

Generalized Lee Bounds

Vira Semenova*

November 24, 2021

Abstract

Lee (2009) is a common approach to bound the average causal effect in the presence of selection bias, assuming the treatment effect on selection has the same sign for all subjects. This paper generalizes Lee bounds to allow the sign of this effect to be identified by pre-treatment covariates, relaxing the standard (unconditional) monotonicity to its conditional analog. Asymptotic theory for generalized Lee bounds is proposed in low-dimensional smooth and high-dimensional sparse designs. The paper also generalizes Lee bounds to accommodate multiple outcomes and non-compliance. The estimated bounds achieve nearly point-identification in JobCorps (Lee (2009)), where unconditional monotonicity is shown to fail, and Oregon Health Insurance Experiment (Finkelstein et al. (2012)) empirical applications.

*First version: August, 31, 2020. This version: November, 24, 2021. Email: semenovavira@gmail.com. I am indebted to Victor Chernozhukov, Michael Jansson, Patrick Kline, Anna Mikusheva, Whitney Newey and Demian Pouzo for their patience and encouragement. I am grateful to my discussants Jorg Stoye (Chamberlain seminar), Xiaoxia Shi (Cowles Econometrics Conference), Isaiah Andrews (NBER Summer Institute (Labor Studies)), whose comments have substantially improved the paper relative to the first version. I am also thankful to Alberto Abadie, Chris Ackerman, Stephane Bonhomme, Kirill Borusyak, Sydnee Caldwell, Matias Cattaneo, Colin Cameron, Xiaohong Chen, Denis Chetverikov, Ben Deaner, Mert Demirer, Bulat Gafarov, Dalia Ghanem, Jerry Hausman, Keisuke Hirano, Peter Hull, Tetsuya Kaji, Michal Kolesar, Ivana Komunjer, Kevin Li, Elena Manresa, Eric Mbakop, David McKenzie, Rachael Meager, Francesca Molinari, Ismael Mourifie, Denis Nekipelov, Alexandre Poirier, Geert Ridder, Jonathan Roth, Oles Shtanko, Cory Smith, Sophie Sun, Takuya Ura, Roman Zarate, Edward Vytlačil and participants in numerous seminars and conferences for helpful comments.

1 Introduction

Randomized controlled trials are often complicated by endogenous sample selection and non-response. This problem occurs when treatment affects the researcher’s ability to observe an outcome (a selection effect) in addition to the outcome itself (the causal effect of interest). For example, being randomized into a job training program affects both an individual’s wage and employment status. Since wages exist only for employed individuals, treatment-control wage difference is contaminated by selection bias. A common way to proceed is to bound the average causal effect from above and below, focusing on subjects whose outcomes are observed regardless of treatment receipt (the always-observed principal strata, Frangakis and Rubin (2002) or the always-takers, Lee (2009)).

Seminal work by Lee (2009) proposes nonparametric bounds assuming the selection effect is non-negative for all subjects (monotonicity). For example, if JobCorps cannot deter employment, basic Lee lower bound is the treatment-control difference in wages, where the top wages in the treated group are trimmed until treated and control employment rates are equal. Furthermore, Lee (2009) shows that the covariate density-weighted conditional Lee bound is weakly tighter than the basic bound that does not involve any covariates. However, only a handful of discrete covariates can be utilized to tighten the bound, since each covariate cell is required to have a positive number of treated and control outcomes.

This paper begins by quantifying the importance of covariates under the same (unconditional) monotonicity assumption as in Lee (2009). Assuming the outcome shock is homoscedastic, I show that the bounds’ width is proportional to the shock’s standard deviation. If covariates perfectly explain the outcome, the upper and the lower bounds collapse into a point. Furthermore, in a special case, the bounds are inversely proportional to the variance of conditional probability of selection in the control state. The link between the sharp width and the first-stage predictive power motivates selecting covariates based on their out-of-sample predictive fit, and, therefore, employing modern regularized and machine learning tools (Athey (2015), Mullainathan and Spiess (2017), Athey and Imbens (2019)).

Next, the paper generalizes Lee bounds to accommodate subjects with differential selection response, relaxing unconditional monotonicity to its conditional analog. This step gives rise to

covariate spaces of positive and negative selection response. I show that the sharp conditional Lee bound has continuous transition through the boundary, determined by zero values of conditional treatment effect on selection. If the boundary is continuously distributed with a bounded density, the misclassification bias is negligible under plausible smoothness and/or sparsity assumptions. This property holds only for the **sharp** version of conditional Lee bound, where the same set of covariates is used to classify subjects and to define the Lee bound on each sub-space. The link between sharpness and misclassification robustness appears to be new.

Overcoming classification challenge, the paper represents the generalized Lee bound via a semiparametric moment equation. The moment-based estimator no longer requires covariates to be discrete. Furthermore, it no longer requires the propensity score (i.e., the probability of treatment) to be known, overcoming a key historical limitation to the widespread adoption of Lee bounds in quasi-experiments. If the conditional selection probability and conditional outcome quantile are smooth functions of covariates, they can be estimated by logistic series regression of Hirano et al. (2003) and quantile series of Belloni et al. (2019), respectively. Alternatively, if these functions have a sparse representation with respect to some basis, one could employ their ℓ_1 -penalized analogs proposed in Belloni et al. (2016) and Belloni and Chernozhukov (2011), Belloni et al. (2017) to Neyman-orthogonal (Neyman (1959), Neyman (1979), Ai and Chen (2003), Newey (1994), Chernozhukov et al. (2018), Chernozhukov et al. (2016)) moments, derived in Supplementary Appendix (Semenova (2020)). The paper's theoretical contribution is to establish the validity of the orthogonal moment-based estimator in the presence of classification step.

In the final part of the paper, I generalize Lee bounds to accommodate multiple outcomes and non-compliance. I show that Lee's Identified set is compact and convex and derive its support function $q \rightarrow \sigma(q)$ describing its boundary (Rockafellar (1997), Beresteanu and Molinari (2008), Beresteanu et al. (2011), Bontemps et al. (2012), Chandrasekhar et al. (2012)). Standardized treatment effect (STE) of multiple outcomes in a related domain is an important special case. Indeed, since the standardization vector (i.e., the support function's argument q) is an unknown population parameter, inference using $\hat{\sigma}(\hat{q})$ calls for uniform Gaussian approximation for the support function. Finally, I generalize Lee's trimming strategy to handle non-compliance (Im-

bens and Angrist (1994), Angrist and Imbens (1995)), which gives bounds for the always-takers' local ATE.

I estimate Lee bounds in two empirical applications. First, I study the effect of the JobCorps training program on wages and wage growth, using data from Schochet et al. (2008). The paper's major finding is to show that the unconditional monotonicity fails for JobCorps program. After accounting for the differential JobCorps effect on employment, I find that the average JobCorps effect on the always-takers' week 90 wages is 4.0–4.6%, which is slightly smaller than Lee's replicated estimate of 4.9%. Furthermore, the average JobCorps effect on wage growth from week 104 to week 208 ranges between -11% and 11% . Thus, the average growth rate is 15% in the control status and ranges between 4% and 26% in the treated status. Finally, I provide evidence of mean reversion of the expected log wage for the always-takers in the control status. This mean reversion corroborates Ashenfelter (1978) pattern and shows that earnings would have recovered even without JobCorps training. Therefore, evaluating JobCorps would have been very difficult without a randomized experiment, as one would need to explicitly model mean reversion in the potential wage in the control status.

In the next application, I study the effect of a Medicaid lottery on applicants' self-reported healthcare utilization and health, as in Finkelstein et al. (2012). After accounting for non-response bias, I find that Medicaid exposure and insurance has had a positive effect on all measures of health, confirming Finkelstein et al. (2012)'s baseline results. The proposed Lee bounds attain nearly point-identification in all applications. In contrast, conventional Lee bounds are too wide to determine the direction of the treatment effect. The proposed bounds are straightforward to compute using the R software package `leebounds`, available at <https://github.com/vsemenova/leebounds>.

This paper combines ideas from various branches of economics and statistics, including bounds on causal effects (Manski (1989), Manski (1990), Horowitz and Manski (1995), Frangakis and Rubin (2002), Angrist et al. (2002), Angrist et al. (2006), Feller et al. (2016), Angrist et al. (2013), Abdulkadiroglu et al. (2020), Honore and Hu (2020), Mogstad et al. (2020a), Mogstad et al. (2020b), Kamat (2021)), partial identification (Chernozhukov et al. (2010), Stoye (2009), Stoye (2010), Kaido et al. (2019), Gafarov (2019), Kaido et al. (2021)), monotonicity

and latent index models (Vytlacil (2002), Kline and Walters (2019), Kamat (2019), Sloczynski (2021)), inference on moment inequalities (Andrews and Shi (2017), Andrews and Shi (2013), Bugni et al. (2017), Shi et al. (2018), Chernozhukov et al. (2019), Hsu et al. (2019), Bai et al. (2019), Shi et al. (2021)), and machine learning approaches for heterogenous treatment effects (Athey and Imbens (2016), Wager and Athey (2018), Chernozhukov et al. (2017), Oprescu et al. (2018), Syrgkanis et al. (2019), Nie and Wager (2020), Farrell et al. (2021a), Athey and Wager (2021), Farrell et al. (2021b)). Merging the ideas from classification analysis, debiased/orthogonal inference and bounds literatures, this paper shows how pre-randomization covariates can make Lee bounds more robust and informative at the same time.

The paper is organized as follows. Section 2 reviews basic Lee bounds and Lee’s estimator under the standard monotonicity assumption. Section 3 establishes the link between the sharp width – the width of the sharp bounds – and the first-stage covariate predictive power. Section 4 generalizes Lee bounds under conditional monotonicity and provides inference results for this parameter. Section 5 allows for multiple outcomes and endogenous treatment receipt. Section 6 gives a finite-sample evidence of robustness to misclassification bias and importance of orthogonality. Section 7 presents empirical applications. The Supplementary Appendix (Semenova (2020)) contains proofs (Appendix A) and supplementary results (Appendix B).

2 Lee (2009) bounds

In this section, I review the Lee (2009) sample selection model. Let $D = 1$ be an indicator for treatment receipt. Let $Y(1)$ and $Y(0)$ denote the potential outcomes if an individual is treated or not, respectively. Likewise, let $S(1) = 1$ and $S(0) = 1$ be dummies for whether an individual’s outcome is observed with and without treatment. The data vector $W = (D, X, S, S \cdot Y)$ consists of the treatment status D , a baseline covariate vector X , the selection status $S = D \cdot S(1) + (1 - D) \cdot S(0)$ and the outcome $S \cdot Y = S \cdot (D \cdot Y(1) + (1 - D) \cdot Y(0))$ for selected individuals. Lee (2009) focuses on the average treatment effect (ATE)

$$\beta_0 = \mathbb{E}[Y(1) - Y(0) \mid S(1) = 1, S(0) = 1] \tag{2.1}$$

for subjects who are selected into the sample regardless of treatment receipt—the *always-takers*.

ASSUMPTION 1 (Assumptions of Lee (2009)). *The following statements hold.*

(1) (*Complete Independence*). *The potential outcomes vector $(Y(1), Y(0), S(1), S(0), X)$ is independent of D .*

(2) (*Monotonicity*). $S(1) \geq S(0)$ *a.s.*

The independence assumption holds by random assignment. In addition, it requires all subjects to have the same probability of being treated. The monotonicity requires all subjects to have the same direction of selection response. In particular, a subject that is selected into the sample when untreated must remain selected if treated:

$$S(0) = 1 \quad \Rightarrow \quad S(1) = 1.$$

As a result,

$$\mathbb{E}[Y(0) \mid S(1) = 1, S(0) = 1] = \mathbb{E}[Y(0) \mid S(0) = 1].$$

By complete independence,

$$\mathbb{E}[Y(0) \mid S(0) = 1] = \mathbb{E}[Y \mid S = 1, D = 0],$$

and $\mathbb{E}[Y(0) \mid S(1) = 1, S(0) = 1]$ is point-identified.

In contrast to the control group, a treated outcome can be either an always-taker's outcome or a complier's outcome. The always-takers' share among the treated is

$$p_0 = \Pr[S(1) = 1, S(0) = 1 \mid S(1) = 1] = \Pr[S(0) = 1 \mid S(1) = 1] = \frac{\Pr[S = 1 \mid D = 0]}{\Pr[S = 1 \mid D = 1]}. \quad (2.2)$$

In the best case, the always-takers comprise the top p_0 quantile of the treated outcomes. The largest possible value of β_0 is

$$\beta_U^{\text{basic}} = \mathbb{E}[Y \mid Y \geq Q_{Y \mid S=1, D=1}(1 - p_0), D = 1, S = 1] - \mathbb{E}[Y \mid S = 1, D = 0], \quad (2.3)$$

where $Q_{Y \mid S=1, D=1}(u)$ is the u -quantile of the treated outcomes and p_0 in (2.2) is the trimming

threshold. Likewise, the smallest possible one is

$$\beta_L^{\text{basic}} = \mathbb{E}[Y \mid Y \leq Q_{Y|S=1,D=1}(p_0), D=1, S=1] - \mathbb{E}[Y \mid D=0, S=1].$$

Lee's identification strategy can be implemented conditional on covariates. Denote the conditional trimming threshold $p_0(x)$ as

$$p_0(x) = \frac{\Pr[S=1 \mid D=0, X=x]}{\Pr[S=1 \mid D=1, X=x]} = \frac{s(0,x)}{s(1,x)} \quad x \in \mathcal{X} \quad (2.4)$$

and the conditional upper bound $\beta_U^{\text{basic}}(x)$ as

$$\beta_U^{\text{basic}}(x) = \mathbb{E}[Y \mid D=1, S=1, Y \geq Q^1(1-p_0(x), x), X=x] - \mathbb{E}[Y \mid D=1, S=0, X=x] \quad (2.5)$$

where $Q^1(u, x) := Q_{Y|S=1,D=1,X=x}(u, x)$ is the conditional u -quantile of Y in $S=1, D=1, X=x$ group. The sharp upper bound is

$$\beta_U = \int_{x \in \mathcal{X}} \beta_U^{\text{basic}}(x) f_X(x \mid S=1, D=0) dx, \quad (2.6)$$

which, as Lee has shown, obeys $\beta_U \leq \beta_U^{\text{basic}}$.

3 Sharp width

In this section, I link the sharp width – the width of sharp Lee bounds – to covariate predictive power in outcome and selection equations. The sharp width is

$$\Delta = \int_{\mathcal{X}} (\beta_U^{\text{basic}}(x) - \beta_L^{\text{basic}}(x)) f_X(x \mid S=1, D=0) dx, \quad (3.1)$$

where $f_X(x \mid S=1, D=0)$ is the always-takers' covariate density.

Suppose Assumption 1 holds. In addition, suppose the treated potential outcome's shock is

homoscedastic, that is

$$\text{Var}(Y(1) - \mathbb{E}[Y(1) | X = x]) = \text{Var}(\sigma_\varepsilon \cdot \varepsilon) = \sigma_\varepsilon^2 \quad \forall x. \quad (3.2)$$

For example, if the additive outcome shock in Heckman (1976, 1979) is independent of covariates, (3.2) holds (see Remark A.1 in Appendix) for i.i.d data. I show that Δ is proportional to σ_ε .

Conditional on X , trimming an outcome Y is equivalent to trimming the unobserved shock ε

$$Y \geq Q_{Y|S=1,D=1,X}(1 - p_0(X), X) \Leftrightarrow \varepsilon \geq Q_{\varepsilon|S=1,D=1,X}(1 - p_0(X), X).$$

For each x , the conditional bound $\beta_U^{\text{basic}}(x)$ is linear in σ_ε

$$\begin{aligned} \beta_U^{\text{basic}}(x) &= \mathbb{E}[Y(1) | X = x] - \mathbb{E}[Y(0) | S(0) = 1, X = x] \\ &\quad + \sigma_\varepsilon(\mathbb{E}[\varepsilon | \varepsilon \geq Q_{\varepsilon|S(1)=1,X=x}(1 - p_0(x), x), S(1) = 1, X = x]). \end{aligned}$$

As a result, the conditional width

$$\beta_U^{\text{basic}}(x) - \beta_L^{\text{basic}}(x)$$

is proportional to σ_ε , and so is the sharp width Δ . If $\sigma_\varepsilon = 0$, the width $\Delta = 0$, and the bounds collapse into a point. The stronger the predictive power of X in the outcome equation, the smaller σ_ε , and the smaller the sharp width Δ .

To quantify the role of selection equation, I assume that $\varepsilon | S = 1, D = 1, X$ is independent of X

$$\Pr(\varepsilon < t | S(1) = 1, X = x) = \Pr(\varepsilon < t | S(1) = 1) \quad \forall t \in \mathbb{R}. \quad (3.3)$$

For one example, (3.3) holds if $(\varepsilon, S(1))$ is independent of X , which would imply $s(1, x)$ is a constant. Indeed, (3.3) holds if there is no selection in the treated group: $S(1) = 1$ a.s. and $\varepsilon \perp X$. For another example, (3.3) holds in Heckman (1976, 1979) where ε and $(S(1), X)$ are independent.

In both cases, the always-takers' ATE remains partially identified.

Define the integrated quantile function $K(p)$ as

$$K(p) = \int_{1-p}^1 Q_{\varepsilon|S(1)=1}(u) du. \quad (3.4)$$

Since $p \rightarrow Q_{\varepsilon|S(1)=1}(1-p)$ is non-increasing, $K(p)$ is globally concave, and $K''(p) < 0$ for all $p \in (0, 1)$. Consider a Taylor expansion of $K(p_0(x))$ around p_0

$$K(p_0(x)) \approx K(p_0) + K'(p_0)(p_0(x) - p_0) + 0.5K''(p_0)(p_0(x) - p_0)^2 + o((p_0(x) - p_0)^2). \quad (3.5)$$

By construction, the first-order term integrates out to zero:

$$\int_x (p_0(x) - p_0) f_X(x | S = 1, D = 1) dx = p_0 - p_0 = 0.$$

Since $K''(p_0) < 0$, the second-order term has a negative effect on width. Plugging (3.5) into (3.1) gives (see Remark A.2) a local (in x) approximation of sharp width

$$\Delta \approx \sigma_\varepsilon p_0^{-1} \left(K(p_0) + K(1-p_0) - K(1) + 0.5(K''(p_0) + K''(1-p_0)) \mathbb{E}(p_0(X) - p_0)^2 \frac{s(1, X)}{\mathbb{E}_S(1, X)} \right). \quad (3.6)$$

In particular, the larger the weighted variance of $p_0(X)$, the smaller the sharp width Δ .

4 Generalized Lee Bounds

4.1 Definitions and Assumptions

In this section, I generalize Lee bounds under conditional monotonicity. Define the conditional average treatment effect on selection as

$$\tau(x) := s(1, x) - s(0, x) \quad (4.1)$$

and the following sets

$$\mathcal{X}_{\text{help}} := \{x : \tau(x) > 0\}, \quad \mathcal{X}_{\text{hurt}} := \{x : \tau(x) < 0\} \quad (4.2)$$

ASSUMPTION 2 (Conditional monotonicity). *The covariate set $\mathcal{X} = \mathcal{X}_{\text{help}} \sqcup \mathcal{X}_{\text{hurt}}$ can be partitioned into $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$ so that*

$$X \in \mathcal{X}_{\text{help}} \Rightarrow S(1) \geq S(0) \text{ a.s.}, \quad X \in \mathcal{X}_{\text{hurt}} \Rightarrow S(1) \leq S(0) \text{ a.s.}$$

Assumption 2 requires the direction of treatment effect on selection to be identified by covariate vector X . However, the sign of the effect can vary along with covariates. When there are no covariates, Assumption 2 reduces to Assumption 1 (2). The larger the covariate set, the weaker Assumption 2 is. The weakest form of Assumption 2, based on the full vector of X , is untestable.

Define the conditional upper bound $\beta_U(x)$ as

$$\beta_U(x) := \begin{cases} \beta_U^{\text{help}}(x) & x \in \mathcal{X}_{\text{help}} \\ \beta_U^{\text{hurt}}(x) & x \in \mathcal{X}_{\text{hurt}}, \end{cases} \quad (4.3)$$

where $\beta_U^{\text{help}}(x) = \beta_U^{\text{basic}}(x)$ in (2.5) for $x \in \mathcal{X}_{\text{help}}$ and $\beta_U^{\text{hurt}}(x)$ is its analog on $\mathcal{X}_{\text{hurt}}$. For a boundary point $x : \tau(x) = 0$,

$$\beta_U(x) = \beta_U^{\text{help}}(x) = \beta_U^{\text{hurt}}(x).$$

Define the aggregate bound

$$\beta_U = \frac{\int_{\mathcal{X}} \beta_U(x) \min(s(0,x), s(1,x)) f_X(x) dx}{\int_{\mathcal{X}} \min(s(0,x), s(1,x)) f_X(x) dx}. \quad (4.4)$$

Lemma 1 (Generalized Lee bound). Under Assumptions 1(1) and 2, the bound β_U in (4.4) is a sharp upper bound on β_0 in (2.1).

ASSUMPTION 3 (Regularity Conditions). *(BO) Bounded Outcome: There exists $M < \infty$ so that $|Y| \leq M$ a.s. (SO) Strict Overlap: There exists $\kappa \in (0, 1/2)$ so that $s(d,x) \in (\kappa, 1 - \kappa) \quad \forall d, x$*

and $\mu_1(x) := \Pr(D = 1 \mid X = x) \in (\kappa, 1 - \kappa)$ for any x . (MA) Margin Assumption: There exist absolute constants $1/2 < \alpha < \infty$ and $0 < \eta \leq 1$ so that

$$\Pr_X(|\tau(X)| \leq t) \leq (t/\eta)^\alpha, \quad 0 \leq t \leq \eta.$$

(REG): For $d \in \{1, 0\}$, the conditions hold. (i) The conditional density $f^d(y \mid x) := f_{Y \mid S=1, D=d, X=x}(y \mid x)$ is bounded from above uniformly over $y \in \mathcal{Y}_x$ by B_f ; (ii) $\inf_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}_x} f^d(y \mid x)$ is bounded away from zero. (iii) The derivative of $y \rightarrow f^d(y \mid x)$ is continuous and bounded in absolute value from above uniformly over $y \in \mathcal{Y}_x$.

Assumption 3 states regularity conditions. Strict overlap (SO) is a standard condition in treatment effects literature. I impose it both for the conditional probability of selection $s(d, x)$ and the propensity score $\mu_1(x)$. The Margin Assumption (MA) has been considered in the literature on classification analysis (Mammen and Tsybakov (1999), Tsybakov (2004)) and empirical welfare maximization (Kitagawa and Tetenov (2018), Mbakop and Tabord-Meehan (2021), Sun (2021)). The parameters η and α characterize the size of population when $\tau(X)$ is close to the margin $\tau(X) = 0$. For example, if $\tau(X)$ is continuously distributed with a bounded density, (MA) holds with $\alpha = 1$ and some $\eta > 0$. The fourth condition (REG) requires the outcome to be continuously distributed without point masses, and is routinely imposed for quantile estimation (e.g., Belloni and Chernozhukov (2011) and Belloni et al. (2019)).

Definition 1 (Selection Rate). There exist a sequence of numbers $\varepsilon_N = o(1)$ and a sequence of sets $S_N^d, d \in \{1, 0\}$ such that the first-stage estimates $\widehat{s}(d, x)$ of the true function $s_0(d, x)$ belong to S_N^d with probability at least $1 - \varepsilon_N$. The sets S_N^d shrink at the following rate

$$\sup_{d \in \{1, 0\}} \sup_{s \in S_N^d} \left(\mathbb{E}_X |s(d, X) - s_0(d, X)|^p \right)^{1/p} \leq s_N^p, \quad 1 \leq p \leq \infty$$

and the functions in S_N^d obey $\inf_{x \in \mathcal{X}} \inf_{d \in \{1, 0\}} s(d, x) > \kappa/2 > 0$. Let s_N and s_N^∞ be the mean square rate and the sup-norm rate, respectively.

Definition 2 (Uniform Quantile Rate). There exist a sequence of numbers $\xi_N = o(1)$ and a sequence of sets Q_N^d so that the first-stage estimate $\widehat{Q}^d(u, x)$ of $Q_0^d(u, x)$ shrinks at the rate uniformly

over $\mathcal{U}_N = [\xi_N, 1 - \xi_N]$

$$\sup_{d \in \{1,0\}} \sup_{Q \in \mathcal{Q}_N^d} \sup_{\mathcal{U}_N} (\mathbb{E}|Q^d(u, X) - Q_0(u, X)|^p)^{1/p} \leq q_N^p, \quad 1 \leq p \leq \infty,$$

where the sets \mathcal{Q}_N^d consist of a.s. M -bounded functions. Let q_N and q_N^∞ be mean square and sup-norm rates, respectively.

ASSUMPTION 4 (First-Stage Rates). (1) For α in Assumption 3(MA), the sequences $\rho_N := \max(2s_N^\infty/\kappa, \xi_N)$ and q_N^∞ obey the following bounds:

$$(\rho_N)^{\alpha+1} = o(N^{-1/2}), \quad (\rho_N)^\alpha q_N^\infty = o(N^{-1/2}). \quad (4.5)$$

(2) The mean square rates obey $\max(s_N, q_N) = o(N^{-1/4})$.

Remark 1 (Plausibility of Assumption 4(1)). Suppose $s_N^\infty = o(N^{-5/16})$. If Assumption 3(MA) holds with $\alpha = 1$, ξ_N can be taken to $\xi_N := \rho_N := N^{-5/16}$. For the quantile functions, a standard practice (e.g., Belloni and Chernozhukov (2011)) is to establish the rates in a compact set $\mathcal{U} \in (0, 1)$ that does not change with N . I conjecture that on $[\xi_N, 1 - \xi_N]$, the rate $q_N^\infty = O(\sqrt{s_Q^2 \log p_Q / N \xi_N}) = O(N^{-11/32} \sqrt{s_Q \log p_Q}) = o(N^{-8/32}) = o(N^{-1/4})$, where s_Q and p_Q are the sparsity index and the total number of covariates, respectively. Then, Assumption 4 (1) holds.

Assumption 4 is stated in a high-level form in order to accommodate various classic nonparametric and modern regularized estimators. Mean square and ℓ_∞ rates are available for logistic series regression (Hirano et al. (2003)) under smoothness and its ℓ_1 -penalized analog under sparsity (Belloni et al. (2016)), respectively. Likewise, mean square and ℓ_∞ rates are available for quantile series regression (Belloni et al. (2019)) and its ℓ_1 -penalized analog (Belloni and Chernozhukov (2011)). Adaptive choices of ℓ_1 -penalty that guard against overfitting are provided in Belloni and Chernozhukov (2011) and Belloni et al. (2017), respectively.

4.2 The estimator

The moment equation. I represent (4.4) as a ratio of two moments. For the sake of brevity, suppose the propensity score $\mu_1(x) = \Pr(D = 1 | X = x) = 1/2 \quad \forall x$. Define the numerator's moment

$$m_U^{\text{help}}(W, \xi) = D \cdot S \cdot Y \cdot 1\{Y \geq Q^1(1 - p(X), X)\} - (1 - D) \cdot S \cdot Y \quad (4.6)$$

and let $m_U^{\text{hurt}}(W, \xi)$ is its counterpart on $\mathcal{X}_{\text{hurt}}$ in (A.46). Here, $W = (D, X, S, S \cdot Y)$ is the data vector and the nuisance parameter ξ_0

$$\xi_0(x) := \{s(0, x), s(1, x), Q^1(u, x), Q^0(u, x)\}, \quad (4.7)$$

where $Q^d(u, x)$ is the u -quantile of $Y | S = 1, D = d, X$ for $d \in \{1, 0\}$. Taking

$$m_U(W, \xi) = 1\{X \in \mathcal{X}_{\text{help}}(\tau)\} m_U^{\text{help}}(W, \xi) + 1\{X \in \mathcal{X}_{\text{hurt}}(\tau)\} m_U^{\text{hurt}}(W, \xi) \quad (4.8)$$

gives a moment-based representation of β_U

$$\beta_U = \frac{\mathbb{E} m_U(W, \xi_0)}{\mathbb{E} \min(s(0, X), s(1, X))}. \quad (4.9)$$

Lemma 2 (Small misclassification bias). Under Assumption 3 (BO, SO, MA),

$$\sup_{s(d, x) \in S_N^d} \mathbb{E} 1\{|\tau_0(X)| \leq \rho_N\} (m_U(W; Q_0; s) - m_U(W; Q_0; s_0)) = O(\rho_N^{\alpha+1}). \quad (4.10)$$

Lemma 2 is an important intermediate result. If the true quantile function $Q_0^d(u, x)$ is known, the numerator's moment is robust to the misclassification mistakes. Indeed, for a covariate X to be misclassified, the $\tau(X)$ must be within ρ_N -distance from the margin. By the margin assumption (MA), this event occurs with probability $(\rho_N)^\alpha$. For example, if $\alpha = 1$ and $s_N^\infty = o(N^{-5/16})$, the misclassification bias $o(N^{-1/2})$ is negligible. A similar result holds for the denominator's moment (Lemma A.2).

Orthogonal moment equation. To enable data-driven covariate selection, for example, via ℓ_1 -penalized logistic and quantile estimators, the moment equation $m_U(W, \xi)$ must be replaced by its orthogonal counterpart $g_U(W, \xi)$. A moment function $g_U(W, \xi)$ is orthogonal (Neyman (1959), Newey (1994), Chernozhukov et al. (2018), Chernozhukov et al. (2016) etc.) if it obeys the zero-derivative property

$$\partial_r \mathbb{E} g_U(W, r(\xi - \xi_0) + \xi_0) [\xi(X) - \xi_0(X)]|_{r=0} = 0, \quad (4.11)$$

which makes it insensitive with respect to the first-order biased estimation error of $\hat{\xi} - \xi_0$. Focusing on $\mathcal{X}_{\text{help}}$, define

$$g_U^{\text{help}}(W, \xi) := m_U^{\text{help}}(W, \xi) + \text{cor}_U^{\text{help}}(W, \xi), \quad (4.12)$$

where the correction term is

$$\begin{aligned} \text{cor}_U^{\text{help}}(W, \xi) = Q^1(1 - p_0(X), X) & \left(2(1 - D) \cdot S - 2 \cdot D \cdot S p_0(X) \right. \\ & \left. + (2D \cdot S \cdot 1\{Y \leq Q^1(1 - p_0(X), X)\} - s(1, X) + s(0, X)) \right). \end{aligned} \quad (4.13)$$

Let $g_U^{\text{hurt}}(W, \xi_0)$ be as in (A.49). The orthogonal moments are

$$\begin{aligned} g_U(W, \xi) & := 1\{X \in \mathcal{X}_{\text{help}}(\tau)\} g_U^{\text{help}}(W, \xi) + 1\{X \in \mathcal{X}_{\text{hurt}}(\tau)\} g_U^{\text{hurt}}(W, \xi) \\ g_D(W, \tau) & := 1\{X \in \mathcal{X}_{\text{help}}(\tau)\} 2(1 - D) \cdot S + 1\{X \in \mathcal{X}_{\text{hurt}}(\tau)\} 2D \cdot S. \end{aligned} \quad (4.14)$$

Definition 3 (Generalized Lee Bounds). Estimate:

1. The conditional selection probabilities $x \rightarrow \hat{s}(d, x)$ for $d \in \{1, 0\}$ and sets

$$\mathcal{X}_{\text{help}}(\hat{\tau}) := \{x : \hat{s}(1, x) - \hat{s}(0, x) > 0\}, \quad \mathcal{X}_{\text{hurt}}(\hat{\tau}) := \{x : \hat{s}(1, x) - \hat{s}(0, x) < 0\},$$

2. Given $\widehat{p}(x) = \widehat{s}(0,x)/\widehat{s}(1,x)$, define the rounded conditional trimming threshold

$$\widehat{p}^{\text{trim}}(x) := \begin{cases} \widehat{p}(x) & \widehat{p}(x) \in [\xi_N, 1 - \xi_N], \\ \xi_N & \widehat{p}(x) \leq \xi_N, \\ 1 - \xi_N & \widehat{p}(x) \geq 1 - \xi_N \end{cases} \quad (4.15)$$

and $\widehat{\xi} := \{\widehat{s}(0,x), \widehat{s}(1,x), \widehat{Q}^1(1 - \widehat{p}^{\text{trim}}(x), x), \widehat{Q}^0(1/\widehat{p}^{\text{trim}}(x), x)\}$.

3. The lower and the upper bound as

$$\widehat{\beta}_L = \frac{\mathbb{E}_{NGL}(W_i, \widehat{\xi}_i)}{\mathbb{E}_{NGD}(W_i, \widehat{\tau}_i)}, \quad \widehat{\beta}_U = \frac{\mathbb{E}_{NGU}(W_i, \widehat{\xi}_i)}{\mathbb{E}_{NGD}(W_i, \widehat{\tau}_i)}. \quad (4.16)$$

Definition 3 introduces generalized Lee bounds. The first step is to classify subjects into the regions of positive and negative selection response. The second one is to round the estimated quantile level $\widehat{p}(x)$ to the closest point of $[\zeta_N, 1 - \zeta_N]$ where the estimated quantile function converges. The third one is to compute sample averages of orthogonal moments. Similar to Chernozhukov et al. (2018), the first and the second stages are performed on different samples, in order to facilitate regularized methods.

Theorem 1 (Generalized Lee bounds). *Suppose Assumptions 2, 3, and 4 hold. Then, the estimator (4.16) is consistent and asymptotically normal,*

$$\sqrt{N} \begin{pmatrix} \widehat{\beta}_L - \beta_L \\ \widehat{\beta}_U - \beta_U \end{pmatrix} \Rightarrow N(0, \Omega).$$

Theorem 1 is my main result. It establishes consistency and asymptotic normality of generalized Lee bounds.

4.3 Discussion

Remark 2 (Sorted Bounds). The bounds $\widehat{\beta}_L, \widehat{\beta}_U$ in Definition 3 are not ordered by construction. Likewise, the endpoints of the $(1 - \gamma)$ -confidence region for the identified set

$$[\widehat{\beta}_L + c_{\gamma/2} \widehat{\Omega}_{LL}^{1/2} N^{-1/2}, \widehat{\beta}_U + c_{1-\gamma/2} \widehat{\Omega}_{UU}^{1/2} N^{-1/2}],$$

where $c_{1-\gamma/2}$ is the $(1 - \gamma/2)$ -quantile of $N(0, 1)$, are not ordered either. Chernozhukov et al. (2013) shows that sorting the estimated bounds and the confidence region

$$\widetilde{\beta}_L = \min(\widehat{\beta}_L, \widehat{\beta}_U), \quad \widetilde{\beta}_U = \max(\widehat{\beta}_L, \widehat{\beta}_U)$$

weakly improves the convergence rate and coverage, respectively. However, $(\widetilde{\beta}_L, \widetilde{\beta}_U)$ may not obey the local super-efficiency assumption of Imbens and Manski (2004) and calls for the Stoye (2009)'s confidence interval instead. If the preliminary bounds cross (i.e., the width $\widehat{\beta}_U - \widehat{\beta}_L$ is negative), the Stoye (2009)'s confidence interval may be empty, indicating the violation of Assumptions 3 or 4.

Remark 3 (Strong¹ separation). Given fixed $\varepsilon > 0$, a separation condition

$$\inf_{x \in \mathcal{X}} |\tau(x)| = \inf_{x \in \mathcal{X}} |s(1, x) - s(0, x)| > \varepsilon \quad (4.17)$$

may be plausible in settings with discrete covariates. If $s_N^\infty = o(1)$, the subjects are correctly classified into $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$ w.p. approaching one. Then, the statement of Theorem 1 holds under Assumptions 3 (SO) and (REG) and Assumption 4 (2).

Remark 4 (Sharpness \Leftrightarrow Robustness to misclassification). When X is high-dimensional, the quantile function $Q^1(u, x)$ requires non-trivial regularization assumptions, such as sparsity, to be estimated consistently. In contrast, the unconditional quantile $u \rightarrow Q_{Y|S=1, D=1}^{\text{help}}(u)$ is straightforward to estimate by a sample analog. Sacrificing sharpness, one may work with the no-covariate

¹The first version of the manuscript was based on this condition. The author thanks two discussants who pointed out its weaknesses.

bound

$$\bar{\beta}_U(x) = \begin{cases} \bar{\beta}_U^{\text{help}} & x \in \mathcal{X}_{\text{help}} \\ \bar{\beta}_U^{\text{hurt}} & x \in \mathcal{X}_{\text{hurt}}, \end{cases} \quad (4.18)$$

where $\bar{\beta}_U^{\text{help}}$ and $\bar{\beta}_U^{\text{hurt}}$ are basic Lee bounds of Section 2, defined on $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$, respectively. Let $\bar{\beta}_U$ is the analog of (4.4) based on $\bar{\beta}_U(x)$. Unless $\bar{\beta}_U^{\text{help}} = \bar{\beta}_U^{\text{hurt}}$, switching from $\mathcal{X}_{\text{help}}$ to $\mathcal{X}_{\text{hurt}}$ involves a discontinuous jump. Thus, a strong separation condition (4.17) is required for the existing analysis to apply.

Remark 5 (Agnostic approach). Consider an intermediate bound

$$\beta_U^A(x) := \begin{cases} \bar{\beta}_U^{\text{help}}(x_A) & x \in \mathcal{X}_{\text{help}} \\ \bar{\beta}_U^{\text{hurt}}(x_A) & x \in \mathcal{X}_{\text{hurt}}, \end{cases} \quad (4.19)$$

where $\bar{\beta}_U^{\text{help}}(x_A)$ and $\bar{\beta}_U^{\text{hurt}}(x_A)$ are defined conditional on subvector $X_A \subset X$. Section 3 motivates selecting X_A by minimizing the amount of unexplained variance in the outcome equation. This problem becomes conceptually similar to selecting X_A for modeling heterogeneous treatment effects, e.g. Athey and Imbens (2016) and Chernozhukov et al. (2017). Conditional inference on β_U^A is facilitated by selecting X_A on an auxiliary sample.

Remark 6 (Unknown propensity score). Assumption 1(1) requires the propensity score $\mu_1(X) = \Pr(D = 1 | X) = \Pr(D = 1)$ to be a constant. Appendix A establishes the statement of Theorem 1 under conditional independence assumption

$$(Y(1), Y(0), S(1), S(0)) \perp D | X. \quad (4.20)$$

To invoke Section 4 results, replace D and $1 - D$ by $D/\mu_1(X)$ and $(1 - D)/(1 - \mu_1(X))$ in (4.6), respectively. If the propensity score needs to be estimated by regularized methods, the correction term (4.13) must include a 4th summand correcting the propensity scores's regularization bias. Unlike other correction terms, this term depends on a truncated conditional mean (a.k.a. conditional value-at-risk) $\mathbb{E}[Y | Y \geq Q^1(1 - p_0(X), X), D = 1, S = 1, X]$.

Detecting unconditional monotonicity failure raises the question of the validity of its conditional analog, especially with few covariates. Nevertheless, assigning a monotonicity direction implied by the data gives a wider bound than forcing the same direction for each value of x . Forcing Assumption 1(2) means that the trimming threshold $p_0(x) := \min(s(0,x)/s(1,x), 1)$ must be capped at one. Since such a bound is always tighter, and, therefore, less robust, than the generalized one, I recommend against this practice.

5 Extensions

In this section, I generalize Lee bounds to handle practical difficulties encountered in empirical applications. Section 5.1 generalizes one-dimensional bounds to a set for a multi-dimensional treatment effect. Section 5.2 introduces fuzzy Lee bounds in the presence of endogenous non-compliance.

5.1 Multiple outcomes

Consider a setup with a multi-valued outcome \mathbf{Y} . As before, the observed sample $(D_i, X_i, \mathbf{S}_i, \mathbf{S}_i \mathbf{Y}_i)_{i=1}^N$ consists of the realized treatment D , the vector of baseline covariates X , the selection outcome

$$\mathbf{S} = D \cdot \mathbf{S}(1) + (1 - D) \cdot \mathbf{S}(0),$$

and outcomes for the selected subjects $\mathbf{S} \cdot \mathbf{Y} = \mathbf{S} \cdot (D \cdot \mathbf{Y}(1) + (1 - D) \cdot \mathbf{Y}(0))$, where $\mathbf{S} \in \mathbb{R}^d$ and $\mathbf{Y} \in \mathbb{R}^d$ are d -vectors. The parameter of interest is the average treatment effect

$$\beta_0 = \mathbb{E}[\mathbf{Y}(1) - \mathbf{Y}(0) \mid \mathbf{S}(1) = \mathbf{S}(0) = \mathbf{1}] \tag{5.1}$$

for a group of subjects who are selected into the sample for each scalar outcome regardless of treatment status.

I reduce the problem to the one-dimensional case. Let $\mathcal{S}^{d-1} = \{q \in \mathbb{R}^d, \|q\| = 1\}$ be a unit sphere. For every $q \in \mathcal{S}^{d-1}$, denote the selection variable $S = \mathbf{1} \cdot \mathbf{S}$ and the outcome variable

$Y_q := q'Y$. The sharp upper bound on $q'\beta_0$ is

$$\sigma(q) = \frac{\mathbb{E}m_U(W_q, \xi_0(q))}{\mathbb{E}\min(s(0, X), s(1, X))} \quad (5.2)$$

and the sharp identified set \mathcal{B} for β_0 is

$$\mathcal{B} = \bigcap_{q \in \mathbb{R}^d: \|q\|=1} \{b \in \mathbb{R}^d : q'b \leq \sigma(q)\}. \quad (5.3)$$

Theorem 2 (Lee's Identified Set). *Under Assumption 2, the set \mathcal{B} in (5.3) is a convex and compact set whose support function is (5.2). It is a sharp identified set for β_0 in (5.1).*

Example 1. Wage Growth Let $\mathbf{S} = (S_{t_1}, S_{t_2})$ be a vector of employment outcomes for $t \in \{t_1, t_2\}$, $\mathbf{Y} = (Y_{t_1}, Y_{t_2})$ be a vector of log wages, and $\beta_0 = (\beta_{t_1}, \beta_{t_2})$ be the effect on log wage in time periods t_1 and t_2 . The sharp upper and lower bounds on the average wage growth effect from t_1 to t_2 , $\beta_{t_2} - \beta_{t_1}$, are given by

$$[-\sqrt{2}\sigma(-q), \sqrt{2}\sigma(q)], \quad q = (1/\sqrt{2}, -1/\sqrt{2}). \quad (5.4)$$

Example 1 demonstrates the use of support function when $q = (1/\sqrt{2}, -1/\sqrt{2})$ is a known vector. To conduct inference on $\sigma(q)$, invoke Theorem 1 with $W_q = (D, X, S, S \cdot Y_q)$.

Example 2. Standardized Treatment Effect Let \mathbf{Y} be a vector of related outcomes and β_0 be a vector of average effects. A common approach for summarizing findings is to consider the *standardized treatment effect*

$$\text{STE} = \frac{1}{d} \sum_{j=1}^d \frac{\beta_j}{\zeta_j}, \quad (5.5)$$

where ζ_j is the standard deviation of the outcome j in the control group. The sharp lower and upper bounds on STE are given by

$$[-C_\zeta \sigma(-q), C_\zeta \sigma(q)], \quad (5.6)$$

where $q = \zeta / \|\zeta\|$ and $C_\zeta = \|\zeta\|/d$. Example 2 demonstrates the use of support functions when

$q = \zeta / \|\zeta\|$ is a population parameter. In contrast to Example 1, the direction $q = \zeta / \|\zeta\|$ is unknown and needs to be estimated. Therefore, it is important for the support function $\sigma(q)$ to be approximated in some neighborhood of q in addition to the point q itself.

5.2 Fuzzy Lee Bounds

This section generalizes Lee (2009)'s results to accommodate endogenous treatment receipt. Let $Z \in \{1, 0\}$ be a binary instrument, such as an offer of participation, that is randomly assigned conditional on X . Let $D(1)$ and $D(0)$ be the binary potential treatment outcomes for D if subject is treated and not treated, respectively, and $D = Z \cdot D(1) + (1 - Z) \cdot D(0)$. Likewise, let $S(1)$ and $S(0)$ be dummies for whether an individual's outcome is observed with and without instrument, and let $S = Z \cdot S(1) + (1 - Z) \cdot S(0)$. The observed data vector $W = (Z, X, D, S, S \cdot Y(D))$ consists of the pre-randomization covariates X , instrument Z , and post-randomization data $(D, S, S \cdot Y)$ where $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$. The object of interest is the ATE for subjects who are always-takers (with respect to selection) and compliers (with respect to the treatment choice).

$$\beta_0 = \mathbb{E}[Y(1) - Y(0) \mid S(1) = S(0) = 1, D(1) > D(0)]. \quad (5.7)$$

ASSUMPTION 5 (Fuzzy Lee Bounds). *The following statements hold.*

(1) *(Independence). The potential outcomes vector is independent of Z*

$$(Y(1), Y(0), S(1), S(0), D(1), D(0)) \perp Z \mid X$$

(2) *(Monotonicity of Choice). $D(1) \geq D(0)$ a.s. with $\Pr(D(1) > D(0)) > 0$.*

(3) *(Independence of Selection and Choice) $(S(1), S(0)) \perp (D(1), D(0)) \mid X$.*

To accommodate endogenous treatment receipt, the outcomes should be trimmed separately for $(Z = 1, D = 1)$ and $(Z = 1, D = 0)$ groups at equal proportions $p_0(X) = s(0, X) / s(1, X) = \Pr(S = 1 \mid Z = 0, X) / \Pr(S = 1 \mid Z = 1, X)$, a distinction that does not exist in the perfect compliance case. Define the new quantile function $Q(u, d, x)$ as

$$Q(u, d, x) : \Pr(Y \leq Q(u, d, x) \mid S = 1, Z = 1, D = d, X = x) = u, \quad u \in [0, 1]. \quad (5.8)$$

For the sake of exposition, suppose $S(1) \geq S(0)$ a.s.. Then, the upper-truncated subjects are

$$\Lambda_U(W) = \left\{ W : (Z = 1 \cap S = 1 \cap Y \geq Q(1 - p_0(X), D, X)) \cup (Z = 0 \cap S = 1) \right\}. \quad (5.9)$$

Theorem 3 (Fuzzy Lee Bounds). *Under Assumptions 1(2) and 5, a sharp upper bound β_U on β_0 is the Wald estimand*

$$\beta_U = \frac{E_{\Lambda_U}[Y | Z = 1] - E_{\Lambda_U}[Y | Z = 0]}{\Pr_{\Lambda_U}[D = 1 | Z = 1] - \Pr_{\Lambda_U}[D = 1 | Z = 0]}. \quad (5.10)$$

6 Simulation Evidence

I build a simulation exercise on the JobCorps data set. The vector $X = (1, X_1, X_2)$ consists of a constant and two binary indicators, one for female gender (X_1) and one for getting away from home being a very important motivation for joining JobCorps (X_2), taken from the JobCorps data. An artificial treatment variable D is determined by an unbiased coin flip. A binary employment indicator S is

$$S = 1\{X'\alpha_0 + D \cdot X'\gamma_0 + U > 0\}, \quad (6.1)$$

where U is an independently drawn logistic shock. Likewise, log wages are generated according to the model

$$Y = (1, X_1)'\kappa_0 + \varepsilon, \quad \varepsilon \sim N(0, \tilde{\sigma}^2), \quad (6.2)$$

where ε is an independent normal random variable. The parameter vector $(\alpha_0, \gamma_0, \kappa_0, \tilde{\sigma}^2)$ is taken to be the estimates of (6.1) and (6.2), where S and Y are week 90 employment and log wages. The sets $\mathcal{X}_{\text{help}} = \{X_1 = 0 \text{ and } X_2 = 0\}$ and $\mathcal{X}_{\text{hurt}} = \{X_1 \neq 0 \text{ or } X_2 \neq 0\}$, as determined by the sign of the parameter γ . The population data set is taken to be 9,145 observations of baseline covariates X and the artificial variables $D, S, S \cdot Y$, generated for each observation. By construction, the average treatment effect on the always-takers β_0 is zero. The true generalized Lee bound $\beta_U = 0.018$. The true no-covariate bound $\bar{\beta}_U$ is 0.035. Because of differential selection response on $\mathcal{X}_{\text{help}}$

Table 1: Finite-sample performance of oracle, basic, naive and ortho methods

N	Bias				St. Dev.				Coverage Rate			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Oracle	Basic	Naive	Ortho	Oracle	Basic	Naive	Ortho	Oracle	Basic	Naive	Ortho
3,000	0.00	0.05	0.04	0.00	0.01	0.02	0.03	0.02	0.94	0.21	0.64	0.93
5,000	0.00	0.04	0.03	-0.00	0.01	0.01	0.02	0.01	0.95	0.25	0.63	0.95
9,000	-0.00	0.03	0.02	0.00	0.01	0.01	0.01	0.01	0.95	0.28	0.65	0.97
10,000	-0.00	0.03	0.02	-0.00	0.01	0.01	0.01	0.01	0.95	0.29	0.64	0.97
15,000	-0.00	0.02	0.01	-0.00	0.00	0.01	0.01	0.01	0.94	0.28	0.64	0.97

Notes. Results are based on 10,000 simulation runs. The true parameter value is 0.035 for the basic method, and 0.018 for all other methods. Bias is the difference between the true parameter and the estimate, averaged across simulation runs. St. Dev. is the standard deviation of the estimate. Coverage Rate is the fraction of times a two-sided symmetric CI with critical values $c_{\alpha/2}$ and $c_{1-\alpha/2}$ covers the true parameter, where $\alpha = 0.95$. N is the sample size in each simulation run. The naive method estimates the first-stage functions (2.4) and $Q(u, x)$ by logistic and quantile regression on all 28 covariates.

and $\mathcal{X}_{\text{hurt}}$, neither β_U nor $\bar{\beta}_U$ coincide with original Lee bound defined under unconditional monotonicity.

I compare the performance of four estimators—oracle, basic, naive and ortho methods—by drawing random samples with replacement from the population data set. To mimic the researcher’s covariate selection problem, I augment this data set with 28 covariates selected by Lee. Although these variables are absent from equations (6.1) and (6.2), they are strongly correlated with X_1 and X_2 , making covariate selection an interesting problem. The oracle method estimates β_U based on the known first-stage parameter. In contrast, all other methods need to learn $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$ from the available sample. The basic method estimates $\mathcal{X}_{\text{help}}$ by logistic and quantile regression on 28 raw covariates. It targets the no-covariate upper bound 0.035. Both the naive and the ortho methods target the generalized Lee bound β_U , where the same covariate set is used to classify subjects into $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$ and to define the bound. The naive method estimates the first-stage functions (2.4) and $Q(u, x)$ by standard regression methods on all 28 covariates and uses a non-orthogonal moment equation. In contrast, the ortho method selects covariates by post-lasso-logistic of Belloni et al. (2016) for the employment equation and by post-lasso of Belloni et al. (2017) for the wage equation and uses an orthogonal moment.

Table 1 reports the finite-sample performance for the oracle, basic, naive and ortho methods. The basic method, which focuses on the no-covariate bound, exhibits substantial coverage dis-

tortion (Column 10), supporting the conclusion of Remark 4. In contrast, the generalized Lee bound (4.4) (Columns 11 and 12) is a more robust target. Switching from a non-orthogonal to an orthogonal moment equation (i.e., from the naive to the ortho method) gives an extra boost of the coverage rate from 64% to 93%.

7 Empirical application

7.1 JobCorps

In this section, I review the basics of JobCorps training program. I then discuss how the direction of JobCorps' effect on employment differs with observed characteristics.

Lee (2009) studies the effect of winning a lottery to attend JobCorps, a federal vocational and training program, on applicants' wages. In the mid-1990s, JobCorps used lottery-based admission to assess its effectiveness. The control group of 5,977 applicants was essentially embargoed from the program for three years, while the remaining applicants (the treated group) could enroll in JobCorps as usual. The sample consists of 9,145 JobCorps applicants and has data on lottery outcome, hours worked and wages for 208 consecutive weeks after random assignment. In addition, the data contain educational attainment, employment, recruiting experiences, household composition, income, drug use, arrest records, and applicants' background information. These data were collected as part of a baseline interview, conducted by Mathematica Policy Research (MPR) shortly after randomization (Schochet et al. (2008)). After converting applicants' answers to binary vectors and adding numeric demographic characteristics, I obtain a total of 5,177 raw baseline covariates, which are summarized in Supplementary Appendix.

7.2 Testing framework

Having access to baseline covariates X means that the monotonicity assumption can be tested. Using the notation of Section 2, let S correspond to employment and Y correspond to log wages. If monotonicity holds, the treatment-control difference in employment rates (4.1) must be either

non-positive or non-negative for all covariate values. Consequently, it cannot be the case that

$$\text{Prob}(\mathcal{X}_{\text{help}}) > 0 \quad \text{and} \quad \text{Prob}(\mathcal{X}_{\text{hurt}}) > 0. \quad (7.1)$$

My first exercise is to estimate $s(1,x)$ and $s(0,x)$ by a week-specific cross-sectional logistic regression

$$s(D,X) = \Lambda(X'\alpha_0 + D \cdot X'\gamma_0), \quad (7.2)$$

where $\Lambda(\cdot) = \frac{\exp(\cdot)}{1 + \exp(\cdot)}$ is the logistic CDF, X is a vector of baseline covariates that includes a constant, $D \cdot X$ is a vector of covariates interacted with treatment, and α and γ are fixed vectors.

Figure 1 reports the share of subjects with positive employment response.

The second exercise is to test monotonicity without relying on logistic approximation. For each week, I select a small number of discrete covariates and partition the sample into discrete cells C_j , $j \in \{1, 2, \dots, J\}$, determined by covariate values. For example, one binary covariate corresponds to $J = 2$ two cells. By monotonicity, the vector of cell-specific treatment-control differences in employment rates, $\mu = (\mathbb{E}[\Delta(X)|X \in C_j])_{j=1}^J$, must be non-negative:

$$H_0: \quad (-1) \cdot \mu \leq 0. \quad (7.3)$$

The test statistic for the hypothesis in equation (7.3) is

$$T = \max_{1 \leq j \leq J} \frac{(-1) \cdot \hat{\mu}_j}{\hat{\sigma}_j}, \quad (7.4)$$

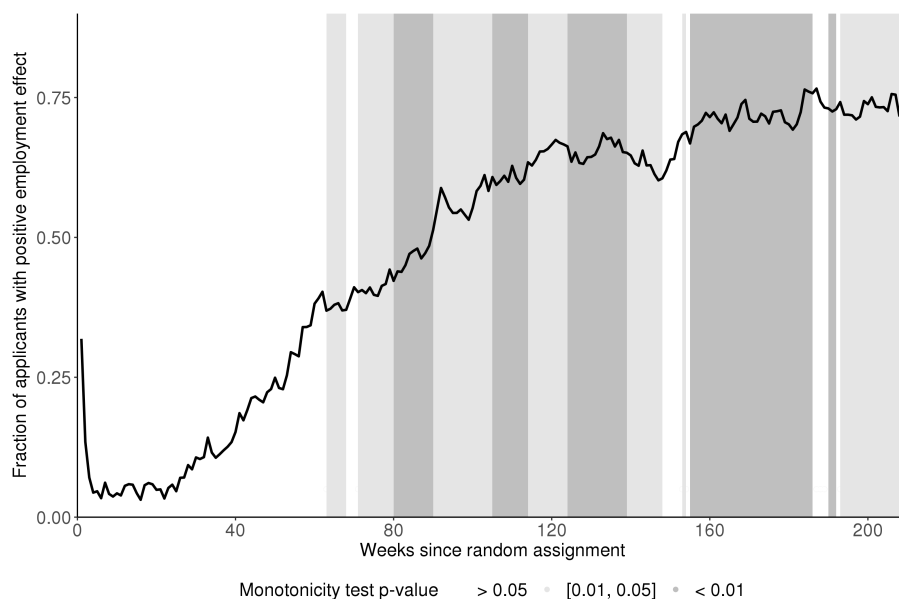
and the critical value is the self-normalized critical value of Chernozhukov et al. (2019). The critical values as in Hsu et al. (2019) and Bai et al. (2019) imply qualitatively similar results.

Figure 1 plots the fraction of subjects with a positive JobCorps effect on employment in each week. In the first weeks after random assignment, there is no evidence of a positive JobCorps effect on employment for any group. By the end of the second year (week 104), JobCorps increases employment for nearly half of the individuals, and this fraction rises to 0.75 by the end of the study period (week 208). This pattern is consistent with the JobCorps program description.

While being enrolled in JobCorps, participants cannot hold a job, which is known as the lock-in effect (e.g., Blanco et al. (2013)). After finishing the program, JobCorps graduates may have gained employment skills that help them outperform the control group.

Figure 1 shows the results of testing the inequality in (7.3) for each week. The direction of the employment effect varies with socio-economic factors. For example, the applicants who received AFDC benefits during the 8 months before RA or who belonged to median income and yearly earnings groups experience a significantly positive ($p \leq 0.05$) employment effect at weeks 60–89, although the average effect is significantly negative. As another example, the applicants who answered “1: Very important” to the question “How important was getting away from community on the scale from 1 (very important) to 3 (not important)?” and who smoke marijuana or hashish a few times each months experience a significantly negative ($p \leq 0.05$) employment effect at week 117–152 despite the average effect being positive. Finally, at week 153–186, the average JobCorps effect is significantly negative for subjects whose most recent arrest occurred less than 12 months ago, despite the average effect being positive.

Figure 1: Fraction of JobCorps applicants with positive conditional employment effect by week.



Notes. The horizontal axis shows the number of weeks since random assignment. The vertical axis shows the fraction of applicants whose conditional employment effect $\tau(x)$ is positive. Following week 60, a week is shaded if the test statistic T exceeds the critical value at the $p = 0.01$ (dark gray) or $p \in [0.05, 0.01)$ (light gray) significance level. For each week, $\tau(x)$ is defined in equation (4.1) and estimated as in equation (7.2), the null hypothesis is as in equation (7.3), the test statistic T is as in equation (7.4), and the test cells and critical values are as defined in Table B.9. Computations use design weights.

Table 2 reports generalized Lee bounds on the JobCorps week 90 wage effect on the always-takers and the confidence region for the identified set. The no-covariate Lee bounds (4.18) cannot determine the direction of the effect (Column (1)). Neither can the generalized bounds defined conditional on a subset of the covariates selected by Lee (Column (2)). If few of the covariates affect week 90 employment and wage, the Column (3) bounds suggest that JobCorps raises week 90 wages by 4.0–4.6% on average, which is slightly smaller than Lee’s original estimate (4.9–5%). Despite numerical proximity, Lee’s basic estimates (Table B.1, Column 1) and generalized Lee estimates (Table 2, Column 3) have substantially different reasons for being tight. Lee’s estimates are tight because one believes JobCorps’s week 90 employment effect is close to zero. In contrast, the generalized bounds are tight because variation in employment is well-explained by reasons for joining JobCorps, highest grade completed, and variation in wages is explained by pre-randomization earnings, household income, gender and other socio-economic factors.

The sparsity assumption of Column (3) may not be economically plausible. In Column (4), the target bounds are defined as the generalized Lee bounds given the 15 covariates, selected for either employment or wage equation in Column (3). The Column (4) are almost the same as the Column (3) ones, suggesting that the bounds are not too sensitive to sparsity violations. However, the Column (4) confidence region does not account for the uncertainty in how these 15 covariates are selected.

To properly quantify the uncertainty of the Column (4) bounds, I invoke the conditional (Column (5)) and variational (Column (6)) agnostic approaches of Chernozhukov et al. (2017). In Column (5), the auxiliary sample is taken to be 6,241 applicants that Lee excluded from consideration due to missing data in weeks other than week 90. The Column (5) bounds target the sharp bounds given the covariates selected on this auxiliary sample. The estimates suggest that JobCorps raises week 90 wages by 4.1–4.3%, which is consistent with the lasso-based findings (Columns (3) and (4)). Furthermore, the 95% confidence region is almost the same as the Column (4) one, suggesting that the Column (4) confidence region adequately captures uncertainty of the Column (4) estimate. Column (6) differs from Column (5) by splitting Lee’s sample into the auxiliary and the main part. The bounds in Column (6) are slightly wider than the Column (5) ones. Overall, the results suggest that JobCorps has had a small positive effect on week 90 log wages, but the estimate is significant only under a sparsity assumption.

7.3 Finkelstein et al. (2012)

Finkelstein et al. (2012) studies the effect of access to Medicaid on self-reported healthcare utilization and measures of health. The data come from the Oregon Health Insurance Experiment (OHIE), which allowed a subset of uninsured low-income applicants to apply for Medicaid in 2008. OHIE used a lottery to determine who was eligible to apply for Medicaid. One year after randomization, a subset of $N = 58,405$ applicants were mailed a survey with questions about recent changes in their healthcare utilization and general well-being. The sample contains the lottery outcome, actual Medicaid enrollment, and survey responses. In addition, the sample has 64 pre-determined characteristics including demographics, enrollment in SNAP and TANF government programs, and pre-existing health conditions. While the number of raw covariates

is moderate, the number of their pairwise interactions $p = 64^2 = 4,096$ is quite large for classic nonparametric methods. Since the survey response rate is close to 50% and the control applicants respond 1.07 more likely than the treated ones, Finkelstein et al. (2012)'s findings are subject to potential nonresponse bias.

Finkelstein et al. (2012) describes the effect of winning the Medicaid lottery using the intent-to-treat (ITT) and Local Average Treatment Effect (LATE) parameters. If an applicant wins the lottery, all members of their household become eligible to enroll. As a result, larger households are more likely to win the lottery than smaller ones. Furthermore, the control applicants were oversampled in the earlier survey waves. To account for the correlation between household size and survey wave fixed effects, the intent-to-treat equation takes the form

$$Y_{ih} = \beta_0 + \beta_1 \text{Lottery}_h + \bar{X}_{ih} \beta_2 + \varepsilon_{ih}, \quad (7.5)$$

where i denotes an individual, h denotes a household, $\text{Lottery}_h = 1$ is a dummy for whether household h was offered access to Medicaid, and \bar{X}_{ih} is a vector of stratification characteristics (survey wave and household size fixed effects). The coefficient β_1 is the main coefficient of interest interpreted as the impact of being able to apply for Medicaid through the Oregon lottery. Finkelstein et al. (2012) also studies the local average treatment effect (LATE) of insurance,

$$Y_{ih} = \pi_0 + \pi_1 \text{Insurance}_{ih} + \bar{X}_{ih} \pi_2 + v_{ih}, \quad (7.6)$$

where Insurance_{ih} is an applicant-specific measure of insurance coverage defined as “ever on Medicaid during study period”, and all other variables are as defined in (7.5). Finkelstein et al. (2012) estimates (7.6) by two-stage least squares (2SLS), using Lottery_h as an instrument for Insurance and including \bar{X}_{ih} in both the first and the second stages of 2SLS. The coefficient π_1 is the main coefficient of interest: it shows the impact of insurance coverage on subjects who enroll in Medicaid if and only if they become eligible. If non-response is exogenous for each household size and survey wave, Medicaid eligibility and enrollment have a positive and significant effect on all measures of health and healthcare utilization (Tables 3, 4, B.15, B.16, Columns (1) and (4)).

I examine whether the Intent-to-Treat (7.5) and Local Average Treatment Effect (7.6) equations are robust to non-response bias. Tables 3 and 4 show the results for self-reported health outcomes. Lee's density-based approach is very conservative and cannot determine the direction of the effect for any of the health outcomes. For each household size and survey wave stratum, the smallest number of the worst-case responses is trimmed in the control group until treatment-control difference in response rate exceeds zero for each strata. Since incorporating the additional 48 baseline covariates requires considering more than 2^{48} discrete cells, it is not possible to incorporate all of them at once. An ad-hoc choice of three demographic indicators: gender, English as preferred language, and urban area residence does not improve standard estimates.

A logistic single-index assumption on the conditional response probability and the conditional probability of zero outcome drastically changes the result. The findings in Tables 3 and 4, Columns (3) and (6), suggest that Medicaid eligibility and insurance has had positive effect on 7 out of 7 health outcomes. Furthermore, Medicaid insurance is associated with at least 0.981 (std. error 0.577) more days in good overall health after accounting for non-response bias. This estimate is 75% of the baseline LATE estimate (1.317 (std. error 0.562)). Likewise, Medicaid eligibility and insurance have a positive effect on all measures of healthcare utilization (Tables B.15 and B.16) except emergency room visits. To conclude, Finkelstein et al. (2012)'s baseline results continue to hold after accounting for non-response bias.

Table 2: Estimated bounds on the JobCorps effect on week 90 log wages

	(1)	(2)	(3)	(4)	(5)	(6)
	[-0.027, 0.111]	[-0.005, 0.091]	[0.040, 0.046]	[0.041, 0.059]	[0.041, 0.043]	[0.024, 0.065]
	(-0.058, 0.142)	(-0.054, 0.135)	(0.001, 0.078)	(-0.019, 0.112)	(-0.023, 0.101)	(-0.05, 0.131)
Selection covs	28	28	5 177	15	13	12-13
Post-lasso-log.	N/A	N/A	9	N/A	N/A	N/A
Wage covs	0	28	470	15	13	12-13
Post-lasso	N/A	N/A	6	N/A	N/A	N/A

Notes. Estimated bounds are in square brackets and the 95% confidence region for the identified set is in parentheses. All subjects are partitioned into the sets $\mathcal{X}_{\text{help}} = \{\hat{p}(X) < 1\}$ and $\mathcal{X}_{\text{hurt}} = \{\hat{p}(X) > 1\}$, where the trimming threshold $\hat{p}(x) = \hat{s}(0, x) / \hat{s}(1, x)$ is estimated as in equation (7.2). (1): no-covariate bounds given 28 Lee's covariates. (2): generalized Lee bounds given 28 Lee's covariates. (3): generalized bounds given all covariates assuming few of them affect employment and wage. (4): generalized bounds given the union of raw covariates selected for the employment and wage equations in Column (3). (5): generalized bounds given the covariates selected on the sample that Lee excluded due to missing data in weeks other than 90. Column (6): variational bounds adapted from Chernozhukov et al. (2017). Covariates are defined in Section A.3. First-stage estimates are given in Table B.10 for Columns (1) and (2), Table B.11 for Columns (3) and (7), Table B.14 for Column (4). Computations use design weights.

Table 3: Estimated lower bound on the effect of access to Medicaid on self-reported binary health outcomes

	ITT			LATE		
	(1)	(2)	(3)	(4)	(5)	(6)
	None	Standard	ML	None	Standard	ML
Health good /very good/excellent	0.039 (0.008)	-0.013 (0.013)	0.032 (0.017)	0.133 (0.026)	-0.067 (0.044)	0.077 (0.058)
Health fair/good/very good/excellent	0.029 (0.005)	-0.052 (0.012)	0.019 (0.010)	0.099 (0.018)	-0.195 (0.038)	0.011 (0.033)
Health same or gotten better	0.033 (0.007)	-0.033 (0.014)	0.015 (0.019)	0.113 (0.023)	-0.138 (0.049)	0.051 (0.065)
Did not screen positive for depression	0.023 (0.007)	-0.045 (0.014)	0.002 (0.010)	0.078 (0.025)	-0.183 (0.049)	0.007 (0.065)
Compulsory covariates (stratification)	N/A	16	16	N/A	16	16
Additional covariates (trimming)	N/A	0	21	N/A	0	21

* Standard errors in parentheses. This table reports results from a Lee bounding exercise on self-reported health outcomes for 3 specifications: no trimming, standard trimming, and the agnostic ML approach. Columns (1)–(3) report the coefficient and standard error on Lottery from estimating equation (7.5) by OLS. Columns (4)–(6) report the coefficient and standard error on Insurance from estimating equation (7.6) by 2SLS with Lottery as an instrument for Insurance. All regressions include household size fixed effects, survey wave fixed effects, and their interactions. Trimming methods. None: exact replicate of Finkelstein et al. (2012), Table IX. Standard: the minimal number of zero outcomes are trimmed in the control group until the treatment-control difference in response rates switches from negative to non-negative for each strata. Agnostic: Step 1. 21 additional covariates are selected on an auxiliary sample of 4,000 households as described in Appendix A.5. Step 2. In the main sample of 46,000 households, a zero outcome with covariate vector x is trimmed in the control group if a flipped coin with success prob. $(1 - p_0(x))/\phi_0(x)$ is success, where the trimming threshold $p_0(x)$ is defined in (B.4) and the zero outcome probability $\phi_0(x)$ is defined in (B.3). Standard errors are estimated by a cluster-robust bootstrap with $B = 1000$ repetitions. Both the trimming and regression steps are bootstrapped. Computations (the first and the second stage) use survey weights. Covariates are described in Table B.17.

Table 4: Estimated lower bound on the effect of access to Medicaid on self-reported number of days in good health

	ITT			LATE		
	(1)	(2)	(3)	(4)	(5)	(6)
	None	Standard	NP	None	Standard	NP
# of days overall health good, past 30 days	0.381 (0.162)	-1.096 (0.349)	0.272 (0.166)	1.317 (0.562)	-4.411 (1.166)	0.981 (0.577)
# of days phys. health good, past 30 days	0.459 (0.174)	-1.230 (0.384)	0.272 (0.170)	1.585 (0.605)	-4.929 (1.308)	0.627 (0.592)
# of days mental health good, past 30 days	0.603 (0.184)	-0.862 (0.374)	0.220 (0.179)	2.082 (0.640)	-3.573 (1.298)	0.750 (0.624)
Compulsory covariates (stratification)	N/A	16	16	N/A	16	16
Additional covariates (trimming)	N/A	0	9	N/A	0	9

* Standard errors in parentheses. This table reports results from a Lee bounding exercise on self-reported health outcomes for 3 specifications: no trimming, standard trimming, and the classic nonparametric (NP) approach. Columns (1)–(3) report the coefficient and standard error on Lottery from estimating equation (7.5) by OLS. Columns (4)–(6) report the coefficient and standard error on Insurance from estimating equation (7.6) by 2SLS with Lottery as an instrument for Insurance. All regressions include household size fixed effects, survey wave fixed effects, and their interactions. Trimming methods. None: exact replicate of Finkelstein et al. (2012), Table IX. Standard: the minimal number of control outcomes are trimmed from below for each value of fixed effect until the treatment-control difference in response rates switches from negative to non-negative for each strata. NP. Step 1. 9 additional covariates are taken as described in Appendix A.5 based on OHIE documentation. Step 2. An outcome with covariate vector x is trimmed if it is less than $Q(1 - 1/p_0(x), x)$, where the trimming threshold $p_0(x)$ is defined in equation (B.4) and the conditional quantile is defined in equation (2.4). Standard errors are estimated by a cluster-robust bootstrap with $B = 1000$ repetitions. Both the trimming and regression steps are bootstrapped. Computations (the first and the second stage) use survey weights. Covariates are described in Table B.17. See Appendix A.5 for details.

Appendix A. Proofs

A.1 Proofs for Section 3

Consider a semiparametric version of Heckman (1976, 1979) selection model

$$S = 1\{h(D, X) + \eta \geq 0\}, \quad (\text{A.7})$$

$$Y = \psi(D, X) + \sigma_\varepsilon \cdot \varepsilon, \quad (\text{A.8})$$

where the outcome Y is observed if and only if $S = 1$. As in Heckman (1976), the shock vector (ε, η) is independent of D and X , but the functional form of $h(d, x)$ and $\psi(d, x)$ may not be parametric.

Remark A.1. Equation (A.7) implies Assumption 2. Take

$$\mathcal{X}_{\text{help}} := \{X : h(1, X) > h(0, X)\}, \quad \mathcal{X}_{\text{hurt}} := \{X : h(1, X) < h(0, X)\}.$$

Then, for any $X \in \mathcal{X}_{\text{help}}$,

$$h(0, X) + \eta \geq 0 \Rightarrow h(1, X) + \eta \geq 0 \Rightarrow S(1) \geq S(0),$$

and a similar argument applies to $\mathcal{X}_{\text{hurt}}$. Equation (A.8) implies (3.2):

$$Y(1) - \mathbb{E}[Y(1) | X] = Y(1) - \psi(1, X) = \sigma_\varepsilon \cdot \varepsilon \perp X \quad \Rightarrow \text{Var}(Y(1) - \mathbb{E}[Y(1) | X]) = \sigma_\varepsilon^2.$$

Remark A.2 (Approximate sharp width). Assumptions (3.2) and (3.3) imply (3.6).

Proof of Remark A.2. Bayes rule implies

$$\frac{f_X(x | S = 1, D = 0)}{f_X(x | S = 1, D = 1)} = \frac{s(0, x)f_X(x)}{s(1, x)f_X(x)} p_0^{-1} = p_0(x)/p_0. \quad (\text{A.9})$$

Plugging (A.9) into the sharp width gives

$$\begin{aligned}
\Delta &= \int_x (\beta_U^{\text{basic}}(x) - \beta_L^{\text{basic}}(x)) f_X(x | S = 1, D = 0) dx \\
&= p_0^{-1} \int_x (\beta_U^{\text{basic}}(x) - \beta_L^{\text{basic}}(x)) p_0(x) f_X(x | S = 1, D = 1) dx \\
&= \sigma_\varepsilon p_0^{-1} \int_x \frac{(K(p_0(x)) + K(1 - p_0(x)) - K(1))}{p_0(x)} p_0(x) f(x | S = 1, D = 1) dx \\
&= \sigma_\varepsilon p_0^{-1} \int_x (K(p_0(x)) + K(1 - p_0(x)) - K(1)) f(x | S = 1, D = 1) dx
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}[\varepsilon | \varepsilon \geq \mathcal{Q}_{\varepsilon|S(1)=1}(1-p), S(1) = 1] &= K(p)/p, \\
\mathbb{E}[\varepsilon | \varepsilon \leq \mathcal{Q}_{\varepsilon|S(1)=1}(p), S(1) = 1] &= (K(1) - K(1-p))/p.
\end{aligned}$$

Plugging (3.5) above gives (3.6). □

A.2 Proof of Section 4

Proof of Lemma 1. Bayes rule for conditional density gives

$$\Pr(X = x | S(1) = S(0) = 1) = (\Pr(S(1) = S(0) = 1))^{-1} \begin{cases} s(0, x) f_X(x) & x \in \mathcal{X}_{\text{help}} \\ s(1, x) f_X(x) & x \in \mathcal{X}_{\text{hurt}}, \end{cases} \quad (\text{A.10})$$

where the denominator is

$$\begin{aligned}
\Pr(S(1) = S(0) = 1) &= \int_{\mathcal{X}_{\text{help}}} s(0, x) f_X(x) dx + \int_{\mathcal{X}_{\text{hurt}}} s(1, x) f_X(x) dx \\
&= \int_x \min(s(0, x), s(1, x)) f_X(x) dx.
\end{aligned}$$

As shown in Lee (2009), $\beta_U(x)$ in (4.3) is a sharp upper bound on $\beta_0(x)$ for each x

$$\beta_0(x) = \mathbb{E}[Y(1) - Y(0) | S(1) = S(0) = 1, X = x] \leq \beta_U(x) \quad \forall x.$$

Integrating the inequality by $\Pr(X = x \mid S(1) = S(0) = 1)$ gives the statement

$$\beta_0 = \frac{\int_X \beta_0(x) \min(s(0,x), s(1,x)) f_X(x) dx}{\int \min(s(0,x), s(1,x)) f_X(x) dx} \leq \beta_U.$$

□

Lemma A.1 (Negligible First-Stage Error). Let $R(W, \xi)$ be a known function of the data vector W and the nuisance parameter ξ_0 . Let $\{\Xi_N : N \geq 1\}$ be a sequence of sets that contain the first-stage estimate $\hat{\xi}$ w.p. approaching one. The sets shrink at the following rates

$$\begin{aligned} \sup_{\xi \in \Xi_N} |\mathbb{E}[R(W, \xi) - R(W, \xi_0)]| &= O(B_N) = o(N^{-1/2}) \\ \sup_{\xi \in \Xi_N} (\mathbb{E}(R(W, \xi) - R(W, \xi_0))^2)^{1/2} &= O(V_N) = o(1). \end{aligned}$$

Then, $\mathbb{E}_N[R(W_i; \hat{\xi}_i) - R(W_i, \xi_0)] = o_P(1)$.

Lemma A.1 is a restatement of Lemma A.3 in Semenova and Chernozhukov (2021).

Lemma A.2 (Denominator). Under Assumptions 3 and 4, for $g_D(W, \tau)$ in (4.14)

$$\begin{aligned} \sup_{\tau \in S_N^1 - S_N^0} |\mathbb{E}[g_D(W, \tau) - g_D(W, \tau_0)]| &\leq (2s_N^\infty/\eta)^{\alpha+1} \\ \sup_{\tau \in S_N^1 - S_N^0} (\mathbb{E}(g_D(W, \tau) - g_D(W, \tau_0))^2)^{1/2} &= O((2s_N^\infty/\eta)^{\alpha/2}). \end{aligned}$$

By Lemma A.1, $\mathbb{E}_N[g_D(W_i; \hat{\tau}_i) - g_D(W_i, \tau_0)] = o_P(1)$.

Proof of Lemma A.2. Define the misclassification events

$$\mathcal{D}_\tau^{\text{help}} := \{X : \tau(X) < 0 < \tau_0(X)\}, \quad \mathcal{D}_\tau^{\text{hurt}} := \{X : \tau_0(X) < 0 < \tau(X)\}. \quad (\text{A.11})$$

The misclassified point must be close to the margin

$$\mathcal{D}_\tau^{\text{help}} \cup \mathcal{D}_\tau^{\text{hurt}} \Rightarrow \{0 < |\tau_0(X)| < |\tau(X) - \tau_0(X)|\} =: \mathcal{D}_\tau^2.$$

For any $\tau(x) = s(1, x) - s(0, x) \in S_N^1 - S_N^0$,

$$\sup_{x \in \mathcal{X}} |\tau(x) - \tau_0(x)| \leq \sup_{x \in \mathcal{X}} (|s(1, x) - s_0(1, x)| + |s(0, x) - s_0(0, x)|) \leq 2s_N^\infty. \quad (\text{A.12})$$

Thus, the first and second moments are bounded as

$$\begin{aligned} |\mathbb{E}[g_D(W, \tau) - g_D(W, \tau_0)]| &\leq \mathbb{E}1\{X \in \mathcal{D}_\tau^2\} |\tau_0(X)| \leq \mathbb{E}1\{X \in \mathcal{D}_\tau^2\} |\tau(X) - \tau_0(X)| \\ &\leq \Pr(X \in \mathcal{D}_\tau^2) 2s_N^\infty \leq (2s_N^\infty/\eta)^{\alpha+1}. \\ \mathbb{E}(g_D(W, \tau) - g_D(W, \tau_0))^2 &\leq \kappa^{-2} \Pr(X \in \mathcal{D}_\tau^2) \leq O((2s_N^\infty/\eta)^\alpha). \end{aligned}$$

□

For $\xi(x) = \{s(0, x), s(1, x), Q^1(1 - p(x), x), Q^0(1/p(x), x)\}$, define

$$\begin{aligned} R_1(X, \xi) &= \begin{cases} \mathbb{E}[S \cdot Y 1\{Y \leq Q^1(1 - p(X), X)\} \mid D = 1, X], & \xi \neq \xi_0 \\ \mathbb{E}[S \cdot Y 1\{Y \leq Q_0^1(1 - p_0(X), X)\} \mid D = 1, X], & \xi = \xi_0 \end{cases} \\ R_0(X, \xi) &= \begin{cases} \mathbb{E}[S \cdot Y 1\{Y \geq Q^0(1/p(X), X)\} \mid D = 0, X], & \xi \neq \xi_0 \\ \mathbb{E}[S \cdot Y 1\{Y \geq Q_0^0(1/p_0(X), X)\} \mid D = 0, X], & \xi = \xi_0 \end{cases} \end{aligned}$$

and the conditional CDF

$$F_0^d(t \mid x) := \Pr(Y \leq t \mid S = 1, D = d, X = x), \quad d \in \{1, 0\}.$$

Lemma A.3 (Bound on remainder terms). Under Assumptions 3 and 4,

$$\sup_{d \in \{1, 0\}} \sup_{\xi \in \Xi_N} |\mathbb{E}\{|\tau_0(X)| \leq \rho_N\} R_d(X, \xi)| = O((\rho_N/\eta)^\alpha (\rho_N \vee q_N^\infty)). \quad (\text{A.13})$$

Proof of Lemma A.3. Step 1. Let $d = 1$. For any ξ , Assumption 3 (BO) implies a bound

$$\begin{aligned} |R_1(X, \xi)| &\leq M |F_0^1(Q^1(1 - p(X), X))|_{s_0(1, X)} \\ &\leq M |F_0^1(Q^1(1 - p(X), X)) - (1 - p(X))|_{s_0(1, X)} + M |(1 - p(X))|_{s_0(1, X)}, \quad (\text{A.14}) \end{aligned}$$

where $p(x) = p^{\text{trim}}(x)$ for $\xi \neq \xi_0$ and $p(x) = p_0(x)$ for $\xi = \xi_0$. By construction, $p^{\text{trim}}(x) \in [\zeta_N, 1 - \zeta_N]$ for any x . Invoking mean value theorem and Assumption 3 (REG) gives

$$\sup_{\xi \in \Xi_N} |F_0^1(Q^1(1 - p(x), x)) - F_0^1(Q_0^1(1 - p(x), x), x)| \quad (\text{A.15})$$

$$\leq \sup_{\xi \in \Xi_N} \sup_{u \in [\zeta_N, 1 - \zeta_N]} \sup_{t \in \mathcal{Y}_x} \sup_{x \in \mathcal{X}} f_1(t | x) |Q^1(u, x) - Q_0^1(u, x)| = O(q_N^\infty). \quad (\text{A.16})$$

For $\xi = \xi_0$, the first summand in (A.14) reduces to zero

$$F_0^1(Q_0^1(1 - p_0(X), X)) - (1 - p_0(X)) = 0.$$

Step 2. (A.12) implies a bound on the second term

$$\begin{aligned} & \sup_{\tau \in S_N^1 - S_N^0} \sup_{x \in \mathcal{X}} 1\{|\tau_0(x)| \leq \rho_N\} |(1 - p^{\text{trim}}(x))|s_0(1, x)| \\ & \leq \sup_{\tau \in S_N^1 - S_N^0} \sup_{x \in \mathcal{X}} 1\{|\tau_0(x)| \leq \rho_N\} \max(\kappa^{-1}|\tau(x)|, \rho_N) \\ & \leq \sup_{\tau \in S_N^1 - S_N^0} \sup_{x \in \mathcal{X}} 1\{|\tau_0(x)| \leq \rho_N\} \max(\kappa^{-1}|\tau_0(x)| + \kappa^{-1}2s_N^\infty, \rho_N) = O(\rho_N). \end{aligned}$$

If $\xi = \xi_0$,

$$\sup_{x \in \mathcal{X}} 1\{|\tau_0(x)| \leq \rho_N\} |R_1(x, \xi_0)| \leq M \sup_{x \in \mathcal{X}} 1\{|\tau_0(x)| \leq \rho_N\} |\tau_0(x)| = O(\rho_N).$$

Assumption 3 (MA) implies $\Pr(|\tau_0(X)| \leq \rho_N) \leq (\rho_N/\eta)^\alpha$. A similar argument applies to $d = 0$, which implies (A.13). □

Proof of Lemma 2. I establish the bound

$$\sup_{\xi \in \Xi_N} |\mathbb{E} 1\{|\tau_0(X)| \leq \rho_N\} (m_U(W, \xi) - m_U(W, \xi_0))| = O((\rho_N/\eta)^\alpha \rho_N \vee q_N^\infty) \quad (\text{A.17})$$

Lemma 2 is a special case of (A.17) with a known quantile function $Q_0(u, x)$, in which case

$q_N^\infty = 0$. Define the function

$$\phi_0(x) := \mathbb{E}[S \cdot Y \mid D = 1, X = x] - \mathbb{E}[S \cdot Y \mid D = 0, X = x]$$

Decompose

$$1 = 1\{Y \geq Q^1(1 - p(X), X)\} + 1\{Y \leq Q^1(1 - p(X), X)\},$$

and observe that

$$\begin{aligned} \mathbb{E}[S \cdot Y \mid D = 1, X] &= \mathbb{E}[S \cdot Y \cdot 1\{Y \geq Q^1(1 - p(X), X)\} \mid D = 1, X] \\ &\quad + \mathbb{E}[S \cdot Y \cdot 1\{Y \leq Q^1(1 - p(X), X)\} \mid D = 1, X], \end{aligned}$$

which implies

$$\mathbb{E}[m_U^{\text{help}}(W, \xi) \mid X] - \phi_0(X) = -R_1(X, \xi), \quad \mathbb{E}[m_U^{\text{hurt}}(W, \xi) \mid X] - \phi_0(X) = R_0(X, \xi).$$

Therefore,

$$\begin{aligned} |\mathbb{E}[m_U(W, \xi) - m_U(W, \xi_0) \mid X]| &= |\mathbb{E}[m_U(W, \xi) \pm \phi_0(X) - m_U(W, \xi_0) \mid X]| \\ &\leq 2 \sup_{d \in \{1, 0\}} |R_d(X, \xi)| + |R_d(X, \xi_0)|. \end{aligned}$$

□

Lemma A.4 (Orthogonal moment at the boundary).

$$\sup_{\xi \in \Xi_N} |\mathbb{E}1\{|\tau_0(X)| \leq \rho_N\} (g_U(W, \xi) - g_U(W, \xi_0))| = O((\rho_N/\eta)^\alpha \rho_N \vee q_N^\infty) \quad (\text{A.18})$$

$$\sup_{\xi \in \Xi_N} \mathbb{E}1\{|\tau_0(X)| \leq \rho_N\} (g_U(W, \xi) - g_U(W, \xi_0))^2 = O((\rho_N/\eta)^\alpha) \quad (\text{A.19})$$

Proof of Lemma A.4. Step 1. Bound on (A.18) By Lemma 2, it suffices to show (A.18), replacing $g_U(W, \xi)$ by $\text{cor}_U(W, \xi)$ in (A.48)-(A.49). Observe that

$$\text{cor}_U^{\text{help}}(W, \xi) := \Lambda_\alpha(X, \xi)R_\alpha(W, \xi) + \Lambda_\beta(X, \xi)R_\beta(W, \xi) + \Lambda_\gamma(X, \xi)R_\gamma(W, \xi),$$

where

$$\begin{aligned}\Lambda_\alpha(X, \xi) &= Q^1(1 - p(X), X), \quad R_\alpha(W, \xi) := (1 - D)S/\mu_0(X) - s(0, X) \\ \Lambda_\beta(X, \xi) &= -Q^1(1 - p(X), X)p(X), \quad R_\beta(W, \xi) := DS/\mu_1(X) - s(1, X) \\ \Lambda_\gamma(X, \xi) &= Q^1(1 - p(X), X)s(1, X), \quad R_\gamma(W, \xi) := \frac{D \cdot S \cdot 1\{Y \leq Q^1(1 - p(X), X)\}}{s(1, X)\mu_1(X)} - 1 + p(X).\end{aligned}$$

and

$$\mathbb{E}[\text{cor}^U(W, \xi_0) | X] = 0.$$

By Assumption 4, $\sup_{j \in \{\alpha, \beta, \gamma\}} \sup_{\xi \in \Xi_N} |\Lambda_j(X, \xi)| \leq M$ a.s. As for residuals $R_\alpha(W, \xi)$ and $R_\beta(W, \xi)$,

$$\sup_{x \in \mathcal{X}} \sup_{\xi \in \Xi_N} |\mathbb{E}[R_j(W, \xi) | X = x]| \leq \sup_{d \in \{0, 1\}} \sup_{x \in \mathcal{X}} |s_0(d, x) - s(d, x)| \leq s_N^\infty \leq \rho_N, \quad j \in \{\alpha, \beta\}.$$

Invoking (A.15) for

$$\mathbb{E}[R_\gamma(W, \xi) | X = x] = F_0^1(Q^1(1 - p(x), x)) - F_0^1(Q_0^1(1 - p(x), x), x).$$

gives (A.18). A similar argument holds for $\text{cor}_U^{\text{hurt}}(W, \xi)$.

Step 2. Bound on (A.19) By Assumptions 3-4, $\sup_{\xi \in \Xi_N} |g_U(W, \xi)| \leq 8M\kappa^{-1}$ a.s. and

$$\mathbb{E}1\{|\tau_0(X)| \leq \rho_N\} (g_U(W, \xi) - g_U(W, \xi_0))^2 \leq (8M\kappa^{-1})^2 \Pr(|\tau_0(X)| \leq \rho_N) = O((\rho_N/\eta)^\alpha).$$

□

Lemma A.5 (Moment Bounds). Under Assumptions 3 and 4,

$$\sup_{\xi \in \Xi_N} |\mathbb{E}1\{|\tau_0(X)| > \rho_N\} (g_U(W, \xi) - g_U(W, \xi_0))| = O(s_N^2 + q_N^2) \quad (\text{A.20})$$

$$\sup_{\xi \in \Xi_N} \mathbb{E}1\{|\tau_0(X)| > \rho_N\} (g_U(W, \xi) - g_U(W, \xi_0))^2 = O(s_N + q_N) \quad (\text{A.21})$$

Proof of Lemma A.5. Step 1. For any $X : |\tau_0(X)| > \rho_N$ and $\xi \in \Xi_N$, the covariate X must be

classified correctly and $p(X) = s(0, X)/s(1, X) \in [\xi_N, 1 - \xi_N]$ without trimming. For any such X , define

$$\phi(r; x, \xi) := \mathbb{E}[g_U^{\text{help}}(W; r(\xi - \xi_0) + \xi_0) - g_U^{\text{help}}(W, \xi_0) \mid X = x]$$

and let $\partial_\alpha, \partial_\beta, \partial_\gamma$ denote derivatives of $\mathbb{E}[g_U(W, \xi_0) \mid X = x]$ with respect to the coordinates of

$$\xi_0(x) = \{s(0, x), s(1, x), Q^1(u, x)\}.$$

The first-order derivative is

$$\begin{aligned} \phi'(0; X, \xi_0) &= \partial_\alpha \mathbb{E}[g_U(W, \xi_0) \mid X][s(0, X) - s_0(0, X)] \\ &\quad + \partial_\beta \mathbb{E}[g_U(W, \xi_0) \mid X][s(1, X) - s_0(1, X)] \\ &\quad + \partial_\gamma \mathbb{E}[g_U(W, \xi_0) \mid X][Q(1 - p_0(X), X) - Q_0(1 - p_0(X), X)] \\ &= 0 + 0 + 0, \end{aligned}$$

since $\Lambda_j(X) := \partial_j \mathbb{E}[m_U(W, \xi_0) \mid X]$ for $j \in \{\alpha, \beta, \gamma\}$ by construction. The second derivative is

$$|\phi''(r; X, \xi_0)| = |[\xi(X) - \xi_0(X)]' B(r; X) [\xi(X) - \xi_0(X)]| \leq \|\xi(X) - \xi_0(X)\|^2 \max \text{eig}(B(r; X)),$$

where $B(r; x) := \nabla_{\xi_0}^2 \mathbb{E}[g_U(W, r(\xi - \xi_0) + \xi_0) \mid X = x]$ is a 3x3 matrix of second derivatives. Step 4 shows that $B(r; x)$ has bounded entries uniformly over x and $r \in (0, 1)$. For some $\tilde{r} \in (0, 1)$, second-order Taylor expansion at each x

$$\phi(1; x, \xi) = \phi(0; x, \xi_0) + \phi'(0; x, \xi_0) + 0.5\phi''(\tilde{r}; x, \xi_0) = 0.5\phi''(\tilde{r}; x, \xi_0).$$

As a result,

$$\sup_{\xi \in \Xi_N} |\mathbb{E}\phi(1; X, \xi) 1\{|\tau_0(X)| > \rho_N\}| = O\left(\sup_{\xi \in \Xi_N} \mathbb{E} 1\{|\tau_0(X)| > \rho_N\} \|\xi(X) - \xi_0(X)\|^2\right) = O(s_N^2 + q_N^2).$$

Step 3. Observe that

$$(g_U^{\text{help}}(W, \xi) - g_U^{\text{help}}(W, \xi_0))^2 \leq 4[(m_U^{\text{help}}(W, \xi) - m_U^{\text{help}}(W, \xi_0))^2 + \sum_{j \in \{\alpha, \beta, \gamma\}} (\Lambda_j(X, \xi)R_j(W, \xi) - \Lambda_j(X, \xi_0)R_j(W, \xi_0))^2].$$

The bound on the first term is

$$\begin{aligned} & \mathbb{E}[(m_U^{\text{help}}(W, \xi) - m_U^{\text{help}}(W, \xi_0))^2 | X] \\ & \leq \kappa^{-1} M^2 \mathbb{E}[(1\{Y \geq Q^1(1 - p(X), X)\} - 1\{Y \geq Q_0^1(1 - p_0(X), X)\})^2 | S = 1, D = 1, X] \end{aligned}$$

Therefore,

$$\begin{aligned} & \mathbb{E}[(m_U^{\text{help}}(W, \xi) - m_U^{\text{help}}(W, \xi_0))^2 1\{|\tau_0(X)| > \rho_N\}] \\ & \leq \kappa^{-1} M^2 (\mathbb{E}1\{|\tau_0(X)| > \rho_N\} (F_0^1(Q^1(1 - p(X), X)) - (1 - p_0(X)))^2)^{1/2} = O(s_N + q_N). \end{aligned}$$

The bound on the second term's multiplier $\Lambda_j(X, \xi)$ is

$$\begin{aligned} & \sup_{\xi \in \Xi_N} \mathbb{E}1\{|\tau_0(X)| > \rho_N\} (\Lambda_j(X, \xi) - \Lambda_j(X, \xi_0))^2 \\ & = \sup_{\xi \in \Xi_N} \mathbb{E}1\{|\tau_0(X)| > \rho_N\} (Q^1(1 - p(X), X) - Q_0^1(1 - p_0(X), X))^2 = O(s_N^2 + q_N^2), \end{aligned}$$

and a similar bound holds for $j \in \{\beta, \gamma\}$. For $j \in \{\alpha, \beta\}$, the bound on $R_j(W; \xi) - R_j(W; \xi_0)$ in mean square sense is

$$\sup_{\xi \in \Xi_N} \mathbb{E}(R_j(W, \xi) - R_j(W, \xi_0))^2 \leq \sup_{d \in \{1, 0\}} \sup_{\xi \in \Xi_N} \mathbb{E}(s(d, X) - s_0(d, X))^2 \leq s_N^2.$$

For $j = \gamma$, the bound on

$$\sup_{\xi \in \Xi_N} \mathbb{E}[(R_j(W, \xi) - R_j(W, \xi_0))^2] = O(s_N + q_N).$$

Step 4. The first partial derivatives of $Q(p(x), x)$ w.r.t. α, β, γ take the form

$$\begin{aligned}\partial_\alpha Q(p(x), x) &= f^{-1}(Q(p_0(x), x)|x)s^{-1}(1, x) \\ \partial_\beta Q(p(x), x) &= f^{-1}(Q(p_0(x), x)|x)s^{-2}(1, x)s(0, x) \\ \partial_\gamma Q(p(x), x) &= 1.\end{aligned}$$

The second partial derivatives of $Q(p(x), x)$ w.r.t. α, β, γ take the form

$$\begin{aligned}\partial_{\alpha\alpha}^2 Q(p(x), x) &= -f^{-2}(Q(p_0(x), x)|x)f'(Q(p_0(x), x)|x)s^{-2}(1, x) \\ \partial_{\beta\beta}^2 Q(p(x), x) &= -f^{-2}(Q(p_0(x), x)|x)f'(Q(p_0(x), x)|x)s^{-4}(1, x)s^2(0, x) \\ &\quad + 2f^{-1}(Q(p_0(x), x)|x)s^{-3}(1, x)s(0, x) \\ \partial_{\alpha\beta}^2 Q(p(x), x) &= -f^{-1}(Q(p_0(x), x)|x)s^{-2}(1, x) \\ \partial_{\beta\alpha}^2 Q(p(x), x) &= -s^{-2}(1, x)f^{-1}(Q(p_0(x), x)|x) - f^{-2}(Q(p_0(x), x)|x)s^{-2}(1, x)s(0, x) \\ \partial_{\gamma\alpha}^2 Q(p(x), x) &= f^{-2}(Q(p_0(x), x)|x)s^{-1}(1, x)f'(Q(p_0(x), x)|x) \\ \partial_{\gamma\beta}^2 Q(p(x), x) &= f^{-2}(Q(p_0(x), x)|x)s^{-2}(1, x)f'(Q(p_0(x), x)|x)s(0, x) \\ \partial_{\alpha\gamma}^2 Q(p(x), x) &= \partial_{\beta\gamma}^2 Q(p(x), x) = \partial_{\gamma\gamma}^2 Q(p(x), x) = 0.\end{aligned}$$

By Assumption 3 (SO) and (REG), all functions of x above are bounded a.s. in X . By definition of S_N^d, Q_N^d in Assumption 4(2), all other elements of $B(r; X)$ are bounded from above a.s. in X and $r \in (0, 1)$. □

Proof of Theorem 1. By Lemmas A.1 and A.2,

$$\sqrt{N}(\mathbb{E}_N g_D(W_i; \hat{\tau}_i) - g_D(W_i, \tau_0)) = o_P(N^{-1/2}).$$

By Lemma A.4 and A.5, the first moments are bounded as

$$\begin{aligned} \sup_{\xi \in \Xi_N} |\mathbb{E}(g_U(W, \xi) - g_U(W, \xi_0))| &= |\mathbb{E}1\{|\tau_0(X)| \leq \rho_N\}(g_U(W, \xi) - g_U(W, \xi_0))| \\ &\quad + |\mathbb{E}1\{|\tau_0(X)| \geq \rho_N\}(g_U(W, \xi) - g_U(W, \xi_0))| \\ &\leq O((\rho_N/\eta)^\alpha(\rho_N \vee q_N^\infty) + s_N^2 + q_N^2). \end{aligned}$$

The second moments are bounded as

$$\sup_{\xi \in \Xi_N} \mathbb{E}(g_U(W, \xi) - g_U(W, \xi_0))^2 \leq O((\rho_N/\eta)^\alpha + s_N + q_N).$$

By Lemma A.1, $\sqrt{N}(\mathbb{E}_N g_U(W_i; \widehat{\xi}_i) - \mathbb{E}_N g_U(W_i, \xi_0)) = o_P(N^{-1/2})$. A similar argument applies to the lower bound, which gives $\sqrt{N}(\mathbb{E}_N g_L(W_i; \widehat{\xi}_i) - \mathbb{E}_N g_L(W_i, \xi_0)) = o_P(N^{-1/2})$. Invoking Delta method for $\psi(x, y, z) := (x/z, y/z)'$ with $x = \mathbb{E}_N g_L(W_i, \xi_0)$; $y = \mathbb{E}_N g_U(W_i, \xi_0)$; $z = \mathbb{E}_N g_D(W_i, \xi_0)$ gives the statement of the Theorem. □

Proof of Theorem 2. Step 1. I show that $\sigma(q)$ is (1) convex, (2) positive homogenous of degree one and (3) lower-semicontinuous function of q . By Corollary 13.2.1 from Rockafellar (1997), the properties (1)-(3) imply that \mathcal{B} in (5.3) is a convex and compact set and $\sigma(q)$ is its support function.

Step 2. Verification of (1). Lemma 1 proves that $\sigma(\lambda q_1 + (1 - \lambda)q_2)$ is a sharp upper bound on $(\lambda q_1 + (1 - \lambda)q_2)'\beta_0$. Furthermore, by Lemma 1,

$$q_1'\beta_0 \leq \sigma(q_1) \quad \text{and} \quad q_2'\beta_0 \leq \sigma(q_2).$$

Therefore, $(\lambda q_1 + (1 - \lambda)q_2)'\beta_0 \leq \lambda \sigma(q_1) + (1 - \lambda)\sigma(q_2)$. By sharpness, $\sigma(\lambda q_1 + (1 - \lambda)q_2)$ is the smallest bound on $(\lambda q_1 + (1 - \lambda)q_2)'\beta_0$. Therefore,

$$\sigma(\lambda q_1 + (1 - \lambda)q_2) \leq \lambda \sigma(q_1) + (1 - \lambda)\sigma(q_2),$$

which implies that $\sigma(q)$ is a convex function of q .

Verification of (2). Let $\lambda > 0$. Observe that the event $\{\lambda Y_q \geq \mathcal{Q}_{\lambda Y_q}(u, X)\}$ holds if and only if $\{Y_q \geq \mathcal{Q}_{Y_q}(u, X)\}$. Since $Y_q = q'Y$ is a linear function of q , $\sigma(q)$ defined in (5.2) is positive homogenous of degree 1.

Verification of (3). Consider a sequence of vectors $q_k \rightarrow q, k \rightarrow \infty$. Suppose $\sigma(q_k) \leq C$. Then, $q'_k \beta_0 \leq \sigma(q_k) \leq C$, which implies that $q' \beta_0 \leq C$ must hold. Therefore, C is a bound on $q' \beta_0$. By sharpness, $\sigma(q)$ is the smallest bound on $q' \beta_0$, which implies $\sigma(q) \leq C$. \square

Define the lower-truncated subjects

$$\Lambda_L(W) = \left\{ W : (Z = 1 \cap S = 1 \cap Y \leq \mathcal{Q}^1(p_0(X), D, X)) \cup (Z = 0 \cap S = 1) \right\}. \quad (\text{A.22})$$

and the upper-truncated ones

$$\Lambda_U(W) = \left\{ W : (Z = 1 \cap S = 1 \cap Y \geq \mathcal{Q}^1(1 - p_0(X), D, X)) \cup (Z = 0 \cap S = 1) \right\}. \quad (\text{A.23})$$

Bayes rule implies

$$\begin{aligned} \Pr(\Lambda_U(W) \mid Z = 1, D(1) = d, X) &= p_0(X) \Pr(S = 1 \mid Z = 1, D(1) = d, X) & (\text{A.24}) \\ &= p_0(X) \Pr(S = 1 \mid Z = 1, X) = p_0(X) s(1, X) \\ &= s(0, X) \quad \forall d \in \{1, 0\}. \end{aligned}$$

Union bound implies

$$\Pr(\Lambda_U(W) \mid Z = 1, X) = \sum_{d=0}^{d=1} p_0(X) \Pr(S = 1 \mid Z = 1, D(1) = d, X) \Pr(D(1) = d \mid Z = 1, X) \quad (\text{A.25})$$

$$= s(0, X) \sum_{d=0}^{d=1} \Pr(D(1) = d \mid Z = 1, X) = s(0, X). \quad (\text{A.26})$$

and

$$\Pr(\Lambda_U(W) \mid X) = s(0, X) \Pr(Z = 1 \mid X) + \Pr(S = 1 \mid Z = 0, X) \Pr(Z = 0 \mid X) = s(0, X). \quad (\text{A.27})$$

Invoking Bayes rule gives equality of covariate densities

$$f_{\Lambda_U}(x) = \frac{\Pr(\Lambda_U(W) | X = x)f(x)}{\mathbb{E}\Pr(\Lambda_U(W) | X)} = \frac{s(0,x)f(x)}{\mathbb{E}s(0,X)}. \quad (\text{A.28})$$

Likewise,

$$\Pr(S(1) = S(0) = 1 | Z = 1, X) = s(0, X) \sum_{d=0}^{d=1} \Pr(D(1) = d | Z = 1, X) = s(0, X), \quad (\text{A.29})$$

which implies $f_{11}(x) = f_{\Lambda_U}(x)$. Assumption 5 (3) implies

$$\Pr(D = 1 | Z = 1, X, S(1) = S(0) = 1) = \Pr(D = 1 | Z = 1, X). \quad (\text{A.30})$$

Proof of Theorem 3. Step 1. As discussed in Lee (2009), for each $d \in \{1, 0\}$ and x , the group $Z = 1, D(1) = d, X = x$ consists of the always-takers and compliers. By Assumption 5 (3), the share of the always-takers is equal to $p_0(X)$

$$\Pr(S(1) = S(0) = 1 | D(1) = d, Z = 1, X) = \Pr(S(1) = S(0) = 1 | Z = 1, X) = \frac{s(0, X)}{s(1, X)}. \quad (\text{A.31})$$

Invoking Corollary 4.1 in Horowitz and Manski (1995) gives an upper bound

$$\mathbb{E}_{\Lambda_U}[Y | D(1) = d, Z = 1, X] \geq \mathbb{E}_{\Lambda_{11}}[Y | D(1) = d, Z = 1, X], \quad d = 1, 0. \quad (\text{A.32})$$

Invoking Bayes rule gives

$$\begin{aligned} \Pr_{\Lambda_U}(D = 1 | Z = 1, X) &= \frac{\Pr(\Lambda_U(W) | Z = 1, D(1) = 1, X) \Pr(D = 1 | Z = 1, X)}{\Pr(\Lambda_U(W) | Z = 1, X)} \\ &\stackrel{i}{=} \frac{s(0, X) \Pr(D = 1 | Z = 1, X)}{s(0, X)} = \Pr(D = 1 | Z = 1, X), \end{aligned} \quad (\text{A.33})$$

where (i) follows from (A.25) and (A.29). Likewise, $\Pr_{\Lambda_U}(D = 0 | Z = 1, X) = \Pr(D = 0 | Z = 1, X)$. Invoking bound (A.32) for each $d = 1$ and $d = 0$ and (A.33) gives

$$\mathbb{E}_{\Lambda_U}[Y | Z = 1, X] \geq \mathbb{E}_{\Lambda_{11}}[Y | Z = 1, X]. \quad (\text{A.34})$$

Step 2. The group $Z = 0, X = x$ is not truncated. For each $d = 1, 0$,

$$\begin{aligned}\mathbb{E}_{\Lambda_U}[Y \mid D(0) = d, Z = 0, X] &= \mathbb{E}[Y(d) \mid S = 1, D(0) = d, Z = 0, X] \\ &= \mathbb{E}[Y(d) \mid S(0) = 1, D(0) = d, X] = \mathbb{E}[Y(d) \mid S(1) = S(0) = 1, D(0) = d, X] \\ &= \mathbb{E}_{\Lambda_{11}}[Y \mid D(0) = d, Z = 0, X].\end{aligned}$$

Summing over $d = 1$ and $d = 0$ gives

$$\mathbb{E}_{\Lambda_U}[Y \mid Z = 0, X] = \mathbb{E}_{\Lambda_{11}}[Y \mid Z = 0, X]. \quad (\text{A.35})$$

Step 3. (A.34) and (A.35) obey the following inequality for the numerators of β_U

$$\mathbb{E}_{\Lambda_U}[Y \mid Z = 1, X] - \mathbb{E}_{\Lambda_U}[Y \mid Z = 0, X] \geq \mathbb{E}_{\Lambda_{11}}[Y \mid Z = 1, X] - \mathbb{E}_{\Lambda_{11}}[Y \mid Z = 0, X]. \quad (\text{A.36})$$

(A.30) and (A.33) imply equality of denominators

$$\Pr_{\Lambda_U}(D = 1 \mid Z = 1, X) = \Pr_{\Lambda_{11}}(D = 1 \mid Z = 1, X).$$

Finally, (A.28) implies that the $\mathbb{E}_{\Lambda_U}[Y(1) - Y(0) \mid S(1) = S(0) = 1, D(1) > D(0), X]$ and $\mathbb{E}_{\Lambda_{11}}[Y(1) - Y(0) \mid S(1) = S(0) = 1, D(1) > D(0), X]$ are integrated with respect to the same covariate densities. Therefore, the statement (1) of Theorem 3 holds. □

General Case Notation. The propensity score is

$$\mu_1(X) = \Pr(D = 1 \mid X), \quad \mu_0(X) = 1 - \mu_1(X) = \Pr(D = 0 \mid X). \quad (\text{A.37})$$

The conditional quantiles in the selected treated and selected control groups are

$$Q^d(u, x) : \Pr(Y \leq Q^d(u, x) \mid S = 1, D = d, X = x) = u, \quad u \in [0, 1], \quad d \in \{1, 0\}. \quad (\text{A.38})$$

Because $Q^1(u, x)$ is invoked only for $x \in \mathcal{X}_{\text{help}}$ and $Q^0(u, x)$ is invoked only for $x \in \mathcal{X}_{\text{hurt}}$, it makes sense to define combined conditional quantile:

$$Q(u, x) = 1\{x \in \mathcal{X}_{\text{help}}\}Q^1(u, x) + 1\{x \in \mathcal{X}_{\text{hurt}}\}Q^0(u, x). \quad (\text{A.39})$$

Likewise, the conditional outcome densities in the selected treated and selected control groups are

$$f^d(t|x) = f_{Y|S=1, D=d, X=x}(y|x), \quad d \in \{1, 0\}$$

and combined conditional density is

$$f(t|x) = 1\{x \in \mathcal{X}_{\text{help}}\}f^1(t|x) + 1\{x \in \mathcal{X}_{\text{hurt}}\}f^0(t|x). \quad (\text{A.40})$$

For $x \in \mathcal{X}_{\text{help}}$, the conditional upper bound is

$$\bar{\beta}_U^{\text{help}}(x) = \mathbb{E}[Y|D = 1, S = 1, Y \geq Q(1 - p_0(x), x), X = x] - \mathbb{E}[Y|D = 0, S = 1, X = x] \quad (\text{A.41})$$

and the conditional lower bound is

$$\bar{\beta}_L^{\text{help}}(x) = \mathbb{E}[Y|D = 1, S = 1, Y \leq Q(p_0(x), x), X = x] - \mathbb{E}[Y|D = 0, S = 1, X = x]. \quad (\text{A.42})$$

For $x \in \mathcal{X}_{\text{hurt}}$, the conditional upper bound is

$$\bar{\beta}_U^{\text{hurt}}(x) = \mathbb{E}[Y|D = 1, S = 1, X = x] - \mathbb{E}[Y|D = 0, S = 1, Y \leq Q(1/p_0(x), x), X = x] \quad (\text{A.43})$$

and the conditional lower bound is

$$\bar{\beta}_L^{\text{hurt}}(x) = \mathbb{E}[Y|D = 1, S = 1, X = x] - \mathbb{E}[Y|D = 0, S = 1, Y \geq Q(1 - 1/p_0(x), x), X = x]. \quad (\text{A.44})$$

The sharp Lee bounds β_L and β_U are:

$$\beta_\star = \frac{\int_{\mathcal{X}} \beta_\star(x) \min(s(0,x), s(1,x)) f_X(x) dx}{\int_{\mathcal{X}} \min(s(0,X), s(1,X)) f_X(x) dx} = \frac{\mathbb{E} m_\star(W, \xi_0)}{\mathbb{E} \min(s(0,X), s(1,X))}, \quad \star \in \{L, U\}. \quad (\text{A.45})$$

The non-orthogonal moment equations for the numerator of β_U are

$$m_U(W, \xi_0) = \begin{cases} \frac{D}{\mu_1(X)} \cdot S \cdot Y 1\{Y \geq Q(1 - p_0(X), X)\} - \frac{(1-D)}{\mu_0(X)} \cdot S \cdot Y, & X \in \mathcal{X}_{\text{help}} \\ \frac{D}{\mu_1(X)} \cdot S \cdot Y - \frac{(1-D)}{\mu_0(X)} \cdot S \cdot Y 1\{Y \leq Q(1/p_0(X), X)\}, & X \in \mathcal{X}_{\text{hurt}} \end{cases} \quad (\text{A.46})$$

and

$$m_L(W, \xi_0) = \begin{cases} \frac{D}{\mu_1(X)} \cdot S \cdot Y 1\{Y \leq Q(p_0(X), X)\} - \frac{(1-D)}{\mu_0(X)} \cdot S \cdot Y, & X \in \mathcal{X}_{\text{help}} \\ \frac{D}{\mu_1(X)} \cdot S \cdot Y - \frac{(1-D)}{\mu_0(X)} \cdot S \cdot Y 1\{Y \geq Q(1 - 1/p_0(X), X)\}, & X \in \mathcal{X}_{\text{hurt}}. \end{cases} \quad (\text{A.47})$$

The bias correction terms are

$$\begin{aligned} \text{cor}_U^{\text{help}}(W, \xi_0) = Q(1 - p_0(X), X) & \left[\left(\frac{(1-D) \cdot S}{\mu_0(X)} - s(0, X) \right) \right. \\ & - p_0(X) \left(\frac{D \cdot S}{\mu_1(X)} - s(1, X) \right) \\ & \left. + s(1, X) \left(\frac{D \cdot S \cdot 1\{Y \leq Q(1 - p_0(X), X)\}}{s(1, X) \mu_1(X)} - 1 + p_0(X) \right) \right], \end{aligned} \quad (\text{A.48})$$

$$\begin{aligned} \text{cor}_U^{\text{hurt}}(W, \xi_0) = Q(1/p_0(X), X) & \left[(1/p_0(X)) \left(\frac{(1-D) \cdot S}{\mu_0(X)} - s(0, X) \right) \right. \\ & + \left(\frac{D \cdot S}{\mu_1(X)} - s(1, X) \right) \\ & \left. + s(0, X) \left(\frac{(1-D) \cdot S \cdot 1\{Y \leq Q(1/p_0(X), X)\}}{s(0, X) \mu_1(X)} - 1/p_0(X) \right) \right] \end{aligned} \quad (\text{A.49})$$

For the lower bound

$$\begin{aligned} \text{cor}_L^{\text{help}}(W, \xi_0) = & Q(p_0(X), X) \left[\left(\frac{(1-D) \cdot S}{\mu_0(X)} - s(0, X) \right) \right. \\ & - p_0(X) \left(\frac{D \cdot S}{\mu_1(X)} - s(1, X) \right) \\ & \left. - s(1, X) \left(\frac{D \cdot S \cdot 1_{\{Y \leq Q(p_0(X), X)\}}}{s(1, X) \mu_1(X)} - p_0(X) \right) \right]. \end{aligned} \quad (\text{A.50})$$

$$\begin{aligned} \text{cor}_L^{\text{hurt}}(W, \xi_0) = & -Q(1 - 1/p_0(X), X) \left[(1/p_0(X)) \left(\frac{(1-D) \cdot S}{\mu_0(X)} - s(0, X) \right) \right. \\ & - \left(\frac{D \cdot S}{\mu_1(X)} - s(1, X) \right) \\ & \left. - \left(\frac{(1-D) \cdot S \cdot 1_{\{Y \leq Q(1-1/p_0(X), X)\}}}{\mu_1(X)} - s(1, X) + s(0, X) \right) \right]. \end{aligned} \quad (\text{A.51})$$

Finally, the correction term for the propensity score for the first summand in moment (4.6) is

$$S_{1\mu} := -\frac{1}{\mu_1(X)} \mathbb{E}[Y \mid Y \geq Q(1 - p_0(X), X), D = 1, S = 1, X] \cdot s(0, X) \cdot (D - \mu_1(X)) \quad (\text{A.52})$$

and for the second one is $S_{0\mu} = \frac{1}{(1 - \mu_1(X))} \mathbb{E}[Y \mid D = 0, S = 1, X] \cdot s(0, X) \cdot (D - \mu_1(X))$, which gives

$$S_{1\mu} + S_{0\mu} = -\bar{\beta}_U^{\text{help}}(X) \cdot s(0, X) \cdot (D - \mu_1(X)).$$

Thus, if the propensity score is estimated, the total bias correction term is $\text{cor}_U^{\text{help}}(W, \xi_0) + S_{1\mu} + S_{0\mu}$.

Appendix B. Supplementary results

Replication. Table B.1 replicates Lee's estimates of basic (Column (1)) and covariate-based (Column (2)) bounds on JobCorps effect on the wages of always-takers. Week 90 is the only horizon where Lee found JobCorps effect on wages to be statistically significant. However,

basic Lee bounds do not overlap with the covariate-based ones. Sharpness fails because one of the five covariate-specific trimming thresholds exceeds 1 and is being capped at 0.999 to impose unconditional monotonicity. Capping corresponds to the researcher’s belief that the covariate-specific threshold exceeded 1 due to sampling noise, the only belief consistent with unconditional monotonicity.

A.3 JobCorps Data description.

In this section, I describe baseline covariates for the JobCorps empirical application. The data is taken from Schochet et al. (2008), who provides covariate descriptions in Appendix L. All covariates describe experiences before random assignment (RA). Most of the covariates represent answers to multiple choice questions; for these covariates I list the question and the list of possible answers. An answer is highlighted in boldface if it is selected by post-lasso-logistic of Belloni et al. (2016) for one of employment equation specifications, described below. Table B.3 lists the covariates selected by Lee (2009). A full list of numeric covariates, not provided here, includes $p = 5,177$ numeric covariates.

Covariates selected by Lee (2009). Lee (2009) selected 28 baseline covariates to estimate parametric specification of the sample selection model. They are given in Table B.3.

Reasons for joining JobCorps (R_X). Applicants were asked a question “How important was reason X on the scale from 1 (very important) to 3 (not important), or 4 (N/A), for joining JobCorps?”. Each reason X was asked about in an independent question.

Table B.2: Reasons for joining JobCorps

Name	description	Name	description
R_HOME	getting away from home	R_COMM	getting away from community
R_GETGED	getting a GED	R_CRGOAL	desire to achieve a career goal
R_TRAIN	getting job training	R_NOWORK	not being able to find work

For example, a covariate R_HOME1 is a binary indicator for the reason R_HOME being ranked as a very important reason for joining JobCorps.

Table B.1: Estimated bounds on the JobCorps effect on log wages under monotonicity.

	Basic (1)	Covariate-based (2)
Week 45	[-0.072, 0.140] (-0.097, 0.170) (-0.096, 0.168)	[-0.074, 0.127] (-0.096, 0.156) (-0.096, 0.155)
Week 90	[0.048, 0.049] (0.011, 0.081) (0.012, 0.081)	[0.036, 0.048] (0.009, 0.075) (0.011, 0.073)
Week 104	[0.017, 0.064] (-0.020, 0.102) (-0.012, 0.095)	[0.017, 0.054] (-0.009, 0.081) (-0.007, 0.079)
Week 135	[-0.007, 0.084] (-0.042, 0.113) (-0.037, 0.109)	[-0.001, 0.075] (-0.032, 0.103) (-0.028, 0.100)
Week 180	[-0.032, 0.087] (-0.063, 0.112) (-0.060, 0.109)	[-0.019, 0.080] (-0.048, 0.107) (-0.044, 0.104)
Week 208	[-0.020, 0.095] (-0.050, 0.118) (-0.047, 0.117)	[-0.014, 0.084] (-0.041, 0.109) (-0.039, 0.107)
Covariates	N/A	5

Notes. The sample ($N = 9,145$) and the time horizons are the same as in Lee (2009). Each panel reports estimated bounds (first row), the 95% confidence region for the identified set (second row) and the 95% Imbens and Manski (2004) confidence interval for the true parameter (third row). Column (1) reports basic Lee bounds. Column (2) reports covariate-based Lee bounds. All bounds assume that JobCorps weakly hurts employment in week 45 and helps employment following week 90. The covariate in Column (2) is a linear combination of 28 baseline covariates, selected by Lee, weighted by the coefficients from a regression of Week 208 wages on all baseline characteristics in the control group. Lee refers to this combination as predicted wage potential. The five groups are formed according to whether the predicted wage is within intervals defined by \$6.75, \$7, \$7.50, and \$8.50. Week 90 is highlighted in bold as the only week where Lee found a statistically significant effect on wages. Computations use design weights.

Table B.3: Baseline covariates selected by Lee (2009).

Name	Description
FEMALE	female
AGE	age
BLACK, HISP, OTHERRAC	race categories
MARRIED, TOGETHER, SEPARATED	family status categories
HASCHILD	has child
NCHILD	number of children
EVARRST	ever arrested
HGC	highest grade completed
HGC_MOTH, HGC_FATH	mother's and father's HGC
HH_INC1 – HH_INC5	five household income groups with cutoffs 3,000, 6,000, 9,000, 18,000
PERS_INC1 – PERS_INC4	four personal income groups with cutoffs 3,000, 6,000, 9,000
WKEARNR	weekly earnings at most recent job
HRSWK_JR	usual weekly work hours at most recent job
MOSINJOB	the number of months employed in past year
CURRJOB	employed at the moment of interview
EARN_YR	total yearly earnings
YR_WORK	any work in the year before RA

Sources of advice about the decision to enroll in JobCorps (IMP_X). Applicants were asked a question “How important was advice of X on the scale from 1 (important) to 0 (not important) ?”. Each source of advice was asked about in an independent question.

Table B.4: Sources of advice about the decision to enroll in JobCorps.

Name	description	Name	description
IMP_PAR	parent or legal guardian	IMP_FRD	friend
IMP_TCH	teacher	IMP_CW	case worker
IMP_PRO	probation officer	IMP_CHL	church leader

Main types of worry about joining JobCorps (TYPEWORR). Applicants were asked to select one main type of worry about joining JobCorps.

Table B.5: Types of worry about joining JobCorps

#	description	#	description
1	not knowing anybody or not fitting in	2	violence / safety
3	homesickness	4	not knowing what it will be like
5	dealing with other people	6	living arrangements
7	strict rules and highly regimented life	8	racism
9	not doing well in classes	10	none

Drug use summary (DRUG_SUMP). Applicants were asked to select one of 5 possible answers best describing their drug use in the past year before RA.

Table B.6: Summary of drug use in the year before RA

#	description	#	description
1	did not use drugs	2	marijuana / hashish only
3	drugs other than marijuana / hashish	4	both marijuana and other drugs

Frequency of marijuana use (FRQ_POT) . Applicants were asked to select one of 5 possible answers best describing their marijuana / hashish use in the past year before RA.

Table B.7: Frequency of marijuana/hashish use in the year before RA

#	description	#	description
1	daily	2	a few times each week
3	a few times each month	4	less often
5	missing	6	N/A

Applicant's welfare receipt history. Applicants were asked whether they ever received food stamps (GOTFS), AFDC benefits (GOTAFDC) or other welfare (GOTOTHW) in the year prior to RA. In case of receipt, they asked about the duration of receipt in months (MOS_ANYW, MOS_AFDC). For example, GOTAFDC=1 and MOS_AFDC=8 describes an applicant who received AFDC benefits during 8 months before RA.

Household welfare receipt history (WELF_KID). Applicants were asked about family welfare receipt history during childhood.

Table B.8: Family was on welfare when growing up

#	description	#	description
1	never	2	occasionally
3	half of the time	4	most or all time

Table B.9: Figure 1 details: monotonicity test results

Weeks (1)	Cell with the largest t -statistic (2)	Average Test Statistic (3)
Weeks 60 – 89	MOS_AFDC=8 or PERS_INC=3 and EARN_YR $\in [720, 3315]$	2.390
Weeks 90 – 116	R_HOME=1 and MARRCAT11=1 or WELF_KID=4 and TYPEWORR=5	2.536
Weeks 117 – 152	R_COMM=1 and IMP_PRO=1 and FRQ_POT=3 or DRG_SUMP=2 and TYPEWORR=5 and IMP_PRO=1	2.690
Weeks 153 – 186	IMP_PRO=1 and MARRCAT11 or REASED_R4 = 1 and R_COMM=1 and DRG_SUMP=2	3.303
Weeks 187 – 208	same as weeks 90–116	2.221

Notes. This table shows the results for the monotonicity test in Figure 1. The test is conducted separately for each week using a week-specific test statistic and p-value. For each test, I partition $N = 9,145$ subjects into $J = 2$ cells C_1, C_2 . Column (2) describes the cell with the largest t -statistic whose value is compared to the critical value. Column (3) shows the average test-statistic across time period in Column (1). The test statistic is $T = \max_{j \in \{1,2\}} \hat{\mu}_j / \hat{\sigma}_j$, where $\hat{\mu}_j$ and $\hat{\sigma}_j$ are sample average and standard deviation of random variable $\xi_j := \mathbb{E}[(2D - 1) \cdot S | X \in C_j]$, weighted by design weights DSGN_WGT. The critical value c_α is the self-normalized critical value of Chernozhukov et al. (2019). For $\alpha = 0.05$, $c_\alpha = 1.960$. For $\alpha = 0.01$, $c_\alpha = 2.577$. Covariates are defined in Section A.3.

A.4 Lee (2009): First-Stage Estimates

The first-stage selection estimates are constructed as in (7.2). Let $\mathcal{U}_N = [\xi_N, 1 - \xi_N]$ be an compact set. Focusing on $Q^1(u, x)$, approximate $Q^1(u, x)$ by a linear function

$$Q(u, x) = Z(x)' \delta_0(u) + R(u, x), \quad (\text{B.1})$$

where $Z(x) \in \mathbb{R}^{p_Q}$ is a vector of basis functions, $\delta_0(u)$ is the pseudo-true parameter value, and $R(u, x)$ is approximation error. Let $N_{11} = \sum_{i=1}^N D_i S_i$. The quantile regression estimate

$$\widehat{Q}(u, x) = Z(x)' \widehat{\delta}(u),$$

where $\widehat{\delta}(u), u \in \mathcal{U}$ is

$$\begin{aligned} \widehat{\delta}(u) &:= \arg \min_{\delta \in \mathbb{R}^{p_Q}} \frac{1}{N_{11}} \sum_{D_i=1, S_i=1}^N \rho_u(Y_i - Z(X_i)' \delta) \\ &= \arg \min_{\delta \in \mathbb{R}^{p_Q}} \frac{1}{N_{11}} \sum_{D_i=1, S_i=1}^N (u - 1_{\{Y_i - Z(X_i)' \delta < 0\}}) \cdot (Y_i - Z(X_i)' \delta), \end{aligned} \quad (\text{B.2})$$

converges at mean square rate $q_N = \sqrt{\frac{p_Q}{N}} = o(N^{-1/4})$ and sup-norm rate $q_N^\infty = o(N^{-1/4})$, as shown in Belloni et al. (2019). Its ℓ_1 -penalized analog converges at mean square rate $q_N = \sqrt{\frac{s_Q \log p_Q}{N}} = o(N^{-1/4})$ and sup-norm rate $q_N^\infty = \sqrt{\frac{s_Q^2 \log p_Q}{N}} = o(N^{-1/4})$, as shown in Belloni and Chernozhukov (2011). The quantile is evaluated in the next 4 steps.

1. The parameter $\delta_0(u)$ is estimated by quantile regression defined in equation (B.2) with $Z(x) = x$ and $u \in \{0.01, 0.02, \dots, 0.99\}$. Likewise, an analog of $\delta_0(u)$ is estimated by quantile regression defined in equation for $S = 1, D = 0$ group.
2. For each covariate value x and quantile level $u \in \{0.01, 0.02, \dots, 0.99\}$, $\widehat{Q}(u, x) := x' \widehat{\delta}(u)$ is evaluated.
3. For each covariate value x , the vector $(\widehat{Q}(u, x))_{u=0.01}^{0.99}$ is sorted. Furthermore, $\widehat{Q}(u, x)$ is capped at the minimal and maximal outcome values.

4. For each covariate value x , the trimming threshold $\hat{p}(x) = \text{round}(\hat{p}(x), 2)$ is rounded to 2 decimal places. $\hat{Q}(\hat{p}(x), x)$ is evaluated.

Table B.11: First-Stage Estimates, Table 2, Columns (4) and (6).

	(1)	Logistic		Quantile	
		Baseline coef. (α)	Interaction coef. (γ)	Control	Treated
	(1)	(2)	(3)	(4)	(5)
1	(Intercept)	-0.518	0.154	2.305	2.561
2	BLACK and R_GETGED=1	-0.200			
3	R_COMM=1 and R_GETGED=1	-0.224			
4	MOS_ANYW and R_GETGED=1	-0.022			
5	HGC : EVWORK	0.044			
6	HGC : HRWAGER	0.001			
7	HGC : MOSINJOB	0.004			
8	HRWAGER : MOSINJOB	0.006			
9	EARN_YR		0.000		
10	R_HOME = 1		-0.260		
11	PAY_RENT = 1			0.054	0.033
12	HRWAGER			0.017	-0.021
13	WKEARNR			0	0.001
14	FEMALE			-0.139	-0.036
15	PERS_INC1			0.011	-0.12
16	HH_INC5			0.073	0.133

Notes. Table shows the first-stage logistic and quantile regression estimates that produce Table 2 post-lasso bounds (Column (4),(6)). Column (2): baseline coefficient α of equation (7.2). Column (3): interaction coefficient γ of equation (7.2). Column (4): $\delta(u)$ of equation (B.2) on wage 90 $u = 0.95$ -quantile in the control group (sample size = 1, 660). Column (5): $\delta(u)$ of equation (B.2) on wage 90 $u = 0.97$ -quantile in the treated group (sample size = 2, 564). Covariates are defined in Section A.3. Computations use design weights.

Health status (HEALTH). Applicants were asked to rate their health at the moment of RA

Table B.12: Health status at RA

#	description	#	description
1	excellent	2	good
3	fair	4	poor

Arrest experience. CPAROLE21=1 is a binary indicator for being on probation or parole at the moment or RA. In addition, arrested applicants were asked about the time past since most recent arrest **MARRCAT**.

Table B.10: First-Stage Estimates, Table 2, Columns (1)-(2).

		Logistic		Quantile	
		Baseline coef. (α)	Interaction coef. (γ)	Control ($S = 1, D = 0$)	Treated ($S = 1, D = 1$)
(1)		(2)	(3)	(4)	(5)
1	(Intercept)	-1.047	0.553	2.669	2.197
2	AGE	0.038	-0.037	-0.003	0.014
3	BLACK	-0.203	-0.109	-0.135	-0.176
4	CPAROLE21	0.028	-0.547	-0.029	0.023
5	CURRJOB	0.201	-0.044	0.036	0.085
6	DRG_SUMP2	-0.085	0.042	0.03	-0.058
7	EARN_YR	0.000	0.000	0.000	0.000
8	EVARRST	-0.123	0.147	-0.024	0.024
9	FEMALE	-0.23	-0.058	-0.113	-0.126
10	HASCHLD	0.425	-0.177	-0.012	0.103
11	HGC	0.036	0.026	-0.011	-0.011
12	HGC_FATH	0.013	-0.001	0.003	0.004
13	HGC_MOTH	-0.004	0.008	0.003	0.000
14	HH_INC2	0.148	-0.186	-0.032	-0.026
15	HH_INC3	0.142	-0.035	-0.013	-0.01
16	HH_INC4	0.373	-0.23	0.007	0.061
17	HH_INC5	0.276	0.036	0.077	0.151
18	HISP	-0.155	0.004	0.095	0.029
19	HRSWK_JR	-0.006	0.003	0.000	-0.003
20	IMP_PRO1	-0.009	-0.04	-0.212	0.029
21	MARRCAT11	0.031	-0.284	0.096	0.075
22	MARRIED	0.339	-0.253	-0.034	-0.021
23	MOSINJOB	0.039	0.007	-0.006	0.000
24	NCHLD	-0.324	0.137	0.067	0.023
25	OTHERRAC	-0.191	-0.284	0.121	0.054
26	PAY_RENT1	-0.137	0.171	0.002	0.013
27	PERS_INC2	0.182	0.007	0.172	-0.059
28	PERS_INC3	0.200	-0.024	0.185	0.044
29	PERS_INC4	0.031	0.419	0.222	-0.14
30	REASED_R4	0.068	-0.226	-0.038	-0.094
31	R_COMM1	-0.136	-0.069	-0.006	0.015
32	R_GETGED1	-0.348	0.032	-0.061	-0.009
33	R_HOME1	-0.214	-0.047	-0.03	0.031
34	SEPARATED	-0.149	-0.165	-0.084	-0.105
35	TOGETHER	-0.199	0.339	-0.026	0.014
36	TYPEWORR5	0.121	-0.631	0.168	-0.057
37	WKEARNR	0.001	-0.001	0.001	0.001
38	YR_WORK	0.260	0.147	-0.070	-0.042

Notes. Table shows the first-stage logistic and quantile regression estimates that produce bounds in Columns (1)-(2) of Table 2. 37 covariates are 28 Lee's covariates (Table B.3) and 9 covariates important for differential employment effect. Column (2): baseline coefficient α in equation (7.2). Column (3): interaction coefficient γ in equation (7.2). Column (4): $\delta(u)$ from equation (B.2) on wage 90 $u = 0.95$ -quantile in the control group (sample size = 1, 660). Column (5): $\delta(u)$ of equation (B.2) on wage 90 $u = 0.97$ -quantile in the treated group (sample size = 2, 564). Covariates are defined in Section A.3. Computations use design weights.

Table B.13: Number of months since most recent arrest

#	description	#	description
1	less than 12	2	12 to 24
3	24 or more	4	N/A

Table B.14: First-Stage Estimates, Table 2, Column (5).

		Logistic		Quantile	
		Baseline coef. (α)	Interaction coef. (γ)	Control	Treated
	(1)	(2)	(3)	(4)	(5)
1	(Intercept)	-0.68	0.34	2.28	0.07
2	EARN_YR	0.00	-0.00	0.00	0.00
3	EVWORKB1	-0.40	-0.09	-0.01	-0.03
4	FEMALE	-0.21	-0.03	-0.12	0.00
5	HGC	0.07	-0.02	0.01	0.00
6	HH_INC5	0.14	0.15	0.02	0.10
7	HRWAGER	0.16	-0.00	0.00	-0.01
8	MOSINJOB	0.04	0.02	0.00	-0.00
9	MOS_ANYW	-0.02	-0.00	0.00	-0.00
10	PAY_RENT1	-0.09	0.14	0.06	0.04
11	PERS_INC1	-0.09	-0.02	-0.01	-0.10
12	RACE_ETH2	-0.15	-0.03	-0.15	0.04
13	R_COMM1	-0.12	-0.06	0.03	-0.07
14	R_GETGED1	-0.27	-0.01	-0.07	0.06
15	R_HOME1	-0.21	-0.06	-0.05	0.05
16	WKEARNR	-0.00	0.00	0.00	0.00
17	R_GETGED1:RACE_ETH2	-0.021			
18	HGC:EVWORKB1	0.081			
19	R_COMM1:R_GETGED1	-0.054			
20	R_GETGED1:MOS_ANYW	0.004			
21	HRWAGER:HGC	-0.014			
22	HGC:MOSINJOB	0.000			
23	HRWAGER:MOSINJOB	0.003			

Notes. Table shows the first-stage logistic and quantile regression estimates that produce bounds in Column (5) of Table 2. Column (2): baseline coefficient α of equation (7.2). Column (3): interaction coefficient γ of equation (7.2). Column (4): $\delta(u)$ of equation B.2 on wage 90 $u = 0.95$ -quantile in the control group (sample size = 1, 660). Column (5): $\delta(u)$ of equation B.2 on wage 90 $u = 0.97$ -quantile in the treated group (sample size = 2, 564). Covariates are defined in Section A.3. Computations use design weights.

Table B.15: Estimated lower bound on the effect of access to Medicaid on self-reported healthcare utilization: extensive margin

	ITT			LATE		
	(1)	(2)	(3)	(4)	(5)	(6)
	None	Standard	ML	None	Standard	ML
Prescription drugs currently	0.025 (0.008)	-0.008 (0.014)	0.017 (0.017)	0.088 (0.029)	-0.036 (0.046)	0.060 (0.060)
Outpatient visits last six months	0.062 (0.007)	0.005 (0.013)	0.042 (0.017)	0.212 (0.025)	0.001 (0.045)	0.146 (0.058)
ER visits last six months	0.006 (0.007)	-0.020 (0.008)	-0.004 (0.011)	0.022 (0.023)	-0.076 (0.030)	-0.015 (0.037)
Hospital admissions last six months	0.002 (0.004)	-0.005 (0.004)	0.002 (0.005)	0.008 (0.014)	-0.020 (0.016)	0.007 (0.016)
Compulsory covariates (stratification)	N/A	16	16	N/A	16	16
Additional covariates (trimming)	N/A	0	21	N/A	0	21

* Standard errors in parentheses. This table reports results from a Lee bounding exercise on self-reported healthcare utilization outcomes for 3 specifications: no trimming, standard trimming, and the agnostic ML approach. Columns (1)–(3) report the coefficient and standard error on Lottery from estimating equation (7.5) by OLS. Columns (4)–(6) report the coefficient and standard error on Insurance from estimating equation (7.6) by 2SLS with Lottery as an instrument for Insurance. All regressions include household size fixed effects, survey wave fixed effects, and their interactions. Trimming methods. None: exact replicate of Finkelstein et al. (2012), Table V. Standard: the minimal number of control outcomes are trimmed from below for each value of fixed effect until the treatment-control difference in response rates switches from negative to non-negative. Agnostic: Step 1. 21 additional covariates are selected on an auxiliary sample of 4,000 households as described in Appendix A.5. Step 2. In the main sample of 46,000 households, a zero outcome with covariate vector x is trimmed in the control group if a flipped coin with success prob. $(1 - p_0(x))/\phi_0(x)$ is success, where the trimming threshold $p_0(x)$ is defined in (B.4) and the zero outcome probability $\phi_0(x)$ is defined in (B.3). Standard errors are estimated by a cluster-robust bootstrap with $B = 1000$ repetitions. Both the trimming and regression steps are bootstrapped. Computations (the first and the second stage) use survey weights. Covariates are described in Table B.17. See Appendix A.5 for details.

Table B.16: Estimated lower bound on the effect of access to Medicaid on self-reported healthcare utilization: total utilization

	ITT			LATE		
	(1)	(2)	(3)	(4)	(5)	(6)
	None	Standard	NP	None	Standard	NP
Prescription drugs currently	0.100 (0.051)	-0.024 (0.066)	0.077 (0.052)	0.347 (0.175)	-0.124 (0.225)	0.270 (0.179)
Outpatient visits last six months	0.314 (0.054)	0.121 (0.065)	0.246 (0.054)	1.083 (0.182)	0.372 (0.228)	0.853 (0.183)
ER visits last six months	0.007 (0.016)	-0.040 (0.019)	-0.008 (0.016)	0.026 (0.056)	-0.152 (0.065)	-0.027 (0.056)
Hospital admissions last six months	0.006 (0.006)	-0.004 (0.007)	0.003 (0.006)	0.021 (0.021)	-0.014 (0.024)	0.010 (0.021)
Compulsory covariates (stratification)	N/A	16	16	N/A	16	16
Additional covariates (trimming)	N/A	0	9	N/A	0	9

* Standard errors in parentheses. This table reports results from a Lee bounding exercise on self-reported health outcomes for 3 specifications: no trimming, standard trimming, and the classic nonparametric approach. Columns (1)–(3) report the coefficient and standard error on Lottery from estimating equation (7.5) by OLS. Columns (4)–(6) report the coefficient and standard error on Insurance from estimating equation (7.6) by 2SLS with Lottery as an instrument for Insurance. All regressions include household size fixed effects, survey wave fixed effects, and their interactions. Trimming methods. None: exact replicate of Finkelstein et al. (2012), Table V. Standard: the minimal number of control outcomes are trimmed from below for each value of fixed effect until the treatment-control difference in response rates switches from negative to non-negative. NP. Step 1. 9 additional covariates are taken as described in Appendix A.5 based on OHIE documentation. Step 2. An outcome with covariate vector x is trimmed if it is less than $Q(1 - 1/p_0(x), x)$, where the trimming threshold $p_0(x)$ is defined in equation (B.4) and the conditional quantile is defined in equation (2.4). Standard errors are estimated by a cluster-robust bootstrap with $B = 1000$ repetitions. Both the trimming and regression steps are bootstrapped. Computations (the first and the second stage) use survey weights. Covariates are described in Table B.17. See Appendix A.5 for details.

A.5 Finkelstein et al. (2012) empirical details

Data source. The data set is the output of OHIE_QJE_Replication_Code/SubPrograms/prepare_data.do file, one of the subprograms of OHIE replication package of Finkelstein et al. (2012). It contains $N = 58,405$ observations, survey wave, household size fixed effects, and their interactions, and 48 optional baseline covariates, summarized in Table B.17.

Agnostic approach: composition of $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$. To estimate the composition of $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$, I invoke post-lasso-logistic of Belloni et al. (2016) with X being equal to 64 baseline covariates and the penalty λ being equal to recommended choice of penalty, on the full sample $N = 58,405$. For each of 15 outcomes in reported in Tables 3, 4, B.15, B.16, the trimming threshold exceeds 1 for at least 99.43% of subjects. For that reason, $\mathcal{X}_{\text{help}}$ is taken to be \emptyset for each outcome under consideration.

Covariate selection for Tables 3 and B.15: Agnostic approach. The main sample M consists of 46,000 randomly selected households, and the auxiliary sample A is its complement. On the auxiliary sample A , my selection equation is (7.2), where $D = 1$ is a binary indicator for winning Medicaid lottery, $X = 1,152$ pairwise covariate interactions, and $S = 1$ is a binary indicator for a non-missing response about receiving any prescription drugs. (Table B.15, Row 1, rx_any_12m). Invoking logistic lasso of Belloni et al. (2016) with $\lambda = 100$ to estimate (7.2), I select 46 pairwise interactions and break them down to 37 raw covariates. They are listed in Table B.18, Column (1).

Covariate selection for Tables 4 and B.16. 9 selected covariates are: female_list, english_list, zip_msa, snap_ever_prenotify_07, tanf_ever_prenotify_07, snap_tot_prenotify_07, tanf_tot_prenotify_07, num_visit_pre_cens_ed, num_out_pre_cens_ed.

For a binary outcome (e.g., a binary outcome in Finkelstein et al. (2012)), the conditional probability of zero outcome in the treated group

$$\phi_0(x) := \Pr(Y = 1|X = x) := \Lambda(x'\delta) \tag{B.3}$$

is estimated by logistic regression. An outcome is trimmed if a coin with head probability $(1 - \widehat{p}(x))/\widehat{\phi}(x)$ turns out head. For a non-binary outcome that has a point mass, a combination of both approaches is applied. In what follows, the quantile regression estimates are reported only for the modal (i.e., most frequent) quantile value.

First-Stage Estimates: Selection Equation. Selection equation is

$$S = 1_{\{X'\alpha + D \cdot Z'\gamma + U > 0\}}, \quad (\text{B.4})$$

where $Z = 1$ is a binary indicator of treatment offer (i.e., “treatment”), X is a vector of covariates, selected on auxiliary sample, and $S = 1$ is a binary indicator for non-missing response. Therefore,

$$\widehat{s}(0, x) := \Lambda(x'\widehat{\alpha}), \quad \widehat{s}(1, x) := \Lambda(x'(\widehat{\alpha} + \widehat{\gamma})).$$

First-Stage Estimates: Outcome Equation for ITT. Outcome equation is

$$Y = 1_{\{X'\kappa_0 + \xi > 0\}}, \quad S = 1, Z = 0, \quad (\text{B.5})$$

where $Y = 1$ is a binary indicator for negative (“No”) answer in Table 3, Row 1. The estimate of $\phi_0(x)$ in equation (B.3) is $\widehat{\pi}(x) := \Lambda(x'\widehat{\delta})$. To construct a trimmed data set for ITT, a zero outcome in the control group is trimmed if a coin with success prob. $(1 - \widehat{p}(x))/\widehat{\phi}(x)$ turns out success. For numerical stability, $\widehat{\phi}(x) := \max(\widehat{\phi}(x), 0.05)$.

First-Stage Estimates for Binary Outcomes: Outcome Equation for LATE. Outcome equation is

$$Y = 1_{\{X'\delta_0 + D \cdot X'\rho_0 + \xi > 0\}}, \quad S = 1, Z = 0, \quad (\text{B.6})$$

where $D = 1$ is a binary indicator of having Medicaid insurance (i.e., “insurance”). Therefore, $\widehat{\pi}(0, x) := \Lambda(x'\widehat{\delta})$ and $\widehat{\pi}(1, x) := \Lambda(x'\widehat{\delta} + x'\widehat{\rho})$. To construct a trimmed data set for LATE, a zero outcome in the control uninsured (insured) group is trimmed if a coin with success prob.

$(1 - \hat{p}(x))/\hat{\phi}(0,x) ((1 - \hat{p}(x))/\hat{\phi}(1,x))$ turns out success.

Table B.17: Baseline covariates in Oregon Health Insurance Experiment.

Name	Description
female_list	female
english_list	requested English materials
zip_msa	zip code is in MSA
visit_pre_ed	ED visit
hosp_pre_ed	ED visit resulting in hospital admission
out_pre_ed	oupatient ED visit
on_pre_ed	ED visit on week-day
off_pre_ed	week-end or nighttime ED visit
edcnpnp_pre_ed	emergent, non-preventable ED visit
edcnpa_pre_ed	emergent, preventable ED visit
unclas_pre_ed	unclassified ED visit
epct_pre_ed	primary care treatable ED visit
ne_pre_ed	non-emergent ED visit
acsc_pre_ed	ambulatory case sensitive ED visit
chron_pre_ed	ED visit for chronic condition
inj_pre_ed	ED visit for injury, pre-randomization
skin_pre_ed	ED visit for skin condition
abdo_pre_ed	abdominal pain visit
back_pre_ed	ED visit for back pain
back_ed	back pain ED visit
heart_pre_ed	chest pain ED visit
depres_pre_ed	mood disorders ED visit
psysub_pre_ed	psych conditions/substance abuse ED visit
hiun_pre_ed	high uninsured volume hospital ED visit
loun_pre_ed	low uninsured volume hospital ED visit
charg_tot_pre_ed	total charges
ed_charg_tot_pre_ed	ED total charges
snap_ever_prenotify_07	ever on SNAP
tanf_ever_prenotify_07	ever on TANF
snap_tot_prenotify_07	total household benefits from SNAP
tanf_tot_prenotify_07	total household benefits from TANF
ddd_numhh_li_j	household size fixed effect for $j = 1, 2, 3$
ddddraw_sur_k	survey wave fixed effect for $k = 1, 2, \dots, 7$
dddraXnum_k_j	interaction of survey wave and household size

Notes. All ED and state program variables summarize events occurring between January, 1, 2007 and lottery notification date. Each health-related ED visit variable is represented by two measures: extensive margin (any_X) and total count (num_X). Covariates ddd_X represent fixed effects for household size and survey waves.

Table B.18: First-Stage Estimates, Table 3, Columns (3) and (6).

			ITT		LATE	
	α	γ	κ	δ	ρ	
(Intercept)	-0.55		0.14	0.08	0.07	
any_acsc_pre_ed	-0.10		-0.19	0.09	-1.96	
any_back_pre_ed	0.13		0.62	0.33	1.25	
any_depres_pre_ed	0.02		-0.11	0.15	-1.59	
any_head_pre_ed	-0.04		0.39	0.02	1.70	
any_hiun_pre_ed	-0.23		0.28	0.26	0.34	
any_hosp_pre_ed	0.09		0.29	0.22	0.38	
any_on_pre_ed	-0.07		-0.16	-0.11	-0.61	
charg_tot_pre_ed	0.00		0.00	0.00	-0.00	
english_list	0.23		-0.44	-0.43	-0.20	
female_list	0.33		-0.07	-0.02	-0.46	
num_epct_pre_ed	-0.02		0.18	0.21	-0.13	
num_ne_pre_ed	-0.04		-0.03	0.13	-0.54	
num_on_pre_cens_ed	0.04		0.11	0.21	-0.05	
num_out_pre_cens_ed	0.11		0.19	0.27	-0.27	
num_skin_pre_cens_ed	-0.01		0.16	0.13	0.79	
num_visit_pre_cens_ed	-0.17		-0.23	-0.44	0.63	
snap_ever_prenotify07	-0.04		0.53	0.47	0.43	
snap_tot_hh_prenotify07	-0.00		-0.00	0.00	-0.00	
tanf_ever_prenotify07	-0.53		-0.95	-1.33	0.83	
tanf_tot_hh_prenotify07	-0.00		0.00	0.00	-0.00	
zip_msa	-0.11		-0.20	-0.15	-0.32	
dddraXnum_2_2	-0.335		-0.009		0.000	
dddraXnum_2_3	0.458		0.147	0.175	-0.694	
dddraXnum_3_2	0.083		0.100	0.065	0.000	
dddraXnum_3_3	0.752				0.274	
dddraXnum_4_2	-0.079	-0.112	-0.009	-0.185	0.606	
dddraXnum_5_2	-0.041		-0.249	-0.271	0.154	
dddraXnum_6_2	-0.057		-0.225	-0.250	0.129	
dddraXnum_7_2	0.162		-0.237	0.553	0.000	
dddrow_sur_2	0.015	0.010	-0.087	-0.104	0.126	
dddrow_sur_3	-0.128	0.133	-0.073	-0.103	0.205	
dddrow_sur_4	-0.040	0.056	0.003	0.024	-0.094	
dddrow_sur_5	-0.053	0.002	0.089	0.093	0.078	
dddrow_sur_6	-0.110	0.047	0.068	0.052	0.213	
dddrow_sur_7	-0.053	0.002	-0.029	-0.021	-0.025	
dddnumhh_li_2	0.147	-0.027	-0.105	-0.070	-0.223	
dddnumhh_li_3	-1.066	0.649	-11.630	-11.667	0.000	
<i>N</i>	53, 646	8, 383	8, 383	8, 383	8, 383	

Notes. Table shows the first-stage estimates for the estimated effect of Medicaid exposure (Column (3)) and insurance (Column (6)) in Table 3, Row 1. Column (2) : baseline coefficient α in equation (B.4). Column (3) : interaction coefficient γ of equation (B.4). Column (4): baseline coefficient κ in equation (B.5) in $S = 1, D = 0$ group (sample size = 8, 383) to estimate ITT bounds. Columns (5)-(6): baseline coefficient δ and interaction coefficient ρ in (B.6) in $S = 1, Z = 0$ group (sample size = 8, 383) to estimate LATE bounds. Computations use survey weights.

References

- Abdulkadiroglu, A., Pathak, P. A., and Walters, C. R. (2020). Do parents value school effectiveness. *American Economic Review*, 110(5):1502–1539.
- Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.
- Andrews, D. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, 81:609–666.
- Andrews, D. and Shi, X. (2017). Inference based on many conditional moment inequalities. *Journal of Econometrics*, 196:275–287.
- Angrist, J., Bettinger, E., Bloom, E., King, E., and Kremer, M. (2002). Vouchers for private schooling in colombia: Evidence from a randomized natural experiment. *The American Economic Review*, 92(5):1535–1558.
- Angrist, J., Bettinger, E., and Kremer, M. (2006). Long-term consequences of secondary school vouchers: Evidence from administrative records in colombia. *The American Economic Review*, 96(3):847–862.
- Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.
- Angrist, J. D., Pathak, P. A., and Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, 5(4):1–27.
- Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, 60:47–50.
- Athey, S. (2015). Machine learning and causal inference for policy evaluation.
- Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7460.

- Athey, S. and Imbens, G. (2019). Machine learning methods economists should know about.
- Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89:133–161.
- Bai, Y., Santos, A., and Shaikh, A. (2019). A practical method for testing many moment inequalities.
- Belloni, A. and Chernozhukov, V. (2011). ℓ_1 -penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.
- Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernandez-Val, I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(260):4–29.
- Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85:233–298.
- Belloni, A., Chernozhukov, V., and Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619.
- Beresteanu, A., Molchanov, I., and Molinari, F. (2011). Sharp identification regions in models with convex moment predictions. *Econometrica*, 79:1785–1821.
- Beresteanu, A. and Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica*, 76(4):763–814.
- Blanco, G., Flores, C. A., and Flores-Lagunes, A. (2013). Bounds on average and quantile treatment effects of job corps training on wages. *The Journal of Human Resources*, 48(3):659–701.
- Bontemps, C., Magnac, T., and Maurin, E. (2012). Set identified linear models. *Econometrica*, 80:1129–1155.

- Bugni, F., Canay, I., and Shi, X. (2017). Inference for functions of partially identified parameters in moment inequality models. *Quantitative Economics*, 8:1–38.
- Chandrasekhar, A., Chernozhukov, V., Molinari, F., and Schrimpf, P. (2012). Inference for best linear approximations to set identified functions. *arXiv e-prints*, page arXiv:1212.5627.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68.
- Chernozhukov, V., Chetverikov, D., and Kato, K. (2019). Inference on causal and structural parameters using many moment inequalities. *Review of Economic Studies*, 86:1867–1900.
- Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2017). Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments. *arXiv e-prints*, page arXiv:1712.04802.
- Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016). Locally Robust Semiparametric Estimation. *arXiv e-prints*, page arXiv:1608.00033.
- Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Biometrics*, 81(6):2205–2268.
- Chernozhukov, V., Rigobon, R., and Stoker, T. (2010). Set identification and sensitivity analysis with Tobin regressors. *Quantitative Economics*, 1(6B):255 – 277.
- Farrell, M., Liang, T., and Misra, S. (2021a). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.
- Farrell, M. H., Liang, T., and Misra, S. (2021b). Deep learning for individual heterogeneity: An automatic inference framework.
- Feller, A., Greif, E., Ho, N., Miratrix, L., and Pillai, N. (2016). Weak separation in mixture models and implications for principal stratification.

- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J., Allen, H., Baicker, K., and Group, O. H. S. (2012). The oregon health insurance experiment: Evidence from the first year. *Quarterly Journal of Economics*, 127(3):1057–1106.
- Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.
- Gafarov, B. (2019). Inference in high-dimensional set-identified affine models.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4):475–492.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.
- Hirano, K., Imbens, G., and Reeder, G. (2003). Efficient estimation of average treatment effects under the estimated propensity score. *Econometrica*, 71(4):1161–1189.
- Honore, B. and Hu, L. (2020). Selection without exclusion. *Econometrica*, 88(88):1007–1029.
- Horowitz, J. L. and Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica*, 63(2):281–302.
- Hsu, Y.-C., Liu, C.-A., and Shi, X. (2019). Testing generalized regression monotonicity. *Econometric Theory*, 35:1146–1200.
- Imbens, G. and Manski, C. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.
- Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.
- Kaido, H., Molinari, F., and Stoye, J. (2019). Confidence intervals for projections of partially identified parameters. *Econometrica*, 87(4):1397–1432.

- Kaido, H., Molinari, F., and Stoye, J. (2021). Constraint qualifications in partial identification. *Econometric Theory*.
- Kamat, V. (2019). On the identifying content of instrument monotonicity.
- Kamat, V. (2021). Identifying the effects of a program offer with an application to head start.
- Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86:591–616.
- Kline, P. and Walters, C. (2019). On heckits, late, and numerical equivalence. *Econometrica*, 87(2):677–696.
- Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102.
- Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808 – 1829.
- Manski, C. F. (1989). Anatomy of the selection problem. *The Journal of Human Resources*, 24(3):343–360.
- Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.
- Mbakop, E. and Tabord-Meehan, M. (2021). Model selection for treatment choice: Penalized welfare maximization. *Econometrica*, 89:825–848.
- Mogstad, M., Torgovitsky, A., and Walters, C. (2020a). The causal interpretation of two-stage least squares with multiple instrumental variables.
- Mogstad, M., Torgovitsky, A., and Walters, C. (2020b). Policy evaluation with multiple instrumental variables.
- Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

- Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):245–271.
- Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics*, 213(57):416–444.
- Neyman, J. (1979). $c(\alpha)$ tests and their use. *Sankhya*, pages 1–21.
- Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*.
- Oprescu, M., Syrgkanis, V., and Wu, Z. S. (2018). Orthogonal random forest for causal inference. *arXiv e-prints*, page arXiv:1806.03467.
- Rockafellar, R. T. (1997). *Convex Analysis*. Princeton University Press.
- Schochet, P. Z., Burghardt, J., and McConnell, S. (2008). Does job corps work? impact findings from the national job corps study. *American Economic Review*, 98(1):1864–1886.
- Semenova, V. (2020). Generalized lee bounds: Supplementary appendix.
- Semenova, V. and Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions.
- Shi, X., Shum, M., and Song, W. (2018). Estimating semi-parametric panel multinomial choice models using cyclic monotonicity. *Econometrica*, 86:737–761.
- Shi, X., Shum, M., and Song, W. (2021). Inference on estimators defined by mathematical programming. *Journal of Econometrics*.
- Sloczynski, T. (2021). When should we (not) interpret linear iv estimands as late?
- Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315.
- Stoye, J. (2010). Partial identification of spread parameters. *Quantitative Economics*, 2:29–51.
- Sun, L. (2021). Empirical welfare maximization with constraints.

- Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., and Lewis, G. (2019). Machine learning estimation of heterogeneous treatment effects with instruments.
- Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135 – 166.
- Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70:331–341.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal Journal of the American Statistical Association*, 113:1228–1242.