, like much of the literature reviewed in section 1, we perform a sensitivity analysis. This entails replacing unconfoundedness with a weaker assumption and investigating how this changes the conclusions we can draw about our parameter of interest. Specifically, we define the following class of assumptions, called *conditional c-dependence* (Masten and Poirier 2018):

Definition 1. Let $x \in \{0, 1\}$. Let $w \in \mathcal{W}$. Let c be a scalar between 0 and 1. Say X is *conditionally c-dependent* with Y_x given W if for all $w \in \mathcal{W}$ the following equation holds:

$$\sup_{y_x \in \text{supp}(Y_x|W=w)} |\mathbb{P}(X = 1 \mid Y_x = y_x, W = w) - \mathbb{P}(X = 1 \mid W = w)| \leq c. \tag{1}$$

When $c = 0$, conditional c-dependence is equivalent to $X \perp\!\!\!\perp Y_x \mid W$. For $c > 0$, however, we allow for violations of unconfoundedness by allowing the unknown conditional probability $\mathbb{P}(X = 1 \mid Y_x = y_x, W = w)$ to differ from the propensity score $\mathbb{P}(X = 1 \mid W = w)$ by at most c. Thus we actually allow for some selection on unobservables, since treatment assignment may depend on Y_x, but in a constrained manner. For sufficiently large c, however, conditional c-dependence imposes no constraints on the relationship between Y_x and X. This happens when $c \geq \overline{C}$ where $\overline{C} = \sup_{w \in \mathcal{W}} \max\{p_{1|w}, p_{0|w}\}$. When $c \in (0, \overline{C})$, conditional c-dependence imposes some constraints on treatment assignment, but it does not require conditional independence to hold exactly. For this reason, we call it a *conditional partial independence* assumption. Thus our sensitivity analysis replaces unconfoundedness with conditional c-dependence of X with Y_1 and Y_0 given W.

Treatment Effect Bounds

By relaxing conditional independence our main parameters of interest—ATE and ATT—are no longer point identified. Instead they are partially identified: We can bound them from above and from below. As c gets close to zero, however, these bounds collapse to a point. The goal of a sensitivity analysis is to understand how the shape and width of these bounds change as c varies from 0 to 1.

These bounds were derived in Masten and Poirier (2018), which we summarize here. We provide more details and intuition in appendix A. Although that paper studied both continuous and binary outcomes, here we only summarize the results for continuous Y_x. These bounds were derived under the assumption that $Y_x \mid X, W$ has a continuous and strictly increasing cdf, that the support of $Y_x \mid X, W$ is an interval independent of X, and that $p_{1|w} \in (0, 1)$. Assumption A1 in that paper formally states this assumption; we also restate it in our appendix A for convenience. All of our parameters of interest can be written in terms of bounds on the quantile regressions $Q_{Y_x|W}(\tau \mid w)$ where $\tau \in (0, 1)$ is the quantile index and $w \in \mathcal{W}$. Under the conditional partial independence assumption stated above and some regularity conditions,

$$[\underline{Q}^c_{Y_x|W}(\tau \mid w), \overline{Q}^c_{Y_x|W}(\tau \mid w)] = \left[Q_{Y|X,W}\left(\underline{t}(\tau, p_{x|w}) \mid x, w\right), Q_{Y|X,W}\left(\overline{t}(\tau, p_{x|w}) \mid x, w\right)\right] \quad (2)$$

are sharp bounds on this quantile regression, uniformly in τ, x, and w, where

$$\begin{aligned}\overline{t}(\tau, p_{x|w}) &= \min\left\{\tau + \frac{c}{p_{x|w}}\min\{\tau, 1-\tau\}, \frac{\tau}{p_{x|w}}, 1\right\}\\ \underline{t}(\tau, p_{x|w}) &= \max\left\{\tau - \frac{c}{p_{x|w}}\min\{\tau, 1-\tau\}, \frac{\tau-1}{p_{x|w}} + 1, 0\right\}.\end{aligned}$$

Taking differences of these bounds for $x = 1$ and $x = 0$ yields sharp bounds on the conditional quantile treatment effect $\text{CQTE}(\tau \mid w)$, uniformly in τ and w:

$$\left[\underline{\text{CQTE}}^c(\tau \mid w), \overline{\text{CQTE}}^c(\tau \mid w)\right] \equiv \left[\underline{Q}^c_{Y_1|W}(\tau \mid w) - \overline{Q}^c_{Y_0|W}(\tau \mid w),\ \overline{Q}^c_{Y_1|W}(\tau \mid w) - \underline{Q}^c_{Y_0|W}(\tau \mid w)\right].$$

Integrating these bounds over τ yields sharp bounds on $\text{CATE}(w)$, uniformly in w:

$$\left[\underline{\text{CATE}}^c(w), \overline{\text{CATE}}^c(w)\right] \equiv \left[\int_0^1 \underline{\text{CQTE}}^c(\tau \mid w)\, d\tau, \int_0^1 \overline{\text{CQTE}}^c(\tau \mid w)\, d\tau\right].$$

Further integrating over the marginal distribution of W yields sharp bounds on ATE:

$$\left[\underline{\text{ATE}}^c, \overline{\text{ATE}}^c\right] \equiv \left[\mathbb{E}\big(\underline{\text{CATE}}^c(W)\big),\ \mathbb{E}\big(\overline{\text{CATE}}^c(W)\big)\right].$$

Note that these bounds are always finite for sufficiently small c, but potentially infinite when Y_x is unbounded and c is sufficiently large. To obtain bounds on ATT, let $\underline{E}^c_x(w) = \int_0^1 \underline{Q}^c_{Y_x|W}(\tau \mid w)\, d\tau$ and $\overline{E}^c_x(w) = \int_0^1 \overline{Q}^c_{Y_x|W}(\tau \mid w)\, d\tau$ denote bounds on $\mathbb{E}(Y_x \mid W = w)$. Averaging these over the marginal distribution of W yields bounds on $\mathbb{E}(Y_x)$, denoted by $\underline{E}^c_x = \mathbb{E}\big(\underline{E}^c_x(W)\big)$ and $\overline{E}^c_x = \mathbb{E}\big(\overline{E}^c_x(W)\big)$. This yields the following bounds on ATT:

$$\left[\mathbb{E}(Y \mid X = 1) - \frac{\overline{E}^c_0 - p_0\mathbb{E}(Y \mid X = 0)}{p_1},\ \mathbb{E}(Y \mid X = 1) - \frac{\underline{E}^c_0 - p_0\mathbb{E}(Y \mid X = 0)}{p_1}\right] \tag{3}$$

where $p_x = \mathbb{P}(X = x)$ for $x \in \{0, 1\}$. Finally, note that all of these bounds are sharp.

Breakdown Points

So far we have discussed sharp bounds on various parameters of interest as a function of the sensitivity parameter c. In addition to the bounds themselves, it is common to analyze *breakdown points* for various conclusions of interest; see Oster (2019) for a prominent application. For example, suppose that under the baseline model ($c = 0$) we find that $\text{ATE} > 0$. We then ask: How much can we relax unconfoundedness while still being able to conclude that the ATE is nonnegative? To answer this question, define the breakdown point for the conclusion that the ATE is nonnegative as

$$c_{\text{BP}} = \sup\{c \in [0,1] : \left[\underline{\text{ATE}}^c, \overline{\text{ATE}}^c\right] \subseteq [0, \infty)\}. \tag{4}$$

This number is a scalar measure of the robustness of the conclusion that ATE is positive to relaxations of the key identifying assumption of unconfoundedness. We analyze them in our empirical application in section 6.

Interpreting Conditional c-Dependence

We conclude this section by giving some suggestions for how to interpret conditional c-dependence in practice. In particular, what values of c are large? What values are small? This interpretation and calibration of c is important because its true value is completely unidentified and therefore is unknown. Here we summarize and extend the discussion in Masten and Poirier (2018). We illustrate these interpretations in our empirical analysis in section 6.

Let W_k denote a component of W and let W_{-k} denote its other components. Denote the propensity score by $p_{1|W}(w_{-k}, w_k) = \mathbb{P}(X = 1 \mid W = (w_{-k}, w_k))$. Let $p_{1|W_{-k}}(w_{-k}) = \mathbb{P}(X = 1 \mid W_{-k} = w_{-k})$ denote the leave-out-variable-k propensity score. This is the proportion of the population who are treated, conditional on only W_{-k}. Consider the random variable

$$\Delta_k = |p_{1|W}(W_{-k}, W_k) - p_{1|W_{-k}}(W_{-k})|.$$

This difference is a measure of the impact on the propensity score of adding W_k, given that we already included W_{-k}. Conditional c-dependence is defined by a similar difference, except there we add the unobservable Y_x given that we already included W. So c reflects the additional impact of an unobservable on treatment while Δ_k captures the additional impact of an observable on treatment. Following Altonji et al. (2005) and Oster (2019), among others, we consider calibrating the magnitude of selection on unobservables using measures of selection on observables. Hence we suggest using the distribution of Δ_k to calibrate values of c. For example, you could first consider the upper bound on its support, $\bar{c}_k = \max \text{supp}(\Delta_k)$, and then the 90th, 75th, and 50th quantiles of Δ_k. You may also find it useful to plot an estimate of the density of Δ_k. All of these reference values can be compared to the breakdown point c_{BP} for a specific conclusion of interest. Specifically, if c_{BP} is larger than the chosen reference value, then the conclusion of interest could be considered robust. In contrast, if c_{BP} is smaller than the chosen reference value, then the conclusion of interest could be considered sensitive. You may also want to see where c_{BP} lies relative to the distribution of Δ_k. This can be done by computing $F_{\Delta_k}(c_{\text{BP}})$.

While this can be done for all covariates k, it may be helpful to restrict attention to covariates that have a sufficiently large impact on the baseline point estimates. For example, suppose we are interested in the ATE. Let ATE_{-k} denote the ATE estimand obtained in the baseline selection on observables model using only the covariates W_{-k}. Let ATE denote the ATE estimand obtained in the baseline model using all the covariates. Then $\left|\frac{\text{ATE}-\text{ATE}_{-k}}{\text{ATE}}\right|$ denotes the effect of omitting covariate k on the ATE point estimand, as a percentage of the baseline estimand that uses all covariates in W. You may want to restrict attention to covariates k for which this ratio is relatively large. We illustrate this approach in our empirical analysis in section 6. Note that we recommend examining ATE_{-k} primarily as a supplement to Δ_k, to obtain a potentially less conservative definition of robustness; percentage changes in ATE that occur from omitting covariate k are not on the same scale as c and hence cannot be used to calibrate c by themselves.

3 Estimation

In the previous section we assumed the entire population distribution of (Y, X, W) was known. In practice we only have a finite sample $\{(Y_i, X_i, W_i)\}_{i=1}^n$ from this distribution. In this section we explain how to use this finite sample data to estimate the population bounds of section 2. We give the corresponding asymptotic theory in section 4 where we obtain the joint limiting distribution of treatment effects bounds.

As shown in section 2, all of our bounds can be constructed from the distribution of (Y, X, W) and the bounds on $Q_{Y_x|W}$ given in equation (2). These bounds on $Q_{Y_x|W}$, in turn, depend on just two features of the data: (a) the conditional quantile function $Q_{Y|X,W}(\tau \mid x, w)$, and (b), the propensity score $p_{1|w} = \mathbb{P}(X = 1 \mid W = w)$. In both cases, we can use parametric, semiparametric, or nonparametric estimation methods. In this paper we focus on flexible parametric approaches. Even in this case the asymptotic distribution theory is non-standard and quite complicated.

In section 3.1 we describe our first step estimators of these two functions. Given these estimators, we then construct sample analog estimates of our bound functions in a second step. We describe these estimators in section 3.2.

3.1 First Step Estimation

We estimate $Q_{Y|X,W}$ by a linear quantile regression of Y on flexible functions of (X, W) that we denote by $q(X, W) \in \mathbb{R}^{d_q}$. For example, $q(x, w)$ could be $(1, x, w)$, $(1, x, w, x \cdot w)$, or could contain additional interactions between the treatment indicator X and functions of the covariates W. For $\tau \in (0, 1)$, let $\widehat{Q}_{Y|X,W}(\tau \mid x, w) = q(x, w)'\widehat{\gamma}(\tau)$ where $\widehat{\gamma}(\tau)$ are the estimated coefficients from a linear quantile regression of Y on $q(X, W)$ at the quantile τ.

We estimate the propensity score by maximum likelihood. In particular, specify the parametric model $p_{1|w} = F(r(w)'\beta_0)$ where F is a known cdf, $r(w)$ is a known vector function, and β_0 is an unknown constant vector. The functions $r(w)$ could simply be

$(1, w)$ or may contain functions of w, like squared or interaction terms. For notational simplicity, we will assume throughout the paper that $r(w) = w$. This specification for the propensity score includes the probit and logit estimators as special cases. Those estimators are commonly used in the literature; e.g., see chapter 13 of Imbens and Rubin (2015). We let $\widehat{p}_{1|w} = F(w'\widehat{\beta})$ denote our propensity score estimator, where $\widehat{\beta}$ is the maximum likelihood estimator of β_0.

3.2 Second Step Estimation of the Bound Functions

Given the first step estimators from section 3.1, we obtain the following sample analog estimators of the CQTE bound functions defined in equation (2):

$$\left(\widehat{\overline{Q}}^c_{Y_x|W}(\tau \mid w), \underline{\widehat{Q}}^c_{Y_x|W}(\tau \mid w)\right) = \left(\widehat{Q}_{Y|X,W}(\overline{t}(\tau, \widehat{p}_{x|w}) \mid x, w), \widehat{Q}_{Y|X,W}(\underline{t}(\tau, \widehat{p}_{x|w}) \mid x, w)\right).$$

As discussed in section 2, averaging these over $\tau \in (0, 1)$ yields sample analog estimates of bounds on CATE(w), which we can then use to get bounds on ATE. This approach requires estimation of extremal quantiles, however—estimation for τ's close to 0 or 1. This is well known to be a delicate problem (see Chernozhukov, Fernández-Val, and Kaji 2017) since extremal quantiles are effectively estimated by a small fraction of observations and thus converge in distribution at nonstandard rates. So in this paper we use a common solution: Fixed trimming of the extremal quantiles. We do this by modifying the quantile bound estimators to ensure that the argument of $\widehat{Q}_{Y|X,W}(\cdot \mid x, w)$ lies in $[\varepsilon, 1 - \varepsilon]$ for some fixed and known $\varepsilon \in (0, 0.5)$. Specifically, this yields the trimmed estimators of the two quantile bounds:

$$\widehat{\overline{Q}}^c_{Y_x|W}(\tau \mid w) = \widehat{Q}_{Y|X,W}\left(\max\{\min\{\overline{t}(\tau, \widehat{p}_{x|w}), 1 - \varepsilon\}, \varepsilon\} \mid x, w\right), \tag{5}$$

$$\underline{\widehat{Q}}^c_{Y_x|W}(\tau \mid w) = \widehat{Q}_{Y|X,W}\left(\max\{\min\{\underline{t}(\tau, \widehat{p}_{x|w}), 1 - \varepsilon\}, \varepsilon\} \mid x, w\right). \tag{6}$$

We use these estimators for the rest of the paper. Common choices of ε are 0.05 or 0.01. In our asymptotic analysis we hold ε fixed with sample size. In principle we could generalize the results to allow $\varepsilon \to 0$ as $n \to \infty$, but this would complicate the analysis of inference, which is already non-standard for other reasons. Since we fix ε throughout, we omit ε from the notation for brevity, except when necessary.

We estimate the CQTE bounds by taking differences of the quantile bound estimators:

$$\left[\underline{\widehat{\text{CQTE}}}^c(\tau \mid w), \widehat{\overline{\text{CQTE}}}^c(\tau \mid w)\right] \equiv \left[\widehat{\underline{Q}}^c_{Y_1|W}(\tau \mid w) - \widehat{\overline{Q}}^c_{Y_0|W}(\tau \mid w),\; \widehat{\overline{Q}}^c_{Y_1|W}(\tau \mid w) - \widehat{\underline{Q}}^c_{Y_0|W}(\tau \mid w)\right].$$

Since our CATE bounds are simply the integral of the CQTE bounds over all the quantiles τ, we can estimate them by

$$\left[\underline{\widehat{\text{CATE}}}^c(w), \widehat{\overline{\text{CATE}}}^c(w)\right] \equiv \left[\int_0^1 \underline{\widehat{\text{CQTE}}}^c(\tau \mid w)\, d\tau, \int_0^1 \widehat{\overline{\text{CQTE}}}^c(\tau \mid w)\, d\tau\right]. \qquad (7)$$

A second integration over w with respect to the marginal distribution of W yields bounds on ATE. Like much of the literature, we use the empirical distribution of W to estimate the marginal distribution of W. This yields the following estimator of our ATE bounds:

$$\left[\underline{\widehat{\text{ATE}}}^c, \widehat{\overline{\text{ATE}}}^c\right] = \left[\frac{1}{n}\sum_{i=1}^n \underline{\widehat{\text{CATE}}}^c(W_i),\; \frac{1}{n}\sum_{i=1}^n \widehat{\overline{\text{CATE}}}^c(W_i)\right].$$

In appendix B, we show that these bounds approximately collapse to a standard regression adjustment estimator of the ATE when $c = 0$. Next consider the estimation of the ATT bounds. Let

$$\widehat{\underline{E}}^c_0 = \frac{1}{n}\sum_{i=1}^n \int_0^1 \widehat{\underline{Q}}^c_{Y_0}(\tau \mid W_i)\, d\tau \qquad \text{and} \qquad \widehat{\overline{E}}^c_0 = \frac{1}{n}\sum_{i=1}^n \int_0^1 \widehat{\overline{Q}}^c_{Y_0}(\tau \mid W_i)\, d\tau.$$

For $x \in \{0,1\}$ let $\widehat{\mathbb{E}}(Y \mid X = x) = \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i = x)}{\sum_{i=1}^n \mathbb{1}(X_i = x)}$ and $\widehat{p}_x = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i = x)$. We can then estimate the ATT bounds by replacing the population quantities in (3) with their estimators that we just defined.

For $c = 0$, our estimated upper and lower bounds are equal and give point estimates of the various parameters of interest. For $c > 0$, our bounds have positive width. To use these bounds in a sensitivity analysis, we recommend producing the following plot: select a grid $\{c_1, \ldots, c_K\} \subseteq [0,1]$ of values for c. Compute our bound estimates on this grid and plot them against these values of c. Then compute and plot confidence bands for these bound estimates against c as well; we describe how to compute these bands in section 5. We illustrate all of these steps in our empirical analysis in section 6.

4 Asymptotic Theory

In this section we provide results on the consistency and limiting distributions of the estimators we described in section 3. We leave a full discussion of the technical details and assumptions to appendix D. In section 5 we show how to use these results to do inference based on a non-standard bootstrap.

4.1 Convergence of the First Step Estimators

Our first step estimators are standard in the literature. Hence we only briefly review the main assumptions and results for these estimators. Throughout this paper we assume that we observe a random sample from the distribution of (Y, X, W). We assume the propensity score and quantile regression functions are correctly specified: i.e., $p_{1|w} = F(w'\beta_0)$ and that $Q_{Y|X,W}(\tau \mid x, w) = q(x, w)'\gamma_0(\tau)$ for $\tau \in [\varepsilon, 1 - \varepsilon]$. We study the role of this assumption in a simulation study in appendix H. Under assumptions stated in appendix D, both $\widehat{\beta}$ and $\widehat{\gamma}(\tau)$ converge in distribution at the $\sqrt{n}$-rate to a Gaussian distribution. Moreover, the convergence of $\widehat{\gamma}(\tau)$ to $\gamma_0(\tau)$ is uniform over $\tau \in [\varepsilon, 1 - \varepsilon]$.

4.2 Convergence of the Second Step Estimators

Next we consider the limiting distribution of our various second step estimators.

The CATE Bounds

Let $\widehat{\theta} = (\widehat{\beta}, \widehat{\gamma}(\cdot))$ denote the first step estimators and let $\theta_0 = (\beta_0, \gamma_0(\cdot))$. The CATE bound estimators of equation (7) can be written as

$$\left[\underline{\widehat{\text{CATE}}}^c(w), \overline{\widehat{\text{CATE}}}^c(w)\right] \equiv \left[\underline{\Gamma}_2(1, w, \widehat{\theta}) - \overline{\Gamma}_2(0, w, \widehat{\theta}),\ \overline{\Gamma}_2(1, w, \widehat{\theta}) - \underline{\Gamma}_2(0, w, \widehat{\theta})\right]$$

for known mappings $(\underline{\Gamma}_2, \overline{\Gamma}_2)$ defined in appendix D. Let the trimmed population CATE bounds be given by their population counterpart

$$\left[\underline{\text{CATE}}^c_\varepsilon(w), \overline{\text{CATE}}^c_\varepsilon(w)\right] = \left[\underline{\Gamma}_2(1, w, \theta_0) - \overline{\Gamma}_2(0, w, \theta_0),\ \overline{\Gamma}_2(1, w, \theta_0) - \underline{\Gamma}_2(0, w, \theta_0)\right].$$

The mappings $(\underline{\Gamma}_2, \overline{\Gamma}_2)$ are generally not Hadamard differentiable in θ because they depend on the minimum and maximum functions. Hence we cannot apply the functional delta method to establish a Gaussian limiting distribution for these estimated bounds. They are, however, Hadamard directionally differentiable (HDD); see definition 1 in appendix D. It turns out that this weaker version of differentiability is sufficient to establish their (non-Gaussian) limiting distribution.

As a technical assumption, we restrict the complexity of the space the quantile regression coefficients $\gamma_0(\cdot)$ lie in. Assumption 5 in appendix D imposes that $\gamma_0(\cdot)$ lies in a Hölder ball. We can then establish the limiting distribution of the CATE bound estimates.

Proposition 1 (CATE convergence)**.** Fix $w \in \mathcal{W}$, $c \in [0, 1]$. Suppose A1–A5 hold. Then

$$\sqrt{n}\left(\widehat{\overline{\text{CATE}}}^c(w) - \overline{\text{CATE}}^c_\varepsilon(w), \widehat{\underline{\text{CATE}}}^c(w) - \underline{\text{CATE}}^c_\varepsilon(w)\right) \xrightarrow{d} \mathbf{Z}_{\text{CATE}}(w),$$

where $\mathbf{Z}_{\text{CATE}}(w)$ is a random vector in $\mathbb{R}^2$ whose distribution is characterized in the proof.

In the statement of this result we deferred the full characterization of $\mathbf{Z}_{\text{CATE}}(w)$ to the proof. To get a brief idea of what it looks like, however, consider the first component. It is $\mathbf{Z}^{(1)}_{\text{CATE}}(w) = \overline{\Gamma}'_{2,\theta_0}(1, w, \mathbf{Z}_1) - \underline{\Gamma}'_{2,\theta_0}(0, w, \mathbf{Z}_1)$ where $\overline{\Gamma}'_{2,\theta_0}(x, w, \mathbf{Z}_1)$ is the Hadamard directional derivative of $\overline{\Gamma}_2$ evaluated at $\mathbf{Z}_1$, the limiting distribution of the first step estimators. See page 35 of the appendix for the expression for $\overline{\Gamma}'_{2,\theta_0}$. Likewise, $\underline{\Gamma}'_{2,\theta_0}(x, w, \mathbf{Z}_1)$ is the Hadamard directional derivative of $\underline{\Gamma}_2$ evaluated at $\mathbf{Z}_1$. Although $\mathbf{Z}_1$ is Gaussian, the HDDs are continuous but generally nonlinear functionals. Hence the distribution of $\mathbf{Z}_{\text{CATE}}(w)$ is generally non-Gaussian. In section 5 we show how to use a non-standard bootstrap to approximate its distribution. In special cases, the limiting distribution is Gaussian.

This can be the case for the ATE and ATT bound estimates as well. In appendix I we provide sufficient conditions under which standard bootstrap inference is also valid.

The ATE Bounds

Next we derive the limiting distribution of our ATE bound estimators. Let $\overline{\Gamma}_3(x,\theta) = \int_{\mathcal{W}} \overline{\Gamma}_2(x,w,\theta)\, dF_W(w)$ and $\underline{\Gamma}_3(x,\theta) = \int_{\mathcal{W}} \underline{\Gamma}_2(x,w,\theta)\, dF_W(w)$. Then

$$[\underline{\text{ATE}}^c_\varepsilon, \overline{\text{ATE}}^c_\varepsilon] = \left[\underline{\Gamma}_3(1,\theta_0) - \overline{\Gamma}_3(0,\theta_0),\, \overline{\Gamma}_3(1,\theta_0) - \underline{\Gamma}_3(0,\theta_0)\right]$$

are the trimmed population ATE bounds. We estimate them by $\left[\widehat{\underline{\text{ATE}}}^c, \widehat{\overline{\text{ATE}}}^c\right]$. Adding a regularity assumption on the propensity score $F(w'\beta)$ (see assumption 6 in appendix D), we can now establish the limiting distribution of the ATE bound estimates.

Theorem 1 (ATE convergence)**.** Suppose A1–A6 hold. Then

$$\sqrt{n}\left(\widehat{\overline{\text{ATE}}}^c - \overline{\text{ATE}}^c_\varepsilon, \widehat{\underline{\text{ATE}}}^c - \underline{\text{ATE}}^c_\varepsilon\right) \xrightarrow{d} \mathbf{Z}_{\text{ATE}},$$

where $\mathbf{Z}_{\text{ATE}}$ is a random vector in $\mathbb{R}^2$ whose distribution is characterized in the proof.

Like the CATE bound estimators, the limiting distribution of the ATE bound estimators is non-Gaussian.

The ATT Bounds

Finally we study the limiting properties of our ATT bound estimators. Our trimmed population ATT bounds are estimated by

$$[\widehat{\underline{\text{ATT}}}^c, \widehat{\overline{\text{ATT}}}^c] = \left[\widehat{\mathbb{E}}(Y \mid X=1) - \frac{\widehat{\overline{E}}{}^c_0 - \widehat{p}_0\widehat{\mathbb{E}}(Y \mid X=0)}{\widehat{p}_1}, \widehat{\mathbb{E}}(Y \mid X=1) - \frac{\widehat{\underline{E}}{}^c_0 - \widehat{p}_0\widehat{\mathbb{E}}(Y \mid X=0)}{\widehat{p}_1}\right].$$

Proposition 2 (ATT convergence)**.** Suppose the assumptions of Theorem 1 hold. Suppose further that $\text{var}(Y\mathbb{1}(X=x)) < \infty$ for each $x \in \{0,1\}$. Then

$$\sqrt{n}\left(\widehat{\overline{\text{ATT}}}^c - \overline{\text{ATT}}^c_\varepsilon, \widehat{\underline{\text{ATT}}}^c - \underline{\text{ATT}}^c_\varepsilon\right) \xrightarrow{d} \mathbf{Z}_{\text{ATT}},$$

where $\mathbf{Z}_{\text{ATT}}$ is a random vector in $\mathbb{R}^2$ whose distribution is characterized in the proof.

Similarly to $\mathbf{Z}_{\text{CATE}}$ and $\mathbf{Z}_{\text{ATE}}$, the asymptotic distribution of our ATT bound estimators is a linear combination of Gaussian and non-Gaussian random variables.

5 Bootstrap Inference

We now show how to conduct inference on our bounds for CATE, ATE, and ATT. Earlier we noted that these bounds are generally not ordinary Hadamard differentiable mappings of the underlying parameters θ_0. By Corollary 3.1 in Fang and Santos (2019), this implies that standard bootstrap approaches cannot be used for these bounds. We instead use the non-standard bootstrap approach developed by Fang and Santos (2019). For brevity we focus on ATE and ATT in this section. We provide analogous results for CQTE and CATE in lemmas S3 and S4 in appendix G. In appendix I we discuss one special case where standard bootstrap inference is also valid. In this section we conduct inference on the identified set, or bound functions, because sensitivity analyses ask how the identified set changes as c changes, rather than focusing on a single identified set or parameter. For inference on the parameter itself, our results will be valid but potentially conservative. For any fixed c, our results in section 4 can likely be used to do more powerful inference on the parameters themselves, although we do not develop this here.

5.1 Inference on Potential Outcome Means

The bounds for ATE and ATT can be written in terms of bounds on $\mathbb{E}(Y_x)$. In this section we describe how to do inference on bounds for these means. We then use these results to do inference on our ATE bounds in the next subsection. Recall that our bounds on $\mathbb{E}(Y_x)$ can be written as a functional of θ. This functional is Hadamard directionally differentiable in θ, but it is generally not ordinary Hadamard differentiable. Theorem 3.1 of Fang and Santos (2019) shows how to do bootstrap inference by consistently estimating the Hadamard directional derivative (HDD). This can be done by using analytical estimators or by using a numerical derivative as described in Hong and Li (2018). Here we use analytical estimates of the HDD. This approach explicitly uses the functional form of the HDD to estimate it. It allows us to avoid picking the numerical derivative step size, although other tuning parameters are used to estimate the HDDs analytically.

Setup

We first define some general notation. Let $Z_i = (Y_i, X_i, W_i)$ and $Z^n = \{Z_1, \ldots, Z_n\}$. Let ϑ_0 denote a parameter and let $\widehat{\vartheta}$ be an estimator of ϑ_0 based on the data Z^n. Let $\mathbf{A}_n^*$ denote $\sqrt{n}(\widehat{\vartheta}^* - \widehat{\vartheta})$ where $\widehat{\vartheta}^*$ is a draw from the nonparametric bootstrap distribution of $\widehat{\vartheta}$. Suppose $\mathbf{A}$ is the tight limiting process of $\sqrt{n}(\widehat{\vartheta} - \vartheta_0)$. Denote bootstrap consistency by $\mathbf{A}_n^* \overset{P}{\rightsquigarrow} \mathbf{A}$ where $\overset{P}{\rightsquigarrow}$ denotes weak convergence in probability, conditional on the data Z^n. Weak convergence in probability conditional on Z^n is defined as

$$\sup_{h \in \mathrm{BL}_1} |\mathbb{E}[h(\mathbf{A}_n^*) \mid Z^n] - \mathbb{E}[h(\mathbf{A})]| = o_p(1)$$

where BL_1 denotes the set of Lipschitz functions into $\mathbb{R}$ with Lipschitz constant no greater than 1. We leave the domain of these functions and its associated norm implicit.

We focus on the choices $\vartheta_0 = \theta_0$ and $\widehat{\vartheta} = \widehat{\theta}$. For these choices, let $\mathbf{Z}_n^* = \sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$. Let $\mathbf{Z}_1$ denote the limiting distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$; see Lemma S1 in appendix D. Theorem 3.6.1 of van der Vaart and Wellner (1996) implies that $\mathbf{Z}_n^* \overset{P}{\rightsquigarrow} \mathbf{Z}_1$. Our parameters of interest are all functionals of θ_0. In particular, in section 4 we showed that $\sqrt{n}(\Gamma(\widehat{\theta}) - \Gamma(\theta_0)) \xrightarrow{d} \Gamma'_{\theta_0}(\mathbf{Z}_1)$ for a variety of functionals Γ. To do inference on these functionals, we therefore want to estimate the distribution of $\Gamma'_{\theta_0}(\mathbf{Z}_1)$. Fang and Santos (2019) show that $\widehat{\Gamma}'_{\theta_0}(\mathbf{Z}_n^*) \overset{P}{\rightsquigarrow} \Gamma'_{\theta_0}(\mathbf{Z}_1)$, where $\widehat{\Gamma}'_{\theta_0}$ is an estimator of the Hadamard directional derivative Γ'_{θ_0} satisfying their regularity conditions. In this section we construct such estimators $\widehat{\Gamma}'_{\theta_0}$ and show that they can be used in this bootstrap. We focus specific functionals, described below, which represent upper and lower bound functions. For different indices $j \geq 1$, we gather these bounds into a vector $\Gamma_j = (\overline{\Gamma}_j, \underline{\Gamma}_j)$.

Main Result

Next, consider the asymptotic expansion in equation (S2) on page 11 of the appendix. As we will show, the second term in this expansion can be approximated using standard bootstrap approaches and replacing θ_0 by $\widehat{\theta}$. The first term requires estimating the HDDs $\overline{\Gamma}'_{3,\theta_0}$ and $\underline{\Gamma}'_{3,\theta_0}$. The formulas for our estimators of these HDDs are long, and so we describe them in appendix E. Denote these estimators by $\widehat{\overline{\Gamma}}'_{3,\theta_0}$ and $\widehat{\underline{\Gamma}}'_{3,\theta_0}$. They require choosing two scalar tuning parameters, κ_n and η_n. κ_n is a slackness parameter and η_n is a step size parameter used to compute numerical derivatives of $\widehat{\gamma}(\cdot)$. Although not used in our proof, the asymptotic independence of the two components implies that approximating their respective marginal distributions is sufficient to obtain their joint distribution.

As we just mentioned, we will use the standard nonparametric bootstrap to approximate the second term of equation (S2) in appendix D. To formalize this, let $\mathbb{G}_n^*$ denote the nonparametric bootstrap empirical process:

$$\mathbb{G}_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (M_{n,i} - 1)\delta_{Z_i}$$

where $(M_{n,1}, \ldots, M_{n,n})$ are multinomially distributed with parameters $(1/n, \ldots, 1/n)$ independently of Z^n, and where δ_{Z_i} is a distribution which assigns probability one to the value $Z_i \in \mathbb{R}^{2+d_W}$. Then for any function g,

$$\mathbb{G}_n^* g(Z) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} g(Z_i^*) - \overline{g(Z)} \right)$$

where Z_i^*, $i = 1, \ldots, n$ are drawn independently with replacement from $\{Z_1, \ldots, Z_n\}$ and $\overline{g(Z)} = \frac{1}{n}\sum_{i=1}^{n} g(Z_i)$. In particular, we will study the asymptotic distribution of

$$\mathbb{G}_n^* \Gamma_2(x, W, \widehat{\theta}) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} \Gamma_2(x, W_i^*, \widehat{\theta}) - \frac{1}{n} \sum_{i=1}^{n} \Gamma_2(x, W_i, \widehat{\theta}) \right).$$

The following proposition is our main bootstrap consistency result.

Proposition 3 (Analytical Bootstrap for Means of Y_x)**.** Suppose the assumptions of Theorem 1 hold. Let $\kappa_n \to 0$, $n\kappa_n^2 \to \infty$, $\eta_n \to 0$, and $n\eta_n^2 \to \infty$ as $n \to \infty$. Then

$$\widehat{\Gamma}'_{3,\theta_0}(x, \sqrt{n}(\widehat{\theta}^* - \widehat{\theta})) + \mathbb{G}_n^* \Gamma_2(x, W, \widehat{\theta}) \overset{P}{\rightsquigarrow} \mathbf{Z}_4(x),$$

where $\mathbf{Z}_4(\cdot)$ is the limiting process of the expression given in equation (S2) in appendix D, as characterized in the proof.

This result shows how to use the bootstrap to approximate the joint limiting distribution of upper and lower bounds of $\mathbb{E}(Y_x)$, $x \in \{0, 1\}$. As we show in section 5.2 below, these approximations can be used to conduct pointwise or uniform-in-c inference on the ATE bounds. As part of the proof, we show that $\widehat{\Gamma}'_{3,\theta_0}(x, \sqrt{n}(\widehat{\theta}^* - \widehat{\theta}))$ weakly converges in probability conditional on the data to $\Gamma'_{3,\theta_0}(x, \mathbf{Z}_1)$, a non-Gaussian vector which reflects the sample variation in the first step estimators. We also show weak convergence in probability conditional on the data of $\mathbb{G}_n^* \Gamma_2(x, W, \widehat{\theta})$ to $\mathbb{G}\Gamma_2(x, W, \theta_0) \sim \mathcal{N}(0, \text{var}(\Gamma_2(x, W, \theta_0)))$, a bivariate Gaussian vector which reflects the variation of the CATE bounds over W. This variation can be approximated using the standard nonparametric bootstrap. Hence the bounds' limiting distribution is approximated by a combination of standard and nonstandard bootstraps. Note that the two bootstrap distributions can be computed from a unique sequence of draws Z_i^* and so has the same computational burden as a single bootstrap.

5.2 Inference on the ATE Bounds

Next we show how to use Proposition 3 to do inference on our ATE bounds, $[\underline{\text{ATE}}_\varepsilon^c, \overline{\text{ATE}}_\varepsilon^c]$. In this section we consider inference pointwise in c. We construct confidence bands that are uniform over c in appendix F.

An immediate corollary of Proposition 3 is

$$\sqrt{n} \begin{pmatrix} \widehat{\overline{\text{ATE}}}^{c,*} - \widehat{\overline{\text{ATE}}}^{c} \\ \widehat{\underline{\text{ATE}}}^{c,*} - \widehat{\underline{\text{ATE}}}^{c} \end{pmatrix} \tag{8}$$

$$
\begin{aligned}
&= \begin{pmatrix} \left(\widehat{\overline{\Gamma}}'_{3,\theta_0}(1, \sqrt{n}(\widehat{\theta}^* - \widehat{\theta})) + \mathbb{G}_n^* \overline{\Gamma}_2(1, W, \widehat{\theta})\right) - \left(\widehat{\underline{\Gamma}}'_{3,\theta_0}(0, \sqrt{n}(\widehat{\theta}^* - \widehat{\theta})) + \mathbb{G}_n^* \underline{\Gamma}_2(0, W, \widehat{\theta})\right) \\ \left(\widehat{\underline{\Gamma}}'_{3,\theta_0}(1, \sqrt{n}(\widehat{\theta}^* - \widehat{\theta})) + \mathbb{G}_n^* \underline{\Gamma}_2(1, W, \widehat{\theta})\right) - \left(\widehat{\overline{\Gamma}}'_{3,\theta_0}(0, \sqrt{n}(\widehat{\theta}^* - \widehat{\theta})) + \mathbb{G}_n^* \overline{\Gamma}_2(0, W, \widehat{\theta})\right) \end{pmatrix} \\
&\overset{P}{\rightsquigarrow} \mathbf{Z}_{\text{ATE}}.
\end{aligned}
$$

Thus we can also use this specific bootstrap to approximate the asymptotic distribution of our ATE bounds estimators. Given this result, we can construct a $100(1-\alpha)\%$ confidence set for the ATE identified set under c-dependence as follows. Let

$$
\text{CI}^c_{\text{ATE}}(1-\alpha) = \left[\widehat{\underline{\text{ATE}}}^c - \frac{\widehat{d}_\alpha}{\sqrt{n}}, \widehat{\overline{\text{ATE}}}^c + \frac{\widehat{d}_\alpha}{\sqrt{n}}\right]
$$

where $\widehat{d}_\alpha = \inf\left\{z \in \mathbb{R} : \mathbb{P}\left(\sqrt{n}(\widehat{\underline{\text{ATE}}}^{c,*} - \widehat{\underline{\text{ATE}}}^c) \leq -z, \sqrt{n}(\widehat{\overline{\text{ATE}}}^{c,*} - \widehat{\overline{\text{ATE}}}^c) \geq z \mid Z^n\right) \geq 1-\alpha\right\}.$ Here we consider a symmetric interval for simplicity, but our results can also be used to construct asymmetric intervals. The probability in the expression for $\widehat{d}_\alpha$ can be approximated by taking a large number of bootstrap draws according to equation (8). Proposition 3 then implies that

$$
\liminf_{n\to\infty} \mathbb{P}\Big(\text{CI}^c_{\text{ATE}}(1-\alpha) \supseteq [\underline{\text{ATE}}^c_\varepsilon, \overline{\text{ATE}}^c_\varepsilon]\Big) \geq 1-\alpha.
$$

Let $d_\alpha = \inf\left\{z \in \mathbb{R} : \mathbb{P}\left(\mathbf{Z}^{(2)}_{\text{ATE}} \leq -z, \mathbf{Z}^{(1)}_{\text{ATE}} \geq z\right) \geq 1-\alpha\right\}$. If $\mathbb{P}(\mathbf{Z}^{(2)}_{\text{ATE}} \leq -z, \mathbf{Z}^{(1)}_{\text{ATE}} \geq z)$ is continuous and strictly increasing in a neighborhood of d_α, Corollary 3.2 in Fang and Santos (2015) yields $\widehat{d}_\alpha = d_\alpha + o_p(1)$ and hence

$$
\lim_{n\to\infty} \mathbb{P}\Big(\text{CI}^c_{\text{ATE}}(1-\alpha) \supseteq [\underline{\text{ATE}}^c_\varepsilon, \overline{\text{ATE}}^c_\varepsilon]\Big) = 1-\alpha.
$$

We show the analogous result for the CQTE, CATE, and ATT bounds in appendix G.

6 Illustration: Costs and Benefits of Supported Work

In this section we illustrate our methods using data on the National Supported Work (NSW) demonstration project (MDRC 1980) studied by LaLonde (1986). Since this is a highly studied and well-known program, we only briefly summarize it here. See, for example, Heckman, LaLonde, and Smith (1999) for further details. We use LaLonde's data as reconstructed by Dehejia and Wahba (1999).

The NSW demonstration project randomly assigned participants to either receive a guaranteed job for 9 to 18 months along with frequent counselor meetings (treatment group) or to be left in the labor market by themselves (control group). We use the Dehejia and Wahba (1999) sample, which are all males in LaLonde's NSW dataset and where earnings are observed in 1974, 1975, and 1978. This dataset has 445 people: 185 in the treatment group and 260 in the control group. Like Imbens (2003), we use this experimental sample primarily as an illustration; in experiments where treatment was truly randomized it is not necessary to assess sensitivity to unconfoundedness. Our results may be useful for assessing the impact of randomization failure in experiments, but that is not our focus here.

In addition to this experimental sample, we construct a sample using observational data. This sample combines the 185 people in the NSW treatment group with 2490 people in a control group constructed from the Panel Study of Income Dynamics (PSID). This control group, called PSID-1 by LaLonde, consists of all male household heads observed in all years between 1975 and 1978 who were less than 55 years old and who did not classify themselves as retired. We further drop observations with earnings above $5,000 in 1974, 1975, or both. This leaves 148 treated units (out of 185) and 242 untreated units (out of 2490). This observational sample was also considered by Imbens (2003).

The outcome of interest is earnings in 1978. There are also nine covariates: Earnings in 1974, earnings in 1975, years of education, age, indicators for race (Black, Hispanic, other), an indicator for marriage, an indicator for having a high school degree, and an indicator for treatment. All earnings variables are measured in 1982 dollars. Table S2 in appendix J, which closely follows table 1 of Imbens (2003), provides summary statistics.

Table 1: Baseline treatment effect estimates (in 1982 dollars).

	ATE	ATT	Sample size
Experimental dataset	1633	1738	445
	(650)	(689)	
Observational dataset	3337	4001	390
	(769)	(762)	

Standard errors in parentheses.

Baseline Estimates

Table 1 shows the baseline point estimates of both ATE and ATT under the unconfoundedness assumption in the two samples we consider. These estimates are all computed by inverse probability weighting (IPW) using a parametric logit propensity score estimator. We do not consider other estimators, since our goal is to illustrate sensitivity to identifying assumptions, rather than finite sample sensitivity to the choice of estimator.

Assessing unconfoundedness

Is unconfoundedness reasonable in this setting? There are at least three reasons to question it. First, as discussed in section 6 of Heckman et al. (1999), in observational data economic theory typically suggests that prospective job training participants will select into training based on some objective function that is related to their potential outcomes. This behavior creates a classic selection bias problem that would violate unconfoundedness.

Second, we can examine balance in the observed covariates. From table S2 we see that almost all of the observed covariates are highly unbalanced across the observational control group and the treatment group. For example, the treated units are much less likely to be married (19% vs 78%), are 13 years younger on average, have one year less education on average, are much more likely to be Black (84% vs 27%), and had more than twice as much earnings in 1974 and 1975. These findings do not violate unconfoundedness, as it is designed to allow for imbalances in the observed covariates. However, it is perhaps implausible to think that the unobserved potential outcomes are perfectly balanced across the two groups, even after adjusting for covariates, when we see such extreme imbalances in the observed covariates.

Third, in this dataset, we can directly check unconfoundedness because we also have experimental data. Indeed, this was part of the original motivation for LaLonde's (1986) study. From table 1 we see that the baseline point estimates are substantially different, suggesting that unconfoundedness does not hold. One caveat to this interpretation, which has been studied extensively in the literature following LaLonde (1986) (e.g., Dehejia and Wahba 1999), is that this comparison conflates nonparametric unconfoundedness with the parametric functional forms used in estimation. Regardless, in most observational studies we do not have experimental data. Instead, the evidence we have presented here that unconfoundedness is unlikely to hold exactly motivates the importance of formal sensitivity analyses. We perform that analysis next.

Relaxing Unconfoundedness

Figure 1 shows our main sensitivity analysis results. These are estimated treatment effect bounds under c-dependence, along with corresponding pointwise confidence bands, as described in sections 2–5. The left plot shows bounds on ATE while the right plot shows bounds on ATT. The solid lines are bounds for the observational dataset while the dashed lines are bounds for the experimental dataset. The light dotted lines are confidence bands for the observational dataset while the light dashed-dotted lines are confidence bands for the experimental dataset. These bands are constructed to have nominal 95% coverage probability pointwise in c based on our non-standard bootstrap results in section 5. For the tuning parameters we use $\varepsilon = 0.05$, and the rule of thumb choice $\eta_n = \kappa_n = \min\{0.05n^{-1/3}, 0.05\}$ that we discuss in appendix H. For each dataset, the point estimates took 16 seconds to compute while the confidence bands took 1.5 hours to compute using 12 cores with

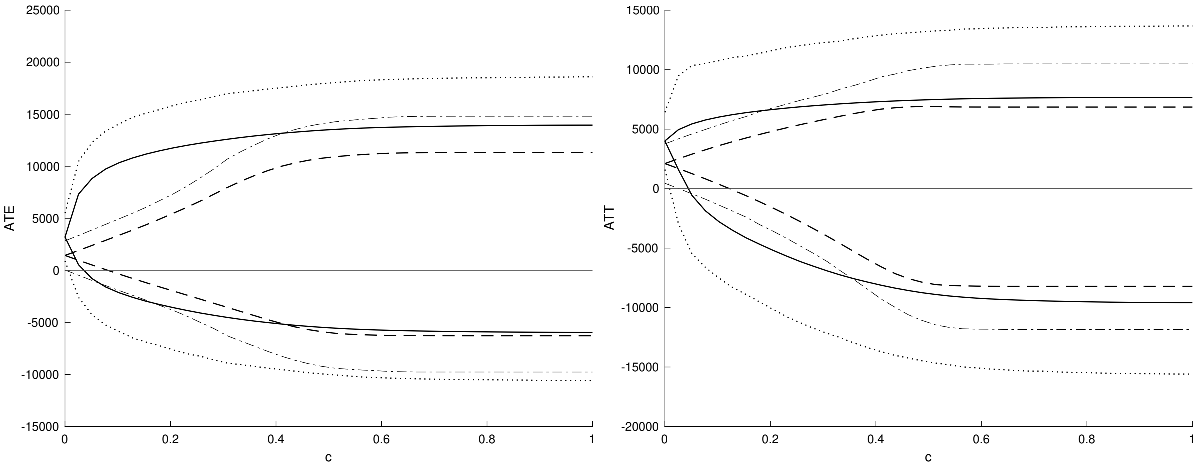

Figure 1: Sensitivity of ATE (left) and ATT (right) estimates to relaxations of the selection on observables assumption. The solid lines are bounds computed using the observational dataset while the dashed lines are bounds computed using the experimental dataset. The light dotted lines are confidence bands for the observational dataset while the light dashed-dotted lines are confidence bands for the experimental dataset.

$B = 1000$ bootstrap draws.

For both datasets, at $c = 0$ the bounds collapse to the baseline point estimate. When $c > 0$, we allow for some selection on unobservables. Comparing the shape of the bounds for both datasets we see that the experimental data are substantially more robust to relaxations from the baseline assumptions than the observational data. Specifically, for most values of c the bounds for the experimental data are substantially tighter than the bounds for the observational data. Even the no assumptions bounds ($c = 1$) are tighter for the experimental data than for the observational data.

A second way to measure robustness uses *breakdown points*. Masten and Poirier (2020) discuss these in detail and give additional references. In the current context, the breakdown point is simply the largest value of c such that we can no longer draw a specific conclusion about some parameter. Specifically, in the next two subsections we consider two conclusions: The conclusion that ATE is nonnegative, and the conclusion that ATT is less than the per participant program cost.

Sign Change Breakdown Points for ATE

The baseline ATE estimate is positive. If unconfoundedness fails, however, it could be that the true ATE is negative. So in this section we ask: How much do we need to relax unconfoundedness before the data and assumptions are consistent with a negative ATE? If this breakdown point is large, meaning that we need a strong amount of unobserved selection to flip the sign of ATE, then researchers can be more confident in their conclusion that the true ATE is positive. In the experimental dataset, the estimated breakdown point is 0.082. This is the value of c such that the lower bound function in figure 1 intersects the horizontal axis. For all $c \leq 0.082$, the estimated identified sets for ATE only contain nonnegative values. For $c > 0.082$, the estimated identified sets contain both positive and negative values. Hence, for such relaxations of unconfoundedness, we cannot be sure that the average treatment effect is positive.

For the observational dataset, the estimated breakdown point for the conclusion that ATE is nonnegative is 0.037. This is less than half the breakdown point for the experimental dataset. Hence again we see that conclusions about ATE from the experimental dataset are substantially more robust than the observational dataset. The same conclusion holds for ATT: The point estimates in both datasets suggest that it is positive. But how robust is that conclusion? The estimated breakdown point for the conclusion that ATT is nonnegative in the experimental data is 0.123 while it is 0.049 for the observational dataset. By this measure, the conclusion that ATT is positive is more than twice as robust using the experimental data compared to the observational data.

Thus far we have compared the robustness of results obtained from the experimental data with results obtained from the observational data. Next we discuss whether either of these results is robust in an absolute sense. To do this, we use the leave-out-variable k propensity score analysis discussed in section 2.

Table 2: Variation in leave-out-variable-k propensity scores, experimental data.

	p50	p75	p90	$\bar{c}_k$
Earnings in 1975	0.001	0.004	0.008	0.053
Black	0.007	0.009	0.014	0.082
Positive earnings in 1974	0.002	0.010	0.018	0.034
Education	0.012	0.022	0.031	0.087
Married	0.006	0.012	0.032	0.042
Age	0.015	0.024	0.034	0.099
Earnings in 1974	0.002	0.011	0.035	0.209
Positive earnings in 1975	0.013	0.017	0.062	0.082
Hispanic	0.007	0.017	0.099	0.124

First consider table 2, which uses data from the experimental sample. For each variable k, listed in the rows of this table, we compute four summary statistics from the estimated distribution of $\Delta_k = |p_{1|W}(W_{-k}, W_k) - p_{1|W_{-k}}(W_{-k})|$. Specifically, we estimate the 50th, 75th, and 90th percentiles of Δ_k, along with the maximum observed value, denoted $\bar{c}_k$. As discussed in section 2, these quantities tell us about the marginal impact of covariate k on treatment assignment. c-dependence constrains the maximum value of the marginal impact of the *unobserved* potential outcome on treatment assignment, above and beyond the *observed* covariates. Thus the values in table 2 can help us calibrate c. Specifically, we will compare the breakdown point to the values in this table. These values could be interpreted as upper bounds on the magnitude of selection on unobservables that we might think is present. Thus, for a given reference value from this table, if the breakdown point is larger than the reference value, we could consider the conclusion of interest to be robust to failure of unconfoundedness. In contrast, if the breakdown point is smaller than the reference value, we could consider the conclusion of interest to be sensitive to failure of unconfoundedness.

Recall that the estimated breakdown point for the conclusion that ATE is nonnegative is 0.082. This is larger than three of the $\bar{c}_k$ values and on the same order of magnitude as four more. If we look at a less stringent comparison, the 90th percentile, we see that the

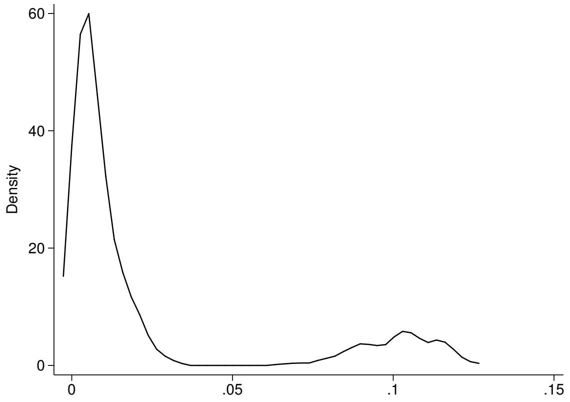

Figure 2: Kernel density estimate of Δ_k, the absolute difference between propensity score and leave-out-variable k propensity score, for k = Hispanic indicator, in the experimental dataset.

estimated breakdown point is now larger than all but one of the rows, corresponding to the indicator for Hispanic. We now examine this variable more closely. Figure 2 plots the density Δ_k for k = Hispanic indicator. Here we see that there is a small proportion of mass who have values larger than 0.082, but most people have values well below the breakdown point. Next suppose we weaken the criterion even more by considering the 75th percentile column in table 2. The breakdown point is larger than all values in this column.

The leave-out-variable k propensity score analysis focuses on the relationship between observed covariates and treatment assignment. It does not use data on outcomes. A less conservative analysis is to only worry about covariates k which have large values in table 2 *and* which also affect our outcomes in some way. Specifically, we next consider leave-out-variable k IPW estimates of ATE under the baseline unconfoundedness assumption. Table 3 shows the effect of leaving out a single variable on the ATE point estimates for both datasets. Continue to consider just the experimental dataset. Here we first see that omitting any covariate at most changes the point estimate by 5.4%. Moreover, recall the main variable we were concerned about before: the indicator for Hispanic. Omitting this variable only changes the ATE point estimate by 1.5%. Therefore, following the discussion on page 11, it may be reasonable to omit this variable when using the values in table 2

Table 3: Magnitude of the effect of omitting a single variable on ATE point estimates (as a percentage of the baseline estimate).

	Experimental dataset	Observational dataset
Earnings in 1975	0.07	0.02
Married	0.21	14.27
Positive earnings in 1974	1.35	10.20
Hispanic	1.51	1.01
Black	2.91	14.11
Positive earnings in 1975	3.32	0.64
Age	3.36	6.49
Earnings in 1974	3.90	0.34
Education	5.39	1.84

to calibrate the magnitude of the breakdown point. Using this less conservative definition of robustness, the estimated breakdown point is now larger than all of the remaining 90th percentiles in table 2, suggesting that the conclusion is robust.

Overall, the leave-out-variable k analysis suggests that, on an absolute scale, the conclusion that ATE is nonnegative using the experimental data is quite robust. A similar analysis applies to conclusions about ATT.

Next consider the observational data. Table 4 shows the leave-out-variable k propensity score analysis. Recall that the estimated breakdown point for the conclusion that ATE is nonnegative in the observational dataset is 0.037. By *any* of these measures the conclusion that ATE is nonnegative is not robust. Suppose we only consider variables which also substantially change the point estimates, as shown in table 3. Even then we still find that the results are sensitive. For example, the indicator for Black changes the ATE point estimate by 14% and also has substantial marginal impact on the propensity score, with its 50th percentile in table 4 about 1.5 times as large as the estimated ATE breakdown point. Thus, using these as absolute measures of robustness, we find that the conclusion that ATE is positive using the observational data is not robust.

In contrast, Imbens (2003) found that the same observational dataset yields relatively

Table 4: Variation in leave-out-variable-k propensity scores, observational data.

	p50	p75	p90	$\bar{c}_k$
Earnings in 1974	0.000	0.001	0.009	0.065
Hispanic	0.003	0.011	0.024	0.214
Education	0.006	0.017	0.042	0.127
Earnings in 1975	0.002	0.010	0.057	0.276
Positive earnings in 1975	0.007	0.019	0.076	0.295
Positive earnings in 1974	0.012	0.028	0.099	0.423
Married	0.028	0.079	0.172	0.314
Age	0.035	0.093	0.205	0.508
Black	0.053	0.143	0.266	0.477

robust results. Imbens' analysis relied importantly on fully parametric assumptions about the joint distribution of the observables and unobservables. In particular, he assumed outcomes were normally distributed, that treatment effects are homogeneous, and that any selection on unobservables arises due to an omitted binary variable. Our identification analysis does not require any of these assumptions. As discussed in section 3, we do impose some parametric assumptions to simplify estimation, but even these assumptions are substantially weaker than those used by Imbens. Since we are making weaker auxiliary assumptions, it is not surprising that our analysis shows the findings to be more sensitive than the analysis in Imbens (2003). However, even with these weaker assumptions, our analysis shows that the experimental dataset results are robust. So for that dataset our conclusion matches Imbens' who also found that the experimental dataset results are robust.

Can Selection Help the Program Pass a Cost-Benefit Analysis?

In the previous subsection we studied the sensitivity of the conclusion that the ATE is nonnegative. In practice, however, this is not necessarily the most policy relevant conclusion. For example, Heckman and Smith (1998, section 3) give a model where the socially optimal decision whether to continue a small scale program or to shut it down can be computed by comparing the ATT with the program's per participant cost. In this subsection we show how our sensitivity analysis can be used in these kinds of cost-benefit analyses. Specifically, we consider the conclusion that the ATT is less than the per participant program cost. Under the model in Heckman and Smith (1998), the program should be shut down when this conclusion holds.

Chapter 8 of MDRC (1980) reports NSW per participant program costs. For males, total costs ranges between \$4,637 and \$5,218 (tables 8-2, 8-3, and 8-4) in 1976 dollars. Our treatment effect estimates are in 1982 dollars. Adjusting these reported costs to 1982 dollars (CPI-U series) gives a range of \$7,865 to \$8,850.

First consider the experimental dataset. The conclusion of interest holds for the baseline estimate: The ATT of \$1,738 is far less than the per participant cost. Suppose, however, that a supporter of the program claims that this baseline estimate is implausible due to selection on unobservables. How strong does selection on unobservables need to be to allow for the possibility that the program is cost effective? Formally, what is the smallest c such that the identified set for ATT includes values that are larger than the per participant cost? If we look at the bounds' point estimates, there are *no* values of c under which the program is cost effective. Accounting for sampling uncertainty by examining the confidence bands, we need c to be at least about 0.5 before it is possible that the program is cost effective. As we argued earlier, these are very large values, so it is unlikely that selection on unobservables is this strong.

Next consider the observational dataset. Here again the conclusion of interest holds for the baseline estimate: The ATT of \$4,001 is smaller than the per participant cost. Next consider the breakdown point for this conclusion. Since the bounds for the observational dataset are larger than those for the experimental dataset, we need less selection on unobservables to allow for possibly large values of the ATT. Despite this, there are still no values of c under which the program is cost effective, based on the bounds' point estimates. This is largely because the uncertainty due to the impact of selection on unobservables is asymmetric in this example: The lower bound varies much faster in c than the upper bound does. Hence conclusions about the largest possible value of the ATT are more robust to relaxations of unconfoundedness than conclusions about the smallest possible value of the ATT. If we account for sampling uncertainty by examining the confidence bands, then we need c to be at least about 0.08 before the confidence intervals contain ATT values larger than the per participant costs. This is a relatively large value, although it is smaller than a decent number of the leave-out-variable k propensity score values in table 4. That, however, likely just reflects the large amount of sampling uncertainty in this data.

Overall, our findings suggest that the program does not pass a cost-benefit analysis, even if we allow for a large amount of selection on unobservables. Hence the conclusion that the ATT is less than the per participant cost—and so the program should be shut down—is quite robust to failures of unconfoundedness.

Finally, note that our analysis here is primarily illustrative. A more comprehensive cost-benefit analysis would require examining many other program outcomes besides just short run post-program earnings. For example, see the analysis in chapter 8 of MDRC (1980) and section 10 of Heckman et al. (1999). Note, however, that given data on these additional outcomes, our methods could then be used to analyze the sensitivity of total program impacts to failures of unconfoundedness.

7 Conclusion

Identification, estimation, and inference on treatment effects under unconfoundedness has been widely studied and applied. This approach uses two assumptions: unconfoundedness and overlap. The overlap assumption is refutable, and many tools have been developed for checking this assumption in practice. In this paper, we provide a suite of tools for assessing the unconfoundedness assumption. There are two key distinctions between our results and the previous literature. First, we begin from fully nonparametric bounds. In contrast, most of the previous literature that takes a super-population perspective relies on parametric assumptions for their identification analysis. Second, we provide tools for inference. This is important because, just like baseline estimators, sensitivity analyses are also subject to sampling uncertainty.

Acknowledgments

This paper was presented at the 2018 Western Economic Association International Conference, the 2019 Stata Conference Chicago, the 2020 World Congress of the Econometric Society, the DC-MD-VA Econometrics Workshop 2020, University of Southern California, University of Toronto, the 2020 SEA Conference, and the 2021 IAAE Conference. We thank participants at those seminars and conferences, as well as Karim Chalak, Toru Kitagawa, and John Pepper. We thank Paul Diegert for excellent research assistance. The accompanying Stata module `tesensitivity` is available on the SSC—type `ssc install tesensitivity` from within Stata—or via Github at `https://github.com/mattmasten/tesensitivity`.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Masten thanks the National Science Foundation for research support under Grant No. 1943138.

References

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, 113, 151–184.

——— (2008): "Using selection on observed variables to assess bias from unobservables when evaluating swan-ganz catheterization," *American Economic Review P&P*, 98, 345–350.

Athey, S. and G. W. Imbens (2017): "The state of applied econometrics: Causality and policy evaluation," *Journal of Economic Perspectives*, 31, 3–32.

Caliendo, M. and S. Kopeinig (2008): "Some practical guidance for the implementation of propensity score matching," *Journal of Economic Surveys*, 22, 31–72.

Chernozhukov, V., I. Fernández-Val, and T. Kaji (2017): "Extremal quantile regression," *Handbook of Quantile Regression*.

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, 39–67.

Dehejia, R. H. and S. Wahba (1999): "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American Statistical Association*, 94, 1053–1062.

Fang, Z. and A. Santos (2015): "Inference on directionally differentiable functions," *Working paper.*

——— (2019): "Inference on directionally differentiable functions," *The Review of Economic Studies*, 86, 377–412.

Fogarty, C. B. (2020): "Studentized sensitivity analysis for the sample average treatment effect in paired observational studies," *Journal of the American Statistical Association*, 115, 1518–1530.

Fogarty, C. B., P. Shi, M. E. Mikkelsen, and D. S. Small (2017): "Randomization inference and sensitivity analysis for composite null hypotheses with binary outcomes in matched observational studies," *Journal of the American Statistical Association*, 112, 321–331.

Fogarty, C. B. and D. S. Small (2016): "Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming," *Journal of the American Statistical Association*, 111, 1820–1830.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999): "The economics and econometrics of active labor market programs," *Handbook of Labor Economics*, 3, 1865–2097.

Heckman, J. J. and J. Smith (1998): "Evaluating the welfare state," in *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*, Cambridge University Press, 31, 241.

Hong, H. and J. Li (2018): "The numerical delta method," *Journal of Econometrics*, 206, 379–394.

Hosman, C. A., B. B. Hansen, and P. W. Holland (2010): "The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder," *The Annals of Applied Statistics*, 4, 849–870.

Ichino, A., F. Mealli, and T. Nannicini (2008): "From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?" *Journal of Applied Econometrics*, 23, 305–327.

Imbens, G. W. (2003): "Sensitivity to exogeneity assumptions in program evaluation," *American Economic Review P&P*, 126–132.

——— (2004): "Nonparametric estimation of average treatment effects under exogeneity: A review," *The Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W. and D. B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Imbens, G. W. and J. M. Wooldridge (2009): "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, 47, 5–86.

Kallus, N., X. Mao, and A. Zhou (2019): "Interval estimation of individual-level causal effects under unobserved confounding," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2281–2290.

Krauth, B. (2016): "Bounding a linear causal effect using relative correlation restrictions," *Journal of Econometric Methods*, 5, 117–141.

LaLonde, R. J. (1986): "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review*, 76, 604–620.

Manpower Demonstration Research Corporation (MDRC) (1980): *Summary and Findings of the National Supported Work Demonstration*, Ballinger Publishing Company.

Manski, C. F. (1990): "Nonparametric bounds on treatment effects," *American Economic Review P&P*, 80, 319–323.

Masten, M. A. and A. Poirier (2018): "Identification of treatment effects under conditional partial independence," *Econometrica*, 86, 317–351.

——— (2020): "Inference on breakdown frontiers," *Quantitative Economics*, 11, 41–111.

Oster, E. (2019): "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 37, 187–204.

Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (2000): "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models," in *Statistical models in epidemiology, the environment, and clinical trials*, Springer, 1–94.

Rosenbaum, P. R. (1984): "Conditional permutation tests and the propensity score in observational studies," *Journal of the American Statistical Association*, 79, 565–574.

——— (1987): "Sensitivity analysis for certain permutation inferences in matched observational studies," *Biometrika*, 74, 13–26.

——— (1988): "Sensitivity analysis for matching with multiple controls," *Biometrika*, 75, 577–581.

——— (1991): "Sensitivity analysis for matched case-control studies," *Biometrics*, 87–100.

——— (1995): *Observational Studies*, Springer.

——— (2018): "Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels," *The Annals of Applied Statistics*, 12, 2312–2334.

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*, 212–218.

van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Science & Business Media.