Better Lee Bounds

Vira Semenova

April 25, 2021

Abstract

This paper develops methods for tightening Lee (2009) bounds on average causal effects when the number of pre-randomization covariates is large, potentially exceeding the sample size. These Better Lee Bounds are guaranteed to be sharp when few of the covariates affect selection and the outcome. If this sparsity assumption fails, the bounds remain valid. I propose inference methods that enable hypothesis testing in either case. My results rely on a weakened monotonicity assumption that only needs to hold conditional on covariates. I show that the unconditional monotonicity assumption that motivates traditional Lee bounds fails for the JobCorps training program. After imposing only conditional monotonicity, Better Lee Bounds are found to be much more informative than standard Lee bounds in a variety of settings.

*Email: semenovavira@gmail.com. I am grateful to Victor Chernozhukov, Michael Jansson, Patrick Kline, Anna Mikusheva, and Whitney Newey for their guidance and encouragement. I am thankful to Alberto Abadie, Chris Ackerman, Sydnee Caldwell, Denis Chetverikov, Ben Deaner, Mert Demirer, Jerry Hausman, Peter Hull, Tetsuya Kaji, Kevin Li, Elena Manresa, Rachael Meager, Francesca Molinari, Denis Nekipelov, Oles Shtanko, Cory Smith, Sophie Sun, Roman Zarate, and the participants at the MIT Econometrics Lunch 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. As a result, wages in the treatment and control groups are not directly comparable since wages only exist for the employed individuals. A common way to estimate the average causal effect is to bound this effect from above and below, focusing on a partially latent group of subjects whose outcomes are observed regardless of their treatment status (the always-observed principal strata, Frangakis and Rubin (2002) or the always-takers, Lee (2009)).

Seminal work by Lee (2009) leverages the monotonicity assumption to bound the average causal effect for always-takers. For example, if job training cannot deter employment, the Lee lower bound is the treatment-control difference in wages, where the top wages in the treated group are trimmed until employment rates in both groups are equal. If pre-randomization covariates are available, Lee bounds can be tightened by averaging covariate-specific bounds over the always-takers' covariate distribution. However, it is hard or impossible to estimate sharp (i.e., the tightest possible) bounds since Lee's method requires a positive number of treated and control outcomes for each covariate value. As a result, empirical researchers spend a lot of energy selecting and discretizing covariates, a process that is subjective, labor-intensive, and prone to erroneous inference.

In this paper, I propose a generalization of Lee bounds—better Lee bounds—and provide theoretical, simulation, and empirical evidence that they substantially outperform standard Lee bounds. First, better Lee bounds are based on a weaker monotonicity assumption that only needs to hold conditional on covariates. Specifically, each subject is allowed to have either a positive or negative selection response, as long as the direction of this response is identified by a covariate vector. In contrast, standard Lee bounds require the same direction of treatment effect on selection response for all subjects. Second, better Lee bounds are asymptotically sharp as long as few of the covariates affect selection and the outcome, permitting the total number of covariates under consideration to exceed sample size. In contrast, standard Lee bounds are sharp only in a model that has a handful of covariates. Finally, better Lee bounds accommodate a broad class of machine learning techniques for estimating the conditional probability of treatment (i.e., the propensity score), overcoming a key historical limitation to the widespread adoption of Lee bounds in quasi-experiments.

As a first step towards sharpness, I represent each bound via a semiparametric moment equation that depends on the conditional outcome quantile and the conditional probability of selection. A naïve approach would be to estimate these functions by quantile and logistic series regressions. If the true functions are sufficiently smooth relative to the covariate vector's dimension, these estimators provide a good approximation to the true outcome quantile. However, the smoothness assumption implicitly restricts the number of covariates (Stone (1982)). This restriction is problematic for JobCorps data set (Schochet et al. (2008)), which has $9,145$ observations and $5,177$ covariates, and calls for selecting covariates in a data-driven way.

The first main contribution of this paper is a method for estimating and conducting inference on sharp Lee bounds with a built-in model selection process based on modern machine learning techniques. For example, if few of the covariates affect selection and the outcome, ℓ_1-regularized logistic (Belloni et al. (2017), Belloni et al. (2016)) and quantile (Belloni and Chernozhukov (2013)) estimators deliver a good approximation to the true functions. An implicit cost of ℓ_1-regularization is bias that converges slower than the parametric rate. To prevent the transmission of this bias into the bounds, I propose a Neyman-orthogonal (Neyman (1959)) moment equation for each bound. Leveraging Neyman-orthogonality and sample splitting ideas, my proposed better Lee bounds permit inference based on the standard normal approximation. The proposed bounds are straightforward to compute using the *R* software package `leebounds`, available at `https://github.com/vsemenova/leebounds`.

In settings where sparsity is not economically plausible, researchers utilize sample splitting strategies to leverage machine learning techniques for model selection, although the results will not be as sharp. To account for the uncertainty generated by the choice of sample split, Chernozhukov et al. (2017) suggests to generate several random splits and aggregate the lower and upper bounds over various partitions. On the one hand, this approach aims at less sharp bounds and leads to conservative inference due to sample splitting. On the other hand, this approach is fully agnostic: it does not require any assumptions on the model selection procedure.

My main result has several extensions. First, I allow the outcome variable to be multi-dimensional and show that the sharp identified set for the treatment effect parameter is compact and convex. Next, I derive an orthogonal moment equation for the identified set's boundary (i.e., support function) and provide a large sample approximation that holds uniformly over the boundary. I also propose a weighted bootstrap procedure for conducting inference on the boundary. In contrast to conventional bootstrap techniques, my algorithm is faster to compute since, by virtue of orthogonality, only the second stage is repeated in the simulation. Second, better Lee bounds accommodate within-cluster dependence and panel data. Third, I derive better Lee bounds for the Intent-to-Treat and Local Average Treatment Effect parameters and provide a complete set of identification, estimation and inference results. Finally, I provide the inference methods accommodating unknown propensity score in quasi-experiments.

The paper builds on a growing literature that incorporates modern regularized machine learning techniques into econometrics, see Mullainathan and Spiess (2017) for a review. A large body of this literature is devoted to establishing convergence properties of ℓ_1-regularized estimators (Belloni et al. (2016), Belloni and Chernozhukov (2013)), as well conducting debiased inference on parameters following Lasso model selection (Belloni et al. (2017), Belloni et al. (2014), Belloni et al. (2016), van der Geer et al. (2014), Javanmard and Montanari (2014), Zhang and Zhang (2014)). Leveraging the work of Belloni et al. (2017), Belloni et al. (2014), Belloni et al. (2016), Chernozhukov et al. (2016), and Chernozhukov et al. (2018), I derive an orthogonal moment equation for better Lee bounds and propose asymptotic theory for conducting inference on these bounds in both one- and multi-dimensional settings. The agnostic approach to inference is an extension of Chernozhukov et al. (2017)'s general machine learning approach to heterogeneous treatment effects, adapted for a partial identification problem.

In the final part of the paper, I estimate Lee bounds in three empirical applications. First, I study the effect of the JobCorps training program on wages and wage growth, using data from Schochet et al. (2008). 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 original 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. Second, I study the effect of private school subsidies on pupils' educational achievement, as in Angrist et al. (2002). I find that the voucher effect on Mathematics, Reading and Writing is smaller than Angrist et al. (2002)'s original estimate that does not account for selection bias, by a factor of 0.5 to 0.75. Finally, 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. better Lee bounds attain nearly point-identification in all three applications. In contrast, conventional Lee bounds are too wide to determine the direction of the treatment effect in any of these settings.

The paper is organized as follows. Section 2 reviews basic Lee bounds and Lee's estimator under the standard monotonicity assumption. Section 3 presents evidence against unconditional monotonicity in the JobCorps training program. Section 4 establishes the asymptotic properties of better Lee bounds, assuming sparsity. Section 5 proposes an agnostic approach for conducting inference on Lee bounds when sparsity fails. Section 6 discusses extensions of my baseline framework to allow for a multi-dimensional outcome, intent-to-treat and local average treatment effect target parameters, clustered or panel data, and the case when the propensity score is unknown. Section 7 presents a simulation study based on JobCorps data. Section 8 presents empirical applications. Section 9 concludes. Appendix A contains additional tables and figures supporting the results from the main text. Appendix B contains supplementary theoretical statements. Appendix C contains proofs. Appendix D contains additional simulations. Appendix E defines JobCorps covariates and contains supplementary results for Section 3. Appendix F contains supplementary results for all empirical applications.

2 Econometric Framework

In this section, I review the Lee (2009) sample selection model and formally define Lee bounds. I describe the bounds' estimator and the confidence region for the identified set. I then discuss how to tighten Lee bounds by conditioning on baseline covariates.

2.1 Model, Estimators and Confidence Region

I use the standard Rubin (1974) potential outcomes framework. 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 random sample $(D_i, X_i, S_i, S_iY_i)_{i=1}^N$ 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. The object of interest is the average treatment effect (ATE)

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

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) (Independence). The random vector $(Y(1), Y(0), S(1), S(0), X)$ *is independent of D.*
(2) (Monotonicity). $S(1) \geq S(0)$ *a.s.*

Suppose Assumption 1 holds. By monotonicity, any outcome observed in the control group must belong to an always-taker. Thus, the always-takers' expected outcome in the control status is identified:

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

In contrast, a treated outcome can be either an always-taker's outcome or a complier's outcome, but it is not possible to distinguish between the two types in the treated group. Nevertheless, by Assumption 1, the proportion of the always-takers in the $\{D = 1, S = 1\}$ group is identified as

$$p_0 = \mathbb{E}[S(1) = 1, S(0) = 1 | S = 1, D = 1] = \frac{\mathbb{E}[S = 1|D = 0]}{\mathbb{E}[S = 1|D = 1]}. \tag{2.2}$$

When the always-takers comprise the top p_0 outcome quantile in the treated group, the ATE (2.1) attains its largest possible value:

$$\bar{\beta}_U = \mathbb{E}[Y|Y \geq Q(1 - p_0), D = 1, S = 1] - \mathbb{E}[Y|D = 0, S = 1],$$

where $Q(1 - p_0)$ is the level-p_0 outcome quantile in the treated selected group. The lower bound $\bar{\beta}_L$ is defined analogously. Lee's estimator $(\widehat{\bar{\beta}}_L, \widehat{\bar{\beta}}_U)$ is defined as follows:

$$\widehat{\bar{\beta}}_U = \frac{\sum_{i=1}^N D_i S_i Y_i 1_{\{Y_i \geq \widehat{Q}(1-\widehat{p})\}}}{\sum_{i=1}^N D_i S_i 1_{\{Y_i \geq \widehat{Q}(1-\widehat{p})\}}} - \frac{\sum_{i=1}^N (1-D_i) S_i Y_i}{\sum_{i=1}^N (1-D_i) S_i}, \tag{2.3}$$

$$\widehat{\bar{\beta}}_L = \frac{\sum_{i=1}^N D_i S_i Y_i 1_{\{Y_i \leq \widehat{Q}(\widehat{p})\}}}{\sum_{i=1}^N D_i S_i 1_{\{Y_i \leq \widehat{Q}(\widehat{p})\}}} - \frac{\sum_{i=1}^N (1-D_i) S_i Y_i}{\sum_{i=1}^N (1-D_i) S_i}, \tag{2.4}$$

$$\widehat{Q}(u) = \min_{y \in \mathcal{Y}} \left\{ y : \frac{\sum_{i=1}^N D_i S_i 1_{\{Y \leq y\}}}{\sum_{i=1}^N D_i S_i} \geq u \in [0,1] \right\}, \tag{2.5}$$

$$\widehat{p} = \frac{(\sum_{i=1}^N S_i (1-D_i)) / \sum_{i=1}^N (1-D_i)}{\sum_{i=1}^N S_i D_i / \sum_{i=1}^N (D_i)}, \tag{2.6}$$

where $\widehat{p}$ and $\widehat{Q}(u)$ are the sample analogs of p_0 and $Q(u)$.

If selection is not exogenous (i.e., $p_0 \neq 1$), there are enough observations from above and below the quantile $Q(p_0)$ for it to be well approximated by its empirical analog $\widehat{Q}(\widehat{p})$ in a large sample. A confidence region for the true identified set $[\beta_L, \beta_U]$ that covers the set with a pre-specified probability α takes the form

$$[\widehat{\bar{\beta}}_L - N^{-1/2} \widehat{\Omega}_{LL} c_{\alpha/2}, \quad \widehat{\bar{\beta}}_U + N^{-1/2} \widehat{\Omega}_{UU} c_{1-\alpha/2}], \tag{2.7}$$

where $\widehat{\Omega}_{LL}$ and $\widehat{\Omega}_{UU}$ are estimates of the asymptotic standard deviations of $\widehat{\bar{\beta}}_L$ and $\widehat{\bar{\beta}}_U$, respectively, and c_α is the critical value based on the standard normal approximation. To conduct inference on the true *parameter* β_0, Imbens and Manski (2004) (IM) propose an adjustment of (2.7) that covers β_0 with a pre-specified probability.

The approach described above can be applied after conditioning on a vector X of baseline, or pre-randomization, covariates. Define the conditional trimming threshold as

$$p_0(x) = \frac{\mathbb{E}[S=1|D=0, X=x]}{\mathbb{E}[S=1|D=1, X=x]} = \frac{s(0,x)}{s(1,x)}, \quad x \in \mathcal{X}. \tag{2.8}$$

The conditional outcome quantile $Q(u,x)$ in the treated group is implicitly defined by

$$\Pr(Y \leq Q(u,x)|D=1,S=1,X=x) = u, \quad u \in [0,1], \quad x \in \mathcal{X}. \tag{2.9}$$

The conditional upper bound is

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

To aggregate $\bar{\beta}_U(x)$ into an average, I need to reweight $\bar{\beta}_U(x)$ by the probability mass function in the always-takers group:

$$\beta_U = \int_{x\in\mathcal{X}} \bar{\beta}_U(x) f(x|S(1)=1,S(0)=1)dx = \int_{x\in\mathcal{X}} \bar{\beta}_U(x) f(x|S=1,D=0)dx. \tag{2.10}$$

Lee (2009) has shown that (2.10) is a sharp (i.e., the smallest possible) upper bound on β_0:

$$\beta_0 \leq \beta_U \leq \bar{\beta}_U. \tag{2.11}$$

Algorithm 1 Standard Lee bounds with covariates

1: Partition the covariate space $\mathcal{X}$ into J discrete cells $\{C_1, C_2, \ldots, C_J\}$.
2: Estimate the vector of cell-specific lower and upper bounds $\{\widehat{\bar{\beta}}_L(j), \widehat{\bar{\beta}}_U(j)\}_{j=1}^J$ and the probability mass function $\{\widehat{f}(j|S=1,D=0)\}_{j=1}^J$ in the selected control group.
3: Estimate bounds as

$$\widehat{\beta}_L = \sum_{j=1}^J \widehat{\bar{\beta}}_L(j)\widehat{f}(j|S=1,D=0), \quad \widehat{\beta}_U = \sum_{j=1}^J \widehat{\bar{\beta}}_U(j)\widehat{f}(j|S=1,D=0). \tag{2.12}$$

Algorithm 1 describes Lee's estimator with covariates. For the estimator (2.12) to be well-defined, each covariate group must contain both treated and control subjects, and a non-zero fraction of control subjects must be selected into the sample. Consequently, the estimator (2.12) can accommodate only coarse partitions of the covariate space. If the vector X contains many informative covariates, Lee (2009)'s covariate-based estimator will not be close to the sharp bound β_U in a large sample.

Lee (2009) argues that including covariates can lead to point identification in extreme cases. First, consider the case where selection is exogenous conditional on covariates X. Then, the conditional probability of selection must be the same in the treatment and control groups:

$$s(0,x) = s(1,x) \quad \forall x \in \mathcal{X}.$$

As a result, the trimming threshold $p_0(x) = 1$ for all covariates x, and

$$\beta_L = \beta_0 = \beta_U. \tag{2.13}$$

Second, consider the case where the outcome is a deterministic function of the covariates. Then, the conditional quantile function of $Q(u,x)$ does not vary within covariate groups, and $Q(p_0(x),x) = Q(1-p_0(x),x) \quad \forall x \in \mathcal{X}$. As a result, (2.13) holds. Thus, the covariates that explain most variation in either selection or outcome are likely to be the most useful for tightening the bounds.

3 JobCorps revisited

In this section, I review the basics of Lee (2009)'s empirical analysis of JobCorps training program and replicate Lee's results. 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 Section C.2.

3.1 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

$$\Delta(x) = s(1,x) - s(0,x) = \Pr(S=1|D=1,X=x) - \Pr(S=1|D=0,X=x), \quad x \in \mathcal{X} \tag{3.1}$$

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

$$\text{Prob}(\Delta(x) > 0) > 0 \quad \text{and} \quad \text{Prob}(\Delta(x) < 0) > 0. \tag{3.2}$$

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

$$s(D,X) = \Lambda(X'\alpha_0 + D \cdot X'\gamma_0), \tag{3.3}$$

where $\Lambda(\cdot) = \dfrac{\exp(\cdot)}{1+\exp(\cdot)}$ 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. I report the average treatment-control difference for the covariate groups $\{\Delta(x) > 0\}$ and $\{\Delta(x) < 0\}$ in Figure 1 and the fraction of subjects in the covariate group $\{\Delta(x) > 0\}$ in Figure 2.

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 $C_j, \quad j \in \{1, 2, \dots, J\}$, 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:

$$H_0 : \quad (-1) \cdot \mu \leq 0. \tag{3.4}$$

The test statistic for the hypothesis in equation (3.4) is

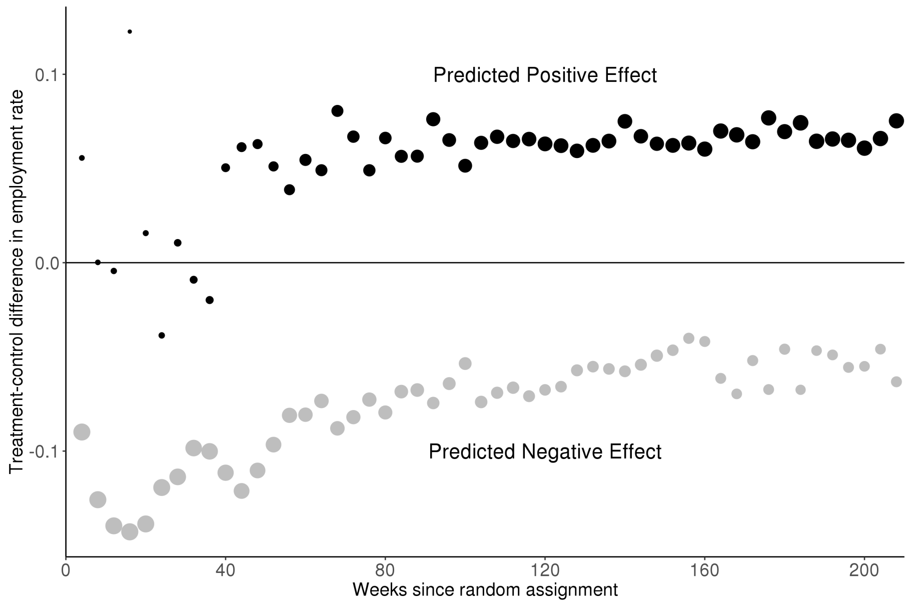
$$T = \max_{1 \leq j \leq J} \frac{(-1) \cdot \widehat{\mu}_j}{\widehat{\sigma}_j}, \tag{3.5}$$

and the critical value is the self-normalized critical value of Chernozhukov et al. (2019).

3.2 Results

Figure 1 shows the treatment-control difference in employment rates for applicant groups whose estimated employment effect is positive (black dots) or negative (gray dots) conditional on covariates. The fraction of applicants with a positive employment effect increases over time. Focusing on week 90, I find that a smaller chance of employment is associated with being female, black, receiving public assistance before RA, such as food stamps or other welfare, being raised in a family that has received welfare most or all the time, being in fair (not excellent or good) health at the moment of RA, and smoking hashish or marijuana a few times each week. In addition, JobCorps is likely to hurt week 90 employment chances for subjects whose most recent arrest occurred less than a year before the baseline interview or who are on probation or parole at the moment of the interview. Week 90 is a special week since it is the first week where the average employment effect switches from negative to positive, and the only one out of five horizons where Lee found the average wage effect on the always-takers to be significant.

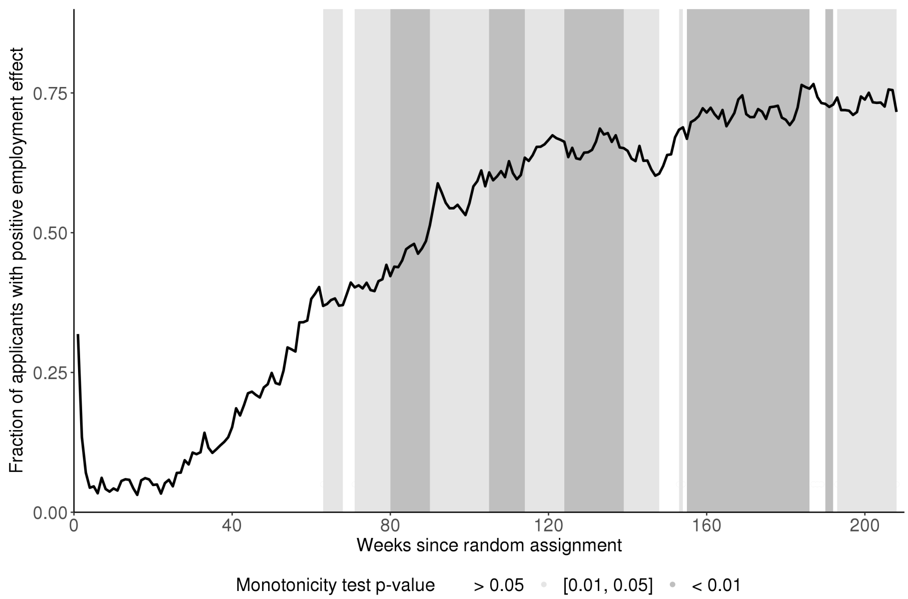
Figure 1: Treatment-control differences in employment rate by week.



Notes. The horizontal axis shows the number of weeks since random assignment. The vertical axis shows the treatment-control difference in employment rate. The black dot represents applicants whose conditional employment effect $\Delta(x)$ is positive, and the gray dot is its complement. (For each week, $\Delta(x)$ is defined as in equation (3.1) and estimated as in equation (3.3)). The size of each dot is proportional to the fraction of applicants. Computations use design weights.

Figure 2 plots the fraction of subjects with a positive JobCorps effect on employment in each week (that is, the fraction of applicants in black dots in Figure 1). 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. After finishing the program, JobCorps graduates may have gained employment skills that help them outperform the control group.

Figure 2: 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 $\Delta(x)$ 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, $\Delta(x)$ is defined in equation (3.1) and estimated as in equation (3.3), the null hypothesis is as in equation (3.4), the test statistic T is as in equation (3.5), and the test cells and critical values are as defined in Table E.9. Computations use design weights.

Figure 2 shows the results of testing the inequality in (3.4) 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.

Table 1 replicates Lee's estimates of basic (Column (1)) bounds on JobCorps effect on the wages of always-takers. In addition, I also compute the covariate-based bounds using the discretized predicted wage potential covariate that Lee proposed (Column (2)). 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. Once this assumption is weakened, basic Lee bounds do not cover zero in any week (Table 3, Column 1).

4 Theoretical results

In this section, I introduce better Lee bounds. Section 4.1 presents bounds under a weakened monotonicity assumption that only needs to hold conditional on covariates. Section 4.2 formulates the statistical assumptions on the data generating process and states asymptotic results under these assumptions.

4.1 Identification

ASSUMPTION 2 (Conditional monotonicity)**.** *The following statements hold.*

(1) (Independence). The vector $(Y(1), Y(0), S(1), S(0))$ *is independent of D conditional on X.*

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

	Basic (1)	Covariate-based (2)
Week 45	[-0.072, 0.140]	[-0.074, 0.127]
	(-0.097, 0.170)	(-0.096, 0.156)
Week 90	**[0.048, 0.049]**	**[0.036, 0.048]**
	(0.011, 0.081)	**(0.011, 0.075)**
Week 208	[-0.020, 0.095]	[-0.014, 0.084]
	(-0.050, 0.118)	(-0.041, 0.109)
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 discourages 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, given in Table E.1. The covariates are weighted by the coefficients from a regression of week 208 wages on all baseline characteristics in the control group. The five discrete 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.

(2) (Monotonicity). There exists a partition of covariate space $\mathcal{X} = \mathcal{X}_{help} \sqcup \mathcal{X}_{hurt}$ so that $S(1) \geq S(0)$ a.s. on $\mathcal{X}_{help}$ and $S(0) \geq S(1)$ a.s. on $\mathcal{X}_{hurt}$.

Assumption 2 allows the sign of the treatment effect on employment to vary along with covariates. A subject with covariate vector X belongs to the covariate group $\mathcal{X}_{\text{help}}$ if and only if his treatment-control difference in selection rate is positive conditional on X:

$$X \in \mathcal{X}_{\text{help}} \quad \Leftrightarrow \quad p_0(X) \leq 1 \quad \Leftrightarrow \quad S(1) \geq S(0) \quad \text{a.s.},$$

where $p_0(x)$ is defined in (2.8). When there are no covariates, Assumption 2 coincides with Assumption 1. The more covariates are available, the weaker the assumption is.

The sharp lower and upper bound are given in equation (B.12) in Appendix B.

ASSUMPTION 3 (Strong Overlap and Endogeneity)**.** *The following conditions hold.*

(1) (Overlap). There exists $0 < \underline{s} < \bar{s} < 1$ *such that* $0 < \underline{s} < s(d,x) < \bar{s} < 1$ *for any x and* $d \in \{1,0\}$.

(2) (Endogeneity). There exists a set $\bar{\mathcal{X}} \subset \mathcal{X}, \Pr(\mathcal{X} \setminus \bar{\mathcal{X}}) = 0$ *and an absolute constant* $\varepsilon > 0$ *such that* $\inf_{x \in \bar{\mathcal{X}}} : |s(0,x) - s(1,x)| > \varepsilon$.

Assumption 3 is the price of relaxing Assumption 1(2) to Assumption 2(2). It ensures that subjects are correctly assigned into $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$ when the sample size is large, with high probability. To attain correct classification, a sufficient condition on the trimming threshold $p_0(x)$ is to have its support bounded away from both zero and one. In particular, the trimming threshold $p_0(x)$ cannot be equal to one (i.e., selection cannot be conditionally exogenous) with positive probability. Lee's model implicitly imposes Assumption 3, requiring that the covariate-specific trimming thresholds are bounded away from one for each discrete cell. Suppose Assumption 1 holds (i.e., $\mathcal{X}_{\text{hurt}} = \emptyset$). Assumption 3 is still required for sharpness, to ensure that there are enough data points above and below the quantile level $u = 1 - p_0(x)$ to estimate the conditional quantile $Q(u,x)$.

4.2 Estimation

4.2.1 First Stage

In this subsection, I state the asymptotic results for better Lee bounds. For the sake of clarity, I derive the results below under Assumption 1 rather than Assumption 2. Appendix B contains the general-case derivations.

Suppose Assumption 1 holds. Applying Bayes rule (see Lemma B.1), I derive a moment equation for β_U:

$$\beta_U = \mu_{10}^{-1}\mathbb{E}\left[\frac{D}{\Pr(D=1)} \cdot S \cdot Y \cdot 1_{\{Y \geq Q(1-p_0(X),X)\}} - \frac{(1-D)}{\Pr(D=0)} \cdot S \cdot Y\right] = \mathbb{E}m_U(W,\xi_0), \tag{4.1}$$

where $W = (D,X,S,S \cdot Y)$ is the data vector, $\mu_{10} = \Pr(S=1|D=0)$,

$$\xi_0 = \{s(0,x), s(1,x), Q(1-p_0(x),x)\}$$

is the first-stage nuisance parameter, and $m_U(W,\xi)$ is a moment function. The nuisance parameter ξ_0 contains the conditional probability of selection $\{s(0,x),s(1,x)\}$ in both the treated and control status, and the conditional quantile function $Q(u,x)$. For simplicity, the population parameters $\Pr(D=0), \Pr(D=1) = 1-\Pr(D=0)$ and $\Pr(D=0,S=1)$ are treated as known; their estimation does not conceptually affect the results.

ASSUMPTION 4 (First-Stage Rate of Selection Equation)**.** *There exist sequences of numbers $\varepsilon_N = o(1)$, $s_N = o(N^{-1/4})$ and a sequence of sets S_N such that the first-stage estimates $\widehat{s}(0,x)$ of the true function $s_0(0,x)$ and $\widehat{s}(1,x)$ of the true function $s_0(1,x)$ belong to S_N with probability at least $1-\varepsilon_N$. Furthermore, the sets S_N shrink sufficiently fast around the true functions:*

$$\begin{aligned} \sup_{s \in S_N} \left(\mathbb{E}_X(s(0,X)-s_0(0,X))^2\right)^{1/2} &\leq s_N = o(N^{-1/4}), \qquad (4.2) \\ \sup_{s \in S_N} \left(\mathbb{E}_X(s(1,X)-s_0(1,X))^2\right)^{1/2} &\leq s_N = o(N^{-1/4}). \end{aligned}$$

Assumption 4 states that the functions $s(0,x)$ and $s(1,x)$ are estimated with sufficient quality. This is a classic assumption in the semiparametric literature (see, e.g., Newey (1994)). This assumption rules out any second-order effects of the estimation error

$\widehat{s}(0,X) - s(0,X)$ and $\widehat{s}(1,X) - s(1,X)$ on the bounds, and allows the researcher to focus on the first-order terms.

Focusing on the function $s(0,x)$, a common approach to estimate $s(0,x)$ is to consider a logistic approximation

$$s(0,x) = \Lambda(B(x)'\gamma_0) + r(x), \tag{4.3}$$

where $\Lambda(\cdot)$ is the logistic CDF, $B(x) = (B_1(x), B_2(x), \ldots B_{p_S}(x))'$ is a vector of basis functions (e.g., polynomial series or splines), $\gamma_0 \in \mathrm{R}^{p_S}$ is the pseudo-true value of the logistic parameter, and $r(x)$ is its approximation error.

Primitive Condition 1 (Smooth Selection Model)**.** *The function $s(0,x)$ is continuously differentiable of the order $\kappa \geq 7 \cdot dim(X)$.*

Primitive Condition 1 is a low-level sufficient condition for Assumption 4. If this condition holds, the logistic series estimator of Hirano et al. (2003) takes the form

$$\widehat{\gamma}_{\mathrm{LSE}} = \arg\max_{\gamma \in \mathrm{R}^{p_S}} \ell(\gamma), \tag{4.4}$$

where $\ell(\gamma)$ is the logistic likelihood function

$$\ell(\gamma) = \frac{1}{N} \sum_{i=1}^{N} 1_{\{D_i = 0\}} \Big(\log(1 + \exp(B(X_i)'\gamma)) - S_i B(X_i)'\gamma \Big). \tag{4.5}$$

If the Primitive Condition 1 holds, plugging $\widehat{\gamma} = \widehat{\gamma}_{\mathrm{LSE}}$ delivers an estimator

$$\widehat{s}(0,x) = \Lambda(B(x)'\widehat{\gamma}), \quad x \in \mathcal{X}, \tag{4.6}$$

which converges at rate $s_N = \sqrt{p_S/N} = o(N^{-1/4})$.

When $\dim(X) \geq \log N$, a consistent estimate of a smooth function does not exist in general case (Stone (1982)). To make progress, we need alternative assumptions on the structure of the function $s(0,x)$. One possible assumption is approximate sparsity, which requires that few of the basis functions in the vector $B(x)$ can approximate $s(0,x)$ sufficiently well.

Primitive Condition 2 (Approximately Sparse Selection Model)**.** *There exists a vector* $\gamma_0 \in \mathrm{R}^{p_S}$ *with only* s_γ *non-zero coordinates such that the approximation error* $r(x)$ *in* (4.3) *decays sufficiently fast relative to the sampling error:*

$$\left(\frac{1}{N}\sum_{i=1}^{N} r^2(X_i)\right)^{1/2} \lesssim_P \sqrt{\frac{s_\gamma^2 \log p_S}{N}} =: s_N.$$

Primitive Condition 2 is a low-level sufficient condition for Assumption 4. If this condition holds, the ℓ_1-regularized logistic series estimator of Belloni et al. (2017) takes the form

$$\widehat{\gamma}_L = \arg\max_{\gamma \in \mathrm{R}^{p_S}} \ell(\gamma) + \lambda \|\gamma\|_1, \tag{4.7}$$

where $\lambda \geq 0$ is a penalty parameter. This penalty term $\lambda\|\gamma\|_1$ prevents overfitting in high dimensions by shrinking the estimate toward zero. Belloni et al. (2017) provides practical choices for the penalty λ that provably guard against overfitting. An imminent cost of applying the penalty λ is regularization, or shrinkage, bias, that does not vanish faster than root-N rate. To prevent this bias from affecting the second stage, I construct a Neyman-orthogonal moment equation for each bound.

If Primitive Condition 2 holds with a sufficiently small s_γ, both the lasso-logistic estimator and its post-penalized analog satisfy Assumption 4 with $s_N = \sqrt{\frac{s_\gamma^2 \log p_S}{N}} = o(N^{-1/4})$. In contrast to the smooth model, the convergence rate s_N depends on $\log p_S$ rather than p_S itself, permitting the number of covariates under consideration to exceed the sample size.

ASSUMPTION 5 (Quantile First-Stage Rate: One-Dimensional Case)**.** *Let* $\bar{\mathcal{U}}$ *be a com-*

pact set in $(0,1)$ *containing the support of* $p_0(X)$ *and* $1-p_0(X)$*. There exist a rate* $q_N = o(N^{-1/4})$*, a sequence of numbers* $\varepsilon_N = o(1)$ *and a sequence of sets* Q_N *such that the first-stage estimate of the quantile function* $Q(u,x) : [0,1] \times \mathcal{X} \to \mathrm{R}$ *belongs to* Q_N *w.p. at least* $1-\varepsilon_N$*. Furthermore, the set* Q_N *shrinks sufficiently fast around the true value* $Q_0(u,x)$ *uniformly on* $\bar{\mathcal{U}}$*:*

$$\sup_{Q \in Q_N} \sup_{u \in \bar{\mathcal{U}}} \left(\mathbb{E}_X (Q(u,X) - Q_0(u,X))^2 \right)^{1/2} \leq q_N = o(N^{-1/4}).$$

Assumptions 5 is an analog of Assumption 4 for the quantile function. It states that the conditional quantile function $Q(u,x)$ is estimated at sufficient quality, as measured by the mean square convergence rate $q_N = o(N^{-1/4})$, and is a classic assumption in non-parametric literature. A classic approach to estimate $Q(u,x)$ is by quantile regression. When the function $Q(u,x)$ is a sufficiently smooth function of x, the quantile series estimator of Belloni et al. (2019) converges at rate $q_N = \sqrt{\frac{p_Q}{N}} = o(N^{-1/4})$ uniformly over $\mathcal{U}$, where p_Q is the number of series terms to approach $Q(u,x)$. Likewise, ℓ_1-penalized quantile regression estimate of Belloni and Chernozhukov (2013) satisfies Assumption 5 with $q_N = \sqrt{s_Q^2 \log p_Q / N} = o(N^{-1/4})$ under the choice of λ proposed in Belloni and Chernozhukov (2013) if the model is sufficiently sparse. Appendix B.2 contains a more technical discussion of sufficient primitive conditions on the conditional quantile function.

4.2.2 Second Stage

In this section, I discuss two well-known ideas: cross-fitting and Neyman-orthogonality (Neyman (1959)). Combining these two ideas, I propose a better Lee bounds estimator of sharp Lee bounds (β_L, β_U).

Cross-fitting. When the first-stage parameter ξ is estimated by ℓ_1-regularized methods, sample splitting is not required for the asymptotic results, but can accommodate larger sparsity indices s_γ and s_Q (Chernozhukov et al. (2018)). For the other machine learning techniques, I rely on the cross-fitting idea of Chernozhukov et al. (2018) to establish theoretical guarantees.

Neyman-orthogonality. A two-stage estimation procedure is orthogonal if the second stage is insensitive (formally, orthogonal) to the first-stage parameter (Neyman (1959)). Lee's moment equation (4.1) is not orthogonal. In particular, its derivative with respect to a local parametrization of the first-stage nuisance parameter ξ at its true value ξ_0 is not equal to zero:

$$\partial_\xi \mathbb{E} m_U(W,\xi_0)[\widehat{\xi}-\xi_0] \neq 0. \tag{4.8}$$

As a result, the biased estimation error of $\widehat{\xi}-\xi_0$ translates into bias in the moment equation (4.1).

To prevent transmission of the bias into the second stage, I derive an orthogonal moment equation for the lower and the upper bound. The proposed moment equations, given in equations (B.23)-(B.24), obey the zero-derivative property

$$\partial_\xi \mathbb{E} g_U(W,\xi_0)[\widehat{\xi}-\xi_0] = 0. \tag{4.9}$$

As a result, the first-stage bias does not affect the asymptotic distribution of the bounds under Assumptions 4 and 5.

Theorem 1 (Asymptotic Theory for Sharp Bounds)**.** *Suppose Assumptions 2–5 hold. In addition, if $\mathcal{X}_{help} \neq \emptyset$ and $\mathcal{X}_{hurt} \neq \emptyset$, suppose $\widehat{s}(d,x)$ converges to $s(d,x)$ uniformly over $\mathcal{X}$ for each $d \in \{1,0\}$. Then, the better Lee bounds estimator $(\widehat{\beta}_L, \widehat{\beta}_U)$ is consistent and* *asymptotically normal,*

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

where Ω *is a positive-definite covariance matrix defined as*

$$\Omega=\begin{bmatrix}\mathbb{E}g_L(W,\xi_0)^2 & \mathbb{E}g_L(W,\xi_0)g_U(W,\xi_0)\\ \mathbb{E}g_L(W,\xi_0)g_U(W,\xi_0) & Eg_U(W,\xi_0)^2\end{bmatrix} \tag{4.10}$$

that can be estimated by a sample analog.

Theorem 1 delivers a root-N consistent, asymptotically normal estimator of (β_L,β_U) assuming the conditional probability of selection and conditional quantile are estimated at a sufficiently fast rate. In particular, this assumption is satisfied when few of the covariates affect selection and the outcome. By orthogonality, the first-stage estimation error does not contribute to the total uncertainty of the two-stage procedure. As a result, the asymptotic variance (4.10) does not have any additional terms due to the first-stage estimation.

Remark 1 (Sorted Bounds). The sorted estimator $(\widetilde{\beta}_L,\widetilde{\beta}_U)$,

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

must converge at least as fast as the original estimator $(\widehat{\beta}_L,\widehat{\beta}_U)$.

Unlike standard Lee bounds, $\widehat{\beta}_L$ and $\widehat{\beta}_U$ are not ordered by construction. Chernozhukov et al. (2013a) suggests that sorting the estimated bounds can only improve the convergence rate. Likewise, sorting the confidence region continues to guarantee coverage if the original (unsorted) confidence region guarantees coverage. However, the Imbens and Manski (2004) local super-efficiency assumption no longer holds (Stoye

(2009)). Instead, one can use Stoye (2009)'s modification of the IM confidence interval,

$$\mathrm{CI}_{\alpha}^{\text{Stoye}} = \begin{cases} \left[\widehat{\beta}_L - \frac{\widehat{\Omega}_{LL}^{1/2} c_{L,\alpha}^S}{\sqrt{N}}, \widehat{\beta}_U + \frac{\widehat{\Omega}_{UU}^{1/2} c_{U,\alpha}^S}{\sqrt{N}}\right], & \widehat{\beta}_L - \frac{\widehat{\Omega}_{LL}^{1/2} c_{L,\alpha}^S}{\sqrt{N}} \leq \widehat{\beta}_U + \frac{\widehat{\Omega}_{UU}^{1/2} c_{U,\alpha}^S}{\sqrt{N}}, \\ \emptyset, & \text{otherwise}, \end{cases}$$

where $c_{L,\alpha}^S$ and $c_{U,\alpha}^S$ are Stoye's critical values. If $\widehat{\beta}_U$ is too far below $\widehat{\beta}_L$, the interval $\mathrm{CI}_{\alpha}^{\text{Stoye}}$ is empty, indicating that Assumptions 4 and 5 are likely to be violated.

5 Agnostic approach

In this section, I relax the sparsity assumption from Section 4. Suppose Assumption 1 holds. Let X_A be a subvector of covariates X. By Lemma B.1,

$$\beta_L^A \leq \beta_0 \leq \beta_U^A,$$

where $[\beta_L^A, \beta_U^A]$ is the sharp identified set for β_0 in the model $(D, X_A, S, S \cdot Y)$ where only X_A covariates are observed.

To select covariates in a data-driven way, randomly split the full sample into an auxiliary sample A and a main sample M. In the auxiliary sample, select the covariates by an arbitrary machine learning method. In the main sample, define the target bounds as the sharp bounds in the model $(D, X_A, S, S \cdot Y)$ with selected covariates X_A. Assuming Primitive Condition 1 holds for the selected model, let $\widehat{\beta}_L^A$ and $\widehat{\beta}_U^A$ be the estimates of β_L^A and β_U^A, based on logistic and quantile series estimates of the first-stage nuisance parameters. Conditional on the auxiliary sample, the $(1-\alpha)$ confidence region for $[\beta_L^A, \beta_U^A]$ is

$$[L_A, U_A] = [\widehat{\beta}_L^A - |M|^{-1/2} \widehat{\Omega}_{A,LL}^{1/2} c_{\alpha/2}, \quad \widehat{\beta}_U^A + |M|^{-1/2} \widehat{\Omega}_{A,UU}^{1/2} c_{1-\alpha/2}].$$

Variational Inference. Different splits (A,M) of the sample $\{1,2,\ldots,N\}$ yield different target bounds (β_L^A,β_U^A) and different approximate distributions of these bounds. If we take the splitting uncertainty into account, the pair of bounds (β_L^A,β_U^A) is random conditional on the full data sample. In practice, one may want to generate several random splits and aggregate various bounds over various partitions. For reporting purposes, I use Chernozhukov et al. (2017)'s adjusted point estimator

$$\widehat{\beta}_L = \text{Med}[\widehat{\beta}_L^A|\text{Data}], \quad \widehat{\beta}_U = \text{Med}[\widehat{\beta}_U^A|\text{Data}].$$

To quantify the uncertainty of the random split, I use Chernozhukov et al. (2017)'s adjusted confidence interval of level $1-2\alpha$:

$$[L,U] = \left[\, \overline{\text{Med}}[L_A|\text{Data}], \quad \underline{\text{Med}}[U_A|\text{Data}]\right], \tag{5.1}$$

where $\underline{\text{Med}}(X) = \inf\{x \in \mathrm{R} : P_X(X \le x) \ge 1/2\}$ is the lower median and $\overline{\text{Med}}(X) = \sup\{x \in \mathrm{R} : P_X(X \ge x) \ge 1/2\}$ is the upper median. A more formal analysis of agnostic approach is given in Section B.3.

6 Extensions

In this section, I discuss extensions of my basic setup. In Section 6.1, I extend my setup to allow for a multi-dimensional outcome, using standardized treatment effect as the main motivation for this extension. Sections 6.2 and 6.3 derive better Lee bounds for the Intent-to-Treat and Local Average Treatment Effect parameters. Sections 6.4 and 6.5 generalize my results to settings with multiple observations for each cross-sectional unit. Section 6.6 considers the case when the propensity score is unknown. The results I present in the main text summarize the insights from a formal analysis in Appendix B.

6.1 Multi-dimensional outcome

In this section, I generalize the Lee (2009) sample selection model to allow for a multi-dimensional outcome variable. 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 $\mathbf{S} \in \mathrm{R}^d$ and $\mathbf{Y} \in \mathrm{R}^d$ are d-vectors. The parameter of interest is the average treatment effect

$$\beta_0 = \mathbb{E}[\mathbf{Y}(1) - \mathbf{Y}(0) | \mathbf{S}(1) = \mathbf{S}(0) = \mathbf{1}] \tag{6.1}$$

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

Proposition B.4 shows that the sharp identified set for β_0 is compact and convex, and thus can be summarized by its projections on various directions of economic interest. For any point q on the unit sphere, the largest admissible value $\sigma(q)$ of $q'\beta_0$ consistent with the observed data, is commonly referred to as the *support function*. In Appendix B, I provide a Gaussian approximation for the support function process that is uniform over the unit sphere and propose a Bayes bootstrap procedure to conduct simultaneous inference uniformly over the sphere. Examples 1 and 2 explain the use of support function in applied work.

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

$$[-\sqrt{2}\sigma(-q), \quad \sqrt{2}\sigma(q)], \quad q = (1/\sqrt{2}, -1/\sqrt{2}). \tag{6.2}$$

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*

$$\text{STE} = \frac{1}{d}\sum_{j=1}^{d}\frac{\beta_j}{\zeta_j}, \tag{6.3}$$

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

$$[-C_\zeta\sigma(q), \quad C_\zeta\sigma(q)], \tag{6.4}$$

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 that that the support function estimator can be approximated uniformly in some neighborhood of q in addition to the point q itself. I establish this approximation in Theorem B.7, and give inference methods for hypothesis testing in Theorem B.9.

6.2 Intent-to-Treat

I consider the standard Intent-to-Treat parameter. Let $D = 1$ be an indicator of being offered treatment and let $\bar{X}$ be a vector of stratification covariates (i.e., fixed effects), so that D is randomly assigned conditional on $\bar{X}$. In addition, $\bar{X}$ is a saturated vector and X is a full covariate vector that includes $\bar{X}$. The object of interest is the intent-to-treat effect (ITT)

$$Y = \beta_0 + \beta_1 D + \bar{X}'\beta_2 + \varepsilon, \quad S(1) = S(0) = 1, \tag{6.5}$$

where β_1 is the main coefficient of interest, interpreted as the average causal effect of being offered treatment to an always-taker. Since only one of $S(1)$ and $S(0)$ is observed, the parameter β_1 is not point-identified. For the sake of simplicity, suppose $\mathcal{X}_{\text{hurt}} = \emptyset$. A sharp upper bound on β_1 is the regression coefficient on D in the truncated regression on the selected outcomes, where the bottom outcomes in the treated group are trimmed until selection response rates in both the $D = 1$ and $D = 0$ groups are equal to each other for each value of strata $\bar{X}$. (Proposition B.13, Appendix B). Furthermore, any subvector of X that contains $\bar{X}$ corresponds to another valid bound on β_1 that may not be sharp. However, a subvector of X that does not contain $\bar{X}$ does not correspond to a valid bound. Lemma B.14 in Appendix B extends the orthogonal and agnostic approaches to the Intent-to-Treat parameter.

6.3 Local Average Treatment Effect

I consider the Local Average Treatment Effect parameter defined in Imbens and Angrist (1994). Let $Z = 1$ be a binary instrument indicator, such as an offer of treatment, and $D = 1$ be a binary treatment, such as actual treatment receipt, and define $\bar{X}$, X, S and Y as in the previous section. Suppose the potential selection outcome

$$S(1,z) = S(0,z) = S(z), \quad \text{for any } z \in \{1,0\}$$

is fully determined by the value of the instrument. The object of interest is the average treatment effect on the subjects who comply with the treatment offer and select into the sample regardless of the treatment offer. Finally, suppose being a complier is independent of being an always-taker given on all observed covariates. Then, the target parameter is identified as the coefficient π_1 in the two-stage least squares (2SLS) regression

$$Y = \pi_0 + \pi_1 D + \bar{X}'\pi_2 + \nu, \quad S(1) = S(0) = 1, \tag{6.6}$$

where the first-stage equation is

$$D = \delta_0 + \delta_1 Z + \bar{X}'\delta_2 + \zeta, \quad S(1) = S(0) = 1. \tag{6.7}$$

As shown in Proposition B.18 in Appendix B, a sharp upper bound on π_1 is the 2SLS effect in the truncated regression on the selected outcomes, where the bottom outcomes in the treated group are trimmed until selection response rates are the same in $D = 1, Z = 1$ and $D = 0, Z = 1$ and $Z = 0$ groups for each value of strata $\bar{X}$. Furthermore, any subvector of X that contains $\bar{X}$ corresponds to another valid bound on π_1 that may not be sharp. However, a subvector of X that does not contain $\bar{X}$ may not correspond to a valid bound.

6.4 Clustered Data

Suppose that the researcher observes data sampled from G clusters: $\{W_{ig}, \quad i = 1, 2, \dots, N_g, \quad g = 1, 2, \dots G\}$. Each cluster size is non-random, and $1 \leq N_g \leq \bar{N} < \infty$ for a constant $\bar{N}$ that does not depend on sample size G. According to Chiang (2020), the post-lasso-logistic estimator with the cluster-robust penalty parameter satisfies the analog of Assumption 4 with $s_G = G^{-1/4}$ rate. Generalizing Chiang (2020)'s arguments, one can establish the analog of Assumption 5 with $q_G = G^{-1/4}$. Under these assumptions, the better Lee bounds estimator is consistent and asymptotically normal. The cluster-robust estimator of asymptotic variance Ω in Theorem 1 takes the form

$$\widehat{\Omega}_{cr} = \begin{bmatrix} 1/G\sum_{g=1}^{G}(\sum_{c=1}^{n_g} g_L(W_{gc}, \widehat{\xi}))^2 & 1/G\sum_{g=1}^{G}(\sum_{c=1}^{n_g} g_L(W_{gc}, \widehat{\xi}))(\sum_{j=1}^{n_g} g_U(W_{gj}, \widehat{\xi})) \\ 1/G\sum_{g=1}^{G}(\sum_{c=1}^{n_g} g_L(W_{gc}, \widehat{\xi}))(\sum_{j=1}^{n_g} g_U(W_{gj}, \widehat{\xi})) & 1/G\sum_{g=1}^{G}(\sum_{c=1}^{n_g} g_U(W_{gc}, \widehat{\xi}))^2. \end{bmatrix} \tag{6.8}$$

6.5 Panel Data

Consider a setting where the units $(D_i, S_{it}, S_{it}Y_{it}, X_i)_{i=1,t=1}^{NT}$ are observed over $t = 1, 2, \ldots, T$ time periods. Using the notation of Section 6.1, let $\mathbf{S}_i := (S_{i1}, S_{i2}, \ldots S_{iT})$ be a vector of selection indicators for an individual i and $\mathbf{Y}_i := (Y_{i1}, S_{i2}, \ldots Y_{iT})$ be a vector of outcomes. The target parameter β_0 is the average treatment effect

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

for subjects who are selected into the sample in each period regardless of treatment status. In contrast to the cross-sectional setup of Section 6.1, it is important to allow the observations to be correlated over time within each individual. In this case, one can group the observations over time into clusters and use the cluster-robust standard error derived in Section 6.4 with $G = N$ and $N_g = T, \quad g \in \{1, 2, \ldots T\}$.

6.6 Unknown propensity score

In this section, I extend better Lee bounds to accommodate the case when the conditional probability of treatment is unknown. The orthogonal moment equation (B.18) involves an additional nuisance parameter

$$\Big\{ \mathbb{E}[Y | S = 1, D = 0, X], \quad \mathbb{E}[Y | Y \leq Q(u, X), S = 1, D = d, X], \quad d \in \{1, 0\} \Big\}$$

that needs to be estimated. If the sparsity assumption holds, Belloni et al. (2017)'s linear lasso estimator of the function $\mathbb{E}[Y | S = 1, D = 0, X]$. The truncated conditional mean function $\mathbb{E}[Y | Y \leq Q(u, X), S = 1, D = 1, X], \quad u \in \{p_0(X), 1 - p_0(X)\}$ can be estimated by Chernozhukov et al. (2018)'s automatic debiasing approach, aimed at generic nuisance functions that emerge as a result of orthogonalization. If few of the covariates affect selection and the outcome, the assumptions of Chernozhukov et al. (2018)' setting hold, and the proposed estimate obeys an analog of Assumptions 4-5.

7 Simulation Evidence

In this section, I compare the performance of basic, naive and better Lee methods, building 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

$$S = 1\{X'\alpha + D \cdot X'\gamma + U > 0\}, \tag{7.1}$$

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

$$Y = (1, X_1)'\kappa + \varepsilon, \quad \varepsilon \sim N(0, \widetilde{\sigma}^2), \tag{7.2}$$

where ε is an independent normal random variable. The parameter vector $(\alpha, \gamma, \kappa, \widetilde{\sigma}^2)$ is taken to be the estimates of (7.1) and (7.2), where S and Y are week 90 employment and log wages, respectively, adjusted as described in Appendix D. 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 sharp identified set is $[-0.011, 0.018]$. Basic Lee bounds are defined as the weighted average of standard Lee bounds on $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$. The true basic identified set is $[-0.014, 0.035]$.

I compare the performance of four estimators—oracle, basic, naive and better Lee 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 (7.1) and (7.2), they are strongly correlated with X_1 and X_2, making covariate selection an interesting problem. The oracle method is the output of Algorithm 1, where the oracle knows the identities of covariates in vector X and the direction of employment effect on $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$. 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 basic identified set $[-0.014, 0.035]$. Both the naive and the better methods target the sharp identified set $[-0.011, 0.018]$. The naive method estimates the first-stage functions (2.8) and (2.9) by standard regression methods on all 28 covariates. In contrast, the better 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. In the second stage, both the naive and the better method rely on orthogonal moment equations (B.18) for the lower and the upper bound, respectively.

Table 2 reports the finite-sample performance for the oracle, basic, naive and better methods. I focus on the lower and the upper bound separately to detect any outward bias or poor coverage of either bound, which is masked when a confidence interval is considered. On average, the width of oracle bounds ranges from 0.028 to 0.034.

Table 2 presents evidence that the better method outperforms all other methods in terms of width, precision, and coverage. Indeed, the standard deviations of the naive and the basic estimates are equal for all sample sizes under consideration. Second, better Lee bounds are substantially tighter than the basic and the naive ones, by a factor of 3 and 2.5 on average. Third, the 95 % confidence interval for the true lower bound based on the better method covers it in at least 88% of simulation runs. For the upper bound, the coverage rate is at least 93%. In contrast, the basic and the naive methods exhibit poor coverage despite that X is always included into the logistic regression together with

Table 2: Finite-sample performance of oracle, basic, naive and better Lee methods

Panel A: Lower Bound

	Bias				St. Dev.				Coverage Rate			
N	Oracle	Basic	Naive	Better	Oracle	Basic	Naive	Better	Oracle	Basic	Naive	Better
3,000	-0.00	-0.05	-0.06	-0.02	0.01	0.02	0.03	0.03	0.94	0.21	0.64	0.88
5,000	-0.00	-0.04	-0.04	-0.01	0.01	0.01	0.02	0.01	0.94	0.24	0.63	0.88
9,000	0.00	-0.03	-0.03	-0.01	0.01	0.01	0.02	0.01	0.95	0.26	0.65	0.93
10,000	0.00	-0.03	-0.03	-0.01	0.01	0.01	0.01	0.01	0.95	0.25	0.64	0.93
15,000	0.00	-0.02	-0.02	-0.01	0.00	0.01	0.01	0.01	0.95	0.23	0.64	0.92

Panel B: Upper Bound

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. In Panel A, the true parameter value is -0.014 for the basic method, and -0.011 for all other methods. In Panel B, 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. Oracle, basic, naive and better estimated bounds cover zero in 100% of the cases. The naive method estimates the first-stage functions (2.8) and (2.9) by logistic and quantile regression on all 28 covariates.

irrelevant covariates. Although the vector X is not been selected for the employment equation in at least 90% of simulation runs (i.e., "perfect" model selection by lasso-logistic rarely occurs), the better estimates prove robust to incorrectly attributing an observation from $\mathcal{X}_{\text{help}}$ into $\mathcal{X}_{\text{hurt}}$ and vice versa. Table D.1 in Appendix D presents additional simulation evidence suggesting that the agnostic version of the better method also outperforms the basic method in terms of width and coverage and has comparable precision.

8 Empirical Applications

In this section, I demonstrate how better Lee bounds can achieve nearly point-identification in three empirical settings. First, I study the effect of JobCorps on wages, as in Lee (2009). Second, I study the effect of PACES voucher tuition subsidy on pupils' test scores, as in Angrist et al. (2002). Finally, I study the effect of Medicaid eligibility and insurance on self-reported healthcare utilization and health, as in Finkelstein et al. (2012).

8.1 Lee (2009)

Lee (2009) studies the effect of winning a lottery to attend JobCorps, a federal vocational and training program, on applicants' wages. The data set is the same as in Section 3.

Table 3 reports estimated bounds on the JobCorps week 90 wage effect on the always-takers and the confidence region for the identified set. The basic Lee bounds cannot determine the direction of the effect (Column (1)). Neither can the sharp bounds given Lee's covariates (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 1, Column 1) and better Lee estimates (Table 3, Column 3) have substantially different reasons for being tight. The former bounds are tight because week 90 employment is interpreted as unaffected by the lottery outcome (i.e., the estimated unconditional trimming threshold $\widehat{p}$ is close to one). In contrast, the better ones account for the differential sign of the JobCorps effect on applicants' employment. The better Lee 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 bounds in Column (3) assume sparsity, which excludes some wage covariates from the employment equation and vice versa. In Column (4), the target bounds are defined as the sharp 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 it is plausible for week 90 employment and wage equations to be sparse. 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. 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.

One may argue that the Column (2) bounds could be more robust than the better estimates because they condition on more covariates in Assumption 2. A closer inspection of Lee's covariates suggests this is unlikely to be the case. Indeed, Lee's covariates contain several equivalent representations of work experience and income. The Column (5) bounds are based on the covariates that capture similar information to the one captured by Lee's covariates, but with fewer covariates. As a result, these bounds are based on more precisely estimated first-stage parameters. To conclude, while the bounds in Columns (4), (5) and (6) are not based on the weakest possible version of Assumption 2, they are based on the same assumption as the Column (2) bounds. Finally, in all bounds in Table 3, the fraction of subjects with positive conditional employment effect is consistent with the Figure 2 estimate.

Table A.7 in Appendix A presents sharp bounds on the average wage effect for all horizons that Lee considered. Across the board, the sparsity-based better Lee bounds (Column (3)) are substantially tighter than the bounds based on Lee's covariates (Columns (1)-(2)), by a factor of 1.7 (week 45) and 18–21.2 (week 104 onwards). In fact, the better Lee bounds are strictly included in the original sharp Lee's estimates (Table 1, Column (2)) where unconditional monotonicity is erroneously imposed. In addition, the better Lee bounds are more precisely estimated.

Let $S_{\text{week}}(1)$ and $S_{\text{week}}(0)$ be the potential employment outcomes in a given week, and $Y_{\text{week}}(1)$ and $Y_{\text{week}}(0)$ be the potential log wage outcomes. Let $S_{\text{week}} = D \cdot S_{\text{week}}(1) + (1-D) \cdot S_{\text{week}}(0)$ be the realized employment and Y_{week} be the realized wage. I focus on a subset of subjects whose treatment effect lower bound, averaged across weeks 80–120, is positive conditional on the full covariate vector X. My subjects of interest are $\mathcal{X}_{\text{PLB}} = \mathcal{X}^{\text{help}}_{\text{PLB}} \cup \mathcal{X}^{\text{hurt}}_{\text{PLB}}$, where

$$\mathcal{X}^{\text{help}}_{\text{PLB}} = \{X \in \mathcal{X}^{\text{help}} : \frac{1}{40} \sum_{\text{week}=80}^{120} \mathbb{E}[Y_{\text{week}} | D=1, S_{\text{week}}=1, Y_{\text{week}} \leq Q_{\text{week}}(p_{week}(X), X), X] \tag{8.1}$$

$$- \mathbb{E}[Y_{\text{week}} | D=0, S_{\text{week}}=1, X] > 0\},$$

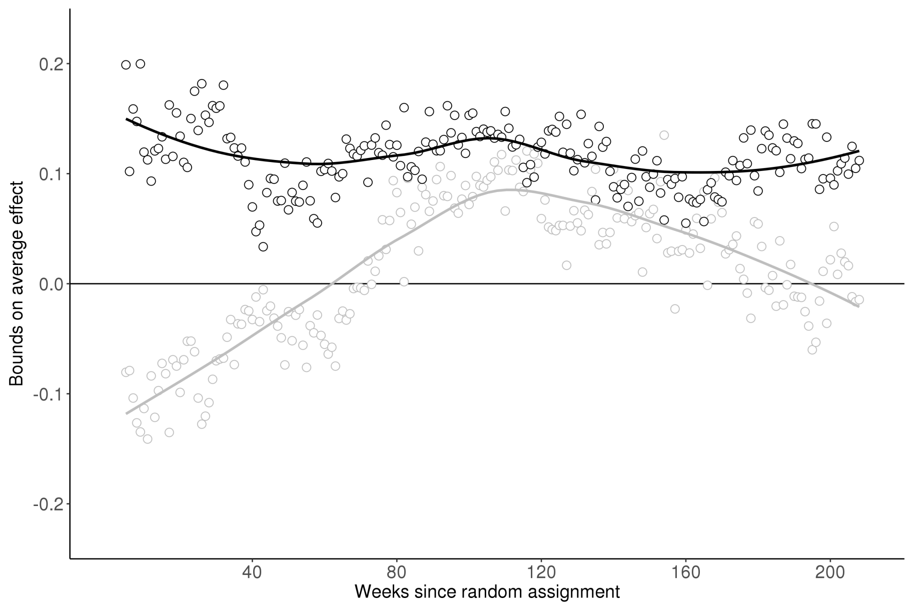
and $\mathcal{X}^{\text{hurt}}_{\text{PLB}}$ is its analog for $\mathcal{X}^{\text{hurt}}$.

Table 3: Estimated bounds on the JobCorps effect on week 90 log wages

	Average Treatment Effect (ATE)						ATE on $\mathcal{X}_{\text{PLB}}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	[-0.027, 0.111]	[-0.005, 0.091]	**[0.040, 0.046]**	[0.041, 0.059]	[0.041, 0.043]	[0.024, 0.065]	**[0.047, 0.061]**
	(-0.058, 0.142)	(-0.054, 0.135)	**(0.001, 0.078)**	(-0.019, 0.112)	(-0.023, 0.101)	(-0.05, 0.131)	**(0.005, 0.100)**
Selection covs	28	28	5 177	15	13	12-13	5 177
Post-lasso-log.	N/A	N/A	9	N/A	N/A	N/A	9
Wage covs	0	28	470	15	13	12-13	470
Post-lasso	N/A	N/A	6	N/A	N/A	N/A	6

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 $\mathcal{X}_{\text{help}} = \{\widehat{p}(X) < 1\}$ and $\mathcal{X}_{\text{hurt}} = \{\widehat{p}(X) > 1\}$, where the trimming threshold $\widehat{p}(x) = \widehat{s}(0,x)/\widehat{s}(1,x)$ is estimated as in equation (3.3). Column (1): basic bounds given 28 Lee's covariates. Column (2): sharp bounds given 28 Lee's covariates (i.e., naive bounds). Column (3): sharp bounds given all covariates assuming few of them affect employment and wage. Column (4): sharp bounds given the union of raw covariates selected for the employment and wage equations in Column (3). Column (5): sharp bounds given the covariates selected on the sample that Lee excluded due to missing data in weeks other than 90. Column (6): variational bounds defined in Section 5 . Column (7): sharp bounds, based on the Column (3) first-stage estimates, for the PLB sample. The PLB sample of $N = 7,735$ is defined as always-takers whose conditional treatment effect lower bound is positive in at least one of six horizons considered by Lee (weeks 45, 90, 104, 135, 180, 208). Covariates are defined in Section C.2. First-stage estimates are given in Table F.12 for Columns (1) and (2), Table F.11 for Columns (3) and (7), Table F.13 for Columns (4), and Table F.14 for Columns (5). Computations use design weights. See Appendix F.3 for details.

Figure 3: Estimated bounds on the JobCorps effect on log wages by week.

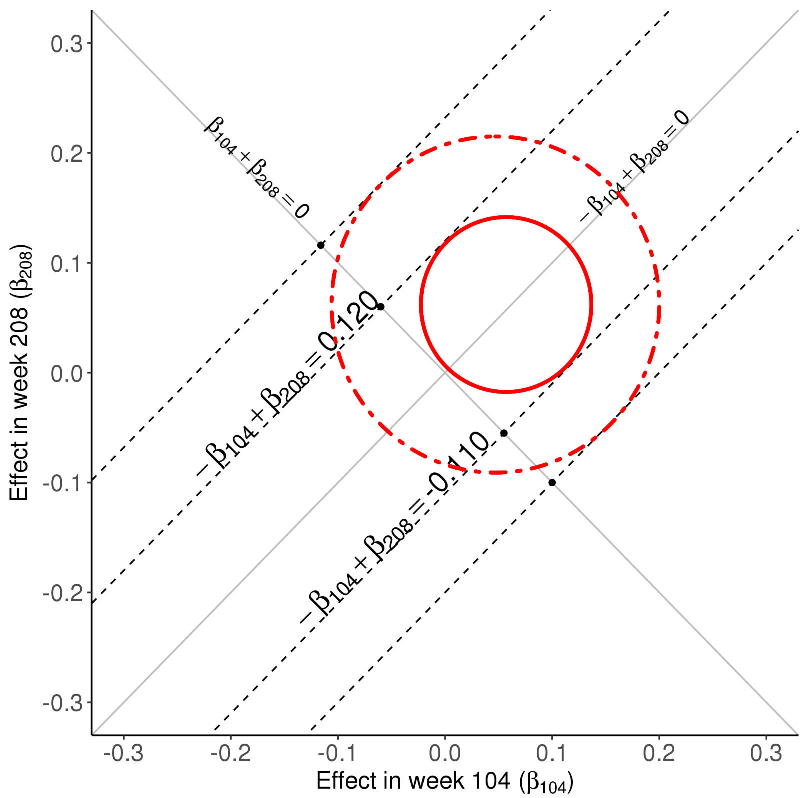


Notes. The horizontal axis shows the number of weeks since random assignment. The black (gray) circles are an estimated upper (lower) bound on the average wage effect. The black (gray) fitted line is estimated by local linear approximation to 201 black (gray) points, respectively. The sample consists of subjects whose average treatment effect lower bound is positive across weeks 80–120 ($N = 4,564$), as defined in equation (8.1). Computations use design weights.

Figure 3 plots the average JobCorps effect on wages. The black and gray lines show the upper and lower bounds on the average wage effect for the always-takers in the $\mathcal{X}_{\text{PLB}}$ subgroup. The lower bound sharply increases from -0.122 at week 5 to 0.072 at week 110 and declines to zero afterwards. The upper bound on the wage effect decreases from 0.2 in week 13 to 0.09 around week 80, after which it fluctuates around 0.10.

JobCorps Effect on Wage Growth Rate. I study the JobCorps effect on long-term wage growth from week 104 to week 208. The 2-year time span between week 104 and week 208 is the longest possible period to consider without encountering short-term effects. Specifically, week 104 is one of the earliest weeks when the lower bound on the wage effect has become positive and stopped growing, according to Figure 3.

Figure 4: Geometric interpretation of the JobCorps effect on wage growth from week 104 to week 208.



Notes. This figure shows the best circular approximation to the estimated identified set (solid red perimeter), the best circular approximation to the 95% pointwise confidence region (dashed red perimeter), the estimated projections of the true sets on the -45-degree line (four black dots), and the intercepts of the corresponding tangent lines. Computations use design weights. See Appendix F.2 for the details.

For the sake of simplicity, I focus on subjects who are always-takers in both 104 and 208 weeks and whose treatment-control difference in the employment rate (3.1) is positive in both weeks. To sum up, my subjects of interest are the always-takers-squared-plus:

$$\mathrm{AT}^{2+} = \{S_{104}(1) = 1, S_{208}(1) = 1, S_{104}(0) = 0, S_{208}(0) = 1, \Delta_{104}(X) > 0, \Delta_{208}(X) > 0\}, \tag{8.2}$$

where "squared" refers to the two time periods under consideration and "plus" refers to the positive sign of the treatment-control difference in the employment rate. For the (AT^{2+}), define the potential wage growth in the treated ($d = 1$) and control ($d = 0$) groups as

$$\rho(d) = \mathbb{E}[Y_{208}(d) - Y_{104}(d) | \mathrm{AT}^{2+}], \quad d \in \{1, 0\}. \tag{8.3}$$

The wage growth rate in the control group is identified as

$$\rho(0) = \mathbb{E}[Y_{208} - Y_{104} | S_{104} = 1, S_{208} = 1, D = 0, \Delta_{104}(X) > 0, \Delta_{208}(X) > 0].$$

The growth rate in the treated group is not identified, but can be bounded using the relation

$$\rho(1) = \rho(0) + \beta_{208} - \beta_{104}, \tag{8.4}$$

where β_{week} stands for the average treatment effect for the AT^{2+} group in a particular week.

A simplistic approach to construct an upper bound on $\beta_{208} - \beta_{104}$ is to subtract the lower bound on β_{104} from the upper bound on β_{208}. Since wages in weeks 104 and 208 are likely to be correlated, this upper bound may not be sustained by any data generating process consistent with observed data. To obtain the sharp bound, project the true identified set on the -45 degree line and take the intercepts of the corresponding tangent lines. In Figure 4, the projection endpoints correspond to the inner black dots, and the tangent lines passing through them have intercepts equal to -0.11 and 0.12, respectively. Adding a sample estimate of $\rho(0) = 0.149$ to the intercepts yields the bounds on $\rho(1)$, $[0.039, 0.260]$.

Figure A.5 reports the upper and lower bounds on the average log wage for the always-takers in the control status, $\mathbb{E}[Y_{\text{week}}(0)|S_{\text{week}}(1) = S_{\text{week}}(0) = 1]$ for each week. The lower (upper) bound grows from 1.45 (1.90) in week 5 to 1.978 (2.036) in week 208. The bounds' width decreases from 0.45 in week 14 to 0.01 in week 208. The gap between the lower and the upper bound shrinks over time as the share of applicants with a positive employment effect, where the average log wage is identified in the control status, increases. One can interpret Figure A.5 as corroborating the Ashenfelter (1978) pattern and showing 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.

8.2 Angrist et al. (2002)

Angrist et al. (2002) studies the effect of winning a voucher from the Colombia PACES program, a voucher initiative established in 1991 to subsidize private school education, on pupils' test scores. In 1999, Angrist et al. (2002) administered a grade-specific test and found the voucher effect on the total test score to be equal to 0.2 standard deviations, a substantial boost equivalent to one additional year of schooling. I examine whether this finding is robust to the endogeneity of test participation. The sample consists of $N = 3,610$ subjects from Bogota's 1995 applicant cohort and has lottery outcome and test scores for Mathematics, Reading, and Writing. In addition, the sample has 25 demographic characteristics, including the applicant's age and gender, their father's and mother's ages, father's and mother's highest grade completed, and a collection of indicators for area of residence. While the number of raw covariates is moderate, the number of their three-way interactions $p = 900$ is quite large for logistic and quantile series methods.

Table 4 reports bounds on the voucher effect on test scores in Mathematics (Panel A), Reading (Panel B), and Writing (Panel C) for various parameters of interest. The sharp bounds based on 25 raw covariates (Columns (4)-(5)) cover zero for all three subjects. Assuming sparsity, the better Lee bounds consider 900 technical covariates and achieve nearly point-identification for all three subjects (Column (6)). In addition, the better Lee bounds are more precisely estimated than the Column (4) ones. The width of 95% confidence region for the true identified set is smaller than its Column (4) analog, by a factor ranging from 0.38 (Reading) to 0.48 (Mathematics). For Mathematics and Writing, the average voucher effect on the always-takers comprises $3/4$ of respective Angrist et al. (2002)'s intent-to-treat estimate that does not account for selection bias.

One may wonder whether it is possible to shorten the 95 %-CR for better Lee bounds by focusing on a non-sharp bound instead of the sharp one. I argue that it is not possible to do so without imposing either unconditional monotonicity or independence. Indeed, the conditional test participation probability must be estimated to account for differential response rates. Since the test participation rate is very low (close to 7%), any estimate of this parameter will introduce substantial noise. Among the three non-parametric specifications in Columns (3)–(6), post-lasso-logistic introduces the smallest amount of noise because it considers more technical covariates but restricts itself to a smaller covariate set.

8.3 Finkelstein et al. (2012)

Finkelstein et al. (2012) studies the effect of access to Medicaid on self-reported health-

Table 4: Estimated bounds on the PACES voucher effect on pupils' test scores

	ITT	Average Treatment Effect (ATE)				
	Exogeneity	Monotonicity	Conditional monotonicity			
	(1)	(2)	(3)	(4)	(5)	(6)
Mathematics	[0.178, 0.178]	[0.075, 0.279]	[-0.432, 0.590]	[-0.274, 0.545]	[-0.169, 0.538]	**[0.056, 0.084]**
	(-0.058, 0.413)	(-0.243, 0.548)	(-0.707, 0.828)	(-0.710, 1.016)	(-0.603, 0.975)	**(-0.298, 0.405)**
Reading	[0.204, 0.204]	[0.029, 0.261]	[-0.386, 0.783]	[-0.333, 0.654]	[-0.252, 0.526]	**[0.163, 0.177]**
	(-0.021, 0.429)	(-0.256, 0.538)	(-0.643, 1.031)	(-0.849, 1.143)	(-0.735, 0.924)	**(-0.140, 0.460)**
Writing	[0.126, 0.126]	[-0.057, 0.222]	[-0.552, 0.433]	[-0.396, 0.380]	[-0.150, 0.474]	**[0.086, 0.094]**
	(-0.101, 0.353)	(-0.346, 0.486)	(-0.808, 0.679)	(-0.855, 0.825)	(-0.586, 0.888)	**(-0.249, 0.396)**
N	282	3610	3610	3610	3610	3610
Test-tak. covs	N/A	N/A	25	25	150	900
Post-lasso-logistic	N/A	N/A	N/A	N/A	N/A	10
Test score covs	N/A	N/A	0	25	25	25
Post-lasso	N/A	N/A	N/A	N/A	N/A	9

Estimated bounds are in square brackets and the 95% confidence region for the identified set is in parentheses. Any test participant (a pupil who arrives at a testing location) is tested in all three subjects. Column (1): ITT estimate from Angrist et al. (2002), Table 5, Column 1. Column (2): basic Lee bounds under unconditional monotonicity. In Columns (3)–(6), all subjects are partitioned into the sets $\mathcal{X}_{\text{help}} = \{\widehat{p}(X) < 1\}$ and $\mathcal{X}_{\text{hurt}} = \{\widehat{p}(X) > 1\}$, where the trimming threshold $\widehat{p}(x) = \widehat{s}(0,x)/\widehat{s}(1,x)$ is estimated as in equation (3.3). Column (3): basic Lee bounds based on 25 raw covariates. Column (4): sharp Lee bounds based on 25 raw covariates. Column (5) differs from Column (4) only by adding second-order interactions of 6 continuous covariates into the test-taking equation. Column (6): sharp Lee bounds, based on the technical covariates selected by post-lasso-logistic for the test participation and on the raw covariates selected by post-lasso for the test score. Baseline covariates are described in Table F.15. First-stage estimates are given in Table F.16. Selected covariates are described in Figure F.7. See Appendix F.4 for details.

care 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 non-parametric 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) studies the effect of winning the Medicaid lottery using the intent-to-treat (ITT) framework. 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

$$Y_{ih} = \beta_0 + \beta_1 \text{Lottery}_h + \bar{X}_{ih}\beta_2 + \varepsilon_{ih}, \tag{8.5}$$

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,

$$Y_{ih} = \pi_0 + \pi_1 \text{Insurance}_{ih} + \bar{X}_{ih}\pi_2 + \nu_{ih}, \tag{8.6}$$

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 (8.5). Finkelstein et al. (2012) estimates (8.6) 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 5, 6, A.8, A.9, Columns (1) and (4)).

I examine whether the Intent-to-Treat (8.5) and Local Average Treatment Effect (8.6) equations are robust to non-response bias. Tables 5 and 6 show the results for self-reported health outcomes. The standard trimming 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 of residence does not improve standard estimates.

A smoothness assumption on the conditional response probability and the outcome quantile drastically changes the result. Tables 5 and 6, 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, which is 75%

of the baseline LATE estimate (1.317 (std. error 0.562)). Overall, Finkelstein et al. (2012)'s baseline results are robust to non-response bias.

9 Conclusion

Lee bounds are a popular empirical strategy for addressing post-randomization selection bias. In this paper, I show that Lee bounds can be improved by incorporating baseline covariates and modern regularized machine learning techniques. First, better Lee bounds accommodate differential selection response. This relaxation is especially important for JobCorps, since the JobCorps effect on employment is unlikely to be in the same direction for everyone. Second, better Lee bounds are sharp if few of very many covariates under consideration affect selection and outcome. In practice, better Lee bounds achieve nearly point-identification for all three empirical examples under interpretable assumptions (i.e., smoothness and sparsity) that are often invoked in point-identified problems. Therefore, better Lee bounds are expected to deliver stronger conclusions that are also more robust to violations of monotonicity.

Table 5: 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	-0.013	**0.032**	0.133	-0.067	**0.077**
	(0.008)	(0.013)	**(0.017)**	(0.026)	(0.044)	**(0.058)**
Health fair/good/very good/excellent	0.029	-0.052	**0.019**	0.099	-0.195	**0.011**
	(0.005)	(0.012)	**(0.010)**	(0.018)	(0.038)	**(0.033)**
Health same or gotten better	0.033	-0.033	**0.015**	0.113	-0.138	**0.051**
	(0.007)	(0.014)	**(0.019)**	(0.023)	(0.049)	**(0.065)**
Did not screen positive for depression	0.023	-0.045	**0.002**	0.078	-0.183	**0.007**
	(0.007)	(0.014)	**(0.010)**	(0.025)	(0.049)	**(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 (8.5) by OLS. Columns (4)–(6) report the coefficient and standard error on Insurance from estimating equation (8.6) 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 F.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 (F.4) and the zero outcome probability $\phi_0(x)$ is defined in (F.2). 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 F.17. The first-stage estimates for Columns (3) and (6) are given in Table F.18. See Appendix F.5 for details.

Table 6: 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	-1.096	**0.272**	1.317	-4.411	**0.981**
	(0.162)	(0.349)	**(0.166)**	(0.562)	(1.166)	**(0.577)**
# of days phys. health good, past 30 days	0.459	-1.230	**0.272**	1.585	-4.929	**0.627**
	(0.174)	(0.384)	**(0.170)**	(0.605)	(1.308)	**(0.592)**
# of days mental health good, past 30 days	0.603	-0.862	**0.220**	2.082	-3.573	**0.750**
	(0.184)	(0.374)	**(0.179)**	(0.640)	(1.298)	**(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 (8.5) by OLS. Columns (4)–(6) report the coefficient and standard error on Insurance from estimating equation (8.6) 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 F.5. 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 (F.4) and the conditional quantile is defined in equation (2.8). 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 F.17. See Appendix F.5 for details.

References

Andrews, D. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, (62):43–72.

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. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

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.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, 60:47–50.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

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 Hansen, C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *Journal of Economic Perspectives*, 28(2):608–650.

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.

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. (2013a). Inference on counterfactual distributions. *Biometrics*, 81(6):2205–2268.

Chernozhukov, V., Lee, S., and Rosen, A. (2013b). Intersection bounds: Estimation and inference. *Econometrica*, 81:667–737.

Chernozhukov, V., Newey, W., and Singh, R. (2018). De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers. *arXiv e-prints*, page arXiv:1802.08667.

Chiang, H. D. (2020). Many average partial effects: With an application to text regression.

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.

Hirano, K., Imbens, G., and Reeder, G. (2003). Efficient estimation of average treatment effects under the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Ichimura, H. and Newey, W. K. (2015). The Influence Function of Semiparametric Estimators. *arXiv e-prints*, page arXiv:1508.01378.

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.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 2(4):2869–2909.

Lee, D. (2005). Trimming for bounds on treatment effects with missing outcomes. *Working Paper*.

Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102.

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*, 62:245–271.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics*, 213(57):416–444.

Rockafellar, R. T. (1997). *Convex Analysis*. Princeton University Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

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.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4):1040–1053.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315.

van der Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202.

van der Vaart, A. (1998). Asymptotic statistics.

Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

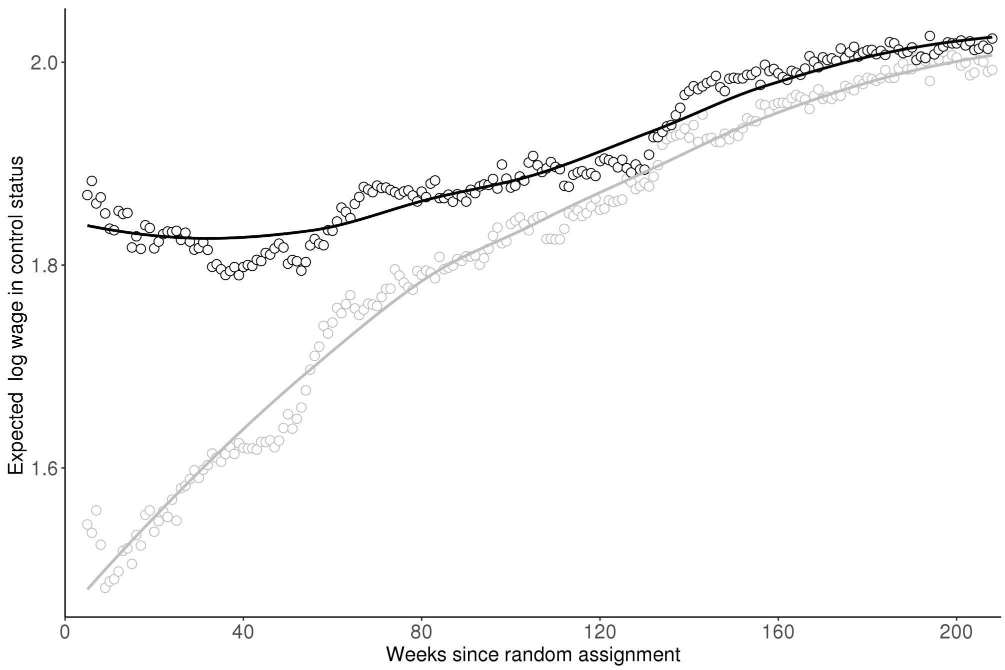
Appendix A: Supplementary Figures and Tables

Table A.7: Estimated bounds on the JobCorps effect on log wages

	Average Treatment Effect (ATE)			ATE on $\mathcal{X}_{\text{PLB}}$
	(1)	(2)	(3)	(4)
Week 45	[-0.085, 0.141]	[-0.081, 0.136]	[-0.048, 0.087]	[-0.043, 0.081]
	(-0.11, 0.172)	(-0.117, 0.172)	(-0.081, 0.121)	(0.000, 0.118)
Week 104	[-0.027, 0.103]	[-0.022, 0.098]	[0.027, 0.032]	[0.044, 0.055]
	(-0.058, 0.133)	(-0.074, 0.148)	(-0.021, 0.078)	(0.000, 0.106)
Week 135	[-0.025, 0.102]	[-0.025, 0.098]	[0.030, 0.037]	[0.059, 0.066]
	(-0.066, 0.136)	(-0.087, 0.149)	(-0.018, 0.076)	(0.005, 0.108)
Week 180	[-0.047, 0.109]	[-0.042, 0.101]	[0.033, 0.063]	[0.047, 0.081]
	(-0.076, 0.133)	(-0.096, 0.149)	(-0.014, 0.103)	(0.000, 0.125)
Week 208	[-0.019, 0.094]	[0.000, 0.091]	[0.030, 0.065]	[0.032, 0.093]
	(-0.051, 0.119)	(-0.054, 0.142)	(-0.016, 0.106)	(0.000, 0.136)
Employment covs	28	28	5 177	5 177
Post-lasso-logistic	N/A	N/A	Varies (10-20)	Varies (10-20)
Wage covs	N/A	28	470	470
Post-lasso covs	N/A	N/A	Varies (6-10)	Varies (6-10)

Notes. Estimated bounds are in square brackets and the 95% confidence region for identified set is in parentheses. All subjects are partitioned into the sets $\mathcal{X}_{\text{help}} = \{\widehat{p}(X) < 1\}$ and $\mathcal{X}_{\text{hurt}} = \{\widehat{p}(X) > 1\}$, where the trimming threshold $\widehat{p}(x) = \widehat{s}(0,x)/\widehat{s}(1,x)$ is estimated as in equation (3.3). Column (1): basic bounds based on Lee's covariates. Column (2): sharp bounds based on Lee's covariates. Columns (3) and (4): sharp bounds given all covariates assuming few of them affect employment and wage, where the full sample $N = 9,145$ is used in Column (3) and the PLB sample $N = 7,735$ is used in Column (4). The PLB sample is defined as the always-takers whose conditional treatment effect lower bound is positive in at least one of six horizons considered by Lee (weeks 45,90,104,135,180,208). Covariates are defined in Section C.2. Computations use design weights. See Appendix F.3 for details.

Figure A.5: Estimated bounds on the average wage in the control status by week.



Notes. The horizontal axis shows the number of weeks since random assignment. The black (gray) circles show the upper (lower) bound on the average untreated log wage for the always-takers. The black (gray) fitted line is estimated by local linear approximation to 201 black (gray) points, respectively. Computations use design weights.

Table A.8: 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	0.017	0.088	-0.036	0.060
	(0.008)	(0.014)	(0.017)	(0.029)	(0.046)	(0.060)
Outpatient visits last six months	0.062	0.005	0.042	0.212	0.001	0.146
	(0.007)	(0.013)	(0.017)	(0.025)	(0.045)	(0.058)
ER visits last six months	0.006	-0.020	-0.004	0.022	-0.076	-0.015
	(0.007)	(0.008)	(0.011)	(0.023)	(0.030)	(0.037)
Hospital admissions last six months	0.002	-0.005	0.002	0.008	-0.020	0.007
	(0.004)	(0.004)	(0.005)	(0.014)	(0.016)	(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 (8.5) by OLS. Columns (4)–(6) report the coefficient and standard error on Insurance from estimating equation (8.6) 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 F.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 (F.4) and the zero outcome probability $\phi_0(x)$ is defined in (F.2). 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 F.17. See Appendix F.5 for details.

Table A.9: 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	-0.024	0.077	0.347	-0.124	0.270
	(0.051)	(0.066)	(0.052)	(0.175)	(0.225)	(0.179)
Outpatient visits last six months	0.314	0.121	0.246	1.083	0.372	0.853
	(0.054)	(0.065)	(0.054)	(0.182)	(0.228)	(0.183)
ER visits last six months	0.007	-0.040	-0.008	0.026	-0.152	-0.027
	(0.016)	(0.019)	(0.016)	(0.056)	(0.065)	(0.056)
Hospital admissions last six months	0.006	-0.004	0.003	0.021	-0.014	0.010
	(0.006)	(0.007)	(0.006)	(0.021)	(0.024)	(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 (8.5) by OLS. Columns (4)–(6) report the coefficient and standard error on Insurance from estimating equation (8.6) 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 F.5. 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 (F.4) and the conditional quantile is defined in equation (2.8). 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 F.17. See Appendix F.5 for details.

Appendix B: Supplementary Statements for Sections 4-6[1]

B.1 Definitions

Sharp Lee Bounds: Definition. In this section, I derive the target parameter — sharp Lee bounds — under Assumption 2. The conditional trimming threshold is

$$p_0(x) = \frac{s(0,x)}{s(1,x)} = \frac{\mathbb{E}[S=1|D=0,X=x]}{\mathbb{E}[S=1|D=1,X=x]}.$$

The covariate groups $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$ are

$$\mathcal{X}_{\text{help}} = \{X: \quad p_0(X) < 1\}, \quad \mathcal{X}_{\text{hurt}} = \{X: \quad p_0(X) > 1\}. \tag{B.1}$$

By Assumption 3, $\Pr(X \in \mathcal{X}_{\text{help}} \cup \mathcal{X}_{\text{hurt}}) = 1$. The conditional probability of treatment (i.e., the propensity score) is

$$\mu_1(X) = \Pr(D=1|X), \quad \mu_0(X) = 1 - \mu_1(X) = \Pr(D=0|X). \tag{B.2}$$

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

$$Q_d(u,x): \quad \Pr(Y \leq Q_d(u,x)|S=1, D=d, X=x) = u, \quad u \in [0,1], \quad d \in \{1,0\}. \tag{B.3}$$

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:

$$Q(u,x) = 1_{x \in \mathcal{X}_{\text{help}}} Q_1(u,x) + 1_{x \in \mathcal{X}_{\text{hurt}}} Q_0(u,x). \tag{B.4}$$

[1] This appendix is for online publication.

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

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

and combined conditional density is

$$f(t|x) = 1_{x \in \mathcal{X}_{\text{help}}} f_1(t|x) + 1_{x \in \mathcal{X}_{\text{hurt}}} f_0(t|x). \tag{B.5}$$

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

$$\bar{\beta}_U^{\text{help}}(x) = \mathbb{E}[Y|D=1, S=1, Y \geq Q(1-p_0(x), x), X=x] - \mathbb{E}[Y|D=0, S=1, X=x] \tag{B.6}$$

and the conditional lower bound is

$$\bar{\beta}_L^{\text{help}}(x) = \mathbb{E}[Y|D=1, S=1, Y \leq Q(p_0(x), x), X=x] - \mathbb{E}[Y|D=0, S=1, X=x]. \tag{B.7}$$

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

$$\bar{\beta}_U^{\text{hurt}}(x) = \mathbb{E}[Y|D=1, S=1, X=x] - \mathbb{E}[Y|D=0, S=1, Y \leq Q(1/p_0(x), x), X=x] \tag{B.8}$$

and the conditional lower bound is

$$\bar{\beta}_L^{\text{hurt}}(x) = \mathbb{E}[Y|D=1, S=1, X=x] - \mathbb{E}[Y|D=0, S=1, Y \geq Q(1-1/p_0(x), x), X=x]. \tag{B.9}$$

Define the treated and control components of each conditional bound as

$$\bar{\beta}_*^\star(x) = \bar{\beta}_{1*}^\star(x) - \bar{\beta}_{0*}^\star(x), \quad * \in \{L,U\}, \quad \star \in \{\text{help}, \text{hurt}\} \tag{B.10}$$

and the normalizing constants as

$$\mu_{10}^{\text{help}} = \mathbb{E}[s(0,X)|X \in \mathcal{X}_{\text{help}}], \quad \mu_{11}^{\text{hurt}} = \mathbb{E}[s(1,X)|X \in \mathcal{X}_{\text{hurt}}]. \tag{B.11}$$

The sharp Lee bounds β_L and β_U are:

$$\begin{aligned} \beta_* = &(\mu_{10}^{\text{help}})^{-1} \Pr(X \in \mathcal{X}_{\text{help}}) \cdot \mathbb{E}[\bar{\beta}_*^{\text{help}}(X) s(0,X)|X \in \mathcal{X}_{\text{help}}] \\ &+ (\mu_{11}^{\text{hurt}})^{-1} \Pr(X \in \mathcal{X}_{\text{hurt}}) \cdot \mathbb{E}[\bar{\beta}_*^{\text{hurt}}(X) s(1,X)|X \in \mathcal{X}_{\text{hurt}}] \quad * \in \{\text{L}, \text{U}\}. \end{aligned} \tag{B.12}$$

Sharp Lee Bounds: Moment Equation If the propensity score $\mu_1(x)$ in (B.2) is known, the first-stage nuisance parameter ξ_0 is

$$\xi_0 = \{s(0,x), s(1,x), Q(u,x)\}. \tag{B.13}$$

Otherwise, ξ_0 is

$$\xi_0 = \{s(0,x), s(1,x), Q(u,x), \mu_1(x)\} \tag{B.14}$$

for a non-orthogonal moment equation $m_*(W, \xi_0), \quad * \in \{L,U\}$ and

$$\xi_0 = \{s(0,x), s(1,x), Q(u,x), \mu_1(x), \beta_{1*}^{\text{help}}(x), \beta_{0*}^{\text{help}}(x), \beta_{1*}^{\text{hurt}}(x), \beta_{0*}^{\text{hurt}}(x)\}, \tag{B.15}$$

for an orthogonal moment equation $g_*(W, \xi_0), \quad * \in \{L,U\}$, described below. Let $\mu_1(X)$ and $\mu_0(X)$ be as in (B.2), μ_{10}^{help} and μ_{11}^{hurt} as in (B.11). The original (i.e., non-orthogonal) moment equation for β_U is

$$m_U(W,\xi_0) = (\mu_{10}^{\text{help}})^{-1} 1_{X\in\mathcal{X}_{\text{help}}} \left(\frac{D}{\mu_1(X)} \cdot S \cdot Y 1_{\{Y \geq Q(1-p_0(X),X)\}} - \frac{(1-D)}{\mu_0(X)} \cdot S \cdot Y \right) \tag{B.16}$$
$$+ (\mu_{11}^{\text{hurt}})^{-1} 1_{X\in\mathcal{X}_{\text{hurt}}} \left(\frac{D}{\mu_1(X)} \cdot S \cdot Y - \frac{(1-D)}{\mu_0(X)} \cdot S \cdot Y 1_{\{Y \leq Q(1/p_0(X),X)\}} \right),$$

and for β_L is

$$m_L(W,\xi_0) = (\mu_{10}^{\text{help}})^{-1} 1_{X\in\mathcal{X}_{\text{help}}} \left(\frac{D}{\mu_1(X)} \cdot S \cdot Y 1_{\{Y \leq Q(p_0(X),X)\}} - \frac{(1-D)}{\mu_0(X)} \cdot S \cdot Y \right) \tag{B.17}$$
$$+ (\mu_{11}^{\text{hurt}})^{-1} 1_{X\in\mathcal{X}_{\text{hurt}}} \left(\frac{D}{\mu_1(X)} \cdot S \cdot Y - \frac{(1-D)}{\mu_0(X)} \cdot S \cdot Y 1_{\{Y \geq Q(1-1/p_0(X),X)\}} \right).$$

Sharp Lee Bounds: Orthogonal Moment Equation. An orthogonal moment function $g_\star(W,\xi_0)$ is

$$g_*(W,\xi_0) = m_*(W,\xi_0) + (\mu_{10}^{\text{help}})^{-1} 1_{X\in\mathcal{X}_{\text{help}}} \alpha_*^{\text{help}}(W,\xi_0) \tag{B.18}$$
$$+ (\mu_{11}^{\text{hurt}})^{-1} 1_{X\in\mathcal{X}_{\text{hurt}}} \alpha_*^{\text{hurt}}(W,\xi_0), \quad * \in \{L,U\}.$$

The bias correction terms $\alpha_U^{\text{help}}(W;\xi)$ and $\alpha_U^{\text{hurt}}(W;\xi)$ are

$$\alpha_U^{\text{help}}(W;\xi_0) = Q(1-p_0(X),X)\left(\frac{(1-D)\cdot S}{\mu_0(X)} - s(0,X)\right) \tag{B.19}$$
$$- Q(1-p_0(X),X)p_0(X)\left(\frac{D\cdot S}{\mu_1(X)} - s(1,X)\right)$$
$$+ Q(1-p_0(X),X)s(1,X)\left(\frac{D\cdot S\cdot 1_{\{Y\leq Q(1-p_0(X),X)\}}}{s(1,X)\mu_1(X)} - 1 + p_0(X)\right)$$
$$- \left(\frac{1}{\mu_1^2(X)}\beta_{1U}^{\text{help}}(X) + \frac{1}{(1-\mu_1(X))^2}\beta_{0U}^{\text{help}}(X)\right)\cdot s(0,X)\cdot(D-\mu_1(X))$$

$$
\begin{aligned}
\alpha_U^{\text{hurt}}(W;\xi_0) = {} & Q(1/p_0(X),X)p_0(X)\left(\frac{(1-D)\cdot S}{\mu_0(X)} - s(0,X)\right) \\
& + Q(1/p_0(X),X)\left(\frac{D\cdot S}{\mu_1(X)} - s(1,X)\right) \\
& + Q(1/p_0(X),X)s(0,X)\left(\frac{(1-D)\cdot S\cdot 1_{\{Y\leq Q(1/p_0(X),X)\}}}{s(0,X)\mu_0(X)} - p_0(X)\right) \\
& - \left(\frac{1}{\mu_1^2(X)}\beta_{1U}^{\text{hurt}}(X) + \frac{1}{(1-\mu_1(X))^2}\beta_{0U}^{\text{hurt}}(X)\right)\cdot s(1,X)\cdot(D-\mu_1(X))
\end{aligned}
\tag{B.20}
$$

$$
\begin{aligned}
\alpha_L^{\text{help}}(W;\xi_0) = {} & Q(p_0(X),X)\left(\frac{(1-D)\cdot S}{\mu_0(X)} - s(0,X)\right) \\
& - Q(p_0(X),X)p_0(X)\left(\frac{D\cdot S}{\mu_1(X)} - s(1,X)\right) \\
& - Q(p_0(X),X)s(1,X)\left(\frac{D\cdot S\cdot 1_{\{Y\leq Q(p_0(X),X)\}}}{s(1,X)\mu_1(X)} - p_0(X)\right) \\
& - \left(\frac{1}{\mu_1^2(X)}\beta_{1L}^{\text{help}}(X) + \frac{1}{(1-\mu_1(X))^2}\beta_{0L}^{\text{help}}(X)\right)\cdot s(0,X)\cdot(D-\mu_1(X)),
\end{aligned}
\tag{B.21}
$$

$$
\begin{aligned}
\alpha_L^{\text{hurt}}(W;\xi_0) = {} & -Q(1-1/p_0(X),X)p_0(X)\left(\frac{(1-D)\cdot S}{\mu_0(X)} - s(0,X)\right) \\
& + Q(1-1/p_0(X),X)\left(\frac{D\cdot S}{\mu_1(X)} - s(1,X)\right) \\
& - Q(1-1/p_0(X),X)s(0,X)\left(\frac{(1-D)\cdot S\cdot 1_{\{Y\leq Q(1-1/p_0(X),X)\}}}{s(0,X)\mu_0(X)} - 1 + p_0(X)\right) \\
& - \left(\frac{1}{\mu_1^2(X)}\beta_{1L}^{\text{hurt}}(X) + \frac{1}{(1-\mu_1(X))^2}\beta_{0L}^{\text{hurt}}(X)\right)\cdot s(1,X)\cdot(D-\mu_1(X)),
\end{aligned}
\tag{B.22}
$$

where $\beta^{\star}_{1*}(x)$ and $\beta^{\star}_{0*}(x)$ are as in (B.10). When Assumption 1 holds and the propensity score is known, the bias correction terms simplify to

$$\begin{aligned}\text{correction}_U(W,\xi_0) = \frac{1}{\Pr(S=1|D=0)}\Bigg[&Q(1-p_0(X),X)\left(\frac{(1-D)\cdot S}{\Pr(D=0)} - s(0,X)\right)\\ &- Q(1-p_0(X),X)p_0(X)\left(\frac{D\cdot S}{\Pr(D=1)} - s(1,X)\right)\\ &+ Q(1-p_0(X),X)s(1,X)\left(\frac{D\cdot S\cdot 1_{\{Y\leq Q(1-p_0(X),X)\}}}{s(1,X)\Pr(D=1)} - 1 + p_0(X)\right)\Bigg],\end{aligned} \tag{B.23}$$

and for the lower bound is

$$\begin{aligned}\text{correction}_L(W,\xi_0) = \frac{1}{\Pr(S=1|D=0)}\Bigg[&Q(p_0(X),X)\left(\frac{(1-D)\cdot S}{\Pr(D=0)} - s(0,X)\right)\\ &- Q(p_0(X),X)p_0(X)\left(\frac{D\cdot S}{\Pr(D=1)} - s(1,X)\right)\\ &- Q(p_0(X),X)s(1,X)\left(\frac{D\cdot S\cdot 1_{\{Y\leq Q(p_0(X),X)\}}}{s(1,X)\Pr(D=1)} - p_0(X)\right)\Bigg].\end{aligned} \tag{B.24}$$

Lemma B.1 (Identification of better Lee bounds)**.** *Under Assumption 2, the following statements hold.*

(a) The bounds β_L and β_U defined in (B.12) *are sharp valid bounds on β_0 in equation* (2.1)*. The moment functions* (B.16)-(B.17) *and* (B.18) *obey*

$$\mathbb{E}g_U(W,\xi_0) = \mathbb{E}m_U(W,\xi_0) = \beta_U, \quad \mathbb{E}g_L(W,\xi_0) = \mathbb{E}m_L(W,\xi_0) = \beta_L.$$

(b) If the conditional densities $f_d(y|x), d\in\{1,0\}$ in (B.5) *have a convex and compact support almost surely in $\mathcal{X}$ and a first derivative that is bounded away from zero and infinity, the interval $[\beta_L,\beta_U]$ is a sharp identified set for β_0.*

(c) If a stronger version of independence assumption (Assumption 2 (1)) holds:

$$D \perp (Y(1), Y(0), S(1), S(0), X) \mid \bar{X},$$

(a) and (b) remain true.

Proposition B.2 (Estimation of Sets $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$)**.** *Suppose Assumption 3 holds. Furthermore, suppose $\widehat{s}(d,x)$ converges uniformly over $\mathcal{X}$:*

$$\sup_{x \in \mathcal{X}} \sup_{d \in \{0,1\}} |\widehat{s}(d,x) - s(d,x)| = o_P(1).$$

Then, for N sufficiently large, $\Pr(\widehat{p}(X_i) < 1 \Leftrightarrow X_i \in \mathcal{X}_{help} \quad \forall i, 1 \leq i \leq N) \to 1$ *as* $N \to \infty$.

Proof of Proposition B.2. **Step 1.** Consider an open set $(\underline{s}/2, 1 - \bar{s}/2) \times (\underline{s}/2, 1 - \bar{s}/2)$. W.p. $1 - o(1)$, the pair of estimated functions $(\widehat{s}(0,\cdot), \widehat{s}(1,\cdot))$ belongs to this set. The function $f(t_1,t_2) = t_1/t_2$ has bounded partial derivatives in any direction on this set. Therefore, $\sup_{x \in \mathcal{X}} |\widehat{p}(x) - p_0(x)| = o_P(1)$ holds.

Step 2. The following statement hold:

$$\begin{aligned}
&\Pr\left(\widehat{p}(X_i) < 1 < p_0(X_i) \text{ or } p_0(X_i) < 1 < \widehat{p}(X_i) \quad \text{for some } i\right) \\
&\leq \Pr\left(X_i \in \mathcal{X} : 0 < |1 - p_0(X_i)| < |\widehat{p}(X_i) - p_0(X_i)| \quad \text{for some } i\right) \\
&\leq \Pr\left(\sup_{x \in \mathcal{X}} |\widehat{p}(x) - p_0(x)| > \frac{\varepsilon}{2\underline{s}}\right) \to 0.
\end{aligned}$$

□

B.2 Supplementary Statements for Section 4

Let $\mathcal{U} \in (0,1)$ be an open set that contains the support of $p_0(X)$ and $1 - p_0(X)$. For the sake of exposition, suppose $\mathcal{X}_{\text{hurt}} = \emptyset$ and $Q(u,x) = Q_1(u,x)$. Suppose that $Q_1(u,x)$

is a sufficiently smooth function of x relative to its dimension for each $u \in \mathcal{U}$, and the smoothness index is the same for all $u \in \mathcal{U}$. Then, $Q(u,x)$ can be approximated by a linear function

$$Q(u,x) = Z(x)'\delta_0(u) + R(u,x), \tag{B.25}$$

where $Z(x) \in \mathrm{R}^{p_Q}$ 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

$$\begin{aligned}
\widehat{\delta}(u) &:= \arg\min_{\delta \in \mathrm{R}^{p_Q}} \frac{1}{N_{11}} \sum_{D_i=1,S_i=1}^{N} \rho_u(Y_i - Z(X_i)'\delta) \\
&= \arg\min_{\delta \in \mathrm{R}^{p_Q}} \frac{1}{N_{11}} \sum_{D_i=1,S_i=1}^{N} (u - 1_{\{Y_i - Z(X_i)'\delta < 0\}}) \cdot (Y_i - Z(X_i)'\delta),
\end{aligned} \tag{B.26}$$

converges at rate $q_N = \sqrt{\dfrac{p_Q}{N}} = o(N^{-1/4})$, as shown in Belloni et al. (2019).

Remark B.1 (ℓ_1-regularized quantile regression of Belloni and Chernozhukov (2013)). For each $u \in \mathcal{U}$, suppose there exists a vector $\delta_0(u) \in \mathrm{R}^{p_Q}$ with only s_Q out of p_Q non-zero coordinates,

$$\sup_{u \in \mathcal{U}} \|\delta_0(u)\|_0 = \sup_{u \in \mathcal{U}} \sum_{j=1}^{p_Q} 1_{\{\delta_{0,j}(u)\}} = s_Q \ll N \tag{B.27}$$

so that the approximation error $R(u,x)$ is sufficiently small relative to the sampling error

$\sqrt{\frac{s_Q \log p_Q}{N}}$:

$$\sup_{u \in \mathcal{U}} \left(\frac{1}{N_{11}} \sum_{D_i=1,S_i=1}^{N} R^2(u, X_i) \right)^{1/2} \lesssim_P \sqrt{\frac{s_Q \log p_Q}{N}}$$

Furthermore, suppose $\delta_0(u)$ is a Lipshitz function of u. The ℓ_1-regularized quantile regression estimator of Belloni and Chernozhukov (2013) minimizes the ℓ_1-regularized check function

$$\widehat{\delta}_{\text{Lasso}}(u) = \arg \min_{\delta \in \mathrm{R}^{p_Q}} \frac{1}{N_{11}} \sum_{D_i=1,S_i=1}^{N} \rho_u(Y_i - Z(X_i)'\delta) + \frac{\lambda \sqrt{u(1-u)}}{N} \sum_{j=1}^{p_Q} \widehat{\rho}_j |\delta_j|, \quad \text{(B.28)}$$

where $\lambda \geq 0$ is a penalty parameter and $\widehat{\rho}_j = \left(\frac{1}{N_{11}} \sum_{D_i=1,S_i=1}^{N} Z_j(X_i)^2 \right)^{1/2}$. If the model is sufficiently sparse, Assumption 5 is satisfied with $q_N = \sqrt{\frac{s_Q^2 \log p_Q}{N}} = o(N^{-1/4})$ under the choice of λ proposed in Belloni and Chernozhukov (2013).

B.3 Supplementary Statements for Section 5

In this section, I provide formal analysis of the agnostic approach summarized in Section 5.

Algorithm 2 Agnostic Bounds

1: Select covariates $X = X_A$ based on the auxiliary sample A.
2: Let $\beta_L = \beta_L^A$ and $\beta_U = \beta_U^A$ be the sharp bounds in the model $(D, X_A, S, S \cdot Y)$. Report $(\widehat{\beta}_L^A, \widehat{\beta}_U^A)$ based on either (1) conventional Lee bounds, (2) non-orthogonal moment function (B.16), or (3) orthogonal moment function (B.18), and logistic and quantile series estimators of the first-stage functions $s_A(d,x)$ and $Q_A(u,x)$.

The key advantage of the agnostic approach is that one is not required to use orthogonal moment equations (B.18) for the bounds. If X_A consists of a few discrete covariates, one can use conventional Lee (2009) bounds without any smoothness assumptions. If smoothness is economically plausible, one can use the original (B.17) moment equation for the bounds and estimate $s_A(d,x)$ and $Q_A(u,x)$ by classic nonparametric estimators with under-smoothing. Example 4 in Chernozhukov et al. (2013b) gives explicit conditions on the function $Q_A(u,x)$ so that a quantile series estimator's bias due to approximation error is asymptotically negligible. Likewise, Hirano et al. (2003) gives explicit conditions on the function $s_A(u,x)$ so that a logistic series estimator's bias due to approximation error is asymptotically negligible.

Conditional Inference. Conditional on the auxiliary sample Data_A, the estimator $(\widehat{\beta}_L^A, \widehat{\beta}_U^A)$ is asymptotically normal,

$$\sqrt{|M|}\begin{pmatrix}\widehat{\beta}_L^A - \beta_L^A \\ \widehat{\beta}_U^A - \beta_U^A\end{pmatrix} \Rightarrow N(0, \Omega_A).$$

A $(1-\alpha)$ conditional CR takes the form

$$[L_A, U_A] = [\widehat{\beta}_L^A - |M|^{-1/2}\widehat{\Omega}_{A,LL}^{1/2} c_{\alpha/2}, \quad \widehat{\beta}_U^A + |M|^{-1/2}\widehat{\Omega}_{A,UU}^{1/2} c_{1-\alpha/2}],$$

where choosing c_α as the α-quantile of $N(0,1)$ delivers a confidence region for the identified set $[\beta_L^A, \beta_U^A]$.

Variational Inference. Different splits (A,M) of the sample $\{1,2,\ldots,N\}$ yield different target bounds (β_L^A, β_U^A) and different approximate distributions of these bounds. If we take the splitting uncertainty into account, the pair of bounds (β_L^A, β_U^A) are random conditional on the full data sample. In practice, one may want to generate several random splits and aggregate various bounds over various partitions. Suppose the following regularity condition holds.

ASSUMPTION 6 (Regularity condition)**.** *Suppose that $\mathcal{A}$ is a set of regular data con-*

figurations such that for all $x \in [0,1]$*, under the null hypothesis*

$$\sup_{P \in \mathcal{P}} |\Pr_P[p_A \leq x] - x| \leq \delta = o(1),$$

and $\inf_{P \in \mathcal{P}} P_P[Data_A \in \mathcal{A}] \leq 1 - \gamma = 1 - o(1)$*. In particular, suppose that this holds for the p-values*

$$p_A := \Phi(\widehat{\Omega}_{A,LL}^{-1/2}(\widehat{\beta}_L^A - \beta_L^A)), \quad p_A := \Phi(\widehat{\Omega}_{A,UU}^{-1/2}(\widehat{\beta}_U^A - \beta_U^A)),$$
$$p_A := 1 - \Phi(\widehat{\Omega}_{A,LL}^{-1/2}(\widehat{\beta}_L^A - \beta_L^A)), \quad p_A := 1 - \Phi(\widehat{\Omega}_{A,UU}^{-1/2}(\widehat{\beta}_U^A - \beta_U^A))$$

Assumption 6 is an extension of PV condition in Chernozhukov et al. (2017). In comparison to the PV condition in Chernozhukov et al. (2017), Assumption 6 involves twice as many p-values, 2 p-values for the lower bound and 2 more for the upper bound. For reporting purposes, I use Chernozhukov et al. (2017)'s adjusted point estimator

$$\widehat{\beta}_L = \text{Med}[\widehat{\beta}_L^A|\text{Data}], \quad \widehat{\beta}_U = \text{Med}[\widehat{\beta}_U^A|\text{Data}].$$

To quantify the uncertainty of the random split, I define the confidence region of level $(1-2\alpha)$:

$$[L,U] = \left[\, \overline{\text{Med}}[L_A|\text{Data}], \quad \underline{\text{Med}}[U_A|\text{Data}] \right], \tag{B.29}$$

where $\underline{\text{Med}}(X) = \inf\{x \in \mathrm{R} : P_X(X \leq x) \geq 1/2\}$ is the lower median and $\overline{\text{Med}}(X) = \sup\{x \in \mathrm{R} : P_X(X \geq x) \geq 1/2\}$ is the upper median. I will also consider a related confidence region of level $1-2\alpha$:

$$CR = \left\{ \beta \in \mathrm{R}, \quad p_L(\beta) > \alpha/2, \quad p_U(\beta) > \alpha/2 \right\}, \tag{B.30}$$

where

$$p_L(\beta) = \Phi(\widehat{\Omega}_{A,LL}^{-1/2}(\widehat{\beta}_L^A - \beta)), \quad p_U(\beta) = \Phi(\widehat{\Omega}_{A,UU}^{-1/2}(\widehat{\beta}_U^A - \beta))$$

Theorem B.3 (Uniform Balidity of Variational Confidence Region)**.** *The confidence region CR in equation* (B.31) *obeys* $CR \subseteq [L,U]$. *Under Assumption 6,*

$$\Pr_P([\beta_L^A, \beta_U^A] \in CR) \geq 1 - 2\alpha - 2(\delta + \gamma) = 1 - 2\alpha - o(1). \tag{B.31}$$

and therefore

$$\Pr_P(\beta_0 \in CR) \geq 1 - 2\alpha - 2(\delta + \gamma) = 1 - 2\alpha - o(1). \tag{B.32}$$

B.4 Supplementary Statements for Section 6.1

Multi-Dimensional Outcome: Definitions. The d-dimensional unit sphere is $\mathcal{S}^{d-1} = \{q \in \mathrm{R}^d, \|q\| = 1\}$. For a point q of interest on a unit sphere, the data vector is

$$W_q = (D, X, S, S \cdot Y_q),$$

where D and X are as defined in the one-dimensional case, $S = 1_{\{\mathbf{S}\}=\mathbf{1}}$ is equal to one if and only if each scalar outcome is selected into the sample, and $Y_q = q'Y$. The conditional quantiles in the selected treated and the selected control groups are

$$Q_d(q,u,x): \quad \Pr(Y_q \leq Q_d(q,u,x) | S = 1, D = d, X = x) = u, \quad u \in [0,1], \quad d \in \{1,0\}$$

and the combined conditional quantile is

$$Q(q,u,x) = 1_{x \in \mathcal{X}_{\text{help}}} Q_1(q,u,x) + 1_{x \in \mathcal{X}_{\text{hurt}}} Q_0(q,u,x).$$

Likewise, the combined conditional density is

$$f(q,t|x) = 1_{x\in\mathcal{X}_{\text{help}}} f_1(q,t|x) + 1_{x\in\mathcal{X}_{\text{hurt}}} f_0(q,t|x),$$

where $f_d(q,t|x)$ is the conditional density of $q'Y$ in $S=1, D=d, X=x$ group. Depending on whether $\mu_1(x)$ is known ot not, the first-stage nuisance parameter $\xi_0(q)$ is as in (B.13)-(B.15), where $Q(u,x)$ is replaced by $Q(q,u,x)$. The sharp upper bound on $q'\beta_0$ is

$$\sigma(q) = \mathbb{E} m_U(W_q, \xi_0(q)) \tag{B.33}$$

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

$$\mathcal{B} = \cap_{q\in\mathrm{R}^d:\|q\|=1}\{b \in \mathrm{R}^d : q'b \leq \sigma(q)\}. \tag{B.34}$$

Denote the sample average of a function $f(\cdot)$ as

$$\mathbb{E}_N[f(W_i)] := \frac{1}{N}\sum_{i=1}^{N} f(W_i)$$

and the centered, root-N scaled sample average as

$$\mathbb{G}_N[f(W_i)] := \frac{1}{\sqrt{N}}\sum_{i=1}^{N}[f(W_i) - \int f(w)dP(w)].$$

Let $\ell^\infty(\mathcal{S}^{d-1})$ be the space of almost surely bounded functions defined on the unit sphere $\mathcal{S}^{d-1}$ and $\mathrm{BL}(\mathcal{S}^{d-1},[0,1])$ be a set of real functions on $\mathcal{S}^{d-1}$ with Lipshitz norm bounded by 1.

Definition 1 (Cross-Fitting). 1. For a random sample of size N, denote a K-fold random partition of the sample indices $[N] = \{1,2,...,N\}$ by $(J_k)_{k=1}^{K}$, where K is the number of partitions and the sample size of each fold is $n = N/K$. For each

$k \in [K] = \{1,2,...,K\}$ define $J_k^c = \{1,2,...,N\} \setminus J_k$.

2. For each $k \in [K]$, construct an estimator $\widehat{\xi}_k(q) = \widehat{\xi}_{(W_{i \in J_k^c})}(q)$ of the nuisance parameter ξ_0 using only the data $\{W_j : j \in J_k^c\}$.

Definition 2 (Support Function Estimator). Define

$$\widehat{\sigma}(q) := \frac{1}{N} \sum_{i=1}^{N} g_U(W_{iq}, \widehat{\xi}(q)), \tag{B.35}$$

where $g(W_{iq}, \widehat{\xi}(q)) := g(W_{iq}, \widehat{\xi}_k(q))$ for any observation $i \in J_k, \quad k = 1,2,\ldots,K$.

Proposition B.4 (Characterization of Identified Set)**.** *Suppose Assumption 2 holds. Then, the set* $\mathcal{B}$*, defined in* (B.34)*, is a sharp identified set for* β_0*. Furthermore,* $\mathcal{B}$ *is a convex and compact set, and* $\sigma(q)$*, defined in* (B.33)*, is its support function.*

Proposition B.4 proves that the sharp identified set $\mathcal{B}$ is compact and convex and proposes a semiparametric moment equation for its support function.

Proposition B.5 (Orthogonal Moment Equation for Support Function)**.** *Suppose Assumption 2 holds and let* $g_U(W, \xi_0)$ *be as in* (B.18)*. Then,*

$$\mathbb{E}[\sigma(q) - g_U(W_q, \xi_0(q))] = 0 \tag{B.36}$$

is an orthogonal moment equation for $\sigma(q)$*.*

Proposition B.5 establishes that the moment equation (B.36) is orthogonal w.r.t the nuisance parameter $\xi_0(q)$ for each $q \in \mathcal{S}^{d-1}$. Define

$$h(W,q) = \sigma(q) - g_U(W_q, \xi_0(q)).$$

ASSUMPTION 7 (Quantile First-Stage Rate: Multi-Dimensional Case)**.** *(1) For each* $q \in \mathcal{S}^{d-1}$*, the conditional density* $f(q,t|x)$ *exists and is bounded from above and away from zero and has bounded derivative, where the bounds do not depend on q, almost* *surely in X. (2) Let $\bar{\mathcal{U}}$ be a compact set in $(0,1)$ containing the support of $p_0(X)$ and $1-p_0(X)$. There exist a rate $q_N = o(N^{-1/4})$, a sequence of numbers $\varepsilon_N = o(1)$ and a sequence of sets Q_N such that the first-stage estimate $\widehat{Q}(q,u,x)$ of the quantile function $Q(q,u,x) : \mathcal{S}^{d-1} \times [0,1] \times \mathcal{X} \to \mathrm{R}$ belongs to Q_N w.p. at least $1-\varepsilon_N$. Furthermore, the set Q_N shrinks sufficiently fast around the true value $Q_0(u,x)$ uniformly on $\bar{\mathcal{U}}$ and $\mathcal{S}^{d-1}$:*

$$\sup_{Q\in Q_N} \sup_{q\in\mathcal{S}^{d-1}} \sup_{u\in(0,1)} (\mathbb{E}_X(\widehat{Q}(q,u,X) - Q(q,u,X))^2)^{1/2} \lesssim q_N = o(N^{-1/4}).$$

Assumption 7 is a generalization of Assumption 5 from one- to multi-dimensional case.

Lemma B.6 (Limit Theory for the Support Function Process)**.** *Suppose Assumptions 2, 4, 7 hold. In addition, if $\mathcal{X}_{help} \neq \emptyset$ and $\mathcal{X}_{hurt} \neq \emptyset$, suppose $\widehat{s}(d,x)$ converges to $s(d,x)$ uniformly over $\mathcal{X}$ for each $d \in \{1,0\}$. The support function process $S_N(q) = \sqrt{N}(\widehat{\sigma}(q) - \sigma(q))$ admits an approximation*

$$S_N(q) = \mathbb{G}_N[h(q)] + o_P(1)$$

in $\ell^\infty(\mathcal{S}^{d-1})$. Moreover, the support function process admits an approximation

$$S_N(q) = \mathbb{G}[h(q)] + o_P(1) \text{ in } \ell^\infty(\mathcal{S}^{d-1}),$$

where the process $\mathbb{G}[h(q)]$ is a tight P-Brownian bridge in $\ell^\infty(\mathcal{S}^{d-1})$ with covariance function

$$\Omega(q_1,q_2) = \mathbb{E}[h(W,q_1)h(W,q_2)], \quad q_1,q_2 \in \mathcal{S}^{d-1}$$

that is uniformly Holder on $\mathcal{S}^{d-1} \times \mathcal{S}^{d-1}$. Furthermore, the canonical distance between the law of the support function process $S_N(q)$ and the law $\mathbb{G}[h(q)]$ in $\ell^\infty(\mathcal{S}^{d-1})$ ap-

proaches zero, namely

$$\sup_{g \in BL(\mathcal{S}^{d-1},[0,1])} |\mathbb{E}[g(S_N)] - \mathbb{E}[g(\mathbb{G}[h])]| \to 0.$$

Lemma B.6 says that the Support Function Estimator is asymptotically equivalent to a Gaussian process and can be used for pointwise and uniform inference about the support function. By orthogonality, the first-stage estimation error does not contribute to the total uncertainty of the two-stage procedure. In particular, orthogonality allows me to avoid relying on any particular first-stage estimator, and to employ modern regularized techniques to estimate the first-stage.

Theorem B.7 (Limit Inference on Support Function Process)**.** *Under the assumptions of Lemma B.6 hold, for any $\widehat{c}_N = c_N + o_P(1)$, $c_N = O_P(1)$ and $f \in \mathcal{F}_c$,*

$$\Pr(f(S_N) \leq \widehat{c}_N) - \Pr(f(\mathbb{G}[h]) \leq c_N) \to 0.$$

If $c_N(1-\tau)$ is the $(1-\tau)$-quantile of $f(\mathbb{G}[h])$ and $\widehat{c}_N(1-\tau) = c_N(1-\tau) + o_P(1)$ is any consistent estimate of this quantile, then

$$\Pr(f(S_N) \leq \widehat{c}_N(1-\tau)) \to 1-\tau.$$

Definition 3 (Weighted Bootstrap). Let B represent a number of bootstrap repetitions. For each $b \in \{1, 2, \ldots, B\}$, repeat

1. Draw N i.i.d. exponential random variables $(e_i)_{i=1}^N : e_i \sim \text{Exp}(1)$.

2. Estimate $\widetilde{\sigma}^b(q) = \mathbb{E}_N e_i g(W_i, q, \widehat{\xi}_i(q))$.

Let $e_i^0 = e_i - 1$, $h^0 = h - \mathbb{E}[h]$, and let P^e be the probability measure conditional on data.

Lemma B.8 (Limit Theory for the Bootstrap Support Function Process)**.** *Suppose assumptions of Lemma B.6 hold. The bootstrap support function process $\widetilde{S}_N(q) = \sqrt{N}(\widetilde{\sigma}(q) - \widehat{\sigma}(q))$ admits the following approximation conditional on the data:*

$$\widetilde{S}_N(q) = \mathbb{G}_N[e_i^0 h_i^0(q)] + o_{P^e}(1)$$

in $L^\infty(\mathbb{S}^{d-1})$. Moreover, the support function process admits an approximation conditional on the data

$$\widetilde{S}_N(q) = \widetilde{\mathbb{G}}[h(q)] + o_{P^e}(1) \text{ in } L^\infty(\mathbb{S}^{d-1}), \text{ in probability } P,$$

where $\widetilde{\mathbb{G}}[h(q)]$ is a sequence of tight P-Brownian bridges in $L^\infty(\mathbb{S}^{d-1})$ with the same distributions as the processes $\mathbb{G}_N[h(q)]$ defined in Lemma B.6, and independent of $\mathbb{G}_N[h(q)]$. Furthermore, the canonical distance between the law of the bootstrap support function process $\widetilde{S}_N(q)$ conditional on the data and the law of $\mathbb{G}[h]$ in $\ell^\infty(\mathbb{S}^{d-1})$ approaches zero, namely

$$\sup_{g \in BL(\mathbb{S}^{d-1},[0,1])} |\mathbb{E}_{P^e}[g(S_N)] - \mathbb{E}[g(\mathbb{G}[h])]| \to_P 0,$$

where $BL(\mathbb{S}^{d-1}, [0,1])$ is a set of real functions on $\mathbb{S}^{d-1}$ with Lipshitz norm bounded by 1.

Lemma B.8 says that weighted bootstrap of Theorem B.7 can be used to approximate the support function process of Theorem B.7 and conduct uniform inference about the support function. By orthogonality, I do not need to re-estimate the first-stage parameter in each bootstrap repetition. Instead, I estimate the first-stage parameter once on an auxiliary sample, and plug the estimate into the bootstrap sampling procedure. Therefore, orthogonal Bayes bootstrap is faster to compute than a non-orthogonal Bayes bootstrap, where both the first and the second stage are re-estimated in each bootstrap repetition.

Theorem B.9 (Bootstrap Inference on the Support Function Process)**.** *Suppose assumptions of Lemma B.6 hold. For any $c_N = O_P(1)$ and $f \in \mathcal{F}$ we have*

$$\Pr(f(S_N) \leq c_N) - \Pr^e(f(\widetilde{S}_N) \leq c_N) \to_P 0.$$

In particular, if $\widetilde{c}_N(1-\tau)$ is the $(1-\tau)$-quantile of $f(\widetilde{S}_N)$ under $\Pr^e$*, then*

$$\Pr(f(S_N) \leq \widetilde{c}_N(1-\tau)) \to_P 1-\tau.$$

Lemmas B.6 and B.8 are a generalization of Theorems 1–4 in Chandrasekhar et al. (2012). Unlike Chandrasekhar et al. (2012), the support function estimator proposed here is based on an orthogonal moment equation (B.36). Therefore, I do not rely on a series estimator of the first-stage nuisance parameter $\xi_0(q)$, but allow any machine learning method obeying Assumption 10 to be used. As a result, this theory can accommodate many covariates.

B.5 Supplementary Statements for Section 6.2

In this section, I derive sharp Lee bounds for the Intent-to-Treat parameter given in equation (8.5). For the sake of exposition, suppose $\mathcal{X}_{\text{hurt}} = \emptyset$. Suppose there exists a vector $\bar{X}$ of saturated covariates so that a stronger version of independence assumption holds.

ASSUMPTION 8 (Conditional Independence)**.** *Conditional on $\bar{X}$, D is independent of* $(Y(1), Y(0), S(1), S(0))$.

The full vector X is

$$X = (\bar{X}, \widetilde{X}),$$

where $\bar{X}$ is stratification covariate vector and $\widetilde{X}$ includes all other baseline covariates. Let A be an event and ξ be a random variable. For a given event A, its probability conditional on being on always-taker is

$$\Pr_{11}(A) = \Pr(A|S(1) = S(0) = 1). \tag{B.37}$$

For a random variable ξ, its expectation conditional on being an always-taker is

$$\mathbb{E}_{11}[\xi] = \mathbb{E}[\xi|S(1) = S(0) = 1]. \tag{B.38}$$

Suppose $\mathcal{X}_{\text{hurt}} = \emptyset$. The lower and upper truncation sets are

$$\mathcal{T}_L(W) := \Big\{\{Y \geq Q(p_0(X), X)\} \cap D = 1\Big\}, \tag{B.39}$$

$$\mathcal{T}_U(W) := \Big\{\{Y \leq Q(1 - p_0(X), X)\} \cap D = 1\Big\}. \tag{B.40}$$

In what follows, I present the argument for the lower truncation set $\mathcal{T}_L(W)$, a symmetric argument holds for $\mathcal{T}_U(W)$. For a given event A, its probability conditional on not being trimmed is

$$\Pr_{\mathcal{T}_L}(A) = \Pr(A|S = 1, W \notin \mathcal{T}_L(W)). \tag{B.41}$$

For a random variable ξ, its expectation conditional on not being trimmed is

$$\mathbb{E}_{\mathcal{T}_L}(A) = \mathbb{E}(A|S = 1, W \notin \mathcal{T}_L(W)). \tag{B.42}$$

For a covariate value x, the density of X conditional on being an always-taker is

$$f_{11}(x) = f(x|S(1) = S(0) = 1).$$

Likewise, the density of X conditional on not being trimmed is

$$f_{\mathcal{T}_L}(x) = f(x|S=1, W \notin \mathcal{T}_L(W)).$$

The target parameter β_1 is the coefficient on D in the infeasible regression (8.5). The proposed bound β_1^L is the coefficient on D in the feasible trimmedregressions

$$Y = \beta_0^L + D\beta_1^L + \bar{X}'\beta_2^L + U, \quad S=1, W \notin \mathcal{T}_L(W), \tag{B.43}$$

and β_1^U is defined similarly, where $\mathcal{T}_L(W)$ is replaced by $\mathcal{T}_U(W)$. An orthogonal moment equation for β_1^L is

$$g_1^L(W, \xi_0) = (\mu_{10}^{-1}) w_{\mathcal{T}_L}(\bar{X})(\beta_L^{\text{help}}(X)s(0,X) + \alpha_L(W;\xi_0)), \tag{B.44}$$

where $\beta_L^{\text{help}}(x)$ is defined in (B.7), $\alpha_L(W;\xi)$ is defined in (B.21), and μ_{10} is as in (B.11).

Lemma B.10 (Intent-to-Treat Theorem, Angrist and Pischke (2009))**.** *Under Assumption 8, the regression coefficient β_1 can be represented as*

$$\beta_1 = \mathbb{E}_{11} w_{11}(\bar{X}) \bigg(\mathbb{E}_{11}[Y|D=1,\bar{X}] - \mathbb{E}_{11}[Y|D=0,\bar{X}] \bigg), \tag{B.45}$$

where the weighting function is

$$w_{11}(\bar{X}) = \frac{\Pr_{11}(D=1|\bar{X})\Pr_{11}(D=0|\bar{X})}{\mathbb{E}_{11}[\Pr_{11}(D=1|\bar{X})\Pr_{11}(D=0|\bar{X})]}.$$

Likewise, β_1^L can be represented as (B.45)*, where* $\Pr_{11}$ *is replaced by* $\Pr_{\mathcal{T}_L}$.

Lemma B.11 (Equal covariate distributions under $\Pr_{11}$ and $\Pr_{\mathcal{T}_L}$)**.** *Under Assumptions 2 and 8, the conditional density $f_{11}(x)$ coincides with $f_{\mathcal{T}_L}(x)$ almost surely in $\mathcal{X}$:*

$$f_{11}(x) = f_{\mathcal{T}_L}(x), \quad x \in \mathcal{X}.$$

Lemma B.12 (Equal propensity scores under $\Pr_{11}$ and $\Pr_{\mathcal{T}_L}$)**.** *Under Assumptions 2 and 8, the conditional propensity score* $\Pr_{11}(D=1|X=x)$ *coincides with* $\Pr_{\mathcal{T}_L}(D=1|X=x)$ *almost surely in* $\mathcal{X}$*:*

$$\Pr_{11}(D=1|X=x) =^{i} \Pr(D=1|X) =^{ii} \Pr(D=1|\bar{X}) =^{iii} \Pr_{\mathcal{T}_L}(D=1|X=x). \quad \text{(B.46)}$$

Proposition B.13 (Sharp Bounds on Intent-to-Treat effect: Identification)**.** *Let* β_1^L *and* β_1^U *be as defined in equation* (B.43)*, and* β_1 *be as defined in equation* (8.5)*. Under Assumptions 2 and 8,* β_1^L *and* β_1^U *are sharp bounds on* β_1*:*

$$\beta_1^L \leq \beta_1 \leq \beta_1^U. \quad \text{(B.47)}$$

Lemma B.14 (Sharp Bounds on Intent-to-Treat effect: Estimation and Inference)**.** *Suppose Assumptions 2 and 8 hold with* $\mathcal{X}_{hurt} = \emptyset$*. Suppose the first-stage parameter* ξ_0 *is estimated by logistic/quantile squares series, and under-smoothing conditions hold. Then, the empirical analog of* (B.43) *based on the estimated* $\widehat{\xi}$ *is asymptotically equivalent to sample average. In addition, if Assumptions 4 and 7 hold, the plug-in cross-fitting orthogonal estimator of* β_L^1 *based on equation* (B.44) *is asymptotically linear:*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_1^L(W_i,\widehat{\xi}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_1^L(W_i,\xi_0) + o_P(1). \quad \text{(B.48)}$$

B.6 Supplementary Statements for Section 6.3

In this section, I define the sharp bounds on the Local Average Treatment Effect parameter, defined in equation (8.6). Let $Z=1$ be a binary instrument, such as the treatment offer, and let $D=1$ be an indicator for treatment receipt. Let $S(d,z)=1$ be a dummy for being selected into the sample when the instrument is z and the treatment is equal to d, and let $Y(d,z)$ be a potential outcome. The observed data $(Z_i,D_i,X_i,S_i,S_iY_i)_{i=1}^N$ consist of the instrument, treatment, a baseline covariate vector X_i, the selection outcome

$S_i = \sum_{d\in\{1,0\}} \sum_{z\in\{1,0\}} 1_{\{Z_i=z\}} 1_{\{D_i=d\}} S_i(d,z)$ and the observed outcomes for the selected subjects, $S_i \cdot Y_i = S_i \cdot (\sum_{d\in\{1,0\}} \sum_{z\in\{1,0\}} 1_{\{Z_i=z\}} 1_{\{D_i=d\}} Y_i(d,z))$.

ASSUMPTION 9 (Assumptions for LATE)**.** *The following statements hold.*

(1) The vector $D(1), D(0), \{(Y(d,z), S(d,z))\}_{d\in\{0,1\}, z\in\{0,1\}}$ is independent of Z conditional on a subset of stratification covariates $\bar{X}$.

(2) Exclusion for outcome. For each $d \in \{1,0\}$, the outcome $Y(d,0) = Y(d,1) = Y(d)$ almost surely.

(3) Monotonicity of treatment w.r.t instrument. The instrument affects treatment in the same direction: $D(1) \geq D(0)$ almost surely.

(4) First Stage. $\Pr(D(1) > D(0) | \bar{X} = \bar{x}) > 0$ *almost surely in $\bar{X}$.*

(5) Exclusion for selection. For each $z \in \{1,0\}$, the outcome $S(0,z) = S(1,z) = S(z)$ almost surely.

(6) Independence of selection and treatment. Treatment potential outcomes $\{D(1), D(0)\}$ are independent of selection potential outcomes $\{S(1), S(0)\}$ conditional on X.

For the sake of exposition, suppose $\mathcal{X}_{\text{hurt}} = \emptyset$. Let $Q(u,x,d)$ be the conditional u-quantile of Y in the $\{S = 1, D = d, Z = 1\}$ group:

$$Q(u,x,d): \quad \Pr(Y \leq Q(u,x,d) | S = 1, D = d, Z = 1, X = x) = u.$$

The lower truncation set is

$$\Lambda_L(W) = \left\{ \cup_{d\in\{1,0\}} \{Y \geq Q(p_0(X), X, d) \cap D = d \cap Z = 1\} \right\}. \tag{B.49}$$

The upper truncation set is

$$\Lambda_U(W) = \left\{ \cup_{d\in\{1,0\}} \{Y \leq Q(1 - p_0(X), X, d) \cap D = d \cap Z = 1\} \right\}. \tag{B.50}$$

For an event A, the probability of A conditional on not being trimmed is

$$\Pr_{\Lambda_L}(A) = \Pr(A|S=1, W \notin \Lambda_L(W)).$$

For a random variable ξ, the expectation of ξ conditional on not being trimmed is

$$\mathbb{E}_{\Lambda_L}[\xi] = \mathbb{E}[\xi|S=1, W \notin \Lambda_L(W)].$$

The target parameter π_1 is the 2SLS coefficient on D in the infeasible regression (8.6)-(6.7). The proposed bound π_L^1 is the 2SLS coefficient on D in the feasible trimmed regression

$$Y = \pi_0^L + D\pi_1^L + \bar{X}'\pi_2^L + \varepsilon, \quad S=1, W \notin \Lambda_L(W), \tag{B.51}$$

where the first-stage equation is

$$D = \delta_0^L + Z\delta_1^L + \bar{X}'\delta_2^L + \xi, \quad S=1, W \notin \Lambda_L(W). \tag{B.52}$$

Lemma B.15 (LATE Theorem, Angrist and Imbens (1995))**.** *Let π_{11} follow the definition in* (8.6)*. If Assumption 9* (1)–(4) *holds,* π_1 *can be represented as*

$$\begin{aligned}
\pi_{11} &= \mathbb{E}_{11}\omega_{11}(\bar{X})\mathbb{E}_{11}[Y(1)-Y(0)|D(1)>D(0),\bar{X}] \\
&= \mathbb{E}_{11}\omega_{11}(\bar{X})\frac{\mathbb{E}_{11}[Y|Z=1,\bar{X}] - \mathbb{E}_{11}[Y|Z=0,\bar{X}]}{\mathbb{E}_{11}[D=1|Z=1,\bar{X}] - \mathbb{E}_{11}[D=1|Z=0,\bar{X}]},
\end{aligned} \tag{B.53}$$

where the weighting function $\omega_{11}(\bar{X})$ takes the form

$$\omega_{11}(\bar{X}) = \frac{V_{11}(\mathbb{E}_{11}[D=1|\bar{X},Z]|\bar{X})}{\mathbb{E}_{11}[V_{11}(\mathbb{E}_{11}[D=1|\bar{X},Z]|\bar{X})]}.$$

Likewise, π_L^1 can be represented as (B.53)*, where* $\Pr_{11}$ *is replaced by* $\Pr_{\Lambda_L}$*.*

Lemma B.16 (Equal covariate distributions under $\Pr_{11}$ and $\Pr_{\Lambda_L}$)**.** *Under Assumptions 2 and 9, the conditional densities are equal*

$$f_{11}(x) = f_{\Lambda_L}(x), \quad \textit{for any } x \in \mathcal{X}.$$

Furthermore, the integrated covariate densities are equal to each other,

$$\int_{\widetilde{x}\in\widetilde{\mathcal{X}}} f_{11}(\bar{x},\widetilde{x})d\widetilde{x} = \int_{\widetilde{x}\in\widetilde{\mathcal{X}}} f_{\Lambda_L}(\bar{x},\widetilde{x})d\widetilde{x}, \quad \textit{for any } \bar{x}.$$

Lemma B.17 (Equal treatment distributions under $\Pr_{11}$ and $\Pr_{\Lambda_L}$)**.** *Under Assumptions 2, 8 and 9 hold, the following equality holds:*

$$\Pr_{11}(D=1|Z=z,X) = \Pr(D=1|Z=z,X) = \Pr_{\Lambda_L}(D=1|Z=z,X), \quad z \in \{1,0\}. \tag{B.54}$$

Proposition B.18 (Sharp Bounds on LATE: Identification)**.** *Let π_1^L and π_1^U be as defined in equation* (B.51)-(B.52)*, and β_1 be as defined in equation* (8.6)-(6.7)*. Under Assumptions 2 and 9, π_1^L and π_1^U are sharp bounds on π_1:*

$$\pi_1^L \leq \pi_1 \leq \pi_1^U. \tag{B.55}$$

Appendix C: Proofs for Sections 4-6

Notation. I use the following standard notation. Let $\mathcal{S}^{d-1} = \{q \in \mathrm{R}^d, \|q\| = 1\}$ be the d-dimensional unit sphere. I use standard notation for numeric and stochastic dominance. For two numeric sequences $\{a_n, n \geq 1\}$ and $\{b_n, n \geq 1\}$, let $a_n \lesssim b_n$ stand for $a_n = O(b_n)$. For two sequences of random variables $\{a_n, n \geq 1\}$ and $\{b_n, n \geq 1\}$, let $a_n \lesssim_P b_n$ stand for $a_n = O_P(b_n)$. Let $a \wedge b = \min\{a,b\}$ and $a \vee b = \max\{a,b\}$. For a random variable ξ, $(\xi)^0 := \xi - \mathbb{E}[\xi]$. Let $\ell^\infty(\mathcal{S}^{d-1})$ be the space of almost surely bounded functions defined on the unit sphere $\mathcal{S}^{d-1}$. Define an $L_{P,c}$ norm of a vector-valued random variable W as: $\|W\|_{L_{P,c}} := \left(\int_{w\in\mathcal{W}} \|W\|^c\right)^{1/c}$. Let $\mathcal{W}$ be the support of the data vector W of the distribution P_W. Let $(W_i)_{i=1}^N$ be an i.i.d sample from the distribution P_W.

I use the standard notation for vector and matrix norms. For a vector $v \in \mathrm{R}^d$, denote the ℓ_2 norm of a as $\|v\|_2 := \sqrt{\sum_{j=1}^d v_j^2}$. Denote the ℓ_1 norm of v as $\|v\|_1 := \sum_{j=1}^d |v_j|$, the ℓ_∞ norm of v as $\|v\|_\infty := \max_{1\leq j\leq d} |v_j|$, and ℓ_0 norm of v as $\|v\|_0 := \sum_{j=1}^d 1_{\{v_j \neq 0\}}$. For a matrix M, denote its operator norm by $\|M\|_2 = \sup_{\alpha\in\mathcal{S}^{d-1}} \|M\alpha\|$. Denote the sample average of a function $f(\cdot)$ as $\mathbb{E}_N[f(W_i)] := \dfrac{1}{N}\sum_{i=1}^N f(W_i)$ and the centered, root-N scaled sample average as $\mathbb{G}_N[f(W_i)] := \dfrac{1}{\sqrt{N}}\sum_{i=1}^N [f(W_i) - \int f(w)dP(w)]$.

Fix a partition k in a set of partitions $[K] = \{1,2,\ldots,K\}$. Define the sample average of a function $f(\cdot)$ within this partition as:

$$\mathbb{E}_{n,k}[f] = \frac{1}{n}\sum_{i\in J_k} f(x_i) \tag{C.1}$$

and the scaled normalized sample average as:

$$\mathbb{G}_{n,k}[f] = \frac{\sqrt{n}}{n}\sum_{i\in J_k}[f(x_i) - \int f(w)dP(w)].$$

For each partition index $k \in [K]$ define an event $\mathcal{E}_{n,k} := \{\widehat{\xi}_k \in \Xi_n\}$ as the nuisance estimate $\widehat{\xi}_k$ belonging to the nuisance realization set $\mathcal{G}_N$. Define

$$\mathcal{E}_N = \cap_{k=1}^K \mathcal{E}_{n,k} \tag{C.2}$$

as the intersection of such events.

Proof of Lemma B.1. Proof of Lemma B.1(a) is shown in Steps 1-3. Proof of Lemma B.1(b) is shown in Steps 4-6.

Step 1. According to Lee (2009) (Proposition 3 in the working paper version Lee

(2005)), (B.6)-(B.7) are valid sharp bounds on $\mathbb{E}[Y(1)-Y(0)|S(1)=S(0)=1,X=x]$ for $x \in \mathcal{X}_{\text{help}}$. Likewise, (B.8)-(B.9) are valid sharp bounds on $\mathbb{E}[Y(1)-Y(0)|S(1)=S(0)=1,X=x]$ for $x \in \mathcal{X}_{\text{hurt}}$.

Step 2. For $x \in \mathcal{X}_*$, Bayes rule implies that

$$f_\star(x|S(1)=S(0)=1) = \frac{\Pr(S(1)=S(0)=1|X=x)f(x)}{\Pr(S(1)=S(0)=1|\mathcal{X}_\star)} \quad \star \in \{\text{help}, \text{hurt}\}.$$

By Assumption 2,

$$\Pr(S(1)=S(0)=1|X=x) = \Pr(S(0)=1|X=x) = s(0,x) \quad \text{for any } x \in \mathcal{X}_{\text{help}},$$
$$\Pr(S(1)=S(0)=1|X=x) = \Pr(S(1)=1|X=x) = s(1,x) \quad \text{for any } x \in \mathcal{X}_{\text{hurt}}.$$

Therefore, the sharp bounds defined as

$$\int_{x\in\mathcal{X}_{\text{help}}} \bar{\beta}_\star^{\text{help}}(x) f_{\text{help}}(x|S(1)=S(0))dx \cdot \text{Prob}(X \in \mathcal{X}_{\text{help}})$$
$$+ \int_{x\in\mathcal{X}_{\text{hurt}}} \bar{\beta}_\star^{\text{hurt}}(x) f_{\text{hurt}}(x|S(1)=S(0))dx \cdot \text{Prob}(X \in \mathcal{X}_{\text{hurt}}), \quad \star \in \{\text{L}, \text{U}\}$$

coincide with the bounds in (B.12).

Step 3. Validity of moment equations. For $X \in \mathcal{X}_{\text{help}}$,

$$\begin{aligned}
& s(0,X)\mathbb{E}[Y|D=1,S=1,X,Y \geq Q(1-p_0(X),X)]] \\
&=^i s(0,X)\mathbb{E}[\frac{Y \cdot 1_{\{Y\geq Q(1-p_0(X),X)\}}}{p_0(X)}|D=1,S=1,X] \\
&= s(0,X)\mathbb{E}\left[D \cdot S \cdot \frac{Y 1_{\{Y\geq Q(1-p_0(X),X)\}}}{p_0(X)s(1,X)\mu_1(X)}|X\right] \\
&=^{ii} \mathbb{E}\left[\frac{D}{\mu_1(X)} \cdot S \cdot Y \cdot 1_{\{Y\geq Q(1-p_0(X),X)\}}|X\right]
\end{aligned}$$

where *i* follows from $\Pr(Y \geq Q(1-p_0(X),X)|D=1,S=1,X) = p_0(X)$ and *ii* follows from $p_0(X) = s(0,X)/s(1,X)$. Likewise,

$$s(0,X)\mathbb{E}[Y|S=1,D=0,X] = \mathbb{E}[\frac{1-D}{\mu_0(X)}S\cdot Y|X], \text{ and}$$
$$\mathbb{E}[1_{X\in\mathcal{X}_{\text{help}}}s(0,X)\bar{\beta}_U^{\text{help}}(X)] = \mathbb{E}[1_{X\in\mathcal{X}_{\text{help}}}m_U(W,\xi_0)].$$

By a similar argument,

$$\mathbb{E}[1_{X\in\mathcal{X}_{\text{hurt}}}s(1,X)\bar{\beta}_U^{\text{hurt}}(X)] = \mathbb{E}[1_{X\in\mathcal{X}_{\text{hurt}}}m_U(W,\xi_0)].$$

Step 4. Proof of Lemma B.1(b). First, suppose there are no covariates (i.e., $X = \emptyset$) and Assumption 1 holds with $p_0 < 1$. It suffices to show that, for each value of $\beta \in [\beta_L, \beta_U]$, there exists a distribution of the always-takers' outcome $Y(1)$, so that $\mathbb{E}[Y(1)-Y(0)|S(1)=S(0)=1] = \beta$. Let $\lambda \in [0,1]$ be a positive number. Consider the following distribution function $F_\lambda(t), t \in (-\infty,\infty)$:

$$F_\lambda(t) = \begin{cases} 0, & t \leq Q_1((1-\lambda)(1-p_0)), \\ p_0^{-1}\int_{Q_1((1-\lambda)(1-p_0)}^{t} f_1(y)dy, & Q_1((1-\lambda)(1-p_0)) \leq t \leq Q_1(\lambda p_0 + (1-\lambda)) \\ 1, & t \geq Q_1(\lambda p_0 + (1-\lambda)), \end{cases}$$

where $Q_1(u)$ is the u-outcome quantile of Y in $S=1, D=1$ group. When $\lambda = 0$, $F_0(t)$ corresponds to the upper tail of the distribution $f_1(y)$, and $\mathbb{E}_{F_0}[Y(1)] - \mathbb{E}[Y(0)] = \beta_U$. When $\lambda = 1$, $F_1(t)$ corresponds to the lower tail of the distribution $f_1(y)$, and $\mathbb{E}_{F_1}[Y(1)] - \mathbb{E}[Y(0)] = \beta_L$. Observe that

$$F_1(t) = p_0 F_\lambda(t) + (1-p_0)G_\lambda(t), \tag{C.3}$$

where $G_\lambda(t)$ is a valid c.d.f.. Therefore, $F_1(t)$ can be represented as a mixture of valid two c.d.f.s, with mixing probabilities p_0 and $1-p_0$. Therefore, $F_\lambda(t)$ is a valid plausible c.d.f for the always-takers.

Step 5. In this step, I show that

$$\beta_\lambda = \mathbb{E}_{F_\lambda}[Y(1)] - \mathbb{E}[Y(0)] : \lambda \in [0,1] \to [\beta_L, \beta_U]$$

is a non-increasing and continuous function of λ. By construction, $F_{\lambda'}(t)$ first-order stochastically dominates $F_\lambda(t)$ for $\lambda' < \lambda$, which implies the first statement. Second, let C be an upper bound $\sup_{t \in \mathrm{R}} |t \cdot f_1(t)|$. By assumption of the Lemma, C is finite. Thus, two real numbers λ' and λ mean value theorem implies:

$$\begin{aligned}
|\beta_\lambda - \beta_\lambda| &\leq p_0^{-1} C |Q_1(\lambda p_0 + (1-\lambda)) - Q_1(\lambda' p_0 + (1-\lambda'))| \\
&\quad + p_0^{-1} C | - Q_1((1-\lambda)(1-p_0)) + Q_((1-\lambda')(1-p_0))| \\
&\leq p_0^{-1} C | \sup_{u \in \mathcal{U}} Q_1'(u) || \lambda - \lambda'|,
\end{aligned}$$

where $Q_1'(u) = f_1^{-1}(Q(u))$ is the conditional quantile's derivative, which exists and is bounded by assumption of the Lemma. Therefore, β_λ is a continuous and monotone function on $[0,1] \to [\beta_L, \beta_U]$. Therefore, each point $\beta \in [\beta_L, \beta_U]$ corresponds to some $\lambda \in [0,1]$. Therefore, Lemma (b) follows.

Step 6. Suppose Assumption 2 holds instead of Assumption 1. By the argument above, $[\beta^L_{\text{help}}(x), \beta^U_{\text{help}}(x)]$ is a sharp identified set for $\mathbb{E}[Y(1) - Y(0) | S(1) = S(0) = 1, X = x], \quad x \in \mathcal{X}_{\text{help}}$. Because the density $f_{\text{help}}(x | S(1) = S(0) = 1) = (\mu_{10}^{\text{help}})^{-1} s(0,x) f(x)$ is an identified function, Lemma B.1(b) under Assumption 2 holds. Furthermore, Lemma B.1(c) holds. □

Proof of Theorem 1. According to Proposition B.2, the sets $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$ are consistently estimated. Therefore, in the proof below, I condition on the event $\widehat{p}(X_i) < 1 \Leftrightarrow X_i \in X_{\text{help}}$ for any $X_i, i = \{1, 2, \ldots, N\}$, which occurs w.p. approaching 1. Theorem 1 is a special case of Theorem B.7 with $\mathcal{S}^{d-1} = \{-1, 1\}$. Here I verify that Assumption

7 for $\mathcal{S}^{d-1} = \{-1, 1\}$. In one-dimensional case, $Q(1,u,x)$ reduces to $Q(u,x)$. Furthermore, the quantile of $-Y$, denoted by $Q(-1,u,x)$, reduces to $-Q(1-u,x)$. Assumption 7(1)-(2) holds by Assumption 5(1)-(2). $\square$

Proof of Theorem B.3. **Step 1.** The probability of non-coverage is bounded as

$$\Pr_P([\beta_L^A, \beta_U^A] \not\in \text{CR}) \leq \Pr_P(p_L(\beta_L^A) < \alpha/2) + \Pr_P(p_U(\beta_U^A) < \alpha/2) \leq 1 - 2\alpha - 2(\delta + \gamma),$$

where the last inequality holds by Lemma 3.1 in Chernozhukov et al. (2017). Therefore, CR is a valid confidence region for the identified set $[\beta_L^A, \beta_U^A]$.

Step 2. By the proof of Theorem 3.2 in Chernozhukov et al. (2017),

$$\{\beta \in \mathrm{R} : p_U(\beta) > \alpha/2\} = \{\beta \in \mathrm{R} : \quad \overline{\text{Med}}[\widehat{\Omega}_{A,UU}^{-1/2}(\beta - \widehat{\beta}_U^A) - \Phi^{-1}(1-\alpha/2)|\text{Data})] < 0]. \tag{C.4}$$

Since the R.H.S. of (C.4) is a monotone increasing function of β, it suffices to show that

$$\overline{\text{Med}}[\widehat{\Omega}_{A,UU}^{-1/2}(U - \widehat{\beta}_U^A) - \Phi^{-1}(1-\alpha/2)|\text{Data})] \geq 0. \tag{C.5}$$

By definition of U,

$$\mathbb{E}[1[U - \widehat{\beta}_U^A - \widehat{\Omega}_{A,UU}^{1/2}\Phi^{-1}(1-\alpha/2) \geq 0]|\text{Data}] \geq 1/2. \tag{C.6}$$

By step 2 of the proof of Theorem 3.2 in Chernozhukov et al. (2017), (C.6) implies (C.5).

Step 3. By definition of Lee bounds, $\beta_0 \in [\beta_L^A, \beta_U^A]$ for any A. Therefore, (B.31) implies (B.32). $\square$

Proof of Proposition B.4. **Step 1.** I invoke Lemma B.1 with data $W_q = (D, X, S, S \cdot Y_q)$, where $S = 1_{\{\mathbf{S}=\mathbf{1}\}}$ and $Y_q = q'Y$. Let β_0 be the true parameter and $\mathcal{B}'$ be its true sharp identified set for β_0. By Lemma B.1 , (B.33) is a sharp upper bound on $q'\beta_0$. Therefore, (B.34) implies that

$$\sigma(q) = \sup_{b \in \mathcal{B}'} q'b. \tag{C.7}$$

Step 2. To show that $\sigma(q)$ defined in (B.33) is a support function of some compact and convex set, I need to show that $\sigma(q)$ is (1) convex, (2) positive homogenous of degree one and (3) lower-semicontinuous function of q. By Theorem 13.2 from Rockafellar (1997), the properties (1)-(3) imply that $\mathcal{B}$ is a convex and compact set and $\sigma(q)$ is its support function. Therefore, $\mathcal{B}$ in (B.34) coincides with the true sharp identified set for β_0.

Verification of (1). Lemma B.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 B.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,

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

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

Verification of (2). Let $\lambda > 0$. Observe that the event $\{\lambda Y_q \leq Q(\lambda q, p_0(X), X)\}$ holds if and only if $\{Y_q \leq Q(q, p_0(X), X)\}$. Since $Y_q = q'Y$ is a linear function of q, $\sigma(q)$ defined in (B.33) is positive homogenous of degree 1.

Verification of (3). Consider a sequence of vectors $q_k \to q, k \to \infty$. 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 non-sharp bound on $q'\beta_0$. By sharpness, $\sigma(q)$ is the smallest bound on $q'\beta_0$, which implies $\sigma(q) \leq C$. By Theorem 13.2 in Rockafellar (1997), $\sigma(q)$ is a support function of a convex, compact set $\mathcal{B}$ defined by a list of linear inequalities (C.7).

$\square$

Proof of Proposition B.5. The moment equation (B.18) is a sum of the original moment equation (B.16) and three bias correction terms, for $s(0,x)$, $s(1,x)$, and $Q(u,x)$, which is an orthogonal moment by Newey (1994). The bias corrections for $s(0,x)$, $s(1,x)$ and $\mu_1(x)$ follow from Proposition 4 in Newey (1994). The bias correction for $Q(u,x)$ follows from Proposition 6 in Ichimura and Newey (2015). According to Newey (1994), a bias correction term for a vector-valued nuisance parameter $\xi_0(x) = \{s(0,x), s(1,x), Q(u,x), \mu_1(x)\}$ is the sum of individual bias correction terms.

$\square$

ASSUMPTION 10 (Quality of the First-Stage Estimation)**.** *There exists a sequence $\{\Xi_N, N \geq 1\}$ of subsets of Ξ (i.e, $\Xi_N \subseteq \Xi$) such that the following conditions hold. (1) The true value ξ_0 belongs to Ξ_N for all $N \geq 1$. There exists a sequence of numbers $\phi_N = o(1)$ such that the first-stage estimator $\widehat{\xi}(q)$ of $\xi(q)$ belongs to Ξ_N with probability at least $1-\phi_N$. There exist sequences $r_N, r'_N, \delta_N: \quad r'_N \log^{1/2}(1/r'_N) = o(1)$, $r_N = o(N^{-1/2})$, and $\delta_N = o(N^{-1/2})$ such that the following bounds hold*

$$\sup_{r\in[0,1)} \sup_{q\in\mathcal{S}^{d-1}} (\partial_r^2 \mathbb{E} g(W, q, r(\xi(q) - \xi_0(q)) + \xi_0(q))) \leq r_N.$$

(2) The following conditions hold for the function class

$$\mathcal{F}_\xi = \{g(W, q, \xi(q)), q \in \mathcal{S}^{d-1}\} \tag{C.8}$$

There exists a measurable envelope function $F_\xi = F_\xi(W)$ that almost surely bounds all elements in the class $\sup_{q\in\mathcal{S}^{d-1}} |g(W,q,\xi(q))| \leq F_\xi(W)$ a.s.. There exists $c > 2$ such that $\|F_\xi\|_{L_{P,c}} := \left(\int_{w\in\mathcal{W}} (F_\xi(w))^c\right)^{1/c} < \infty$. There exist constants a, v that do not depend *on N such that the uniform covering entropy of the function class $\mathcal{F}_\xi$ is bounded*

$$\log \sup_Q N(\varepsilon \|F_\xi\|_{Q,2}, \mathcal{F}_\xi, \|\cdot\|_{Q,2}) \leq v \log(a/\varepsilon), \quad \textit{for all } 0 < \varepsilon \leq 1. \tag{C.9}$$

Lemma C.1 (Verification of Assumption 10 (1))**.** *Under Assumptions 3, 4 and 7, the orthogonal moment equation* (B.18) *obeys Assumption 10(1).*

Proof of Lemma C.1. **Step 1** . Consider the case when $\mathcal{X}_{\text{hurt}} = \emptyset$ and the case of lower bound. Define the following quantities:

$$\begin{aligned}
\Lambda_1(W,\xi_0) &:= \frac{D}{\mu_1(X)} \cdot S \cdot Y 1_{\{Y \leq Q(p_0(X),X)\}} - \frac{1-D}{\mu_0(X)} \cdot S \cdot Y, \\
\Lambda_2(W,\xi_0) &:= Q(p_0(X),X)\left(\frac{(1-D)\cdot S}{\mu_0(X)} - s(0,X)\right) \\
\Lambda_3(W,\xi_0) &:= -Q(p_0(X),X)p_0(X)\left(\frac{D\cdot S}{\mu_1(X)} - s(1,X)\right) \\
\Lambda_4(W,\xi_0) &:= -Q(p_0(X),X)s(1,X)\left(\frac{D\cdot S\cdot 1_{\{Y\leq Q(p_0(X),X)\}}}{s(1,X)\mu_1(X)} - p_0(X)\right)
\end{aligned}$$

and observe that $g_L(W,\xi_0) = (\mu_{10}^{\text{help}})^{-1}(\sum_{k=1}^4 \Lambda_k(W,\xi_0) - \dfrac{1-D}{\mu_0(X)} \cdot S \cdot Y)$.

Step 2 . Let $\xi_0 =:= \{s_0(0,x), s_0(1,x), Q_0(u,x)\}$ be the true value of the nuisance parameter and $\xi := \{s(0,x), s(1,x), Q(u,x)\}$ be a candidate value in the neighborhood of ξ_0. Let $\partial_\alpha \mathbb{E}[g(W,\xi)|X=x], \partial_\beta \mathbb{E}[g(W,\xi)|X=x], \partial_\gamma \mathbb{E}[g(W,\xi)|X=x]$ denote the partial derivatives of $\mathbb{E}[g(W,\xi)|X=x]$ w.r.t the output of the functions $s(0,x), s(1,x)$ and $Q(u,x)$, respectively. Let $\xi_r := r(\xi - \xi_0) + \xi_0$.

Step 3 . By construction of an orthogonal moment $g_L(W,\xi)$, $\partial_t \mathbb{E}[\Lambda_1(W,\xi)|X = x], \quad t \in \{\alpha,\beta,\gamma\}$ coincides with the multipliers in bias correction terms:

$$\begin{aligned}
\partial_\alpha \mathbb{E}[\Lambda_1(W,\xi_0)|X=x] &= Q_0(p_0(x),x) \\
\partial_\beta \mathbb{E}[\Lambda_1(W,\xi_0)|X=x] &= Q_0(p_0(x),x)p_0(x) \\
\partial_\gamma \mathbb{E}[\Lambda_1(W,\xi_0)|X=x] &= Q_0(p_0(x),x)s_0(1,x)f(Q(p(x),x)|x).
\end{aligned}$$

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

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

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

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

By Assumptions 3 and 7(1), all functions of x above are bounded a.s. in X. For $k=2$, $\Lambda_2(W,\xi)$ is a product of $Q(p(x),x)$ and a linear function of $s(0,x)$ at $X=x$. For $k=3$, $\Lambda_3(W,\xi)$ is a product of $Q(p(x),x)$, $p(x)$, and a linear function of $s(1,x)$ at $X=x$. Therefore,

$$
\sup_{t_1,t_2\in\{\alpha,\beta,\gamma\}} \sup_{x\in\mathcal{X}} |\partial^2_{t_1t_2}\mathbb{E}[\Lambda_k(W,\xi)|X=x] \leq, \quad k\in\{1,2,3\}. \text{ is bounded a.s.}
$$

Step 4 . Let $\rho_0(x) := Q(p_0(x),x)$ and $\rho(x)$ be some function in the neighborhood of $\rho_0(x)$. Denote $\rho_r(x) := r(\rho(x)-\rho_0(x))+\rho_0(x)$ Consider the following function

$$
\lambda(r,x) := \mathbb{E}[1_{\{Y\leq r(\rho(X)-\rho_0(X))+\rho_0(X)\}} - 1_{\{Y\leq\rho_0(X)\}}|X=x] = F(\rho_r(x)|x) - F(\rho_0(x)|x).
$$

Therefore, the first derivative $\lambda'(r) = f(\rho_r(x)|x)(r(x) - r_0(x))$ and $\lambda''(r,x) = f'(\rho_r(x)|x)(r(x) - r_0(x))^2$. By Assumption 7, $f'(\rho_r(x)|x)$ is bounded a.s. in X.

Step 5 . Conclusion. By Steps 1-4 and Assumptions 4 and 7,

$$\sup_{r\in[0,1)} |\partial^2_{rr}\mathbb{E}[g(W, r(\xi - \xi_0) + \xi_0)]| \lesssim \sup_{t_1,t_2\in\{\alpha,\beta,\gamma\}} \sup_{x\in\mathcal{X}} |\partial^2_{t_1 t_2}\mathbb{E}[g(W,\xi)|X = x](q_N^2 + q_N s_N + s_N^2) = o(N^{-1/2}).$$

Finally, the functions $\Lambda_k(W, \xi_0)$ are smooth, infinitely differentiable functions of the output of $\mu_1(x)$ and $\mu_0(x)$ on some open set in $(0,1)$ that contains the support of $\mu_1(X)$ and $\mu_0(X)$. □

Denote the conditional c.d.f as

$$F(q,t|x) := \Pr(q'Y \leq t|S = 1, D = 1, X = x).$$

and the conditional density as $f(q,t|x) = \partial_t F(q,t|x)$. The argument is given under assumption $\mathcal{X}_{\text{hurt}} = \emptyset$.

Lemma C.2 (Verification of Assumption 10(2))**.** *(1) Suppose $Y \in \mathrm{R}^d$ is a random vector with a.s. bounded coordinates. (2) There exists an integrable function $m(x)$, so that*

$$\sup_{t\in\mathrm{R}^d} |F(q_1,t|x) - F(q_2,t|x)| \leq m(x)\|q_1 - q_2\|$$

and $\inf_{q\in\mathcal{S}^{d-1}} \inf_{t\in\mathcal{R}} |f(q,t|x)| \geq \underline{C} > 0$. *Then, Assumption 10(2) holds.*

Proof of Lemma C.2. **Step 1.** The function class $\mathcal{M}$ is P-Donsker:

$$\mathcal{M} := \{X \to Q(q, p(X), X), \quad q \in \mathcal{S}^{d-1}\}$$

Since $Y \in \mathrm{R}^d$ is a.s. bounded random vector with $\|Y\| \leq C$, one can take C to be the envelope. Invoking the identity

$$
\begin{aligned}
&F(q_1, Q(q_1, p_0(x))|x) - F(q_2, Q(q_1, p_0(x))|x) \\
&+ F(q_2, Q(q_1, p_0(x))|x) - F(q_2, Q(q_2, p_0(x))|x) = p_0(x) - p_0(x) = 0,
\end{aligned}
$$

and the mean value theorem for $t_1 = Q(q_1, p_0(x)$ and $t_2 = Q(q_2, p_0(x)$:

$$F(q_2, t_1|x) - F(q_2, t_2|x) = f(q_2, \widetilde{t}|x)(t_1 - t_2)$$

observe that

$$
\begin{aligned}
|Q(q_1, p_0(x), x) - Q(q_2, p_0(x), x)| &\leq \sup_{t \in \mathcal{R}} |F(q_1, t|x) - F(q_2, t|x)| \sup_{t \in \mathcal{R}} f^{-1}(q_2, t|x) \\
&\leq C^{-1} m(x) \|q_1 - q_2\| \qquad \text{(C.10)}
\end{aligned}
$$

By Example 19.7 from van der Vaart (1998), the covering numbers of the function class $\mathcal{M}$ obey

$$N_{[]}(\varepsilon \|m\|_{P,r}, \mathcal{M}, L_r(P)) \lesssim \left(\frac{2}{\varepsilon}\right)^d, \text{ every } 0 < \varepsilon < 2.$$

Finally, since $Y \in \mathrm{R}^d$ is an a.s. bounded vector, each element of the class $\mathcal{M}$ is bounded by $\|Y\| \leq C$ a.s. , and C can be taken as the envelope of $\mathcal{M}$. Therefore, $\mathcal{M}$ is P-Donsker and obeys (C.9) with $v = d$ and $a = 2$.

Step 2. By Step 1, the function class $\mathcal{H}' = \left\{ W \to q'Y - Q(q, p_0(X), X), \quad q \in \mathcal{S}^{d-1} \right\}$ is the sum of 2 VC classes. Therefore, by Andrews (1994), $\mathcal{H}'$ is a VC class itself. Therefore, the class of indicators

$$\mathcal{H} := \left\{ W \to 1_{\{q'Y - Q(q, p_0(X), X) \leq 0\}}, \quad q \in \mathcal{S}^{d-1} \right\}.$$

is also a VC class with a constant envelope, and, therefore, P-Donsker.

Step 3 . The function class

$$\mathcal{H}_1 = \left\{ W \to \frac{D \cdot S \cdot 1_{\{q'Y \leq Q(q, p_0(X), X))\}}}{\mu_1(X)} \right\}$$

is obtained by multiplying each element of $\mathcal{H}$ by an a.s. bounded random variable $D \cdot S / \mu_1(X)$. The function class

$$\mathcal{H}_2 = \left\{ W \to Q(q, p(X), X) \left(\frac{(1-D)}{\mu_0(X)} - s(0, X \right) \right\}$$

is obtained from $\mathcal{M}$ by multiplying each element of $\mathcal{M}$ by an a.s. bounded random variable $\left(\frac{(1-D)}{\mu_0(X)} - s(0, X \right)$. The same argument applies to the function class

$$\mathcal{H}_3 = \left\{ W \to Q(q, p(X), X) p_0(X) \left(\frac{D}{\mu_1(X)} - s(1, X) \right) \right\}.$$

The function class

$$\mathcal{H}_4 = \left\{ W \to Q(q, p(X), X) s(1, X) \left(\frac{D \cdot S 1_{\{q'Y \leq Q(q, p_0(X), X))\}}}{\mu_1(X) s(1, X)} - p_0(X) \right) \right\}$$

is obtained as a product of function classes $\mathcal{M}$ and $\mathcal{H}$, multiplied by a random variable $s(1, X)$. Finally, the function class $\mathcal{F}_\xi$ in (C.8) is obtained by adding the elements of $\mathcal{H}_k, \quad k = 1, 2, 3, 4$. Since entropies obey the rules of addition and multiplication by a random variable (Andrews (1994)), the argument follows.

□

Proof of Lemma B.6. Let $\mathbb{E}_{n,k}[\cdot]$ and $\mathcal{E}_N$ be as defined in (C.1) and (C.2). Steps 1 and 2 establish the statement of the theorem under Assumption 10. Steps 3 and 4 verify the statements (1) and (2) of Assumption 10, respectively. Step 5 concludes.

$$
\begin{aligned}
\sqrt{n}|\mathbb{E}_{n,k}[g(W_i,q,\widehat{\xi}(q))-g(W_i,q,\xi_0(q))]| &\le \sqrt{n}|\mathbb{E}\big[g(W_i,q,\widehat{\xi}(q))-g(W_i,q,\xi_0(q))\big]| \\
&\quad + |\mathbb{G}_{n,k}[g(W_i,q,\widehat{\xi}(q))-g(W_i,q,\xi_0(q))]| \\
&=: |i(q)|+|ii(q)|.
\end{aligned}
$$

Step 1. Introduce the function

$$
\lambda(r) := \mathbb{E}[g(W_i,q,r(\widehat{\xi}(q)-\xi_0(q))+\xi_0(q))|\mathcal{E}_N \cup (W_i)_{\{i\in J_k^c\}}] - \mathbb{E}[g(W_i,q,\xi_0(q))].
$$

By Taylor's expansion,

$$
\lambda(r) = \lambda(0)+\lambda'(0)+\lambda''(\widetilde{r})/2, \quad \text{for some } \widetilde{r} \in (0,1).
$$

By construction, $\lambda(0)=0$. Because $g(w,q,\xi(q))$ is an orthogonal moment function, $\lambda'(0)=0$. By Assumption 10(1),

$$
|\lambda''(\widetilde{r})| \le \sup_{r\in(0,1)} |\lambda''(r)| \le r_N.
$$

Therefore, $|i(q)|$ converges to zero conditionally on data $(W_i)_{\{i\in J_k^c\}}$ and the event $\mathcal{E}_N$:

$$
\begin{aligned}
|i(q)| &:= \sup_{q\in\mathcal{S}^{d-1}} \sqrt{n}|\mathbb{E}[g(W_i,q,\widehat{\xi}(q))-g(W_i,q,\xi_0(q))|\mathcal{E}_N \cup (W_i)_{\{i\in J_k^c\}}]| \\
&\le \sup_{q\in\mathcal{S}^{d-1}} \sup_{\xi\in\Xi_n} \sqrt{n}|\mathbb{E}[g(W_i,q,\xi(q))-g(W_i,q,\xi_0(q))|\mathcal{E}_N \cup (W_i)_{\{i\in J_k^c\}}]| \\
&\le \sup_{q\in\mathcal{S}^{d-1}} \sup_{\xi\in\Xi_n} \sqrt{n}|\mathbb{E}[g(W_i,q,\xi(q))-g(W_i,q,\xi_0(q))] \\
&\le^i \sqrt{n}r_n = o(1).
\end{aligned}
$$

By Lemma 6.1 of Chernozhukov et al. (2018), the term $i(q)=O(r_n)=o(1)$ unconditionally.

Step 2. To bound the second quantity, consider the function class

$$\mathcal{F}_{\widehat{\xi}\xi_0} = \{g(W_i, q, \widehat{\xi}(q)) - g(W_i, q, \xi_0(q)), \quad q \in \mathcal{S}^{d-1}\}.$$

for some fixed $\widehat{\xi}$. By definition of the class,

$$\mathbb{E} \sup_{q \in \mathcal{S}^{d-1}} |ii(q)| := \mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{G}_{n,k}[f]|.$$

We apply Lemma 6.2 of Chernozhukov et al. (2018) conditionally on data $(W_i)_{\{i \in J_k^c\}}$ and the event $\mathcal{E}_N$ so that $\widehat{\xi}(q) = \widehat{\xi}_k$ can be treated as a fixed member of Ξ_n. The function class $\mathcal{F}_{\widehat{\xi}\xi_0}$ is obtained as the difference of two function classes: $\mathcal{F}_{\widehat{\xi}\xi_0} := \mathcal{F}_{\widehat{\xi}} - \mathcal{F}_{\xi_0}$, each of which has an integrable envelope and bounded logarithm of covering numbers. In particular, one can choose an integrable envelope as $F_{\widehat{\xi}\xi_0} := F_{\widehat{\xi}} + F_{\xi_0}$ and bound the covering numbers as:

$$\begin{aligned}\log \sup_Q N(\varepsilon \|F_{\widehat{\xi}\xi_0}\|_{Q,2}, \mathcal{F}_{\widehat{\xi}\xi_0}, \|\cdot\|) &\leq \log \sup_Q N(\varepsilon \|F_{\widehat{\xi}}\|_{Q,2}, \mathcal{F}_{\widehat{\xi}}, \|\cdot\|) + \log \sup_Q N(\varepsilon \|F_{\xi_0}\|_{Q,2}, \mathcal{F}_{\xi_0}, \|\cdot\|) \\ &\leq 2v \log(a/\varepsilon), \quad \text{for all } 0 < \varepsilon \leq 1.\end{aligned}$$

Finally, we can choose the speed of shrinkage $(r_n')^2$ such that

$$\sup_{q \in \mathcal{S}^{d-1}} \sup_{\xi \in \Xi_n} \left(\mathbb{E}[g(W_i, q, \xi(q)) - g(W_i, q, \xi_0(q))]^2\right)^{1/2} \leq r_n',$$

the application of Lemma 6.2 of Chernozhukov et al. (2018) gives with $M := \max_{i \in I_k^c} F_{\widehat{\xi}\xi_0}(W_i)$

$$\begin{aligned}\sup_{q \in \mathcal{S}^{d-1}} |ii(q)| &\leq \sup_{q \in \mathcal{S}^{d-1}} |\mathbb{G}_{n,k}[g(W_i, q, \widehat{\xi}(q)) - g(W_i, q, \xi_0(q))]| \\ &\leq \sqrt{v(r_n')^2 \log(a\|F_{\widehat{\xi}\xi_0}\|_{P,2}/r_n')} + v\|M\|_{P,c'}/\sqrt{n} \log(a\|F_{\widehat{\xi}\xi_0}\|_{P,2}/r_n') \\ &\lesssim_P r_n' \log^{1/2}(1/r_n') + n^{-1/2+1/c'} \log^{1/2}(1/r_n')\end{aligned}$$

where a constant $\|M\|_{P,c'} \leq n^{1/c'} \|F\|_{P,c'}$ for the constant $c' \geq 2$.

Step 3. Lemma C.1 verifies Assumption 10(1).

Step 4. Lemma C.2 verifies Assumption 10(2).

Step 5. Asymptotic Normality. In Lemma C.2, we have shown that the function class $\mathcal{F}_{\xi_0} = \{g(W, q, \xi(q)), \quad q \in \mathcal{S}^{d-1}\}$ is P-Donsker. By Theorem 19.14 from van der Vaart (1998), the asymptotic representation follows from the Skorohod-Dudley-Whichura representation, assuming the space $L^\infty(\mathcal{S}^{d-1})$ is rich enough to support this representation. □

Proof of Theorem B.7. The proof of Theorem B.7 follows from Lemma B.6 and the proof of Theorem 2 in Chandrasekhar et al. (2012). □

Lemma C.3. *Let Assumption 10 hold. Then*

$$\sqrt{N}\mathbb{E}_N\big(g(W_i, q, \widehat{\xi}(q)) - g(W_i, q, \xi_0(q))\big)e_i = o_P(1).$$

Proof. **Step 1.** Decompose the sample average into the sample averages within each partition:

$$\mathbb{E}_N\big(g(W_i, q, \widehat{\xi}(q)) - g(W_i, q, \xi_0(q))\big)e_i = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{n,k}\big(g(W_i, q, \widehat{\xi}(q)) - g(W_i, q, \xi_0(q))\big)e_i.$$

Since the number of partitions K is finite, it suffices to show that the bound holds on every partition:

$$\mathbb{E}_{n,k}\big(g(W_i, q, \widehat{\xi}(q)) - g(W_i, q, \xi_0(q))\big)e_i = o_P(1).$$

Let $\mathcal{E}_N := \cap_{k=1}^{K}\{\widehat{\xi}_{I_k^c} \in \Xi_n\}$. By Assumption 10 $\Pr(\mathcal{E}_N) \geq 1 - K\phi_N = 1 - o(1)$. The analysis below is conditionally on $\mathcal{E}_N$ for some fixed element $\widehat{\xi}_k \in \Xi_n$. Since the probability of $\mathcal{E}_N$ approaches one, the statements continue to hold unconditionally, which follows from the Lemma 6.1 of Chernozhukov et al. (2018).

Step 2. Consider the function class $\mathcal{F}^e_{\xi\xi_0} := \{(g(W_i,q,\xi(q)) - g(W_i,q,\xi_0(q)))e_i, \quad q \in \mathcal{S}^{d-1}\}$. The function class is obtained by the multiplication of a random element of class $\mathcal{F}_{\xi\xi_0}$ by an integrable random variable e_i. Therefore, $\mathcal{F}^e_{\xi\xi_0}$ is also P-Donsker and has bounded uniform covering entropy. The expectation of the random element of the class $\mathcal{F}^e_{\xi\xi_0}$ is bounded as:

$$\begin{aligned}\sqrt{n} \sup_{q\in\mathcal{S}^{d-1}} |\mathbb{E}[g(W_i,q,\widehat{\xi}(q)) - g(W_i,q,\xi(q))e_i|\mathcal{E}_N]| &\lesssim \sup_{\xi\in\Xi_n} \mathbb{E}[g(W_i,q,\xi(q)) - g(W_i,q,\xi_0(q))] \\ &\lesssim \sqrt{n}\mu_n = o(1).\end{aligned}$$

The variance of each element of the class $\mathcal{F}^e_{\xi\xi_0}$ is bounded as:

$$\begin{aligned}\sup_{q\in\mathcal{S}^{d-1}} \sup_{\xi\in\Xi_n} \mathbb{E}((g(W_i,q,\xi(q)) - g(W_i,q,\xi_0(q)))^0 e_i)^2 &= \sup_{q\in\mathcal{S}^{d-1}} \sup_{\xi\in\Xi_n} \mathbb{E}\big((g(W_i,q,\xi(q)) - g(W_i,q,\xi_0(q)))^0\big)^2 \mathbb{E}e_i^2 \\ &\leq 2 \sup_{q\in\mathcal{S}^{d-1}} \sup_{\xi\in\Xi_n} \mathbb{E}((g(W_i,q,\xi(q)) - g(W_i,q,\xi_0(q))))^2 \\ &\lesssim r''_n,\end{aligned}$$

where the bound follows from the conditional independence of e_i from W_i, $\mathbb{E}e_i^2 = 2$ for $e_i \sim \text{Exp}(1)$, and Assumption 10. $\square$

Proof of Lemma B.8. The difference between the bootstrap and the true support function as follows:

$$\begin{aligned}\sqrt{N}(\widetilde{\sigma}(q,\mathcal{B}) - \sigma(q,\mathcal{B})) &= \underbrace{\frac{1}{\sqrt{N}}\sum_{i=1}^N e_i\big(g(W_i,q,\widehat{\xi}(q)) - g(W_i,q,\xi_0(q))\big)}_{K_{\xi\xi_0}(q)} \\ &+ \frac{1}{\sqrt{N}}\sum_{i=1}^N e_i(g(W_i,q,\xi_0(q)) - \sigma_0(q))\end{aligned}$$

By Lemma C.3 $\sup_{q\in\mathbb{S}^{d-1}}|K_{\xi\xi_0}| = o_P(1)$. The remainder of the Theorem B.9 follows from Steps 2 and 3 of the Proof of Theorem 3 Chandrasekhar et al. (2012).

□

Proof of Theorem B.9. The proof of Theorem B.9 follows from Lemma B.8 and the proof of Theorem 4 in Chandrasekhar et al. (2012). □

C.1 Proofs for Sections 6.3

I present the argument for the lower truncated measure $f_{\mathcal{T}_L}$ defined in equation (B.39); a similar argument applies for $f_{\mathcal{T}_U}$. I will make use of the following statements. By definition of $\mathcal{T}_L(W)$,

$$\begin{aligned}\Pr(S=1 \text{ and } W\notin\mathcal{T}_L(W)|X) &= \Pr(S=1 \text{ and } W\notin\mathcal{T}_L(W)|X,D=1)\mu_1(X)\\ &\quad+\Pr(S=1|X,D=0)\mu_0(X)\\ &= p_0(X)s(1,X)\mu_1(X)+s(0,X)\mu_0(X)=s(0,X). \quad \text{(C.11)}\end{aligned}$$

For any event $D=1$, Bayes rule and (C.11) imply

$$\Pr(D=1|S=1 \text{ and } W\notin\mathcal{T}_L(W),X) = \frac{p_0(X)s(1,X)}{s(0,X)}\Pr(D=1|X)=\Pr(D=1|X) \quad \text{(C.12)}$$

Proof of Lemma B.11. Invoking Bayes rule, (C.11), and Assumption 8 gives

$$f_{\mathcal{T}_L}(x) = \mathbb{E}^{-1}[s(0,X)]s(0,x)f(x) = f_{11}(x).$$

□

Proof of Lemma B.12. By Assumption 1, D is independent of $S(1)$ and $S(0)$ given $\bar{X}$, which implies *i*. Invoking (C.12) gives *iii*. By Assumption 8, $\Pr(D=1|X)$ is a function of $\bar{X}$, and *ii* holds. Therefore, the numerators of functions $w_{11}(\bar{X})$ and $w_{\mathcal{T}_L}(\bar{X})$ are equal to each other. By Lemma B.11, the denominators of functions $w_{11}(\bar{X})$ and $w_{\mathcal{T}_L}(\bar{X})$ are equal to each other.

$\square$

Proof of Proposition B.13. **Step 1.** By Lemma B.1 (c), the functions $\beta_L(X)$ and $\beta_U(X)$ are sharp bounds on $\mathbb{E}[Y(1)-Y(0)|S(1)=S(0)=1,X]$:

$$\begin{aligned}\beta_L(X) &= \mathbb{E}_{\mathcal{T}_L}[Y|D=1,X]-\mathbb{E}_{\mathcal{T}_L}[Y|D=0,X] \leq \mathbb{E}_{11}[Y(1)-Y(0)|X] \\ &\leq \mathbb{E}_{\mathcal{T}_U}[Y|D=1,X]-\mathbb{E}_{\mathcal{T}_U}[Y|D=0,X] = \beta_U(X),\end{aligned}$$

where $\beta_L(X)$ is defined in (B.7) and $\beta_U(X)$ is defined in (B.6). By the Law of Iterated Expectations (LIE),

$$\mathbb{E}_{11}[\beta_L(X)|\bar{X}] \leq \mathbb{E}_{11}[Y(1)-Y(0)|\bar{X}] \leq \mathbb{E}_{11}[\beta_U(X)|\bar{X}].$$

Step 2. Let $w_{11}(\bar{X})$ and $w_{\mathcal{T}_*}(\bar{X})$ be the weighting functions defined in Lemma B.12. The following statements hold:

$$\begin{aligned}\beta_L^1 &=^{i} \mathbb{E}_{\mathcal{T}_L} w_{\mathcal{T}_L}(\bar{X})\left(\mathbb{E}_{\mathcal{T}_L}\left[\mathbb{E}_{\mathcal{T}_L}[Y|D=1,X]-\mathbb{E}_{\mathcal{T}_L}[Y|D=0,X]\right]\middle|\bar{X}\right) \\ &=^{ii} \mathbb{E}_{\mathcal{T}_L} w_{\mathcal{T}_L}(\bar{X})\mathbb{E}_{\mathcal{T}_L}[\beta_L(X)|\bar{X}] \\ &=^{iii} \mathbb{E}_{11} w_{\mathcal{T}_L}(\bar{X})\mathbb{E}_{11}[\beta_L(X)|\bar{X}] \\ &\leq^{iv} \mathbb{E}_{11} w_{11}(\bar{X})\mathbb{E}_{11}[Y(1)-Y(0)|\bar{X}] \\ &\leq^{v} \mathbb{E}_{11} w_{11}(\bar{X})\mathbb{E}_{11}[\beta_U(X)|\bar{X}]\end{aligned}$$

$$\begin{aligned}&=^{vi} \mathbb{E}_{\mathcal{T}_U} w_{\mathcal{T}_U}(\bar{X})\mathbb{E}_{\mathcal{T}_U}[\beta_U(X)|\bar{X}] \\ &=^{vii} \mathbb{E}_{\mathcal{T}_U} w_{\mathcal{T}_U}(\bar{X})\left(\mathbb{E}_{\mathcal{T}_U}[Y|D=1,\bar{X}]-\mathbb{E}_{\mathcal{T}_U}[Y|D=0,\bar{X}]\right) =^{viii} \beta_U^1,\end{aligned}$$

where *i* and *viii* follow from Lemma B.10, *ii* and *vii* follow from (B.7) and (B.6), *iii* and *vi* follow from Lemmas B.11 and B.12, and *iv* and *v* come from Step 1.

□

Proof of Lemma B.14. Step 1. Validity of (B.44). Observe that

$$\begin{aligned}\mathbb{E}[g_1^L(W,\xi_0)] &=^{i} (\mu_{10}^{-1})\mathbb{E}[[w_{\mathcal{T}_L}(\bar{X})\beta_L(X)s(0,X)]+0\\ &=^{ii} \mathbb{E}_{\mathcal{T}_L}[w_{\mathcal{T}_L}(\bar{X})\beta_L(X)]\\ &=^{iii} \beta_1^L,\end{aligned}$$

where *i* follows from $\alpha_L(W;\xi_0)$ being the sum of three functions that are mean zero conditional on X, *ii* follows from Lemma B.11, and *iii* follows from Lemma B.10.

Step 2. Orthogonality of (B.44). Observe that $(\mu_{10}^{-1})\alpha_L(W;\xi)$ is the bias correction term for $(\mu_{10}^{-1})\beta_L(X)s(0,X)$, the original (non-orthogonal) component of (B.18). Therefore, the bias correction term for $(\mu_{10}^{-1})w_{\mathcal{T}_L}(\bar{X})\beta_L(X)s(0,X)$ is equal to $(\mu_{10}^{-1})w_{\mathcal{T}_L}(\bar{X})\alpha_L(W;\xi)$ (Newey (1994)).

Step 3. Asymptotic linearity of (B.44). From the proof of Lemmas C.1 and C.2, the moment equation (B.18) obeys Assumption 10. The weighting function $w_{\mathcal{T}_L}(\bar{X})$ is bounded a.s.. Therefore, (B.44) obeys Assumption 10 and equation (B.48) holds.

Step 4. From the proof of Lemmas C.1 and C.2, the moment equation (B.18) obeys Assumption 10. As shown in Newey (1994) and Chernozhukov et al. (2016), the series estimators have the low-bias property that implies the last statement of the Lemma.

□

Proof of Lemma B.16. **Step 1**. In the proof of Lemma B.16, I showed that

$$f_{11}(x) = (\mathbb{E}[s(0,X)])^{-1}s(0,x)f(x).$$

Step 2. Observe that

$$
\begin{aligned}
&\Pr(S=1 \text{ and } W\notin\Lambda_L(W)|X)\\
&=\Pr(S=1 \text{ and } W\notin\Lambda_L(W)|X,Z=1)\mu_1(X)+\Pr(S=1|X,Z=0)\mu_0(X)\\
&=\Pr(S=1 \text{ and } W\notin\Lambda_L(W)|X,Z=1,D=1)\Pr(D=1|Z=1,X)\mu_1(X)\\
&+\Pr(S=1 \text{ and } W\notin\Lambda_L(W)|X,Z=1,D=0)\Pr(D=0|Z=1,X)\mu_1(X)\\
&+\Pr(S=1|X,Z=0)\mu_0(X)\\
&=p_0(X)s(1,X)(\Pr(D=1|Z=1,X)+\Pr(D=0|Z=1,X))\mu_1(X)+s(0,X)\mu_0(X)\\
&=s(0,X)
\end{aligned}
\tag{C.13}
$$

Invoking Bayes rule gives $f_{\Lambda_L}(x)=\mathbb{E}^{-1}[s(0,X)]s(0,x)f(x)=f_{11}(x)$. □

Proof of Lemma B.17. **Step 1.** The following equality holds:

$$
\begin{aligned}
\Pr(S(1)=S(0)=1|D=1,Z=z,X)&=\Pr(S(1)=S(0)=1|D(z)=1,Z=1,X)\\
&=^{i}\Pr(S(1)=S(0)=1|D(z),X)\\
&=^{ii}\Pr(S(1)=S(0)=1|X),
\end{aligned}
$$

where i holds by Assumption 8 and ii holds by Assumption 9(6) . Likewise, $\Pr(S(1)=S(0)=1|D=1,Z=z,X)=\Pr(S(1)=S(0)=1|X)$. Bayes rule implies

$$
\Pr_{11}(D=1|Z=z,X)=\frac{\Pr(S(1)=S(0)=1|D=1,Z=z,X)\Pr(D=1|Z=z,X)}{\Pr(S(1)=S(0)=1|Z=z,X)}=\Pr(D=1|Z=z,X),
$$

which establishes a. b follows from Assumption 8.

Step 2. When $Z = 0$, observations are not truncated. Therefore,

$$\begin{aligned}
&\Pr(S = 1, W \notin \mathcal{T}_L(W) | D = d, Z = 0, X) \\
&=^{i} \Pr(S = 1 | D = d, Z = 0, X) =^{ii} \Pr(S = 1 | Z = 0, X) \\
&=^{iii} \Pr(S(0) = 1 | Z = 0, X) = s(0, X),
\end{aligned}$$

where i follows from the definition of $\Lambda_L(W)$ in (B.49), ii follows from Assumption 9 (5), and iii follows from Assumption 9 (1).

Step 3. When $Z = 1$, observations are truncated at equal proportions in $D = 1, Z = 1$ and $D = 0, Z = 1$ groups. Therefore,

$$\begin{aligned}
&\Pr(S = 1, W \notin \Lambda_L(W) | D = d, Z = 1, X) \\
&= \Pr(W \notin \Lambda_L(W) | D = d, Z = 1, S = 1, X) \Pr(S = 1 | D = d, Z = 1, X) \\
&= p_0(X) s(1, X) = s(0, X), \quad d \in \{1, 0\}.
\end{aligned}$$

Thus $\Pr(W \notin \Lambda_L(W) | D = d, Z = z, X) = \Pr(W \notin \Lambda_L(W) | X)$ does not depend on either d or z. Bayes rule implies

$$\begin{aligned}
\Pr_{\Lambda_L}(D = 1 | Z = z, X) &= \frac{\Pr(W \notin \Lambda_L(W) | D = 1, Z = z, X) \Pr(D = 1 | Z = z, X)}{\Pr(W \notin \Lambda_L(W) | Z = z, X)} \\
&= \Pr(D = 1 | Z = z, X).
\end{aligned}$$

Step 4. Steps 1, 2, and 3 imply that

$$\begin{aligned}
&V_{11}(\mathbb{E}_{11}[D = 1 | \bar{X}, Z] | \bar{X}) = \mathbb{E}_{11}[D = 1 | \bar{X}, Z = 1] \cdot \mathbb{E}_{11}[D = 1 | \bar{X}, Z = 0] \\
&= \Pr[D = 1 | \bar{X}, Z = 1] \cdot \Pr[D = 1 | \bar{X}, Z = 0] \\
&= \Pr_{\Lambda_L}[D = 1 | \bar{X}, Z = 1] \cdot \Pr_{\Lambda_L}[D = 1 | \bar{X}, Z = 0] \\
&= V_{\Lambda_L}(\mathbb{E}_{\Lambda_L}[D = 1 | \bar{X}, Z] | \bar{X}).
\end{aligned}$$

Therefore, the numerators of $\omega_{\Lambda_L}(\bar{X})$ and $\omega_{11}(\bar{X})$ are equal to each other. By Lemma B.16, their denominators are also equal to each other. $\square$

Proof of Proposition B.18. **Step 1**. By Lemma B.1(c),

$$\begin{aligned}\beta_L(X) &= \mathbb{E}_{\Lambda_L}[Y|Z=1,X] - \mathbb{E}_{\Lambda_L}[Y|Z=0,X] \leq \mathbb{E}_{11}[(D(1)-D(0))\cdot(Y(1)-Y(0))|X] \\ &\leq \mathbb{E}_{\Lambda_U}[Y|Z=1,X] - \mathbb{E}_{\Lambda_U}[Y|Z=0,X] = \beta_U(X).\end{aligned}$$

Furthermore, $[\beta_L(X), \beta_U(X)]$ is a sharp identified set for $\mathbb{E}[Y(1)-Y(0)|X]$. **Step 2**. Conclusion.

$$\begin{aligned}\pi_L^1 &=^{i} \mathbb{E}_{\Lambda_L}\omega_{\Lambda_L}(\bar{X})\mathbb{E}_{\Lambda_L}\left[\frac{\mathbb{E}_{\Lambda_L}[Y|Z=1,X]-\mathbb{E}_{\Lambda_L}[Y|Z=0,X]}{\mathbb{E}_{\Lambda_L}[D=1|Z=1,X]-\mathbb{E}_{\Lambda_L}[D=1|Z=0,X]}\middle|\bar{X}\right] \\ &=^{ii} \mathbb{E}_{\Lambda_L}\omega_{\Lambda_L}(\bar{X})\mathbb{E}_{\Lambda_L}\left[\frac{\beta_L(X)}{\mathbb{E}_{\Lambda_L}[D=1|Z=1,X]-\mathbb{E}_{\Lambda_L}[D=1|Z=0,X]}\middle|\bar{X}\right] \\ &=^{iii} \mathbb{E}_{11}\omega_{11}(\bar{X})\mathbb{E}_{11}\left[\frac{\beta_L(X)}{\mathbb{E}_{11}[D=1|Z=1,X]-\mathbb{E}_{11}[D=1|Z=0,X]}\middle|\bar{X}\right] \\ &\leq^{iv} \mathbb{E}_{11}w_{11}(\bar{X})\mathbb{E}_{11}\left[Y(1)-Y(0)\middle|D(1)>D(0),\bar{X}\right] \\ &\leq^{v} \mathbb{E}_{11}w_{11}(\bar{X})\mathbb{E}_{11}\left[\frac{\beta_U(X)}{E_{\Lambda_U}[D=1|Z=1,X]-E_{\Lambda_U}[D=1|Z=0,X]}\middle|\bar{X}\right] \\ &=^{vi} \mathbb{E}_{\Lambda_U}\omega_{\Lambda_U}(\bar{X})\left[\frac{\mathbb{E}_{\Lambda_U}[Y|Z=1,X]-\mathbb{E}_{\Lambda_U}[Y|Z=0,X]}{\mathbb{E}_{\Lambda_U}[D=1|Z=1,X]-\mathbb{E}_{\Lambda_U}[D=1|Z=0,X]}\middle|\bar{X}\right] =^{vii} \pi_U^1,\end{aligned}$$

where *i* and *vii* follow from Lemma B.15, *ii* and *vi* follow from Lemmas B.1(c) and B.11 , *iii* follows from Lemma B.17, and *iv* follows by Step 1.

$\square$

Appendix D: Additional Simulations

Definition of the parameters. Consider the parameters in equations (7.1)-(7.2). The parameters α and $\widetilde{\sigma}$ are multiplied by 3.5 and 0.1, respectively, so that the artificial

Table D.1: Finite sample performance of oracle, basic and better Lee bounds

Panel A: Lower Bound									
	Bias			St. Dev.			Coverage Rate		
N	Oracle	Basic	Better	Oracle	Basic	Better	Oracle	Basic	Better
9,000	0.00	-0.03	-0.01	0.01	0.01	0.01	0.95	0.26	0.90
10,000	0.00	-0.03	-0.01	0.01	0.01	0.01	0.95	0.25	0.90
15,000	0.00	-0.02	-0.01	0.00	0.01	0.01	0.95	0.23	0.90
Panel B: Upper Bound									
9,000	-0.00	0.03	0.00	0.01	0.01	0.01	0.95	0.28	0.95
10,000	-0.00	0.03	0.00	0.01	0.01	0.01	0.95	0.29	0.95
15,000	-0.00	0.02	0.00	0.00	0.01	0.01	0.94	0.28	0.95

Notes. Results are based on 10,000 simulation runs. In Panel A, the true parameter value is -0.014 for the basic method, and -0.011 for all other methods. In Panel B, 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. Oracle, basic, naive and better estimated bounds cover zero in 100% of the cases.

wage's interquantile range matches its week 90 counterpart. Second, $(\gamma)_1$—the first coefficient of γ—is multiplied by 0.1 to make the classification problem into $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$ sufficiently difficult, as it could be in the real JobCorps example with 5,177 covariates. Finally, α is multiplied by 3, to make the artificial employment rate match its week 90 counterpart. The true basic identified set is the weighted average of basic Lee bounds (2.3) and (2.4), defined separately for $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$. The true sharp identified set is the output of Algorithm 1, where the direction of the treatment effect on employment is positive if the treatment-control difference in employment rates exceeds zero and is negative otherwise.

The basic method. The basic method is defined as the weighted average of basic Lee bounds, estimated on $\mathcal{X}_{\text{help}}$ and $\mathcal{X}_{\text{hurt}}$ separately. The trimming threshold is estimated as described in equation (F.1) based on all covariates.

The better method: orthogonal approach. In Section 7 of the main text, the better method is the sample average of the orthogonal moment equations (B.18). The first-stage parameters are estimated as described in Section F.1. For the employment equation, covariates are selected by post-lasso-logistic of Belloni et al. (2016). For the wage equation, covariates are selected by post-lasso of Belloni et al. (2017).

The better method: agnostic approach. In this section, I consider an alternative version of the better method based on agnostic approach. To construct the bounds, I randomly split the sample into the auxiliary part with $N/100$ observations and the main part with $99/100N$ observations. On the auxiliary part, I select three covariates. The first covariate is the one with the largest absolute value of the coefficient in the wage equation estimated by linear lasso of Belloni et al. (2017). The next two covariates are the top 2 covariates according to the importance measure of the random forest `ranger` *R* command in the employment equation. For each covariate, its importance shows the reduction in variance explained by random forest if is a given covariate is excluded. Thus, the target bounds are the sharp bounds conditional on three covariates.

Appendix E: Additional details for Section 3

C.2 JobCorps empirical application.

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 is selected by post-lasso-logistic of Belloni et al. (2016) for one of employment equation specifications, described below. Table E.1 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 E.1.

Table E.1: 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	ususal 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

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 E.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.

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 E.3: 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 E.4: 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 E.5: 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 E.6: 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 E.7: Family was on welfare when growing up

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

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

Table E.8: Health status at RA

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

Table E.9: Figure 2 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	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 2. 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 $T = \max_{j \in \{1,2\}} \widehat{\mu}_j / \widehat{\sigma}_j$, where $\widehat{\mu}_j$ and $\widehat{\sigma}_j$ 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 C.2.

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 E.10: Number of months since most recent arrest

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

Appendix F: Empirical applications: additional details

F.1 General Description of better Lee bounds

In this section, I describe how to compute the first stage estimates for better Lee bounds, which applies to all three applications under consideration. All subjects are partitioned into the sets $\widehat{\mathcal{X}}^{\text{help}} = \{X: \quad \widehat{p}(X) < 1\}$ (JobCorps helps employment) and $\widehat{\mathcal{X}}^{\text{hurt}} = \{X: \widehat{p}(X) > 1\}$ (JobCorps hurts employment) by plugging covariate vector X into either logistic or post-lasso-logistic estimate of trimming threshold $p_0(x) = s(0,x)/s(1,x)$ defined as

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

where $\Lambda(t) = \exp(t)/(1+\exp(t))$ is the logistic function, $\widehat{\alpha}$ is the baseline coefficient and $\widehat{\gamma}$ is the interaction coefficient. Treatment variable D is always included into the final logistic regression regardless of being selected by post-lasso-logistic.

For a continuous outcome, the quantile estimate $\widehat{Q}(\widehat{p}(x),x)$ is evaluated in four steps:

1. The parameter $\delta_0(u)$ from equation (B.25) is estimated by quantile regression defined in equation (B.26) 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 $\widehat{p}(x) = \mathrm{round}(\widehat{p}(x), 2)$ is rounded to 2 decimal places. $\widehat{Q}(\widehat{p}(x), x)$ is evaluated.

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

$$\phi_0(x) := \Pr(Y=1|X=x) := \Lambda(x'\delta) \tag{F.2}$$

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.

F.2 Details of Figure 4.

Observe that the support function of a circle centered at β_0 with radius R takes the form $\sigma(q) = q'\beta_0 + R$. Given an estimate of the support function $\widehat{\sigma}(q)$ evaluated at $q \in \{q_1, \dots, q_J\}$, the best circle approximation for an identified set is defined as the circle with radius $(\widetilde{\beta}, \widetilde{R})$ chosen as the minimizers of the Ordinary Least Squares problem:

$$(\widetilde{\beta}, \widetilde{R}) = \arg\min_{R,\beta} \frac{1}{J} \sum_{j=1}^{2J} (\widehat{\sigma}(q_j) - q_j'\beta - R)^2. \tag{F.3}$$

F.3 Lee (2009) empirical details

Table F.11: First-Stage Estimates, Table 3, Columns (3) and (7).

		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 bounds in Columns (3) and (7) of Table 3 . Column (2): baseline coefficient α of equation (F.1). Column (3): interaction coefficient γ of equation (F.1). Column (4): $\delta(u)$ of equation (B.26) on wage 90 $u = 0.95$-quantile in the control group (sample size = 1, 660). Column (5): $\delta(u)$ of equation (B.26) on wage 90 $u = 0.97$-quantile in the treated group (sample size = 2, 564). Covariates are defined in Section C.2. Computations use design weights.

Table F.12: First-Stage Estimates, Table 3, 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	CURRJOB	0.201	-0.044	0.036	0.085
5	EARN_YR	0.000	0.000	0.000	0.000
6	EVARRST	-0.123	0.147	-0.024	0.024
7	FEMALE	-0.23	-0.058	-0.113	-0.126
8	HASCHLD	0.425	-0.177	-0.012	0.103
9	HGC	0.036	0.026	-0.011	-0.011
10	HGC_FATH	0.013	-0.001	0.003	0.004
11	HGC_MOTH	-0.004	0.008	0.003	0.000
12	HH_INC2	0.148	-0.186	-0.032	-0.026
13	HH_INC3	0.142	-0.035	-0.013	-0.01
14	HH_INC4	0.373	-0.23	0.007	0.061
15	HH_INC5	0.276	0.036	0.077	0.151
16	HISP	-0.155	0.004	0.095	0.029
17	HRSWK_JR	-0.006	0.003	0.000	-0.003
18	MARRIED	0.339	-0.253	-0.034	-0.021
19	MOSINJOB	0.039	0.007	-0.006	0.000
20	NCHLD	-0.324	0.137	0.067	0.023
21	OTHERRAC	-0.191	-0.284	0.121	0.054
22	PERS_INC2	0.182	0.007	0.172	-0.059
23	PERS_INC3	0.200	-0.024	0.185	0.044
24	PERS_INC4	0.031	0.419	0.222	-0.14
25	SEPARATED	-0.149	-0.165	-0.084	-0.105
26	TOGETHER	-0.199	0.339	-0.026	0.014
27	WKEARNR	0.001	-0.001	0.001	0.001
28	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 3 . Column (2): baseline coefficient α in equation (F.1). Column (3):interaction coefficient γ in equation (F.1). Column (4): $\delta(u)$ from equation (B.26) on wage 90 $u = 0.95$-quantile in the control group (sample size = 1, 660). Column (5): $\delta(u)$ of equation (B.26) on wage 90 $u = 0.97$-quantile in the treated group (sample size = 2, 564). Covariates are defined in Section C.2. Computations use design weights.

Table F.13: First-Stage Estimates, Table 3, Column (4).

		Logistic		Quantile	
		Baseline coef. (α)	Interaction coef. (γ)	Control	Treated
	(1)	(2)	(3)	(4)	(5)
1	(Intercept)	-0.68	0.38	2.21	0.14
2	EARN_YR	0.00	-0.00	0.00	0.00
3	EVWORK	-0.40	-0.08	-0.02	-0.01
4	FEMALE	-0.22	-0.06	-0.13	0.01
5	HGC	0.07	-0.02	0.01	-0.00
6	HH_INC5	0.14	0.16	0.04	0.09
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	0.13	0.06	0.04
11	PERS_INC1	-0.09	-0.02	-0.01	-0.10
12	RACE_ETH2	-0.15	-0.04	-0.15	0.05
13	R_COMM1	-0.11	-0.05	0.02	-0.07
14	R_GETGED1	-0.27	-0.01	-0.05	0.04
15	R_HOME1	-0.21	-0.06	-0.04	0.04
16	WKEARNR	-0.00	0.00	0.00	0.00
17	R_GETGED1:RACE_ETH2	-0.021			
18	HGC:EVWORK	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 (4) of Table 3. Column (2): baseline coefficient α of equation (F.1). Column (3): interaction coefficient γ of equation (F.1). Column (4): $\delta(u)$ of equation (B.26) on wage 90 $u = 0.95$-quantile in the control group (sample size = 1, 660). Column (5): $\delta(u)$ of equation (B.26) on wage 90 $u = 0.97$-quantile in the treated group (sample size = 2, 564). Covariates are defined in Section C.2. Computations use design weights.

Table F.14: First-Stage Estimates, Table 3, Column (5).

		Logistic		Quantile	
		Baseline coef. (α)	Interaction coef. (γ)	Control	Treated
	(1)	(2)	(3)	(4)	(5)
1	(Intercept)	-1.31	0.46	2.14	0.17
2	AGE	0.06	-0.02	0.01	-0.01
3	BLACK	-0.22	-0.02	-0.17	0.03
4	EARN_CMP	0.00	0.00	0.01	-0.00
5	EARN_YR	0.00	-0.00	-0.00	0.00
6	FEMALE	-0.14	-0.01	-0.11	0.00
7	HGC_FATH	0.02	-0.00	0.00	0.00
8	HRWAGER	0.06	-0.01	0.00	-0.00
9	MONINED	0.00	0.00	-0.01	0.00
10	MOSINJOB	0.06	0.02	0.01	-0.01
11	MOS_ANYW	-0.02	0.00	-0.00	0.00
12	PERS_INC1	-0.13	-0.05	-0.00	-0.07
13	WKEARNR	-0.00	0.00	0.00	0.00

Notes. Table shows the first-stage logistic and quantile regression estimates that produce bounds in Column (5) of Table 3. Column (2): baseline coefficient α of equation (F.1). Column (3): interaction coefficient γ of equation (F.1). Column (4): $\delta(u)$ of equation (B.26) on wage 90 $u = 0.95$-quantile in the control group (sample size = 1, 660). Column (5): $\delta(u)$ of equation (B.26) on wage 90 $u = 0.97$-quantile in the treated group (sample size = 2, 564). Covariates are defined in Section C.2. Computations use design weights.

F.4 Angrist et al. (2002) empirical details

Table F.15: Baseline covariates in Angrist et al. (2002)

Name	Description	Name	Description
AGE2	pupil's age	SEX_NAME	gender by first name
MOM_AGE	mother's age	DAD_AGE	father's age
MOS_SCH	mother's HGC	DAD_SCH	father's HGC
MOM_AGE_IS_NA	missing mom age	DAD_AGE_IS_NA	missing dad age
MOS_SCH_IS_NA	missing mom HGC	DAD_SCH_IS_NA	missing dad HGC
DAREA$4-7,11,15-19$	zip codes	STRATA$1-4$	

Notes. For each of the four parental characteristics, missing values are replaced by zero and an additional indicator variable for having a (non)-missing record is generated. Imputation by median rather than zero leads to quantitatively similar results. Out of 19 zip codes, 9 zip codes that are not perfectly multi-collinear are selected. HGC stands for highest grade completed.

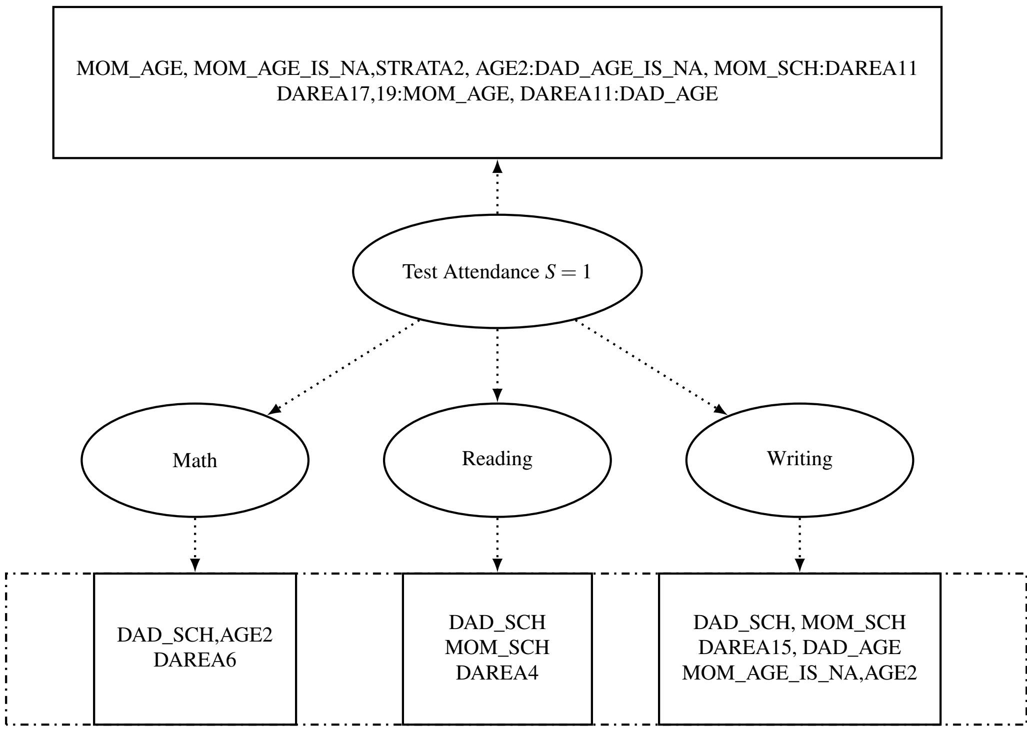
Table F.16: First-Stage Estimates, Table 4, Column (6).

		Logistic		Quantile		
		Baseline coef. (α)	Interaction coef. (γ)	Math	Reading	Writing
	(1)	(2)	(3)	(4)	(5)	(6)
1	(Intercept)	-2.041	0.442	4.884	3.123	4.651
2	STRATA2	0.669				
3	MOM_AGE_IS_NA	-3.992		1.35	0.165	-0.528
4	MOM_AGE		-0.005			
5	AGE2:DAD_AGE_IS_NA	-0.012				
6	MOM_SCH:DAREA11	0.072				
7	MOM_AGE:DAREA17	0.036				
8	MOM_AGE:DAREA19	0.016				
9	DAREA11:DAD_AGE	0.013				
10	AGE2			-0.27	-0.141	-0.199
11	DAD_SCH			0.034	0.018	0.017
12	DAREA6			-1.354	-0.14	-0.228
13	MOM_SCH			-0.032	0.013	-0.065
14	DAREA4			0.135	-0.635	-0.962
15	MOM_SCH_IS_NA			-1.239	-0.612	-1.873
16	DAD_AGE			0.012	-0.001	-0.003
17	DAREA15			-2.073	-1.381	-1.07

Notes. Table shows the first-stage logistic and quantile regression estimates that produce better Lee bounds in Table 4 (Column (6)). Column (2): estimate of baseline coefficient α of equation (F.1). Column (3): estimate of interaction coefficient γ of equation (F.1). Column (4)-(6): covariate effect $\delta(u)$ of equation (B.26) on wage 90 $u = 0.95$-quantile in the treated group for Math, Reading, and Writing.

Data description. The data set for analysis is obtained by merging the baseline dataset `aerdat4.sas7bdat` with the test score data set `tab5v1.sas7bdat` $N = 271$ on the `ID` and filtering in cases from Bogota 1995 applicant cohort that have non-missing pupil's age and gender records. The final data set has $N = 3,610$ cases. Since baseline covariates were collected three years after randomization, I focus only on the covariates whose values are pre-determined at the moment of randomization and are deemed considered exogenous by Angrist et al. (2002). Voucher is randomly assigned w.p. 0.588 that does not depend on covariates (i.e., Assumption 1(1) holds). Table 2 in Angrist et al. (2002) examines that voucher is indeed balanced across baseline characteristics.

Figure F.6: Graphical representation of covariate selection for Table 4, Column (6).



Notes. *Test participation equation*. The covariates for test participation are selected by post-lasso-logistic of Belloni et al. (2016) by regressing $S = 1$ (test attendance) on voucher receipt D, interacted with $p = 900$ covariates, obtained by interacting all 25 baseline covariates with 6^2 pairwise interactions of continuous covariates with default choice of penalty $\lambda/N = 0.038$. The selected interactions are decomposed into raw covariates to form pairwise interactions with 25 baseline covariates. Then, test participation equation is estimated by post-lasso-logistic of Belloni et al. (2016) with default choice of $\lambda/N = 0.018$. The selected covariates are listed in the top rectangle. *Test score equation*. For each subject, the covariates for test scores are selected by post-lasso of Belloni et al. (2017) by regressing Y (test score) on the baseline covaraites in the treated and the control group separately. The resulting set of covariates is the union of all covariates, across 3 subjects and 2 possibilities for treated and control groups.

F.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 F.17.

Agnostic approach: composition of $\mathcal{X}_{\mathbf{help}}$ and $\mathcal{X}_{\mathbf{hurt}}$. 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 5, 6, A.8, A.9, 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 5 and A.8: 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 (3.3), 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 A.8, Row 1, rx_any_12m). Invoking logistic lasso of Belloni et al. (2016) with $\lambda = 100$ to estimate (3.3), I select 46 pairwise interactions and break them down to 21 raw covariates. They are listed in Table F.18, Column (1).

Covariate selection for Tables 6 and A.9. 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`.

First-Stage Estimates: Selection Equation. Selection equation is

$$S = 1_{\{X'\alpha + D \cdot Z'\gamma + U > 0\}}, \tag{F.4}$$

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

$$Y = 1_{\{X'\kappa + \xi > 0\}}, \quad S = 1, Z = 0, \tag{F.5}$$

where $Y = 1$ is a binary indicator for negative ("No") answer in Table 5, Row 1. The estimate of $\phi_0(x)$ in equation (F.2) 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

$$Y = 1_{\{X'\delta + D \cdot X'\rho + \xi > 0\}}, \quad S = 1, Z = 0, \tag{F.6}$$

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 group is trimmed if a coin with success prob. $(1 - \widehat{p}(x))/\widehat{\phi}(0,x)$ turns out success. Likewise, a zero outcome in the control insured group is trimmed if a coin with success prob. $(1 - \widehat{p}(x))/\widehat{\phi}(1,x)$ turns out success.

Table F.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
edcnnp_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 $k = 1,2,\ldots,7$
ddddraXnum_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 F.18: First-Stage Estimates, Table 5, 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
ddddraXnum _2_2	-0.335		-0.009		0.000
ddddraXnum_2_3	0.458		0.147	0.175	-0.694
ddddraXnum_3_2	0.083		0.100	0.065	0.000
ddddraXnum_3_3	0.752				0.274
ddddraXnum_4_2	-0.079	-0.112	-0.009	-0.185	0.606
ddddraXnum_5_2	-0.041		-0.249	-0.271	0.154
ddddraXnum_6_2	-0.057		-0.225	-0.250	0.129
ddddraXnum_7_2	0.162		-0.237	0.553	0.000
ddddraw_sur_2	0.015	0.010	-0.087	-0.104	0.126
ddddraw_sur_3	-0.128	0.133	-0.073	-0.103	0.205
ddddraw_sur_4	-0.040	0.056	0.003	0.024	-0.094
ddddraw_sur_5	-0.053	0.002	0.089	0.093	0.078
ddddraw_sur_6	-0.110	0.047	0.068	0.052	0.213
ddddraw_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
N	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 5, Row 1. Column (2) : baseline coefficient α in equation (F.4). Column (3) : interaction coefficient γ of equation (F.4). Column (4): baseline coefficient κ in equation (F.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 (F.6) in $S = 1, Z = 0$ group (sample size = 8, 383) to estimate LATE bounds. Computations use survey weights.