

Identification and Estimation of Treatment Effects with Instrumental Variables under Data Combination

Ryan Lee*

Northwestern University
Job Market Paper

November 16, 2019

[Link to Current Version](#)

Abstract

In this paper I characterize sharp bounds on treatment effects under data combination with instrumental variables. Data combination in this paper refers to having multiple samples drawn from the same population in which observations cannot be linked across samples. I allow for subsets of the outcome, treatment, instrument and covariates to be observed across these samples. The parameters I can bound include the average treatment effect and certain policy relevant treatment effects. The sharp identified upper and lower bounds for the parameter of interest can each be expressed as the optimal value of the objective function in a linear programming problem where the coefficients are probabilities identified from the samples, under certain conditions. These conditions include standard instrumental variables assumptions allowing for heterogeneous effects, finite range of random variables, and a condition regulating which combinations of variables can be observed across samples. This identification strategy forms the basis for estimation, although estimation is not as simple as replacing the identified coefficients with sample estimates. The application to algorithmic bail reform in Philadelphia suggests that, if a freely available algorithm were used to determine pretrial release, the incarceration rate would decrease under the commonly used monotonicity assumption. The results of this application are dependent on the choice of shape restrictions one is willing to make.

*Thank you to Ivan Canay, Joel Horowitz, and Eric Auerbach for their guidance and support. This research was supported in part through the computational resources provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. All errors are my own. Email: ryanlee@u.northwestern.edu. Job Market Website: sites.northwestern.edu/ral773/

1 Introduction

Suppose a researcher is interested in the effect of a treatment on an outcome. It is common in empirical settings for this treatment not to be randomly assigned. Therefore, an instrumental variable is used to identify a local average treatment effect, which turns out to be equal to the instrumental variables estimand under some conditions. This estimand equals the covariance of the outcome with the instrument divided by the covariance of treatment with the instrument. It is immediately clear that it is not necessary to observe all three variables in the same sample to identify this ratio. If one sample contains the outcome, another sample contains the treatment, and both samples contain the instrument, then both of the covariances are identified and can be estimated using two-sample instrumental variables. This realization from [Klevmarken \(1982\)](#), [Angrist and Krueger \(1992\)](#), and [Arellano and Meghir \(1992\)](#), has allowed for empirical research when two samples are drawn from the population and for which the observations cannot be linked across samples and the observations in each sample are as just described. Here I focus on this type of data combination, where different subsets of variables are observed in multiple samples and the observations cannot be linked across samples.

In addition to two-sample instrumental variables, what else can be said about treatment effects under this type of data combination? For example, one sample has observations of the outcome and treatment and the second sample has observations of the treatment and instrument. Or, the second sample could contain observations of the outcome and the instrument. These two scenarios, although similar to the two-sample instrumental variables setting described previously, are different because the instrument is not observed in both samples. I develop general identification and estimation results for which the settings previously mentioned, including two-sample instrumental variables, are specific cases. Given the prevalence of two-sample instrumental variables in empirical settings, it is likely that researchers often encounter, and are impeded by, these settings with no identification or estimation results to apply.

Consider the following *simplified*¹ setting where the treatment, D , and instrument, Z , are in $\{0, 1\}$ and the outcome, Y , takes finitely many values in \mathbb{R} . The following equations relate the potential outcomes, Y_1 and Y_0 , the potential treatments, D_1 and D_0 , and the instrument to the outcome and treatment:

$$\begin{aligned} Y &= DY_1 + (1 - D)Y_0 \\ D &= ZD_1 + (1 - Z)D_0. \end{aligned}$$

¹I consider a more general setting in [Section 2](#) and further generalizations in the appendix.

Under a standard independence assumption of instrumental variables that $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$ in, for example, [Imbens and Angrist \(1994\)](#) and under rather unrestrictive conditions about which variables are observed in each of the samples, I first characterize the sharp identified set of the distribution of (Y_1, Y_0, D_1, D_0) . Notice that it is equivalent to characterize the sharp set for the vector of all probability masses of points in the range of (Y_1, Y_0, D_1, D_0) . [Theorem 1](#) characterizes the sharp identified set for this vector as the set of vectors satisfying a system of linear equalities and inequalities where the coefficients are either known or identified. This is the first step in showing that sharp identified upper and lower bounds for parameters, such as the average treatment effect, can each be expressed as the optimal value of the objective function in a linear programming problem.

The main contribution of this paper is to show when the possible distributions of (Y_1, Y_0, D_1, D_0) can be characterized as a system of linear equations with identified coefficients. In doing so, I allow for subsets of Y , D , and Z to be observed in multiple samples. This identification is a generalization of [Balke and Pearl \(1997\)](#). In the case where Y , D , and Z are all observed in the same sample, implying the joint distribution of (Y, D, Z) is identified, [Balke and Pearl \(1997\)](#) characterize the set of possible distributions of (Y_1, Y_0, D_1, D_0) as a system of linear equations and identify sharp bounds on the average treatment effect with a linear programming problem.

I base estimation on the identification result. Since all coefficients in the linear programming problem are known or identified, it seems natural to replace identified coefficients with sample estimates and use this new uninformed linear programming problem to estimate the parameter of interest. This approach would be problematic in important scenarios. In particular, I want to allow for the samples to have overlap in the variables observed in each sample. Since I assume that the samples are drawn from the same population, the identified marginal of the subset of variables in this overlap must be the same in both samples. It turns out that the empirical distributions of this subset of variables must also be identical across samples for this uninformed linear program to have a feasible solution. If the samples are independent, this occurs with probability going to zero. My solution is a simple two step procedure. The first step is to solve a quadratic programming problem and perform one matrix multiplication. The second step is to find the optimal value of the objective function in a new linear programming problem. I then demonstrate the consistency of this estimate.

I apply the econometric results to estimate bounds on the change in incarceration rate if felony defendants in Philadelphia were released pretrial via a freely available algorithm as opposed to the status quo of allowing for judicial discretion. This stands in contrast to recent research that estimates the effect of pretrial detention on case and future outcomes. Research on this issue typically estimates a local average treatment effect. However, if a

policy is implemented that changes how individuals are allocated to treatment, the subpopulation of individuals who are included in the local average treatment effect are not the same as those whom the policy adds to/subtracts from treatment. Therefore, the point identified parameter is not necessarily indicative of how average outcomes would change under a new policy. To assess this bail reform policy, I use court records from Philadelphia similar to [Dobbie et al. \(2018\)](#) and [Stevenson \(2018\)](#), which contain information on whether the individual is sentenced to incarceration (the outcome), whether the individual was released or detained pretrial (the treatment), and to which bail judge the defendant was quasi-randomly assigned (the instrument). In contrast to other research, I also use the State Court Processing Statistics. Observations in this sample likewise include the outcome and treatment, but not the instrument. The observations include additional covariates that are inputs to the algorithm that are not observed in the court records. The State Court Processing Statistics are anonymized to protect individuals' identities, are prohibited from being linked to other data sources, and the felony cases in the samples do not overlap.² I use an additional policy invariance assumption for the policy relevant treatment effect to satisfy the assumption on the parameter of interest. In simple terms, this states that the only way the policy changes the incarceration rate is by changing who is treated, not by changing the distribution of covariates or potential outcomes.

I present results for simulations that illustrate the benefit of using the estimation procedure developed in this paper. I discuss the estimation procedure in a more familiar setting where Y , D , and Z are all observed in the same sample and in the context of estimating the local average treatment effect, and so the identification results of this paper are not needed. However, this is a useful setting for discussing the estimation procedure because it allows for comparison to the two-stage least squares estimate. This serves not only to compare my estimation procedure to a procedure that is well understood by the reader, but also to illustrate that the estimate imposes any shape restrictions that the researcher chooses to make. For example, the monotonicity assumption $\mathbf{P}(D_1 \geq D_0) = 1$ used in [Imbens and Angrist \(1994\)](#) to identify the local average treatment effect may be imposed if it is credible in the empirical setting. The idea is that certain distributions of (Y, D, Z) are not possible given the instrumental variables assumptions, including the shape restrictions. It is possible, however, that the empirical distribution of (Y, D, Z) is one such distribution. My estimation procedure uses this additional information. The simulations suggest the estimation from this paper improves upon the precision of two-stage least squares.

²I use observations from the SCPS from May 2004, and I use observations from May 2010 in the court records.

The remainder of the paper is structured as follows: In Section 2 I show when the sharp identified bounds on a parameter of interest can be characterized as optimal values of objective functions in linear programming problems. In Section 3 I show how to consistently estimate these sharp identified bounds. In Section 4 I apply the econometrics of Sections 2 and 3 to algorithmic bail reform. In Section 5 I provide simulation evidence for the precision of the estimates. In Section 6 I conclude.

Literature Review In addition to the two-sample instrumental variables work of [Klevenmarken \(1982\)](#), [Angrist and Krueger \(1992\)](#), and [Arellano and Meghir \(1992\)](#), this paper is related to the broader literature on data combination wherein there are multiple samples and the observations are not linked across samples. Most recently, [Fan et al. \(2014\)](#) identify counterfactual distributions and treatment effects under data combination. They make the selection on observables assumption and consider a setting with two samples where the treatment is observed in each sample, but the outcome and conditioning covariates are observed separately in one of two samples. This is done through a monotone rearrangement inequality. [Cross and Manski \(2002\)](#) are interested in characterizing the “long regression,” the expectation of the outcome given two covariates, when one sample contains observations on the outcome and one covariate, and the other sample contains observations on both covariates. Assuming the conditional expectation of the outcome given the covariates is structural, i.e. the conditional mean function does not change when the distribution of the covariates is changed, [Cross and Manski \(2002\)](#) discuss characterizing the identified set for a counterfactual mean where the distribution of the covariates is altered. The application in Section 4 of this paper to policy relevant treatment effects does not rely on such an assumption. The results of [Cross and Manski \(2002\)](#) rely on earlier results from [Horowitz and Manski \(1995\)](#) for contaminated and corrupted data. For a review of the econometrics of data combination see [Ridder and Moffitt \(2007\)](#).

This paper is also related to the growing literature in econometrics using linear programming problems for identification and estimation. Most related is the previously mentioned paper by [Balke and Pearl \(1997\)](#) who characterize the sharp identified set for the average treatment effect in an instrumental variables setting using linear programming. Recent work by [Mogstad et al. \(2018\)](#) uses linear programming to characterize and estimate bounds on parameters, such as policy relevant treatment effects, that can be written as the sum of two functionals of marginal treatment response functions. [Kamat \(2018\)](#) characterizes parameters related to program access when an instrument exogenously changes choice sets with a linear fractional program.

2 Identification

In this section I establish conditions for when a linear programming problem can be used to characterize the sharp bounds on a parameter, θ , in a heterogeneous effects model with an instrumental variable under data combination. The type of data combination that I allow for is when a researcher has multiple samples each with i.i.d. observations, which cannot be linked across samples. For ease of notation, this section presents the identification argument when only one instrument is available and for two samples. It is possible that more than two samples are observed and that there are multiple instruments. Appendix A discusses these generalizations.

The observable variables in this setting include Y , D , Z , and $\{X_k\}_{k=1}^K$, which denote the outcome, treatment, instrument, and K covariates, respectively. Ideally, a single sample would have i.i.d. observations of all of these variables. This would imply that their joint distribution is identified. Then, the researcher would be able to identify the local average treatment effect or use established procedures to partially identify the average treatment effect. However, the researcher may be constrained by the realities of the available data. In particular, two data sets may be available each containing different subsets of these variables with no way of linking observations across samples. This could be the case if, for example, data are anonymized to protect individuals' identities or the data are sampled in such a way that there are no individuals/units with observations in both samples. The application in Section 4 falls into both of these categories. One data source contains anonymized observations for which it is against the data use agreement to link to observations in other samples. Additionally, the observations in the samples come from the same city, but different years, so the units do not overlap. When this is the case the marginal distributions of the variables in each sample are identified, which can still be informative about a parameter such as the average treatment effect. In this section I discuss when this type of setting can be used to bound a parameter with a linear programming problem.

Let \mathcal{D} and \mathcal{Z} denote the range of D and Z , respectively, each of which is assumed to be finite. In this heterogeneous effects framework with instrumental variables, for each value of the treatment there is a variable representing the outcome that would be observed for an individual had they received that particular value of the treatment; these variables are the potential outcomes. There is similarly a potential treatment for each value of the instrumental variable. Let $\{Y_d\}_{d \in \mathcal{D}}$ denote the potential outcomes and $\{D_z\}_{z \in \mathcal{Z}}$ denote the potential treatments. The following two equations relate the potential outcomes, potential

treatments, and instrument to the treatment and outcome:

$$\begin{aligned} Y &= \sum_d Y_d \mathbb{1}[D = d] \\ D &= \sum_z D_z \mathbb{1}[Z = z]. \end{aligned} \tag{1}$$

Theorem 1 gives conditions under which the possible distributions of potential outcomes, potential treatments, and covariates can be characterized as the set of vectors satisfying a system of linear equations with identified coefficients. With this set of linear equations, one can characterize the identified minimum or maximum of parameters such as the average treatment effect as the optimal value of the objective function in a linear programming problem in which the coefficients are identified. In addition to the conditions for Theorem 1, a researcher may want to include other assumptions to improve their bounds. This linear programming approach can allow for assumptions such as the monotonicity assumption used to identify a local average treatment effect, but these assumptions are not needed. I now introduce the assumptions needed for Theorem 1.

Assumption 1. $(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K) \perp\!\!\!\perp Z$

Assumption 2. $Y \in \mathcal{Y}$, $D \in \mathcal{D}$, $Z \in \mathcal{Z}$, $X_k \in \mathcal{X}_k$ for each k . $|\mathcal{Y}|$, $|\mathcal{D}|$, $|\mathcal{Z}|$, and $|\mathcal{X}_k|$ for each k are finite.

Assumption 3.

1. $(Y, D, Z, X_1, \dots, X_K)$ is identically distributed across samples.
2. Let ϕ_1 and ϕ_2 be functions of $(Y, D, Z, X_1, \dots, X_K)$ that select a subset of these variables. Sample 1 is of i.i.d. observations of $\phi_1(Y, D, Z, X_1, \dots, X_K)$, and Sample 2 is of i.i.d. observations of $\phi_2(Y, D, Z, X_1, \dots, X_K)$.
3. Z is observed in at least one sample.

The functions ϕ_1 and ϕ_2 dictate which subset of variables are observed in each sample. An example of such a function is $\phi_1 : \mathcal{Y} \times \mathcal{D} \times \mathcal{Z} \times \mathcal{X}_1 \times \dots \times \mathcal{X}_K \rightarrow \mathcal{Y} \times \mathcal{D} \times \mathcal{X}_2$ where $\phi_1(y, d, z, x_1, \dots, x_K) = (y, d, x_2)$.

Assumption 3.3 is the only restriction on which variables are observed in the samples. This may seem like too weak of an assumption to the reader as there is no mention of when Y or D are observed. It may help to remember that the identification result states when a linear programming problem *can* be used to characterize the sharp identified set of a parameter;

some of the cases this allows for have trivial bounds, i.e. bounds that could be obtained from the assumptions and knowing the range of the random variables without observing any data. For example, Assumption 3.3 allows for Z to be the only variable observed. If this were the case and the parameter of interest is the average treatment effect, the linear programming characterization is still valid, but yields these trivial bounds. The examples at the end of this section provide intuition for when this characterization can be useful.

Assumption 1 is a generalization of the standard assumption that $(\{Y_d\}_d, \{D_z\}_z) \perp\!\!\!\perp Z$ to additionally include covariates that may be independent of the instruments. Assumption 1 is more general because $K = 0$ (the case with no covariates that can be included as independent of the instrument), $(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K)$ collapses to $(\{Y_d\}_d, \{D_z\}_z)$. In the application in Section 4 the covariates include inputs to a pretrial risk assessment, such as age and prior interactions with the criminal justice system. Since the instrument uses which judge the individual is assigned to, and the judges serve on a rotating schedule, it is reasonable that these covariates and the instrument are independent of each other.

Assumption 3.1 and 3.2 must be carefully interpreted. Assumption 3.1 is saying that the samples are drawn from the same population distribution. Assumption 3.2 says that observations within a sample are i.i.d., but this says nothing of the independence or dependence of observations across samples. For example, if one sample contains 100 observations of (Y, D) and one contains 100 observations of (D, Z) , then it is possible for all 200 observations to be mutually independent. However, I allow for dependence of the observations across samples. For example, the observations in each sample could be from the same 100 individuals, but it is not known which observations correspond to other observations across samples. Assumption 3.2 allows for each of Y, D, Z, X_1, \dots, X_K to be observed or unobserved in each sample. The same variables are observed for each individual within a sample.

Assumption 3.3 implies that the distribution of Z is identified. This is the main requirement that needs attention when moving from one to multiple instruments. With one instrument this assumption is innocuous. When moving to a case with multiple instruments, the joint distribution of all instruments needs to be identified. Appendix A gives a discussion for when there are multiple instruments. Appendix A allows for an arbitrary number of samples and an arbitrary number of instruments across those samples.

An instrument relevance assumption about the covariance of the treatment with the instrument being positive is not needed. The linear programming characterization will be valid without this type of assumption, but how the instrument varies with the treatment affects the bounds.

Theorem 1. *[Linearity] If assumptions 1, 2, 3.1, 3.2 and 3.3 are satisfied, then the sharp identified set for the vector of probabilities masses $\mathbf{P}(\{Y_d = y_d\}_d, \{D_z = d_z\}_z, X_1 = x_1, \dots, X_K = x_K)$ are the vectors p satisfying*

$$\begin{aligned} A_1 p &= b_1 \\ \mathbf{1}' p &= 1 \\ p &\geq 0 \end{aligned}$$

where A_1 is a matrix, b_1 is a vector and the entries of each are identified probabilities.

In the theorem I state that there exists A_1 and b_1 that are identified and do not specify exactly what the entries of these matrices are. This is because I am allowing for various combinations of variables to be observed in each of the samples and do not specify the range of all of the random variables. The dimension and entries of A_1 and b_1 vary based on this. In general, b_1 is the vector which includes all identified probability masses from each of the samples. For example, if one sample contains observations of (Y, Z) and one contains observations of (Y, D) , b_1 is a vector of all probability masses of (Y, D) followed by all probability masses of (Y, Z) . The variables observed in each sample are a function of Z and $(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K)$. Clearly, the instrument and covariates can be written as the identity function of themselves. Y and D are each a function of $\{Y_d\}_d, \{D_z\}_z$ and Z using (1). A row of b_1 represents a probability of a point and the same row of $A_1 p$ represents the probability of the preimage of that point. The probability of this preimage can be written as the sum of probability masses and the independence assumption implies $\mathbf{P}(Z = z, \{Y_d = y_d\}_d, \{D_z = d_z\}_z, X_1 = x_1, \dots, X_K = x_K) = \mathbf{P}(Z = z) \mathbf{P}(\{Y_d = y_d\}_d, \{D_z = d_z\}_z, X_1 = x_1, \dots, X_K = x_K)$. Therefore, the elements of A_1 are in terms of the $\mathbf{P}(Z = z)$'s. Following the proof, I give an example of such A_1 and b_1 .

The set $\{p : A_1 p = b_1, \mathbf{1}' p = 1, p \geq 0\}$ is the sharp set for p . Fixing a p in this set and the identified distribution of Z the independence assumption implies the joint distribution of Z and $(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K)$. Using (1), this must imply the marginal distributions that are identified from each of the samples. This is by constructions of A_1 and b_1 .

Proof (of Theorem 1) *The distributions of $\phi_1(Y, D, Z, X_1, \dots, X_K)$ and $\phi_2(Y, D, Z, X_1, \dots, X_K)$ are identified because i.i.d. samples of each are observed. Let v be a point in the range of ϕ_1 .*

Let $\phi_1^{-1}(v)$ denote the preimage of v . As ϕ_1 is not one-to-one, $\phi_1^{-1}(v)$ is a set.

$$\begin{aligned}
& \underbrace{\mathbf{P}\left(\phi_1(Y, D, Z, X_1, \dots, X_K) = v\right)}_{\text{a coefficient in } b_1} \\
&= \sum_{y, d, z, x_1, \dots, x_K} \mathbb{1}\left[(y, d, z, x_1, \dots, x_K) \in \phi_1^{-1}(v)\right] \mathbf{P}(Y = y, D = d, Z = z, X_1 = x_1, \dots, X_K = x_K) \\
&= \sum_{y, d, z, x_1, \dots, x_K} \mathbb{1}\left[(y, d, z, x_1, \dots, x_K) \in \phi_1^{-1}(v)\right] \mathbf{P}(Y_d = y, D_z = d, Z = z, X_1 = x_1, \dots, X_K = x_K) \\
&= \sum_{y, d, z, x_1, \dots, x_K} \mathbf{P}(Z = z) \mathbb{1}\left[(y, d, z, x_1, \dots, x_K) \in \phi_1^{-1}(v)\right] \mathbf{P}(Y_d = y, D_z = d, X_1 = x_1, \dots, X_K = x_K) \\
&= \sum_{\{y_{d'}\}_{d'}, \{d_{z'}\}_{z'}, y, d, z, x_1, \dots, x_K} \underbrace{\mathbf{P}(Z = z) \mathbb{1}\left[(y_d, d_z) = (y, d) \ \& \ (y, d, z, x_1, \dots, x_K) \in \phi_1^{-1}(v)\right]}_{\text{coefficients in } A_1} \\
&\quad \times \underbrace{\mathbf{P}(\{Y_{d'} = y_{d'}\}_{d'}, \{D_{z'} = d_{z'}\}_{z'}, X_1 = x_1, \dots, X_K = x_K)}_{\text{element of the vector } p}
\end{aligned}$$

The second equality holds because $(Y_d = y, D_z = d, Z = z) \iff (Y = y, D = d, Z = z)$, the third equality holds by independence $(\{Y_{d'}\}_{d'}, \{D_{z'}\}_{z'}, X_1, \dots, X_K) \perp\!\!\!\perp Z$, and the final equality uses that $\mathbf{P}(Y_d = y, D_z = d, X_1 = x_1, \dots, X_K = x_K)$ is a marginal probability of $\mathbf{P}(\{Y_{d'} = y_{d'}\}_{d'}, \{D_{z'} = d_{z'}\}_{z'}, X_1 = x_1, \dots, X_K = x_K)$. The same steps hold for all v 's in the range of ϕ_1 and similarly for ϕ_2 . The left-hand-side of these equations constitute the elements of b_1 , and the elements of A_1 are sums of $\mathbf{P}(Z = z)$'s, and the rows of b_1 and A_1 correspond to the equations from above. Define b_1 to be the vector that includes $\mathbf{P}(\phi_1(Y, D, Z, X_1, \dots, X_K) = v)$ for each v and similarly for ϕ_2 , and define A_1 accordingly by the above equation with entries in terms of the $\mathbf{P}(Z = z)$'s.

The set $\{p : A_1 p = b_1, \mathbf{1}'p, p \geq 0\}$ is the sharp identified set for p . To see this, notice that given p in this set and the identified distribution of Z imply the identified distribution of each of $\phi_1(Y, D, Z, X_1, \dots, X_K)$ and $\phi_2(Y, D, Z, X_1, \dots, X_K)$. The set of identified joint distributions of $(Y, D, Z, X_1, \dots, X_K)$ are those that have the identified distributions $\phi_1(Y, D, Z, X_1, \dots, X_K)$ and $\phi_2(Y, D, Z, X_1, \dots, X_K)$. This is because I assume nothing about how the observations across samples are related. All observations across both samples could be mutually independent, in which case, nothing more about the joint distribution of $(Y, D, Z, X_1, \dots, X_K)$ than the marginals $\phi_1(Y, D, Z, X_1, \dots, X_K)$ and $\phi_2(Y, D, Z, X_1, \dots, X_K)$ is identified. \square

Example 1. Suppose that Y, D and Z are all in $\{0, 1\}$ and there are no covariates. Suppose that one sample contains i.i.d. observations of (Y, D) , and another unlinked sample contains i.i.d. observations of (Y, Z) . In terms of Assumption 3.2, $\phi_1(y, d, z) = (y, d)$ and $\phi_2(y, d, z) = (y, z)$. Therefore, the probabilities $\mathbf{P}(Y = y, D = d)$ for all $(y, z) \in \{0, 1\}^2$ and $\mathbf{P}(Y = y, Z = z)$ for all $(y, d) \in \{0, 1\}^2$ are identified. Notice that the marginal of Z is identified.

$$\underbrace{\begin{pmatrix} 1 & q_1 & q_0 & 0 & 0 & 0 & 0 & 0 & 1 & q_1 & q_0 & 0 & 0 & 0 & 0 & 0 \\ 0 & q_0 & q_1 & 1 & 0 & q_0 & q_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & q_1 & q_0 & 0 & 0 & 0 & 0 & 1 & q_1 & q_0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q_0 & q_1 & 1 & 0 & q_0 & q_1 & 1 \\ q_0 & q_0 & q_0 & q_0 & 0 & q_0 & 0 & q_0 & q_0 & 0 & q_0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & q_0 & 0 & q_0 & 0 & 0 & q_0 & 0 & q_0 & q_0 & q_0 & q_0 & q_0 \\ q_1 & q_1 & q_1 & q_1 & 0 & 0 & q_1 & q_1 & q_1 & q_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & q_1 & q_1 & 0 & 0 & 0 & 0 & q_1 & q_1 & q_1 & q_1 & q_1 & q_1 \end{pmatrix}}_{A_1} p = \underbrace{\begin{pmatrix} \mathbf{P}(Y = 0, D = 0) \\ \mathbf{P}(Y = 0, D = 1) \\ \mathbf{P}(Y = 1, D = 0) \\ \mathbf{P}(Y = 1, D = 1) \\ \mathbf{P}(Y = 0, Z = 0) \\ \mathbf{P}(Y = 1, Z = 0) \\ \mathbf{P}(Y = 0, Z = 1) \\ \mathbf{P}(Y = 1, Z = 1) \end{pmatrix}}_{b_1}$$

where q_0 and q_1 are $\mathbf{P}(Z = 0)$ and $\mathbf{P}(Z = 1)$, respectively. Additionally, p is as follows:

$$p = \begin{pmatrix} \mathbf{P}(Y_1 = 0, Y_0 = 0, D_1 = 0, D_0 = 0) \\ \mathbf{P}(Y_1 = 0, Y_0 = 0, D_1 = 0, D_0 = 1) \\ \mathbf{P}(Y_1 = 0, Y_0 = 0, D_1 = 1, D_0 = 0) \\ \mathbf{P}(Y_1 = 0, Y_0 = 0, D_1 = 1, D_0 = 1) \\ \mathbf{P}(Y_1 = 0, Y_0 = 1, D_1 = 0, D_0 = 0) \\ \mathbf{P}(Y_1 = 0, Y_0 = 1, D_1 = 0, D_0 = 1) \\ \mathbf{P}(Y_1 = 0, Y_0 = 1, D_1 = 1, D_0 = 0) \\ \mathbf{P}(Y_1 = 0, Y_0 = 1, D_1 = 1, D_0 = 1) \\ \mathbf{P}(Y_1 = 1, Y_0 = 0, D_1 = 0, D_0 = 0) \\ \mathbf{P}(Y_1 = 1, Y_0 = 0, D_1 = 0, D_0 = 1) \\ \mathbf{P}(Y_1 = 1, Y_0 = 0, D_1 = 1, D_0 = 0) \\ \mathbf{P}(Y_1 = 1, Y_0 = 0, D_1 = 1, D_0 = 1) \\ \mathbf{P}(Y_1 = 1, Y_0 = 1, D_1 = 0, D_0 = 0) \\ \mathbf{P}(Y_1 = 1, Y_0 = 1, D_1 = 0, D_0 = 1) \\ \mathbf{P}(Y_1 = 1, Y_0 = 1, D_1 = 1, D_0 = 0) \\ \mathbf{P}(Y_1 = 1, Y_0 = 1, D_1 = 1, D_0 = 1) \end{pmatrix}.$$

△

So, the form of A_1 and b_1 depend on the support of the variables and which variables are observed in each sample. Notice that in A_1 the entries are the probabilities that Z takes a particular value. Assumption 3.3 implied that these are identified. I return to Example 1 later in the paper.

Assumption 4. The parameter of interest, θ , can be written as $\mathbb{E}(f(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K))$, where $f : \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{D}^{|\mathcal{Z}|} \times \mathcal{X}_1 \times \dots \times \mathcal{X}_K \rightarrow \mathbb{R}$.

This assumption says that the parameter θ can be written as $\theta = c'p$, where c is the vector of values that f takes over its finite domain, and p is the vector of probabilities of the point

in the domain. This assumption allows for many parameters including the average treatment effect, $\mathbf{E}[Y_1 - Y_0]$, the average outcome under treatment, $\mathbf{E}(Y_1)$, and the average outcome under no treatment, $\mathbf{E}(Y_0)$. f could also be an indicator, and the parameter could be the probability of an event involving $(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K)$. In Subsection 2.1 I discuss how parameters such as the local average treatment effect can be bounded. For the application in Section 4, I show how certain assumptions allow one to identify bounds on a particular policy relevant treatment effect. In Example 2 I show how the average treatment effect fits into Assumption 4.

Example 2. Suppose that Y , D and Z are all in $\{0, 1\}$ and there are no covariates as in Example 1. p is also as in Example 1. Suppose $\theta = \mathbf{E}[Y_1 - Y_0]$. Let

$$c = \left(0 \ 0 \ 0 \ 0 \ -1 \ -1 \ -1 \ -1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \right)'$$

Then, $c'p = \mathbf{E}[Y_1 - Y_0]$. △

Beyond the assumptions already made, it may be the case that the researcher is willing to make stronger additional assumptions such as monotonicity. While such assumptions are not *needed* in order to use linear programming, they can be incorporated into the linear programming problem, and doing so can improve identified bounds.

Assumption 5. Additional assumptions the researcher makes are equivalent to linear equalities in p with known coefficients, i.e. are equivalent to $A_2p = b_2$ where A_2 & b_2 are known a known matrix and vector, respectively.

Assumption 5 accommodates various shape restrictions made in the heterogeneous effects literature such as monotonicity and monotone treatment response. This assumption can be extended to include inequalities, but for simplicity I keep the discussion to equalities. Demuyneck (2015) and Lafférs (2019) both include discussions of types of restrictions can be incorporated into linear programming. I do not cover these details fully as they are not central to the ideas covered in this paper. Note that linear programming problems with inequality constraints can be restated in standard form. The linear programming problem would then be in the vector p and slack variables.

Example 3. The common monotonicity assumption that $\mathbf{P}(D_{z'} > D_z) = 0$ for $z \geq z'$ can be written as a linear equation in p . Similarly, the monotone treatment response assumption is $\mathbf{P}(Y_{d'} > Y_d) = 0$ for $d \geq d'$ can be written as a linear equation in p . In this example I represent these two assumption in the form of Assumption 5. Additionally, I include that

the elements of p sum to one. As in Example 1, Y , D and Z are all in $\{0, 1\}$ and there are no covariates. The elements of p are ordered as in Example 1. The first row of what follows represents that the probabilities sum to one, the second row represents the monotonicity assumption, and the third row represents monotone treatment response.

$$\begin{aligned} & \begin{pmatrix} \sum_{y_1, y_0, d_1, d_0} \mathbf{P}(Y_1 = y_1, Y_0 = y_0, D_1 = d_1, D_0 = d_0) \\ \sum_{y_1, y_0} \mathbf{P}(Y_1 = y_1, Y_0 = y_0, D_1 = 0, D_0 = 1) \\ \sum_{d_1, d_0} \mathbf{P}(Y_1 = 0, Y_0 = 1, D_1 = d_1, D_0 = d_0) \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ \mathbf{P}(D_1 = 0, D_0 = 1) \\ \mathbf{P}(Y_1 = 0, Y_0 = 1) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \begin{cases} \text{Probabilities sum to one} \\ \text{Monotonicity} \\ \text{Monotone Treatment Response.} \end{cases} \end{aligned}$$

The following represents these three restrictions as it relates to Assumption 5:

$$\underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}}_{A_2} p = \underbrace{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}}_{b_2}.$$

△

Theorem 2. *Under assumptions 1, 2, 3.1, 3.2, 3.3, 4, and 5 the sharp identified lower bound for θ , θ_{LB} , can be characterized as follows:*

$$\begin{aligned} \theta_{LB} &= \min_p c'p \\ \text{such that } A_1 p &= b_1 \\ A_2 p &= b_2 \\ p &\geq 0 \end{aligned}$$

where A_1 and b_1 are identified from the data and A_2 and b_2 are known. The upper bound, θ_{UB} , can be written similarly with a max instead of a min.

This result follows immediately from Theorem 1 and Assumptions 4 and 5. Notice additionally that the set $\{p : A_1 p = b_1, A_2 p = b_2, p \geq 0\}$ is a convex set. Because θ_{LB} and θ_{UB} characterize the sharp identified upper and lower bounds, the sharp identified set is the interval $[\theta_{LB}, \theta_{UB}]$.

2.1 Extensions

Conditioning Covariates Suppose that Assumption 1 that $(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K) \perp\!\!\!\perp Z$ is not plausible, but the assumption that the instrument is independent when conditioning on covariates W , i.e. $(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K) \perp\!\!\!\perp Z|W$, holds. I can extend to this setting if I assume that W takes finitely many values, W is observed in every sample, and change the assumption that the distribution of Z is identified to the conditional distribution of $Z|W$ being identified. The parameter of interest is still $\mathbf{E}[f(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K)]$. Since W is observed in each sample, sharp bounds for $\mathbf{E}[f(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K)|W = w]$ can be characterized as the optimal value of the objective function in a linear programming problem for each value of W . The distribution of W is identified, so these bounds can then be used to characterize the identified set for $\mathbf{E}[f(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K)]$.

Conditional Expectations I can allow for conditional expectations such as the local average treatment effect, $\mathbf{E}[Y_1 - Y_0|D_1 = 1, D_0 = 0]$, using linear-fractional programming similar to the approach of [Kamat \(2018\)](#), he considers other parameters in a different setting. This conditional expectation can be written as $\frac{\mathbf{E}[(Y_1 - Y_0)\mathbb{1}[D_1=1, D_0=0]]}{\mathbf{E}[\mathbb{1}[D_1=1, D_0=0]]}$. Instead of Assumption 4, if the parameter can be written as a fraction, where the numerator and denominator individually satisfy the condition in Assumption 4, then the identified upper and lower bounds can be characterized as a linear-fractional program with an additional relevance assumption, so that the denominator is strictly positive. Linear fractional programs can be rearranged into linear programs as first shown in [Charnes and Cooper \(1962\)](#), and therefore, can be easily solved.

2.2 Examples

In this subsection I enumerate some of the possible combinations of variables that could be observed in unlinked samples to which the identification approach applies. The following table enumerates all possible combinations of variables that could occur across two samples where each sample contains a subset of Y , D , and Z . The idea is to highlight some of the more interesting settings. Remember that the only restriction on the subsets of variables in the samples is that Z must be observed at least once. Some of the possibilities this allows for are less interesting than others. For example, if the variables in one sample are a subset of the variables in the other sample, using both samples does not improve the identified bounds over just using the sample with more variables. Also, if Z is the only variable in a sample, that sample would not improve the identified bounds over the just using the other sample. Additionally, one would like each of Y , D , and Z to be observed at least once if

the parameter to be bounded is the average treatment effect. Five settings, as shown in Table 1, do not fall into any of these three categories. In Example 4 I focus on some of these settings. When allowing for more than two samples, allowing for more than one instrument, and allowing for covariates, the number of combinations can grow quite large. This gives a sense for how many settings a researcher may come across and may be interested in using the results of this paper.

		Sample 1						
		(Y)	(D)	(Z)	(Y, D)	(Y, Z)	(D, Z)	(Y, D, Z)
Sample 2	(Y)	a, c						
	(D)	c	a, c					
	(Z)	b, c	b, c	a, b, c				
	(Y, D)	a, c	a, c	b	a, c			
	(Y, Z)	a, c		a, b, c		a, c		
	(D, Z)	a, c	a, c	a, b, c			a, c	
	(Y, D, Z)	a	a	a, b	a	a	a	a

Table 1. The section is blacked out due to symmetry. a) Denotes that the variables in one sample are a subset of the other. b) Denotes that Z is the only variable in one of the samples. c) Denotes that at least one of the three variables is not observed

I explain several points with these examples. Example 4 illustrates how the identified set shrinks when both samples are used as opposed to just one sample. Examples in Appendix F illustrates other points, such as how the full independence assumption improves upon bounds over the mean independence assumption with multiple samples with and without the monotonicity assumption.

Example 4. Suppose that each of Y , D , and Z are in $\{0, 1\}$, $\mathbf{P}(Y_1=1, Y_0=0, D_1=1, D_0=0) = 1$, and $\mathbf{P}(Z=0) = \mathbf{P}(Z=1) = \frac{1}{2}$. Table 2 illustrates the improvement of the identified bounds when both samples are used together as opposed to using each sample individually. This is evidenced by the fact that under this distribution the average treatment effect is point identified under each of the three combinations of samples considered in Table 2. If only the information in one of the two samples is used, the bounds are larger and often trivial. Details for Table 2 can be found in Appendix F.2.

	Identified Set for ATE		
Sample(s)	$(Y, Z), (D, Z)$	(Y, Z)	(D, Z)
Identified Set	$\{1\}$	$[-1, 1]$	$[-1, 1]$
Sample(s)	$(Y, D), (Y, Z)$	(Y, D)	(Y, Z)
Identified Set	$\{1\}$	$[0, 1]$	$[-1, 1]$
Sample(s)	$(Y, D), (D, Z)$	(Y, D)	(D, Z)
Identified Set	$\{1\}$	$[0, 1]$	$[-1, 1]$

Table 2. The first column of this table show the identified set for various combinations of variables observed in two different i.i.d. samples. The second and third columns show the identified set when only one of the two samples is observed.

△

Example 4 is quite idealized; the identified set will not be a point in typical examples. However, one can anticipate an improvement in many settings. I chose this example for the simplicity of the data generating process.

3 Estimation

In this section I develop an estimation procedure based on the identification result in the previous section. It may seem natural to define a linear program which replaces identified probabilities in Theorem 2 with sample analogues. However, this is problematic as I illustrate subsequently. Before I do so, I introduce some notation.

Notation 1. Let n_1 and n_2 denote the sample sizes of each of the two samples. All \xrightarrow{p} in this section should be understood as $n_1, n_2 \rightarrow \infty$.

Definition 1. Let \hat{A}_1 and \hat{b}_1 be the matrix and vector whose elements replace the probabilities in A_1 and b_1 with empirical probabilities.

Considering Example 1, the entries of \hat{A}_1 replace $\mathbf{P}(Z = 0)$ and $\mathbf{P}(Z = 1)$ in A_1 with $\hat{\mathbf{P}}(Z = 0)$ and $\hat{\mathbf{P}}(Z = 1)$, the empirical probabilities of $(Z = 0)$ and $(Z = 1)$ from the sample where (Y, Z) are observed. \hat{b}_1 similarly replaces the probabilities in b_1 with empirical

probabilities.

$$\hat{b}_1 = \left. \begin{array}{l} \hat{\mathbf{P}}(Y = 0, D = 0) \\ \hat{\mathbf{P}}(Y = 0, D = 1) \\ \hat{\mathbf{P}}(Y = 1, D = 0) \\ \hat{\mathbf{P}}(Y = 1, D = 1) \\ \hat{\mathbf{P}}(Y = 0, Z = 0) \\ \hat{\mathbf{P}}(Y = 1, Z = 0) \\ \hat{\mathbf{P}}(Y = 0, Z = 1) \\ \hat{\mathbf{P}}(Y = 1, Z = 1) \end{array} \right\} \begin{array}{l} \text{Empirical probabilities from the sample of } (Y, D) \\ \\ \text{Empirical probabilities from the sample of } (Y, Z) \end{array}$$

Remark: In some circumstances, Definition 1 leaves some ambiguity in exactly how the entries of A_1 and b_1 should be estimated, which did not manifest in the previous example. If Z is observed in both samples, then using all $n_1 + n_2$ observations to create the empirical probabilities estimating $\mathbf{P}(Z = z)$ suffices since $\hat{\mathbf{P}}(Z = z) \xrightarrow{p} \mathbf{P}(Z = z)$. If it is believed that all observations across both samples are mutually independent, this is the natural choice to estimate $\mathbf{P}(Z = z)$. If it is believed that all of the individuals of the first sample are observed in the second sample, and thus each observation of Z in the first sample is observed in the second sample, a better estimate would be to use only the second sample to estimate $\mathbf{P}(Z = z)$. If Z is only observed in one sample, the empirical probabilities for $Z = z$ from that sample are the natural estimators of the $\mathbf{P}(Z = z)$'s.

I illustrate here why replacing A_1 and b_1 with \hat{A}_1 and \hat{b}_1 would be problematic. To understand this, notice that $(Y, D, Z, X_1, \dots, X_K)$ is a function of $(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K)$ and Z . Each distribution of $(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K)$ and distribution of Z , implies a distribution of $(Y, D, Z, X_1, \dots, X_K)$, which in turn implies the distributions of $\phi_1(Y, D, Z, X_1, \dots, X_K)$ and $\phi_2(Y, D, Z, X_1, \dots, X_K)$. This is represented by the linear transformation of p to $A_1 p$. Replacing A_1 with \hat{A}_1 results in a similar transformation and $\hat{A}_1 p$ are the probability masses of the marginals of a distribution represented by ϕ_1 and ϕ_2 . Therefore, if there is any overlap in the variables that ϕ_1 and ϕ_2 keep, the sample marginals of these variables must be *equal* for there to be a solution to $\hat{A}_1 p = \hat{b}_1$. If the samples are independent, the probability that this is true goes to zero. In the following example I show precisely why this is true for a particular setting.

Example 1. (Continued) Make the same assumptions as in Example 1 from Section 2. Suppose one sample has n_1 i.i.d. observations of (Y, D) , and another sample has n_2 i.i.d. observations of (Y, Z) . Suppose all $n_1 + n_2$ observations are mutually independent. A necessary condition for $\{p : \hat{A}_1 p = \hat{b}_1, A_2 p = b_2, \geq 0\}$ to be nonempty, would be that the sample distribution of Y in both samples are identical—each sample distribution of Y converges to the

same population distribution of Y , but that is not enough. This set being nonempty is necessary for an estimator based on replacing identified coefficients with estimated coefficients to exist. The probability of this goes to zero as $n_1, n_2 \rightarrow \infty$ if $\text{var}(Y) \neq 0$.

For this set to be nonempty, it is clearly necessary for $\hat{A}_1 p = \hat{b}_1$ to have a solution where

$$\hat{A}_1 = \begin{pmatrix} 1 & \hat{q}_1 & \hat{q}_0 & 0 & 0 & 0 & 0 & 0 & 1 & \hat{q}_1 & \hat{q}_0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \hat{q}_0 & \hat{q}_1 & 1 & 0 & \hat{q}_0 & \hat{q}_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \hat{q}_1 & \hat{q}_0 & 0 & 0 & 0 & 0 & 0 & 1 & \hat{q}_1 & \hat{q}_0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{q}_0 & \hat{q}_1 & 1 & 0 & \hat{q}_0 & \hat{q}_1 & 1 \\ \hat{q}_0 & \hat{q}_0 & \hat{q}_0 & \hat{q}_0 & 0 & \hat{q}_0 & 0 & \hat{q}_0 & \hat{q}_0 & 0 & \hat{q}_0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{q}_0 & 0 & \hat{q}_0 & 0 & 0 & \hat{q}_0 & 0 & \hat{q}_0 & \hat{q}_0 & \hat{q}_0 & \hat{q}_0 & \hat{q}_0 \\ \hat{q}_1 & \hat{q}_1 & \hat{q}_1 & \hat{q}_1 & 0 & 0 & \hat{q}_1 & \hat{q}_1 & \hat{q}_1 & \hat{q}_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{q}_1 & \hat{q}_1 & 0 & 0 & 0 & 0 & \hat{q}_1 & \hat{q}_1 & \hat{q}_1 & \hat{q}_1 & \hat{q}_1 & \hat{q}_1 \end{pmatrix}.$$

When rows 1 & 2, rows 3 & 4, rows 5 & 7, and rows 6 & 8 are added together this new matrix is formed:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & \hat{q}_0 & \hat{q}_1 & 1 & 1 & \hat{q}_1 & \hat{q}_0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \hat{q}_1 & \hat{q}_0 & 0 & 0 & \hat{q}_0 & \hat{q}_1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & \hat{q}_0 & \hat{q}_1 & 1 & 1 & \hat{q}_1 & \hat{q}_0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \hat{q}_1 & \hat{q}_0 & 0 & 0 & \hat{q}_0 & \hat{q}_1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Notice that the 1st & 3rd rows are equal, and the 2nd & 4th rows are equal. Performing the same operation on \hat{b}_1 , the 1st row is the empirical probability of $(Y = 0)$ from the sample of (Y, D) and the 3rd row is the empirical probability of $(Y = 0)$ from the sample of (Y, Z) , and similarly for $(Y = 1)$ and the 2nd and 4th rows. Therefore, for $\hat{A}_1 p = \hat{b}_1$ to have a solution, it is necessary for the sample distribution of Y to be the same from both samples.

△

The solution to this problem is to project \hat{b}_1 onto $\{\bar{b}_1 : \hat{A}_1 p = \bar{b}_1, A_2 p = b_2, p \geq 0\}$. This new vector, when used in lieu of \hat{b}_1 does not suffer from the problem outlined previously, by construction.

Definition 2. Define $\tilde{b}_1 \equiv \hat{A}_1 \tilde{p}$ where

$$\begin{aligned} (\tilde{p}, \tilde{b}_1) &\in \arg \min_{(p, \bar{b}_1)} \left(\hat{b}_1 - \bar{b}_1 \right)' \left(\hat{b}_1 - \bar{b}_1 \right) \\ \text{such that } \hat{A}_1 p &= \tilde{b}_1 \\ A_2 p &= b_2 \\ p &\geq 0. \end{aligned}$$

Notice that substituting $\hat{A}_1 p$ for \bar{b}_1 in $(\hat{b}_1 - \bar{b}_1)' (\hat{b}_1 - \bar{b}_1)$ yields the term $(\hat{b}_1 - \hat{A}_1 p)' (\hat{b}_1 - \hat{A}_1 p) = p' \hat{A}_1' \hat{A}_1 p - 2\hat{b}_1' \hat{A}_1 p + \hat{b}_1' \hat{b}_1$, which is quadratic in p with $\hat{A}_1' \hat{A}_1$ clearly positive semidefinite. Because $\{p: A_2 p = b_2, p \geq 0\}$ is a convex set and $(\hat{b}_1 - \bar{b}_1)' (\hat{b}_1 - \bar{b}_1)$ is strictly convex in \bar{b}_1 (a choice variable), it must be that each \tilde{p} in the arg min have the same vector $\hat{A}_1 \tilde{p}$. Therefore, calculating \tilde{b}_1 amounts to solving a quadratic programming problem (picking any arbitrary minimizer \tilde{p}) and multiplying the resulting arg min by the matrix \hat{A}_1 to find \tilde{b}_1 . The approach that I take is to replace A_1 and b_1 with \hat{A}_1 and \tilde{b}_1

Definition 3. Define $\hat{\theta}_{LB}$, the estimate of θ_{LB} , as

$$\begin{aligned} \hat{\theta}_{LB} &= \min_p c' p \\ \text{s.t. } \hat{A}_1 p &= \tilde{b}_1 \\ A_2 p &= b_2 \\ p &\geq 0. \end{aligned}$$

Define $\hat{\theta}_{UB}$ similarly with a max replacing the min.

To recap the estimation procedure, using sample estimates to replace the population parameters in the linear programming problem from Theorem 2 may result in an inconsistent system of equations. I solve this problem by projecting the vector \hat{b}_1 onto the space of vectors, such that there is a solution when this vector is used in the linear programming problem. This is a simple two step procedure where the first step is to find the optimal solution in a quadratic programming problem and perform one matrix multiplication and the second step is to perform a new linear programming problem. This is related to the generalized method of moments. In GMM the true value of the parameter sets a vector of population moments equal to zero. In my setting there exists a vector p such that

$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} p - \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The difference is that p is not the parameter we are interested in and there is not a unique minimizer p in Definition 2, so we need to go through the extra linear programming step.

The prior approach of [Freyberger and Horowitz \(2015\)](#) could not be used as they give conditions when sample moments can replace population moments and the estimates exist with probability going to one. As evidenced in Example 1, empirical probabilities cannot replace population probabilities. The advantage the estimator in Definition 3 is that it surely produces an estimate, not just with probability going to one.

The problem that I address in this section is not new. [Mogstad et al. \(2018\)](#) have a similar issue in estimation. In comparison to [Mogstad et al. \(2018\)](#), the approach I take does not require the choice of a tuning parameter. In my approach I make use of the fact that my linear program is bounded, since the choice variable is positive and sums to one, which helps to avoid any tuning parameter. Both my approach and that of [Mogstad et al. \(2018\)](#) are computationally simple. Where their approach requires solving two linear programming problems, my approach requires solving a quadratic programming problem, followed by a linear programming problem. Both approaches also allow the researcher to impose shape restrictions.

Before introducing the additional assumptions used to show consistency of these estimates, consider the assumptions already made in the identification section and how those help estimation. b_1 is composed of identified probabilities and \hat{b}_1 are empirical analogues from i.i.d. samples, and so it is immediate that $\hat{b}_1 \xrightarrow{p} b_1$. Similarly, the elements of A_1 are identified probabilities and the elements of \hat{A}_1 are the empirical analogues from i.i.d. samples, so the elements of \hat{A}_1 converge in probability to the elements of A_1 . As a result of both of these convergences and that the linear programs for identification and estimation are bounded, $\tilde{b}_1 \xrightarrow{p} b_1$. These linear programs are bounded because $p \geq 0$ and $A_2 p = b_2$ is assumed to include $\mathbf{1}'p = 1$ for Theorem 3. This is shown early in the proof of Theorem 3.

Assumption 6. Let $m \equiv \text{rank}(A)$. Let $A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$ and $\hat{A} = \begin{pmatrix} \hat{A}_1 \\ A_2 \end{pmatrix}$. Let \mathcal{I} index rows of A and define $A_{\mathcal{I}}$ to be the submatrix of A that keeps that the rows in \mathcal{I} . Define $\hat{A}_{\mathcal{I}}$ similarly. Assume the following

1. $\text{rank}(\hat{A}) = m$ with probability going to one.
2. There exists \mathcal{I} such that $A_{\mathcal{I}}$ and $\hat{A}_{\mathcal{I}}$ each have m rows and $\text{rank}(A_{\mathcal{I}}) = \text{rank}(\hat{A}_{\mathcal{I}}) = m$ with probability going to one.
3. Let r be the number of rows of A_2 and b_2 that assign elements of p to zero. There exists an optimal basic feasible solution with $m - r$ strictly positive elements to the population linear programming problem in Theorem 2. ³

The r rows referred Assumption 6.3 refer to the rows such that elements of A_2 in that row are all ≥ 0 and the element of b_2 is zero. In Example 3 A_2 and b_2 implies $r = 2$.

Assumption 6 is unnecessary for the consistency result in specific cases. This assumption is needed because the entries of A_1 are estimated. However, in some cases, the linear equa-

³Because A can have rank less than the number of rows of A and the basic feasible solutions only make sense when the number rows of A and the rank of A are the same, this is referring the linear programming problem that keeps the rows of A and $\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ indexed by \mathcal{I} .

tions can be expressed in such a way that A_1 is known and only b_1 needs to be estimated. This includes when Z is observed in every sample and one is willing to assume instead that $\mathbf{P}(Z = z) > 0$. Appendix Subsection D.1 discusses this in more detail.

In Appendix G I discuss why Assumption 6.1 and 6.2 would be satisfied. In this setting much can be said about the form of A and \hat{A} . I decompose A and \hat{A} into the product of matrices, of which, properties can be demonstrated.

In Appendix D.2 I provide an alternative to Assumption 6.3 with boundedness of the dual program. In the proof of Theorem 3 I construct an upper and lower bound on $\hat{\theta}_{LB}$ and show that each converges in probability to θ_{LB} . Showing that the lower bound of $\hat{\theta}_{LB}$ converges to θ_{LB} does not require Assumption 6, but uses that the primal linear programming problem is bounded. If the dual linear programming problem is also bounded, then a similar idea holds for the upper bound on $\hat{\theta}_{LB}$ converging to θ_{LB} . This can be repeated for θ_{UB} and $\hat{\theta}_{UB}$.

Theorem 3. *If Assumptions 1, 2, 3, 4, 5, and 6 are satisfied and assuming $A_2p = b_2$ includes the restriction $\mathbf{1}'p = 1$, then $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$ and $\hat{\theta}_{UB} \xrightarrow{p} \theta_{UB}$.*

Remember that the identified set is the closed interval $[\theta_{LB}, \theta_{UB}]$. Because $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$ and $\hat{\theta}_{UB} \xrightarrow{p} \theta_{UB}$, $[\hat{\theta}_{LB}, \hat{\theta}_{UB}]$ converges in probability to $[\theta_{LB}, \theta_{UB}]$ in the Hausdorff metric.

See Appendix B for the proof of Theorem 3. This proof uses, in part, classical results of sensitivity analysis in linear programs. These results invoke Strong Duality to show that the optimal value of a linear programming problem is continuous piecewise linear in b (the vector of constants in the restrictions). Such results can be found in [Bertsimas and Tsitsiklis \(1997\)](#). One could invoke this immediately if the coefficients in the matrix A_1 were known. Because the coefficients in A_1 need to be estimated, the result is not immediate, but the proof still makes use of this result.

I should also note, this approach imposes any shape restrictions made in Assumption 5. As an illustration of this benefit, I use this estimation procedure in simulations in the Section 5 to compare it to regression estimates in a setting where a local average treatment effect is being estimated.

In Result 1 I give a sufficient condition for Assumption 6.3. $\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ cannot lie in the union of a finite number of linear spaces of dimension $m - r - 1$ or less and the dimension of the span of A is m . The solution is immediate from the definition of a basic feasible solution.

Result 1. *If $\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ is not a linear combination of strictly less than $m - r$ columns of A then 6.3 is satisfied.*

4 Application

In this section I apply the econometric theory of Sections 2 and 3 to algorithmic bail reform. The idea is to use information on felony defendants that is available at the time of arrest to determine who is released and who is detained pretrial. The algorithm is a function that maps from these variables to a decision of pretrial release or pretrial detention. The status quo is that individuals charged with a crime see a bail judge who determines, among other things, whether to release, detain, or set bail for an individual. I am interested in how average outcomes change if a freely available algorithm replaces judicial discretion in pretrial release. The outcome I focus on is whether or not an individual is sentenced to incarceration. This application follows in the previous literature using the random assignment individuals to bail judges and the tendency of the judge to detain individuals as an instrument.

This section is structured as follows: First, I formalize the parameter being estimated. Next, I discuss the algorithm used to determine pretrial release. Then, I discuss the data setting and the instrument. Finally, before presenting and discussing the results, I discuss the related literature.

As mentioned previously, I am interested in the change to the incarceration rate if judicial discretion—the status quo—were replaced with an algorithm to determine which defendants are released pretrial. This is a distinct parameter from the average treatment effect or the local average treatment effect. The policy setting that the average treatment effect would be relevant for is if one wants to know how the average outcome will change if the status quo is to detain everyone pretrial and the new policy is to release everyone pretrial, or vice versa. This is clearly not the case as a set of individuals are released under the status quo, and a different set of individuals would be released under the new policy. A similar statement can be made for the local average treatment effect. The change in the average outcomes from the status quo to the new policy is a parameter that would be helpful for a policymaker in determining whether the policy should be implemented.

Let the superscript a denote the status quo policy and let the superscript b denote the policy of algorithmic pretrial release. Y^a denotes the outcome (incarceration) under the status quo and Y^b denotes the outcome when the algorithm is used to determine pretrial release. The parameter of interest is the policy relevant treatment effect:⁴

$$\text{PRTE} = \mathbf{E}[Y^b - Y^a].$$

⁴The policy relevant treatment effect has been defined differently in different papers. For example, this follows the given definition in Heckman and Vytlacil (2001). Mogstad et al. (2018) define the policy relevant treatment effect as this same parameter, except divided by the change in the proportion of individuals that are treated.

First, notice that in this setting Y^a , the status quo outcome, is observed, so $\mathbf{E}[Y^a]$ is point identified. Therefore, it is sufficient to consider the identified set for $\mathbf{E}[Y^b]$. Before showing how this parameter fits into the form discussed in Section 2, I need to introduce additional notation and one further assumption.

Notation 2. Let $Y^a, D^a, Z^a, Y_1^a, Y_0^a, \{D_z^a\}_z, Z^a, X_1^a, \dots, X_K^a$ denote the variables under the status quo. Define the variables under the alternative policy similarly instead with the superscript b .

Notation 3. Let $d_b : \mathcal{X}_1, \dots, \mathcal{X}_K \rightarrow \{0, 1\}$ denote the algorithm to determine pretrial release.

The status quo retains the same equations relating unobserved variables to observed variables:

$$\begin{aligned} Y^a &= D^a Y_1^a + (1 - D^a) Y_0^a \\ D^a &= \sum_z D_z^a \mathbb{1}[Z^a = z]. \end{aligned}$$

Under the alternative policy, using the algorithm to determine pretrial release, the following equations relate the potential outcomes and covariates to treatments and outcomes:

$$\begin{aligned} Y^b &= D^b Y_1^b + (1 - D^b) Y_0^b \\ D^b &= d_b(X_1^b, \dots, X_K^b). \end{aligned}$$

In order to be able to say anything about the outcomes under this alternative policy b , an assumption about how the distributions of the variables are related under the different policies is needed.

Assumption 7 (Policy Invariance). $(Y_1^a, Y_0^a, X_1^a, \dots, X_K^a) \stackrel{d}{=} (Y_1^b, Y_0^b, X_1^b, \dots, X_K^b) \stackrel{d}{=} (Y_1, Y_0, X_1, \dots, X_K)$

Similar policy invariance assumptions have been made in, for example, [Heckman and Vytlacil \(2005\)](#). In terms of what this means for $\mathbf{E}[Y^b]$, this assumptions says that the only way that the mean of the outcome is affected is by changing who is treated, i.e. who is detained pretrial. The distribution of potential outcomes and covariates does not change depending on which policy is implemented.

This results in the following:

$$\begin{aligned}\mathbf{E}[Y^b] &= \mathbf{E}[d_b(X_1^b, \dots, X_K^b)Y_1^b + (1 - d_b(X_1^b, \dots, X_K^b))Y_0^b] \\ &= \mathbf{E}[d_b(X_1, \dots, X_K)Y_1 + (1 - d_b(X_1, \dots, X_K))Y_0].\end{aligned}$$

Recalling Assumption 4, $\mathbf{E}[d_b(X_1, \dots, X_K)Y_1 + (1 - d_b(X_1, \dots, X_K))Y_0]$ satisfies the conditions for parameters that can be bounded. $d_b(X_1, \dots, X_K)Y_1 + (1 - d_b(X_1, \dots, X_K))Y_0$ is a function of Y_1 , Y_0 , and X_1, \dots, X_K .

The algorithm I consider is the Public Safety Assessment (PSA), which is freely available from the Laura and John Arnold Foundation.⁵ The PSA scores individuals on how likely they are to fail to appear for a court date, how likely they are to commit a new crime, and whether they are likely to commit a new violent crime based on information that could be made available to a bail judge just after arrest, such as age and prior interactions with the criminal justice system. The score for whether an individual is likely to commit a new violent crime maps to $\{0, 1\}$ —1 indicating that the individual is likely to commit a new crime. The algorithm I consider for determining pretrial release, d_b , is this mapping for violent crime.⁶
⁷ In other words, the policy that I compare to the status quo would be to detain those that the PSA deems likely to commit a new violent crime, while releasing those that the PSA does not flag as likely to commit a new violent crime. The algorithm takes as inputs whether the current offense is violent, the age of the defendant, whether the defendant has a pending charge at the time of the offense, and whether the defendant has any prior conviction (misdemeanor or felony) and the number of prior violent felonies.

In the literature using the tendencies of judges as an instrument there are a few ways of calculating the value of the instrument. The instrument in this paper is the mean pretrial detention rate for the first bail judge that the defendant sees. In Philadelphia, there are six judges that oversee the majority of bail hearings at any one time. I include only observations that went to one of these six bail judges. Therefore, the instrument takes six unique values, and all individuals assigned to the same bail judge have the same value of their instrument. Additionally, without any shape restrictions being made, such as monotonicity, the results

⁵For full details see: [LJAF \(2019\)](#)

⁶This application could be repeated using a cutoff rule over the three scores that are given by the PSA. The scores for failure to appear, and new criminal activity take values 1, 2, ...6, so using new violent criminal activity is somewhat less arbitrary.

⁷The LJAF states that this is not intended as a replacement for judicial discretion, but to assist judges in making their decisions. However, in order to say anything about how this affects outcomes, one must make an assumption about how the algorithm is used. This sentiment is shared in [Kleinberg et al. \(2017\)](#). This section could be repeated using similar algorithms.

of identification and estimation are equivalent to treating the instrument as a categorical variable for which judge the defendant was assigned. When I incorporate shape restrictions, such as monotonicity, all that matters is the order of the values of the instrument. The reader may be concerned that the values of the instrument are estimated. However, if no two bail judges have the exact same true propensity to detain individuals pretrial, then because the estimates of this true propensity are consistent, the ordering is correct with probability going to one. The results are reported for various combinations of shape restrictions.

The application relies on two samples: the first sample is a set of court records from felony defendants in Philadelphia and the second sample is the State Court Processing Statistics (SCPS) maintained by the Bureau of Justice Statistics [BJS \(2014\)](#). The SCPS are anonymized and prohibited through a data use agreement from being linked to other data sources. Additionally, the SCPS statistics and court records do not have any overlap in observation as the court records available begin in 2009 and the most recent year Philadelphia is sampled in the SCPS is in May of 2004.⁸ So that the gap in years is as short as possible between the two samples, I use the May 2004 observations from Philadelphia and observations in the court records of those arrested in May 2010—using only May in case there is any seasonality in criminal activity. The court records contain observations at the case level for those charged with a felony. Among other variables, I observe the name of the bail judge for each bail hearing, type of bail and the defendants’ bail statuses, and information on sentencing for each of the charges. Additionally, the court records contain only some of the inputs into the algorithm determining pretrial release d_b , such as the age of the defendant. However, not all of the inputs into the algorithm are in the court records. This is where the SCPS are useful. The SCPS contain the inputs into the algorithm as well as the treatment and outcome, but not the instrument. Mapping this to the notation of Assumption 3.2 of Section 2, the observations in the court records are of $\phi_1(Y, D, Z, X_1, \dots, X_K)$ and the observations in the SCPS are of $\phi_2(Y, D, Z, X_1, \dots, X_K)$. $\phi_1(y, d, z, x_1, \dots, x_K) = (y, d, z, x_1, \dots, x_{K_1})$ where K_1 is the number of covariates in the court records. $\phi_2(y, d, z, x_1, \dots, x_K) = (y, d, x_1, \dots, x_K)$. The covariates, being information about the defendant that is determined before the bail judge is assigned such as age and prior criminal history, should independent of which judge the defendant sees as they serve on a rotating schedule.

The application is most similar to that of [Kleinberg et al. \(2017\)](#). They use machine learning to predict crime, as proxied by failure to appear and/or rearrest, of those arrested in New York. There are two important econometric distinctions between my application and

⁸In even years from 1990 to 2006 and in 2009 the BJS sampled large urban counties and within those selected counties sampled some of the felony defendants arrested in May of that year. For Pennsylvania court records, the standard for bulk electronic data requests is limited to the most recent 10 year time frame.

that of [Kleinberg et al. \(2017\)](#). First, I am taking the algorithm to make bail decisions as given. Second, in addition to using court records, I am using the State Court Processing Statistics, which cannot be linked. If the methodology of [Kleinberg et al. \(2017\)](#) were applied to my data setting in Philadelphia, one would only use the court records and not use the State Court Processing Statistics, which would limit which algorithms are possible, i.e. an algorithm would not be able to take as inputs the covariates observed only in the State Court Processing Statistics. The Pretrial Safety Assessment uses such covariates.

Table 3 displays the results using the estimation procedure described in Section 3. Upper and lower bounds for $\mathbb{E}[Y^b - Y^a]$, which is the change in the incarceration rate if the algorithm described in Appendix C where to replace the status quo, are reported.

Table 3 reports four sets of results. Each set of results imposes a different combination of the monotonicity and monotone treatment response assumptions. In words, the monotonicity assumption says that when judges are ordered from lowest to highest propensity to detain individuals pretrial that if an individual were to be detained by any particular judge, they would also be detained by any judge with a higher propensity to detain. This assumption is dubious in this setting as judges likely place varying importance on various factors relating to the individual and of the case. For example, one judge may only take into account the defendant’s criminal history, whereas another judge may also consider the defendant’s demeanor in court. Monotone treatment response, on the other hand, is more credible. In the context of this application, monotone treatment response says that with probability zero an individual is sentenced to incarceration when detained pretrial but is not sentenced to incarceration when released pretrial. One mechanism for the effect of this treatment on the outcomes discussed in [Dobbie et al. \(2018\)](#) is that the prosecution has leverage for an individual to accept an unfavorable plea deal while being detained until their trial begins.

	(1)	(2)	(3)	(4)
Assumptions				
Monotonicity		✓		✓
Monotone Treatment Response			✓	✓
Estimates				
Lower Bound	-0.3621	-0.2740	-0.2258	-0.0882
Upper Bound	0.1378	-0.0230	0.0758	-0.0193

Table 3. This table reports the estimated upper and lower bounds of the change in the incarceration rate of those charged with a felony in Philadelphia if the algorithm described in this section were to replace the status quo of judicial discretion in determining pretrial release. Each column presents estimates using different combinations of shape restrictions.

Both sets of results for which the monotonicity assumption was not made, (1) and (3), include 0 in the estimate of the identified set. 0 being in the identified set would mean that it cannot be concluded whether replacing the status quo of allowing judges to make pretrial release decisions with this algorithm would increase or decrease the incarceration rate for felony defendants in Philadelphia county. In contrast, both estimated sets for which the monotonicity assumption is made, (2) and (4), do not contain 0 in the estimate for the identified set since the estimated upper bounds are negative. This does not necessarily mean that 0 is not in the *true* identified set because these are just consistent estimates of the bounds, and could be below 0 as a result of random sampling error. If 0 were not in the identified set, this would mean one could conclude that the algorithm would decrease the incarceration rate.

5 Use of Estimator for LATE and Simulations

In this section I compare the estimator discussed in the previous section via simulations. I compare the use of the estimator defined in Definition 3 to estimate parameters in a local average treatment effects (LATE) setting because there exists well established estimates that I can compare to, namely the ordinary least squares estimate for the reduced form and the two stage least squares estimate (TSLS) for the LATE. Before presenting the results of simulations, I discuss how the use of the estimate in Definition 3 relates to that of the regression estimate in this context. I refer to the estimate in Definition 3 in this section as the shape restricted estimate. The parameters being estimated are point identified and the estimates are points (the min and max in Definition 3 are equal). In this section I assess how imposing shape restriction in LATE via Definition 3 improves the precision of estimates relative to OLS or TSLS.

For this section I take D and Z to be in $\{0, 1\}$, Y takes finitely many values, and I observe an i.i.d. sample of (Y, D, Z) . I additionally make the monotonicity assumption of LATE and that $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$. Notice the following:

$$\text{LATE} = \mathbf{E}[Y_1 - Y_0 | D_1 = 1, D_0 = 0] = \frac{\mathbf{E}[(Y_1 - Y_0)\mathbb{1}[D_1 = 1, D_0 = 0]]}{\mathbf{P}[D_1 = 1, D_0 = 0]}. \quad (2)$$

Notice that both the numerator and the denominator are linear functionals of the vector of probabilities p of $\mathbf{P}(Y_1 = y_1, Y_0 = y_0, D_1 = d_1, D_0 = d_0)$. Therefore, $\text{LATE} = \frac{c'_{rf}p}{c'_{fs}p}$ where c_{rf} and c_{fs} are vectors. The sharp set of feasible p can be written as $\{p : A_1p = b_1, A_2p = b_2, p \geq 0\}$, since the results of Section 2 can be applied. Additionally, because the first

stage and reduced form of LATE are point identified, $\{c'_{rf}p: A_1p=b_1, A_2p=b_2, p \geq 0\}$ and $\{c'_{fs}p: A_1p=b_1, A_2p=b_2, p \geq 0\}$ are each a singleton.

Consider the regression estimate of the denominator of (2), the reduced form estimate:

$$\begin{aligned} \hat{\mathbf{E}}[Y|Z=1] - \hat{\mathbf{E}}[Y|Z=0] &= \sum_y y \left(\hat{\mathbf{P}}(Y=y|Z=1) - \hat{\mathbf{P}}(Y=y|Z=0) \right) \\ &= \sum_y y \left[\sum_d \left(\hat{\mathbf{P}}(Y=y, D=d|Z=1) - \hat{\mathbf{P}}(Y=y, D=d|Z=0) \right) \right]. \end{aligned} \quad (3)$$

It turns out that this estimate is equivalent to replacing the coefficients in the linear programming problem with the empirical probabilities *and dropping* the restriction that $p \geq 0$, i.e. it is equivalent to

$$\begin{aligned} & c'_{rf}p \\ \text{s.t. } & \hat{A}_1p = \hat{b}_1 \\ & A_2p = b_2 \end{aligned}$$

where $A_2p=b_2$ represents the monotonicity condition and that the elements of p sum to one.

Imposing the shape restrictions that $A_2p=b_2$ and $p \geq 0$ should improve the precision of the estimate, and the estimation outlined previously does impose this shape restriction. This is equivalent to imposing the sharp testable implication for which a test was developed in Kitagawa (2015). This implies the sign of the population version of the terms in the inner sum in (3). In the sample the terms in could have either sign, despite the population sign.

The following result gives conditions for which the regression estimate and the shape restricted estimate of the reduced form and first stage are equal with probability going to one

Result 2. *The regression estimate of $\mathbf{E}[(Y_1 - Y_0)\mathbb{1}[D_1=1, D_0=0]]$ is equal to the use of estimation procedure in Definition 3 with probability going to one if for each value of y*

1. $\mathbf{P}(Y_1=y, D_1=1, D_0=0) > 0$ or $\mathbf{P}(Y_1=y, D_1=1, D_0=1) = 0$, and
 2. $\mathbf{P}(Y_0=y, D_1=1, D_0=0) > 0$ or $\mathbf{P}(Y_0=y, D_1=0, D_0=0) = 0$
- in addition to the assumptions that (Y, D, Z) are i.i.d., $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$, and the monotonicity assumption $\mathbf{P}(D_1=0, D_0=1) = 0$.*

A similar result holds for $\mathbf{P}(D_1=1, D_0=0)$ replacing $\mathbf{E}[(Y_1 - Y_0)\mathbb{1}[D_1=1, D_0=0]]$. In Result 2 give conditions for when, if the estimation procedure in Section 3 were implemented,

the estimate would be the same as the regression estimate of the reduced form with probability going to one. It follows that if $\mathbf{P}(D_1=1, D_0=0) > 0$, then the ratio of these two estimates is the same as the two-stage least squares estimate of the LATE with probability going to one.

With this previous result in mind I present results comparing the root-mean-squared error of the regression estimate and shape restricted estimate.

In Table 4 I show the data generating process for the Monte Carlo simulations. For all specifications $\mathbf{P}(Z=0) = \mathbf{P}(Z=1) = 1/2$. For each specification the population reduced form $E(Y|Z=1) - E(Y|Z=0) = 0$, and therefore LATE = 0.

		$\mathbf{P}(Y=0, D=0 Z=0)$	$\mathbf{P}(Y=0, D=0 Z=1)$	$\mathbf{P}(Y=0, D=1 Z=0)$	$\mathbf{P}(Y=0, D=1 Z=1)$	$\mathbf{P}(Y=1, D=0 Z=0)$	$\mathbf{P}(Y=1, D=0 Z=1)$	$\mathbf{P}(Y=1, D=1 Z=0)$	$\mathbf{P}(Y=1, D=1 Z=1)$	$\mathbf{P}(Y=2, D=0 Z=0)$	$\mathbf{P}(Y=2, D=0 Z=1)$	$\mathbf{P}(Y=2, D=1 Z=0)$	$\mathbf{P}(Y=2, D=1 Z=1)$	$\mathbf{P}(Y=3, D=0 Z=0)$	$\mathbf{P}(Y=3, D=0 Z=1)$	$\mathbf{P}(Y=3, D=1 Z=0)$	$\mathbf{P}(Y=3, D=1 Z=1)$
Specification	1	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$
	2	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$
	3	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$

Table 4. Data generating processes for Monte Carlo simulations

For the first specification $\mathbf{P}(Y_1=y, D_1=1, D_0=0) = \mathbf{P}(Y_0=y, D_1=1, D_0=0) = 1/8 > 0$. Therefore, the estimates of the reduced form and first stage will be equal to the shape restricted estimate with probability going to one according to Result 2. In the second and third specification the data generating process is adjusted so that more shape restrictions hold with equality in the population. In the second specification I changed two of the probabilities to $\mathbf{P}(Y_1=0, D_1=1, D_0=0) = \mathbf{P}(Y_0=0, D_1=1, D_0=0) = 0$. In the third specification I additionally change the two probabilities to $\mathbf{P}(Y_1=1, D_1=1, D_0=0) = \mathbf{P}(Y_0=1, D_1=1, D_0=0) = 0$.

In Tables 5 and 6 I present Monte Carlo simulation results for the three specifications from Table 4. Table 5 presents the root-mean-square error for the estimate of the reduced form, $\mathbf{E}[Y|Z=1] - \mathbf{E}[Y|Z=0]$, and Table 6 presents results for estimates of the local average treatment effect. All simulations were performed with 10,000 Monte Carlo draws.

	Specification 1		Specification 2		Specification 3	
n	Reg.	Shape Restricted	Reg.	Shape Restricted	Reg.	Shape Restricted
100	0.2245	0.2195	0.2241	0.1773	0.2246	0.1644
500	0.0992	0.0992	0.0992	0.0792	0.0993	0.0728
1000	0.0707	0.0707	0.0705	0.0564	0.0706	0.0519
5000	0.0318	0.0318	0.0317	0.0254	0.0318	0.0233

Table 5. RMSE of the reduced form estimates. The columns labeled “Reg.” indicate the regression estimate, and the columns labeled “Shape Restricted” indicate the estimate discussed in this paper.

	Specification 1		Specification 2		Specification 3	
n	2SLS	Shape Restricted	2SLS	Shape Restricted	2SLS	Shape Restricted
500⁹	0.2004	0.2004	0.2698	0.2047	0.4199	0.2625
1000	0.1422	0.1422	0.1900	0.1465	0.2893	0.1914
5000	0.0636	0.0636	0.0847	0.0667	0.1278	0.0896

Table 6. RMSE of the local average treatment effect estimates. The columns labeled “2SLS” indicate the two-stage least squares estimate, and the columns labeled “Shape Restricted” indicate the estimate discussed in this paper.

The first specification of Tables 5 and 6 support Result 2. The estimates of the root-mean-squared error are identical for the first specification for sample sizes of 500 or larger. The first specification satisfies the assumptions for the two estimates to be identical with probability approaching one. With a small sample size there is a small improvement in RMSE by imposing the shape restrictions.

The second and third specifications do not satisfy the conditions for Result 2, so the estimates do not need to be equal with increasing probability. Additionally, some shape restrictions hold with equality in the population, so imposing the shape restriction should improve the precision of the estimate. The RMSE is about 20% less when the shape restrictions are imposed for the reduced form estimates for each sample size in specification 2. The improvement increases to slightly more than 25% for each sample size in Specification 3.

⁹ $n = 100$ simulation performed but not included in Table 6. There is a weak instrumental variables for the 2SLS estimates. One of the Monte Carlo draws had a first stage of exactly zero for $n = 100$ in Specification 3. Even after dropping that draw, the RMSE of the estimate for the LATE many orders of magnitude greater than the rest of the RMSEs in Table 6. The shape restricted estimate improves on the 2SLS estimate for each specification.

Additionally, these shape restrictions seem to help with the weak instruments problem. The first stage $\mathbf{E}(D|Z = 1) - \mathbf{E}(D|Z = 0)$ decreases when moving from Specification 1 to Specification 3. The RMSE of the estimate with shape restrictions shows a 2% to 5% increase between Specifications 1 and 2 for all samples sizes, while that of the two-stage least squares estimate increases between 33% and 35% depending on the sample size.

6 Conclusion

In this paper I have addressed an instrumental variables setting with multiple samples drawn from the same population that cannot be linked. I have shown that in this setting the sharp identified upper and lower bounds for parameters such as the average treatment effect can be characterized as the optimal value of the objective function in a linear programming problem with coefficients that are identified from the data. I have shown consistency of an estimate which only requires solving a quadratic programming problem followed by a linear programming problem. The methods developed in Sections 2 and 3 are applied to algorithmic bail reform. I estimate the identified set for the change in the incarceration rate if judicial discretion were replaced by a freely available algorithm. I provide estimates under a variety of combinations of shape restrictions, and whether zero is in this estimate of the identified set depends on the shape restrictions made.

References

- ANGRIST, J. D. AND A. B. KRUEGER (1992): “The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples,” *Journal of the American Statistical Association*, 87, 328–336.
- ARELLANO, M. AND C. MEGHIR (1992): “Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets,” *The Review of Economic Studies*, 59, 537–559.
- BALKE, A. AND J. PEARL (1997): “Bounds on Treatment Effects from Studies with Imperfect Compliance,” *Journal of the American Statistical Association*, 92, 1171–1176.
- BERTSIMAS, D. AND J. N. TSITSIKLIS (1997): *Introduction to linear optimization*, vol. 6, Athena Scientific Belmont, MA.
- UNITED STATES DEPARTMENT OF JUSTICE. OFFICE OF JUSTICE PROGRAMS. BUREAU OF JUSTICE STATISTICS (2014): “State Court Processing Statistics, 1990-2009: Felony Defendants in Large Urban Counties,” Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2014-06-24. <https://doi.org/10.3886/ICPSR02038.v5>.
- CHARNES, A. AND W. W. COOPER (1962): “Programming with linear fractional functionals,” *Naval Research Logistics Quarterly*, 9, 181–186.
- CROSS, P. J. AND C. F. MANSKI (2002): “Regressions, Short and Long,” *Econometrica*, 70, 357–368.
- DEMUYNCK, T. (2015): “Bounding average treatment effects: A linear programming approach,” *Economics Letters*, 137, 75 – 77.
- DOBBIE, W., J. GOLDIN, AND C. S. YANG (2018): “The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges,” *American Economic Review*, 108, 201–40.
- FAN, Y., R. SHERMAN, AND M. SHUM (2014): “Identifying Treatment Effects Under Data Combination,” *Econometrica*, 82, 811–822.
- FREYBERGER, J. AND J. L. HOROWITZ (2015): “Identification and shape restrictions in nonparametric instrumental variables estimation,” *Journal of Econometrics*, 189, 41 – 53.

- HECKMAN, J. J. AND E. VYTLACIL (2001): “Policy-Relevant Treatment Effects,” *American Economic Review*, 91, 107–111.
- (2005): “Structural Equations, Treatment Effects, and Econometric Policy Evaluation1,” *Econometrica*, 73, 669–738.
- HOROWITZ, J. L. AND C. F. MANSKI (1995): “Identification and Robustness with Contaminated and Corrupted Data,” *Econometrica*, 63, 281–302.
- IMBENS, G. W. AND J. D. ANGRIST (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–475.
- KAMAT, V. (2018): “Identification with Latent Choice Sets,” .
- KITAGAWA, T. (2015): “A Test for Instrument Validity,” *Econometrica*, 83, 2043–2063.
- KLEINBERG, J., H. LAKKARAJU, J. LESKOVEC, J. LUDWIG, AND S. MULLAINATHAN (2017): “Human Decisions and Machine Predictions*,” *The Quarterly Journal of Economics*, 133, 237–293.
- KLEVMARKEN, N. A. (1982): “Missing Variables and Two-Stage Least-Squares Estimation from More than One Data Set,” in *1981 Proceedings of the American Statistical Association, Business and Economic Statistics Section*.
- LAFFÉRS, L. (2019): “Bounding average treatment effects using linear programming,” *Empirical Economics*, 57, 727–767.
- LAURA AND JOHN ARNOLD FOUNDATION (2019): “Risk factors and formula,” Public Safety Assessment www.psapretrial.org/about/factors, Accessed 9/17/2019.
- MOGSTAD, M., A. SANTOS, AND A. TORGOVITSKY (2018): “Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters,” *Econometrica*, 86, 1589–1619.
- RIDDER, G. AND R. MOFFITT (2007): “Chapter 75 The Econometrics of Data Combination,” Elsevier, vol. 6 of *Handbook of Econometrics*, 5469 – 5547.
- STEVENSON, M. T. (2018): “Distortion of Justice: How the Inability to Pay Bail Affects Case Outcomes,” *The Journal of Law, Economics, and Organization*, 34, 511–542.

A Identification Generalizations

In this section I extend the setting of Section 2 to the case when there are an arbitrary number of samples, J , and an arbitrary number of instruments, L . Let Z_1, Z_2, \dots, Z_L denote the instruments and let $Z = (Z_1, \dots, Z_L)$ be the vector of all instruments. Here I present assumptions that are the counterparts to Assumptions 1, 2, and 3 of Section 2.

Assumption 1a. $(\{Y_d\}_d, \{D_z\}_z, X_1, \dots, X_K) \perp\!\!\!\perp (Z_1, \dots, Z_L)$

Assumption 2a. $Y \in \mathcal{Y}, D \in \mathcal{D}, Z_l \in \mathcal{Z}_l$ for each $l, X_k \in \mathcal{X}_k$ for each k . $|\mathcal{Y}|, |\mathcal{D}|, |\mathcal{Z}_l|$ for each l , and $|\mathcal{X}_k|$ for each k are finite.

Assumption 3a.

1. $(\{Y_d\}_d, \{D_z\}_z, Z_1, \dots, Z_L, X_1, \dots, X_K)$ is identically distributed across samples.
2. Let $\phi_1, \phi_2, \dots, \phi_J$ be functions of $(\{Y_d\}_d, \{D_z\}_z, Z_1, \dots, Z_L, X_1, \dots, X_K)$ that select a subset of these variables. The observations in sample 1 are i.i.d. observations of $\phi_1(\{Y_d\}_d, \{D_z\}_z, Z_1, \dots, Z_L, X_1, \dots, X_K)$, the observations in sample 2 are i.i.d. observations of $\phi_2(\{Y_d\}_d, \{D_z\}_z, Z_1, \dots, Z_L, X_1, \dots, X_K)$, and so forth through sample J and ϕ_J .
3. The distribution of (Z_1, \dots, Z_L) is identified.

Theorem 4. *[Linearity] If Assumptions 1a, 2a, 3a.1, 3a.2, and 3a.3 are satisfied, then the sharp identified set for the vector of probabilities masses $\mathbf{P}(\{Y_d = y_d\}_d, \{D_z = d_z\}_z, X_1 = x_1, \dots, X_K = x_K)$ are the vectors p satisfying*

$$\begin{aligned} A_1 p &= b_1 \\ \mathbf{1}' p &= 1 \\ p &\geq 0 \end{aligned}$$

where A_1 is a matrix, b_1 is a vector and the entries of each are identified probabilities.

Remark: The assumption that the distribution of Z is identified would be satisfied if Z_1, \dots, Z_L are all in the same sample. Alternatively, if Z_1, \dots, Z_L are mutually independent

and each Z_l is observed in at least one sample would also satisfy the requirement that the distribution of Z is identified. In Example 5 I illustrate how this can be generalized when there are an arbitrary number of instruments spread across an arbitrary number of samples. In Lemma 1 I formalize these conditions.

Example 5. Assume $Z = (Z_1 \ Z_2 \ Z_3)'$. Suppose there are two samples and the observations are i.i.d. within each sample. The observations in one sample are of the pair (Z_1, Z_2) and observations in the second sample are of the pair (Z_2, Z_3) . Mutual independence of Z_1 , Z_2 , and Z_3 is not needed. It is sufficient for either $(Z_1, Z_2) \perp\!\!\!\perp Z_3$, or $(Z_2, Z_3) \perp\!\!\!\perp Z_1$. This assumption is clearly weaker than assuming mutual independence. \triangle

Lemma 1. Suppose $Z = (Z_1 \ Z_2 \ \dots \ Z_L)'$ takes finitely many values. Assume:

1. there exists a partition $(P_j)_{j \in J}$ of $\{1, \dots, L\}$ such that the $\{Z_{P_j}\}_{j \in J}$ are mutually independent where Z_{P_j} is a subvector of $(Z_1 \ Z_2 \ \dots \ Z_L)'$, and
2. for all $j \in J$, there exists an i.i.d. sample such that for each $l \in P_j$, Z_l is observed in sample j .

Then the distribution of Z is identified.

Proof. (of Lemma 1) $\mathbf{P}(Z = z) = \mathbf{P}(\bigwedge_{j \in J} Z_{P_j} = z_{P_j}) = \prod_{j \in J} \mathbf{P}(Z_{P_j} = z_{P_j})$ \square

Proof . (of Theorem 4) The distributions of $\phi_1(Y, D, Z_1, \dots, Z_L, X_1, \dots, X_K)$ through $\phi_J(Y, D, Z_1, \dots, Z_L, X_1, \dots, X_K)$ are identified because i.i.d. samples of each are observed. Let v be a point in the range of ϕ_1 . Let $\phi_1^{-1}(v)$ denote the preimage of v . As ϕ_1 is not one-to-one, $\phi_1^{-1}(v)$ is a set.

$$\begin{aligned} & \underbrace{\mathbf{P}(\phi_1(Y, D, Z_1, \dots, Z_L, X_1, \dots, X_K) = v)}_{\text{coefficients in } b_1} \\ &= \sum_{\underbrace{\{y_{d'}\}_{d'}, \{d_{z'}\}_{z'}, z_1, \dots, z_L, x_1, \dots, x_K}_{\text{coefficients in } A_1}} \mathbf{P}(Z = z) \mathbb{1}\left[(y_d, d_z) = (y, d) \ \& \ (y, d, z_1, \dots, z_L, x_1, \dots, x_K) \in \phi_1^{-1}(v)\right] \\ & \quad \times \underbrace{\mathbf{P}(\{Y_{d'} = y_{d'}\}_{d'}, \{D_{z'} = d_{z'}\}_{z'}, X_1 = x_1, \dots, X_K = x_K)}_{\text{element of the vector } p} \end{aligned}$$

This equality holds as in the proof of Theorem 1. These steps can be repeated for ϕ_2 through ϕ_J . \square

The left-hand-side of these equations constitute the elements of b_1 , and the elements of A_1 are in terms of the $\mathbf{P}(Z = z)$'s. b_1 and A_1 are defined analogously to the definitions of b_1 and A_1 in Section 2 where additional rows are added for the additional samples.

Theorem 5. Under Assumptions 1a, 2a, 3a.1, 3a.2, 3a.3, 4, and 5 the sharp identified lower bound for θ , θ_{LB} , can be characterized as follows:

$$\begin{aligned}\theta_{LB} &= \min_p c'p \\ \text{such that } A_1p &= b_1 \\ A_2p &= b_2 \\ p &\geq 0\end{aligned}$$

where A_1 and b_1 are identified from the data and A_2 and b_2 are known. The upper bound, θ_{UB} , can be written similarly with a max instead of a min.

This result follows immediately from Theorem 4 and Assumptions 4 and 5.

B Estimation Proof

In this section I present the proof of Theorem 3.

Proof. (of Theorem 3) I only show $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$. $\hat{\theta}_{UB} \xrightarrow{p} \theta_{UB}$ follows similarly. The proof is broken into two steps.

1. Construct a sequence that is a lower bound on $\hat{\theta}_{LB}$. The lower bound converges in probability to θ_{LB}
2. Construct a sequence that is an upper bound on $\hat{\theta}_{LB}$. The upper bound converges in probability to θ_{LB}

These two steps imply $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$.

Before these two steps, first notice that Assumptions 1, 2, and 3 imply that $\hat{b}_1 \xrightarrow{p} b_1$ and that $\text{vec}(\hat{A}_1) \xrightarrow{p} \text{vec}(A_1)$ because the entries of A_1 and b_1 are probabilities and the entries of \hat{A}_1 and \hat{b}_1 are empirical probabilities. Notice, also, that $\tilde{b}_1 \xrightarrow{p} b_1$. Letting $p_0 \in \{p : A_1p = b_1, A_2p = b_2, p \geq 0\}$ the following equalities/inequalities imply $\tilde{b}_1 \xrightarrow{p} b_1$:

$$\begin{aligned}\|\tilde{b}_1 - b_1\| &\leq \|\tilde{b}_1 - \hat{b}_1\| + \|\hat{b}_1 - b_1\| = \|\hat{A}_1\tilde{p} - \hat{b}_1\| + \|\hat{b}_1 - b_1\| \\ &\leq \|\hat{A}_1p_0 - \hat{b}_1\| + \|\hat{b}_1 - b_1\| \\ &\leq \|\hat{A}_1p_0 - A_1p_0\| + \|A_1p_0 - \hat{b}_1\| + \|\hat{b}_1 - b_1\| \\ &\leq \|\hat{A}_1p_0 - A_1p_0\| + 2\|\hat{b}_1 - b_1\|\end{aligned}$$

where the first inequality is a triangle inequality, the first equality uses the definition of $\tilde{b}_1 = \hat{A}_1\tilde{p}$ for some \tilde{p} in Definition 2, the second line uses the definition of \tilde{p} and \tilde{b} as

the minimizers in Definition 2, the third line uses the triangle inequality, and the last line uses that $A_1 p_0 = b_1$. Additionally, $\|\hat{A}_1 p_0 - A_1 p_0\| \leq \|\text{vec}(\hat{A}_1) - \text{vec}(A_1)\| \xrightarrow{p} 0$ where the inequality holds because p_0 is bounded by 1 as $p_0 \geq 0$ and the elements of p_0 sum to one. Knowing, additionally, that $\hat{b}_1 \xrightarrow{p} b_1$ implies $\tilde{b}_1 \xrightarrow{p} b_1$.

Part 1 Let \underline{p} be as follows:

$$\begin{aligned} \underline{p} &\in \arg \min_p c'p \\ \text{s.t. } \hat{A}_1 \underline{p} &= \tilde{b}_1 \\ A_2 \underline{p} &= b_2 \\ \underline{p} &\geq 0. \end{aligned}$$

\underline{p} can be any arbitrary choice in this set. Notice that $c'\underline{p} = \hat{\theta}_{LB}$. Define $\underline{b}_1 = A_1 \underline{p}$, and define $\underline{\theta}_{LB}$ as

$$\begin{aligned} \underline{\theta}_{LB} &= \min_p c'p \\ \text{s.t. } A_1 p &= \underline{b}_1 \\ A_2 p &= b_2 \\ p &\geq 0. \end{aligned}$$

Because $\underline{p} \in \{p : A_1 p = \underline{b}_1, A_2 p = b_2, p \geq 0\}$, $\underline{\theta}_{LB} \leq \hat{\theta}_{LB}$. Additionally, $\underline{b}_1 - \tilde{b}_1 \xrightarrow{p} 0$ because $\underline{p} \geq 0$, $\mathbf{1}'\underline{p} = 1$, and $\text{vec}(\hat{A}_1) \xrightarrow{p} \text{vec}(A_1)$. $\|\tilde{b}_1 - \underline{b}_1\| = \|\hat{A}_1 \underline{p} - A_1 \underline{p}\| \leq \|\text{vec}(\hat{A}_1) - \text{vec}(A_1)\| \xrightarrow{p} 0$. Therefore, $\underline{b}_1 \xrightarrow{p} b_1$. By Strong Duality,

$$\begin{aligned} \underline{\theta}_{LB} &= \max_x \begin{pmatrix} \underline{b}_1 \\ b_2 \end{pmatrix}' x \\ \text{s.t. } A'x &\leq c \end{aligned}$$

Let $\{x_1, \dots, x_s\}$ be the extreme points of $A'x \leq c$. Note that these points only depend on populations parameters.

Therefore, $\underline{\theta}_{LB} = \max_{\{x \in x_1, \dots, x_s\}} \begin{pmatrix} \underline{b}_1 \\ b_2 \end{pmatrix}' x$, which is continuous in $\begin{pmatrix} \underline{b}_1 \\ b_2 \end{pmatrix}$, and $\begin{pmatrix} \underline{b}_1 \\ b_2 \end{pmatrix} \xrightarrow{p} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$. Additionally, $\theta_{LB} = \max_{\{x \in x_1, \dots, x_s\}} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}' x$, again by Strong Duality. Therefore, $\underline{\theta}_{LB} \xrightarrow{p} \theta_{LB}$.

Part 2 To start, notice that Assumptions 6.1 and 6.2 allow us keep the linear constraints corresponding to \mathcal{I} in the linear programs in Theorem 2 and Definition 3 without changing the set of feasible solutions in either linear programming problem, and thus the optimal values of the objective do not change. For this part of the proof I consider these linear programs with

fewer constraints, i.e. $\begin{pmatrix} \hat{A}_1 \\ A_2 \end{pmatrix}_{\mathcal{I}} p = \begin{pmatrix} \tilde{b}_1 \\ b_2 \end{pmatrix}_{\mathcal{I}}$ replaces $\begin{pmatrix} \hat{A}_1 \\ A_2 \end{pmatrix} p = \begin{pmatrix} \tilde{b}_1 \\ b_2 \end{pmatrix}$ in Definition 3. Let p^* be the length m vector corresponding to the p that satisfies Assumption 6.3. Therefore, there exists an $m \times m$ matrix, B , consisting of m columns of $A_{\mathcal{I}}$ such that $p^* = B^{-1} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}_{\mathcal{I}}$, by the definition of a basic feasible solution. Replacing B and b_1 with \hat{B} and \tilde{b}_1 and we have a basic solution in the sample linear programming problem. To see this, first note that the entries of \hat{B} converge in probability to the entries of B . B is invertible, so \hat{B} is invertible with probability going to one. The r rows of $\hat{B}^{-1} \begin{pmatrix} \tilde{b}_1 \\ b_2 \end{pmatrix}_{\mathcal{I}}$ corresponding to the r rows of p^* that are zero are zero. $\hat{B}^{-1} \begin{pmatrix} \tilde{b}_1 \\ b_2 \end{pmatrix}_{\mathcal{I}}$ converges in probability to $B^{-1} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}_{\mathcal{I}}$ by the Continuous Mapping Theorem. The other $m - r$ rows of p^* are strictly positive, so the other $m - r$ rows of $\hat{B}^{-1} \begin{pmatrix} \tilde{b}_1 \\ b_2 \end{pmatrix}_{\mathcal{I}}$ are strictly positive with probability going to one. Therefore, with probability going to one $\hat{B}^{-1} \begin{pmatrix} \tilde{b}_1 \\ b_2 \end{pmatrix}_{\mathcal{I}} \geq 0$. Thus, this has constructed a sample basic feasible solution in the sample which converges to an optimal basic feasible solution in the population. Let $\bar{\theta}_{LB}$ denote the value of the objective at this sample basic feasible solution. $\hat{\theta}_{LB} \leq \bar{\theta}_{LB}$ since $\bar{\theta}_{LB}$ is a basic feasible solution in the sample, and I have already established $\bar{\theta}_{LB} \xrightarrow{p} \theta_{LB}$.

Part 3 Therefore, we have that $\underline{\theta}_{LB} \leq \hat{\theta}_{LB} \leq \bar{\theta}_{LB}$ with probability going to one. $\underline{\theta}_{LB} \xrightarrow{p} \theta_{LB}$ and $\bar{\theta}_{LB} \xrightarrow{p} \theta_{LB}$, so $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$. \square

C Details of Pretrial Safety Assessment

This section details the algorithm discussed in Section 4. For full details on the Public Safety Assessment see [Laura and John Arnold Foundation \(2019\)](#).

The formula calculates how many “points” an individual has based on the value of the covariates. For new violent criminal activity the number of points an individual has can be integers values between 0 and 7. If an individual has at least 4 points, they are flagged as being likely to commit a violent crime while on pretrial release. The algorithm detains these high risk individuals and release all other individuals pretrial. Define the following variables:

current_violent_of_fense: an indicator equal to one if the current offense is violent
age_less_equal_20: an indicator for whether the defendant is 20 years old or less
pending_charge: an indicator for whether the defendant has a pending charge at the time of the offense
prior_conviction: an indicator for whether the defendant has either a misdemeanor or felony prior conviction
prior_violent_convictions: number of prior violent convictions
NVCA_points: points discussed previously—if this is ≥ 4 the PSA flags the individual as being likely to commit a violent crime, and the algorithm detains that individual

$$\begin{aligned}
NVCA_points = & 2 \times current_violent_of_fense \\
& + current_violent_of_fense \times age_less_equal_20 \\
& + pending_charge \\
& + prior_conviction \\
& + \mathbb{1}[1 \leq prior_violent_convictions \leq 2] \\
& + 2 \times \mathbb{1}[prior_violent_convictions \geq 3]
\end{aligned}$$

The PSA flags an individual for having a high probability of committing a violent crime while released pretrial if this *NVCA_points* is above 4. The algorithm, d_b , detains defendants that have been flagged by the PSA for their risk of committing a new violent crime and release defendants otherwise.

This does not fit *perfectly* into setting, since I do not observe the total number of prior violent convictions in either data set. However, I do observe in the SCPS whether or not an individual has *any* prior violent conviction, as well as the total number of prior felony convictions. With this information I am able to infer exactly the *NVCA_points* for most defendants—if an individual has more than 2 prior felony convictions and I know they have at least one prior violent conviction they could have between 1 and 2 prior violent convictions or have 3 or more prior violent convictions. This would change their *NVCA_points* by 1. This only affects whether they are treated under the algorithm if this changes their *NVCA_points* from 3 to 4. To address this, when estimating the bounds on $\mathbf{E}[Y^b - Y^a]$, I can use a “worst case” choice of the vector c , or include an unobserved variable for prior violent convictions,

and include the relation between number of violent felonies, the indicator for violent felonies, and number of prior felonies as shape restrictions in $A_2p = b_2$.

D Estimation Proofs Under Alternative Assumptions

D.1 A_1 known

The scenario where A_1 is known and only b_1 needs to be estimated would occur if Z is observed in every sample. In the proof of Theorem 1, if Z is observed in a sample with $(Y, D, Z, X_1, \dots, X_K)$, the linear equation is

$$\mathbf{P}(Y = y, D = d, Z = z, X_1, \dots, X_K) = \mathbf{P}(Y_d = y, D_z = d, X_1, \dots, X_K) \mathbf{P}(Z = z)$$

This is equivalent to the following when $\mathbf{P}(Z = z) > 0$:

$$\mathbf{P}(Y = y, D = d, X_1, \dots, X_K | Z = z) = \mathbf{P}(Y_d = y, D_z = d, X_1, \dots, X_K)$$

Where the left-hand-side is identified and can be consistently estimated. A similar rearrangement can be performed when a subset of Y, D, X_1, \dots, X_K is observed. The left-hand-side of these new equations would constitute b_1 . The matrix A_1 would then be a matrix of 0's and 1's, and the entries would be known, and not need to be estimated.

Theorem 6. *Suppose the following conditions are met:*

I) $\hat{b}_1 \xrightarrow{p} b_1$

II) $A_2p = b_2$ includes the restriction $\mathbf{1}'p = 1$

III) $\hat{A}_1 = A_1$

Then, $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$ and $\hat{\theta}_{UB} \xrightarrow{p} \theta_{UB}$

Proof. (of Theorem 6) I show $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$. $\hat{\theta}_{UB} \xrightarrow{p} \theta_{UB}$ follows similarly.

By the construction of \tilde{b}_1 , the set $\{p : A_1p = \tilde{b}_1, A_2p = b_2, p \geq 0\}$ is not empty. By II) $\hat{\theta}_{LB}$ is finite. Therefore, I can invoke Strong Duality and have

$$\hat{\theta}_{LB} = \max_x \begin{pmatrix} \tilde{b}_1 \\ b_2 \end{pmatrix}' x$$

$$s.t. \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}' x \leq c.$$

Let $\{x_1, \dots, x_s\}$ be the extreme points of $\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}' x \leq c$.

Then, $\hat{\theta}_{LB} = \max_{x \in \{x_1, \dots, x_k\}} \begin{pmatrix} \tilde{b}_1 \\ b_2 \end{pmatrix}' x$, which is continuous in $\begin{pmatrix} \tilde{b}_1 \\ b_2 \end{pmatrix}$.

Because $\begin{pmatrix} \tilde{b}_1 \\ b_2 \end{pmatrix} \xrightarrow{p} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$, this implies $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$. To see that $\tilde{b}_1 \xrightarrow{p} b_1$ notice that $\|\tilde{b}_1 - \hat{b}_1\| \leq \|b_1 - \hat{b}_1\|$ by the definition of \tilde{b}_1 . Therefore, $\|\tilde{b}_1 - b_1\| \leq \|\tilde{b}_1 - \hat{b}_1\| + \|\hat{b}_1 - b_1\| \leq 2\|\hat{b}_1 - b_1\|$ \square

D.2 Bounded Dual Problem

The following result replaces Assumption 6.3 with an assumption about the boundedness of the dual linear programming problem.

Theorem 7. *Make Assumptions 1, 2, 3, 4, 5, 6.1 and 6.2 and assume $A_2 p = b_2$ includes the restriction $\mathbf{1}' p = 1$. Additionally, assume that here exists M such that for some $\tilde{x} \in \arg \max_x$ of the dual problem in Definition 3, such that $\|\tilde{x}\|_\infty \leq M$ with probability going to one. Then, $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$ and $\hat{\theta}_{UB} \xrightarrow{p} \theta_{UB}$.*

Drawing a deeper connection between the boundedness condition in Theorem 7 and the setting of this paper is an area for future work. I suspect that there is a connection between M and the smallest value of $\mathbf{P}(Z = z)$ when $\mathbf{P}(Z = z) > 0$ for all z .

Proof. (of Theorem 7) I only show $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$. $\hat{\theta}_{UB} \xrightarrow{p} \theta_{UB}$ follows similarly. The proof is broken into two steps.

1. Construct a sequence that is a lower bound on $\hat{\theta}_{LB}$. The lower bound converges in probability to θ_{LB}
2. Construct a sequence that is an upper bound on $\hat{\theta}_{LB}$. The upper bound converges in probability to θ_{LB}

These two steps imply $\hat{\theta}_{LB} \xrightarrow{p} \theta_{LB}$.

This initial section of the proof is the same as that of the proof of Theorem 3. Before these two steps, first notice that Assumptions 1, 2, and 3 imply that $\hat{b}_1 \xrightarrow{p} b_1$ and that $\text{vec}(\hat{A}_1) \xrightarrow{p} \text{vec}(A_1)$ because the entries of A_1 and b_1 are probabilities and the entries of \hat{A}_1 and \hat{b}_1 are empirical probabilities. Notice, also, that $\tilde{b}_1 \xrightarrow{p} b_1$. Letting $p_0 \in \{p : A_1 p =$

$b_1, A_2p = b_2, p \geq 0\}$ the following equalities/inequalities imply $\tilde{b}_1 \xrightarrow{p} b_1$:

$$\begin{aligned}
\|\tilde{b}_1 - b_1\| &\leq \|\tilde{b}_1 - \hat{b}_1\| + \|\hat{b}_1 - b_1\| = \|\hat{A}_1\tilde{p} - \hat{b}_1\| + \|\hat{b}_1 - b_1\| \\
&\leq \|\hat{A}_1p_0 - \hat{b}_1\| + \|\hat{b}_1 - b_1\| \\
&\leq \|\hat{A}_1p_0 - A_1p_0\| + \|A_1p_0 - \hat{b}_1\| + \|\hat{b}_1 - b_1\| \\
&\leq \|\hat{A}_1p_0 - A_1p_0\| + 2\|\hat{b}_1 - b_1\|
\end{aligned}$$

where the first inequality is a triangle inequality, the first equality uses the definition of $\tilde{b}_1 = \hat{A}_1\tilde{p}$ for some \tilde{p} in Definition 2, the second line uses the definition of \tilde{p} and \tilde{b} as the minimizers in Definition 2, the third line uses the triangle inequality, and the last line uses that $A_1p_0 = b_1$. Additionally, $\|\hat{A}_1p_0 - A_1p_0\| \leq \|\text{vec}(\hat{A}_1) - \text{vec}(A_1)\| \xrightarrow{p} 0$ where the inequality holds because p_0 is bounded by 1 as $p_0 \geq 0$ and the elements of p_0 sum to one. Knowing, additionally, that $\hat{b}_1 \xrightarrow{p} b_1$ implies $\tilde{b}_1 \xrightarrow{p} b_1$.

Part 1 (Same as Part 1 in proof of Theorem 3) Let \underline{p} be as follows:

$$\begin{aligned}
\underline{p} &\in \arg \min_p c'p \\
\text{s.t. } \hat{A}_1\underline{p} &= \tilde{b}_1 \\
A_2\underline{p} &= b_2 \\
\underline{p} &\geq 0.
\end{aligned}$$

\underline{p} can be any arbitrary choice in this set. Notice that $c'\underline{p} = \hat{\theta}_{LB}$. Define $\underline{b}_1 = A_1\underline{p}$, and define $\underline{\theta}_{LB}$ as

$$\begin{aligned}
\underline{\theta}_{LB} &= \min_p c'p \\
\text{s.t. } A_1p &= \underline{b}_1 \\
A_2p &= b_2 \\
p &\geq 0.
\end{aligned}$$

Because $\underline{p} \in \{p : A_1p = \underline{b}_1, A_2p = b_2, p \geq 0\}$, $\underline{\theta}_{LB} \leq \hat{\theta}_{LB}$. Additionally, $\underline{b}_1 - \tilde{b}_1 \xrightarrow{p} 0$ because $\underline{p} \geq 0$, $\mathbf{1}'\underline{p} = 1$, and $\text{vec}(\hat{A}_1) \xrightarrow{p} \text{vec}(A_1)$. $\|\tilde{b}_1 - \underline{b}_1\| = \|\hat{A}_1\underline{p} - A_1\underline{p}\| \leq \|\text{vec}(\hat{A}_1) - \text{vec}(A_1)\| \xrightarrow{p} 0$. Therefore, $\underline{b}_1 \xrightarrow{p} b_1$. By Strong Duality,

$$\begin{aligned}
\underline{\theta}_{LB} &= \max_x \begin{pmatrix} \underline{b}_1 \\ b_2 \end{pmatrix}' x \\
\text{s.t. } A'x &\leq c
\end{aligned}$$

Let $\{x_1, \dots, x_s\}$ be the extreme points of $A'x \leq c$. Note that these points only depend on populations parameters.

Therefore, $\underline{\theta}_{LB} = \max_{\{x \in x_1, \dots, x_s\}} \left(\begin{array}{c} b_1 \\ b_2 \end{array} \right)' x$, which is continuous in $\left(\begin{array}{c} b_1 \\ b_2 \end{array} \right)$, and $\left(\begin{array}{c} b_1 \\ b_2 \end{array} \right) \xrightarrow{p} \left(\begin{array}{c} \tilde{b}_1 \\ b_2 \end{array} \right)$. Additionally, $\theta_{LB} = \max_{\{x \in x_1, \dots, x_s\}} \left(\begin{array}{c} b_1 \\ b_2 \end{array} \right)' x$, again by Strong Duality. Therefore, $\underline{\theta}_{LB} \xrightarrow{p} \theta_{LB}$.

Part 2 (Different than Part 2 in Theorem 3) Let

$$\begin{aligned} \bar{x} \in \arg \max_x \left(\begin{array}{c} \tilde{b}_1 \\ b_2 \end{array} \right)' x \\ \text{such that } \left(\begin{array}{c} \hat{A}_1 \\ A_2 \end{array} \right)' x \leq c \end{aligned} \quad (4)$$

such that $\|\bar{x}\|_\infty \leq M$, which exists by the assumption made in Theorem 7 with probability going to 1. $\hat{\theta}_{LB} = \left(\begin{array}{c} \tilde{b}_1 \\ b_2 \end{array} \right)' \bar{x}$ because (4) is the dual linear programming problem of the linear programming problem in Definition 3. (5) holds because $\|\bar{x}\|_\infty \leq M$ and the elements of \hat{A}_1 converge to the elements of A_1 .

$$\left(\left(\begin{array}{c} \hat{A}_1 \\ A_2 \end{array} \right) - \left(\begin{array}{c} A_1 \\ A_2 \end{array} \right) \right)' \bar{x} \xrightarrow{p} 0 \quad (5)$$

Define

$$\bar{c} = \max \left\{ \left(\begin{array}{c} A_1 \\ A_2 \end{array} \right)' \bar{x}, c \right\}$$

where the max is taken elementwise across the two vectors. Then, $\bar{c} \xrightarrow{p} c$ by the choice of \bar{x} in (4).

Define:

$$\begin{aligned} \bar{\theta}_{LB} = \min_p \bar{c}' p & \quad \text{such that } A_1 p = \tilde{b}_1 \\ A_2 p = b_2 & \quad p \geq 0 \end{aligned} \quad \& \quad \begin{aligned} \ddot{\theta}_{LB} = \min_p c' p & \quad \text{such that } A_1 p = \tilde{b}_1 \\ A_2 p = b_2 & \quad p \geq 0 \end{aligned} \quad (6)$$

Case II Suppose $\mathbf{P}(Y_1=y, D_1=1, D_0=0) > 0$. Therefore,

$$\begin{aligned}
& \hat{\mathbf{P}}(Y=y, D=1|Z=1) - \hat{\mathbf{P}}(Y=y, D=1|Z=0) \\
& \xrightarrow{p} \mathbf{P}(Y=y, D=1|Z=1) - \mathbf{P}(Y=y, D=1|Z=0) \\
& = \mathbf{P}(Y_1=y, D_1=1) - \mathbf{P}(Y_1=y, D_0=1) \\
& = (\mathbf{P}(Y_1=y, D_1=1, D_0=1) + \mathbf{P}(Y_1=y, D_1=1, D_0=0)) \\
& \quad - \mathbf{P}(Y_1=y, D_1=1, D_0=1) \\
& = \mathbf{P}(Y_1=y, D_1=1, D_0=0) \\
& > 0.
\end{aligned}$$

Therefore, $\hat{\mathbf{P}}(Y=y, D=1|Z=1) - \hat{\mathbf{P}}(Y=y, D=1|Z=0) \geq 0$ with probability going to one. The result for $\hat{\mathbf{P}}(Y=y, D=0|Z=0) - \hat{\mathbf{P}}(Y=y, D=0|Z=1) \geq 0$ follows similarly. \square

F Additional Examples and Details of Examples

In this section I provide the details for the example in Subsection 2.2 as well as additional examples.

F.1 Additional Examples

Example 6 illustrates that in a two sample instrumental variables setting (one sample the observations are of (Y, Z) and in another sample observations are of (D, Z)), if the monotonicity assumption is not made, the identified set is smaller when the independence assumption $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$ is made as opposed to the weaker assumption $\mathbf{E}[Y_d|D_1, D_0, Z] = \mathbf{E}[Y_d|D_1, D_0]$ and $(D_1, D_0) \perp\!\!\!\perp Z$. When the monotonicity assumption is maintained, the identified set for the ATE under either set of assumptions are equal. Example 7 demonstrates that the identified bounds with full independence can improve over the identified bounds when the only the mean independence assumption is made, when the monotonicity assumption holds.

Example 6. In this example the assumption that $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$ is satisfied. I compare the bounds on the average treatment effect the assumption $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$ is made against when the assumption that $\mathbf{E}[Y_d|D_1, D_0, Z] = \mathbf{E}[Y_d|D_1, D_0]$ and $(D_1, D_0) \perp\!\!\!\perp Z$ is made. The marginal distribution of (Y, Z) and the marginal distribution of (D, Z) are identified. $Y \in \{0, 1, \dots, 5\}$ and D and Z are in $\{0, 1\}$. No shape restrictions are assumed.

The distributions of Z and (Y_1, Y_0, D_1, D_0) are:

$$Z = \begin{cases} 0 & \text{w/ prob } 1/2 \\ 1 & \text{w/ prob } 1/2 \end{cases} \quad \& \quad (Y_1, Y_0, D_1, D_0) = \begin{cases} (5, 2, 1, 0) & \text{w/ prob } 1/4 \\ (3, 0, 1, 0) & \text{w/ prob } 1/4 \\ (4, 1, 0, 1) & \text{w/ prob } 1/2 \end{cases} .$$

This yields the following marginal distributions of (Y, Z) and (D, Z) :

	Y	0	1	2	3	4	5
Z	0	1/8	0	1/8	0	1/4	0
	1	0	1/4	0	1/8	0	1/8

Table 7. Distribution of (Y, Z) for Example 6.

	D	0	1
Z	0	1/4	1/4
	1	1/4	1/4

Table 8. Distribution of (D, Z) for Example 6.

Under the assumption that $Y = DY_1 + (1 - D)Y_0$ and $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$, the marginal distribution of (Y, Z) says that $\mathbf{P}(D_1 = 0, D_0 = 0) = 0$ and $\mathbf{P}(D_1 = 1, D_0 = 1) = 0$. Therefore, $\mathbf{P}(D_1 = 1, D_0 = 0) = \frac{1}{2}$ and $\mathbf{P}(D_1 = 0, D_0 = 1) = \frac{1}{2}$. This implies that $-3 \leq \text{ATE} \leq 3$.

However, if the full independence assumption is replaced with $\mathbf{E}[Y_d | D_1, D_0, Z] = \mathbf{E}[Y_d | D_1, D_0]$ and $(D_1, D_0) \perp\!\!\!\perp Z$, distribution such that $\mathbf{E}(Y_1 - Y_0) = 4$ can be constructed. This is outside of the set $[-3, 3]$ with the stronger assumption. The full independence assumption is stronger, so the identified set is a strict subset. Proofs for this example can be found in Appendix F.3. \triangle

In Example 6 the bounds on the average treatment effect were compared when the monotonicity assumption is dropped in a two sample instrumental variables setting.

Example 7. In this example the data generating process satisfies $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$. I compare the bounds on the average treatment effect when the assumption that $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$ is made against when the assumptions that $\mathbf{E}[Y_d | D_1, D_0, Z] = \mathbf{E}[Y_d | D_1, D_0]$ and $(D_1, D_0) \perp\!\!\!\perp Z$ are made. In this example $Y \in \{0, 1, \dots, 6\}$ and D and Z are in $\{0, 1\}$. (Y, D) is observed in one sample and (Y, Z) is observed in another sample. I make the monotonicity assumption that $\mathbf{P}(D_1 = 0, D_0 = 1) = 0$.

Suppose the distributions of Z and (Y_1, Y_0, D_1, D_0) are

$$Z = \begin{cases} 0 & \text{w/ prob } 1/2 \\ 1 & \text{w/ prob } 1/2 \end{cases} \quad \& \quad (Y_1, Y_0, D_1, D_0) = \begin{cases} (3, 1, 1, 0) & \text{w/ prob } 1/4 \\ (3, 5, 1, 0) & \text{w/ prob } 1/4 \\ (2, 2, 1, 0) & \text{w/ prob } 1/4 \\ (4, 4, 1, 0) & \text{w/ prob } 1/4 \end{cases}.$$

This yields the following marginal distributions of (Y, D) and (Y, Z)

	Y	0	1	2	3	4	5	6
D	0	0	1/8	1/8	0	1/8	1/8	0
	1	0	0	1/8	1/4	1/8	0	0

Table 9. Distribution of (Y, D) for Example 7

	Y	0	1	2	3	4	5	6
Z	0	0	1/8	1/8	0	1/8	1/8	0
	1	0	0	1/8	1/4	1/8	0	0

Table 10. Distribution of (Y, Z) for Example 7

Making the full independence assumption $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$ and knowing $Y = DY_1 + (1 - D)Y_0$ results in the bounds $-\frac{3}{2} \leq \mathbf{E}(Y_1 - Y_0) \leq \frac{3}{2}$.

Now, make the assumptions that $(D_1, D_0) \perp\!\!\!\perp Z$ and $\mathbf{E}(Y_d|D_1, D_0, Z) = \mathbf{E}(Y_d|D_1, D_0)$. There exists a distribution such that $\mathbf{E}(Y_1 - Y_0) = 3$. This is outside of the interval $[-\frac{3}{2}, \frac{3}{2}]$ that was implied by the stronger assumption of $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$ and $Y = DY_1 + (1 - D)Y_0$. Proofs for this example can be found in Appendix F.4. \triangle

F.2 Example 4

Details (*Example 4*)

The marginals of (Y, D, Z) are as follows:

1. (Y, Z) : $\mathbf{P}(Y = 1, Z = 1) = \frac{1}{2}$ and $\mathbf{P}(Y = 0, Z = 0) = \frac{1}{2}$
2. (D, Z) : $\mathbf{P}(D = 1, Z = 1) = \frac{1}{2}$ and $\mathbf{P}(D = 0, Z = 0) = \frac{1}{2}$
3. (Y, D) : $\mathbf{P}(D = 1, Z = 1) = \frac{1}{2}$ and $\mathbf{P}(D = 0, Z = 0) = \frac{1}{2}$
4. (D) : $\mathbf{P}(D = 1) = \frac{1}{2}$ and $\mathbf{P}(D = 0) = \frac{1}{2}$
5. (Y) : $\mathbf{P}(Y = 1) = \frac{1}{2}$ and $\mathbf{P}(Y = 0) = \frac{1}{2}$

1. $(Y, Z), (D, Z)$

The distribution of (D, Z) implies that $\mathbf{P}(D_1 = 1, D_0 = 0) = 1$. Now using the

distribution of (Y, Z) and that $\mathbf{P}(D_1 = 1, D_0 = 0) = 1$, $\mathbf{P}(Y_0 = 0, D_1 = 1, D_0 = 0, Z = 0) = \mathbf{P}(Y_0 = 0, D_1 = 1, D_0 = 0)\mathbf{P}(Z = 0) = \frac{1}{2} \Rightarrow \mathbf{P}(Y_0 = 0, D_1 = 1, D_0 = 0) = 1$. Similarly, $\mathbf{P}(Y_1 = 1, D_1 = 1, D_0 = 0) = 1$, so $\boxed{\mathbf{E}(Y_1 - Y_0) = 1}$.

2. $(Y, D), (Y, Z)$

These two marginals imply $\mathbf{P}(Y = 1, D = 1, Z = 1) = \frac{1}{2}$ and $\mathbf{P}(Y = 0, D = 0, Z = 0) = \frac{1}{2}$. This says that $\mathbf{P}(D_1 = 1, D_0 = 0) = 1$. Therefore, $\mathbf{P}(Y_1 = 1, D_1 = 1, D_0 = 0, Z = 1) = \mathbf{P}(Y_1 = 1, D_1 = 1, D_0 = 0)\mathbf{P}(Z = 1) = \frac{1}{2} \Rightarrow \mathbf{P}(Y_1 = 1, D_1 = 1, D_0 = 0) = 1$. Similarly, $\mathbf{P}(Y_0 = 1, D_1 = 1, D_0 = 0) = 1$, so $\boxed{\mathbf{E}(Y_1 - Y_0) = 1}$

3. $(Y, D), (D, Z)$

The distribution of (D, Z) says that $\mathbf{P}(D_1 = 1, D_0 = 0) = 1$. Now using the distribution of (Y, D) and that $\mathbf{P}(D_1 = 1, D_0 = 0) = 1$, $\mathbf{P}(Y_0 = 0, D_1 = 1, D_0 = 0, Z = 0) = \mathbf{P}(Y_0 = 0, D_1 = 1, D_0 = 0)\mathbf{P}(Z = 0) = \frac{1}{2} \Rightarrow \mathbf{P}(Y_0 = 0, D_1 = 1, D_0 = 0) = 1$. Similarly, $\mathbf{P}(Y_1 = 1, D_1 = 1, D_0 = 0) = 1$, so $\boxed{\mathbf{E}(Y_1 - Y_0) = 1}$.

4. (Y, Z)

$\mathbf{P}(Y_1 = 0, Y_0 = 1, D_1 = 1, D_0 = 1) = \frac{1}{2}$ and $\mathbf{P}(Y_1 = 0, Y_0 = 1, D_1 = 0, D_0 = 0) = \frac{1}{2}$ is possible, resulting in the trivial bounds $\boxed{[-1, 1]}$.

5. (D, Z)

Nothing about the distribution of Y_0 and Y_1 is known, so the bounds are trivial $\boxed{[-1, 1]}$.

6. (Y, D)

$\mathbf{E}(Y_1 - Y_0) = \mathbf{E}[Y_1|D = 1]\mathbf{P}(D = 1) + \mathbf{E}[Y_1|D = 0]\mathbf{P}(D = 0) - \mathbf{E}[Y_0|D = 1]\mathbf{P}(D = 1) - \mathbf{E}[Y_0|D = 0]\mathbf{P}(D = 0) = \frac{1}{2} + \mathbf{E}[Y_1|D = 0]\frac{1}{2} - \mathbf{E}[Y_0|D = 1]\frac{1}{2} - 0$ and $\mathbf{E}[Y_1|D = 0]$ and $\mathbf{E}[Y_0|D = 1]$ are in $[0, 1]$. Therefore, $\boxed{\mathbf{E}(Y_1 - Y_0) \in [0, 1]}$. 0 is obtained if $\mathbf{P}(Y_1 = 1, Y_0 = 1, D_1 = 1, D_0 = 1) = \frac{1}{2}$ and $\mathbf{P}(Y_1 = 0, Y_0 = 0, D_1 = 0, D_0 = 0) = \frac{1}{2}$. \square

F.3 Example 6

Details (Example 6)

If the full independence assumption is replaced with $\mathbf{E}[Y_d|D_1, D_0, Z] = \mathbf{E}[Y_d|D_1, D_0]$ and $(D_1, D_0) \perp\!\!\!\perp Z$, the following is possible:

$$\begin{aligned} \mathbf{P}(Z = 0) &= \mathbf{P}(Z = 1) = \frac{1}{2} \\ \mathbf{P}(D_1 = 1, D_0 = 1) &= \frac{1}{2}, \quad \mathbf{P}(D_1 = 0, D_0 = 0) = \frac{1}{2} \\ (Y_1|D_1 = 1, D_0 = 1, Z = 1) &= \begin{cases} 3 & \text{w.p. } \frac{1}{2} \\ 5 & \text{w.p. } \frac{1}{2} \end{cases} \\ (Y_1|D_1 = 1, D_0 = 1, Z = 0) &= 4 \quad \text{w.p. } 1 \\ (Y_0|D_1 = 0, D_0 = 0, Z = 0) &= \begin{cases} 0 & \text{w.p. } \frac{1}{2} \\ 2 & \text{w.p. } \frac{1}{2} \end{cases} \\ (Y_0|D_1 = 0, D_0 = 0, Z = 1) &= 1 \quad \text{w.p. } 1 \end{aligned}$$

Knowing $Y \in \{0, 1, \dots, 5\}$, then it is possible for $\mathbf{E}(Y_0|D_1 = 1, D_0 = 1) = 0$ and $\mathbf{E}(Y_1|D_1 = 0, D_0 = 0) = 5$. This would mean $\mathbf{E}(Y_1 - Y_0) = 4$. This is outside of the set $[-3, 3]$ with the stronger assumption. \square

F.4 Example 7

Details (Example 7)

Making the full independence assumption $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$ and knowing $Y = DY_1 + (1 - D)Y_0$, implies

$$\mathbf{P}(Y = 1, D = 0, Z = 0) = \frac{1}{8}, \quad \mathbf{P}(Y = 5, D = 0, Z = 0) = \frac{1}{8}, \quad \mathbf{P}(Y = 3, D = 1, Z = 0) = \frac{1}{4}$$

this implies

$$\mathbf{P}(Y_0 = 1, D_1 = 1, D_0 = 0) = \frac{1}{4}, \quad \mathbf{P}(Y_0 = 5, D_1 = 1, D_0 = 0) = \frac{1}{4}, \quad \mathbf{P}(Y_1 = 3, D_1 = 1, D_0 = 0) = \frac{1}{2}$$

Therefore, $-\frac{3}{2} \leq \mathbf{E}(Y_1 - Y_0) \leq \frac{3}{2}$ The following distribution would satisfy these assumptions and would result in the previous distribution of (Y, D) and (Y, Z) .

$$\begin{aligned} \mathbf{P}(D_1 = 1, D_0 = 1) &= \frac{1}{2} \quad \&\mathbf{P}(D_1 = 0, D_0 = 0) = \frac{1}{2} \\ (Y_0|D_1 = 0, D_0 = 0, Z = 0) &= \begin{cases} 1 & w/ \text{prob. } \frac{1}{2} \\ 5 & w/ \text{prob. } \frac{1}{2} \end{cases} \\ (Y_0|D_1 = 0, D_0 = 0, Z = 1) &= \begin{cases} 2 & w/ \text{prob. } \frac{1}{2} \\ 4 & w/ \text{prob. } \frac{1}{2} \end{cases} \\ (Y_1|D_1 = 0, D_0 = 0) &= 6 \quad w/ \text{prob. } 1 \\ (Y_1|D_1 = 1, D_0 = 1, Z = 0) &= \begin{cases} 2 & w/ \text{prob. } \frac{1}{2} \\ 4 & w/ \text{prob. } \frac{1}{2} \end{cases} \\ (Y_1|D_1 = 1, D_0 = 1, Z = 1) &= 3 \quad w/ \text{prob. } 1 \\ (Y_0|D_1 = 1, D_0 = 1) &= 0 \quad w/ \text{prob. } 1 \end{aligned}$$

This results in $\mathbf{E}(Y_1 - Y_0) = 3$. This is outside of the interval $[-\frac{3}{2}, \frac{3}{2}]$ that was implied by the stronger assumption of $(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z$ and $Y = DY_1 + (1 - D)Y_0$. \square

G Discussion of Assumption 6 for Estimation

Example 8 demonstrates why the rank assumptions 1 and 2 in Theorem 3 should be satisfied.

Example 8. Let $q_0, q_1, \hat{q}_0,$ and \hat{q}_1 denote $\mathbf{P}(Z = 0), \mathbf{P}(Z = 1), \hat{\mathbf{P}}(Z = 0),$ and $\hat{\mathbf{P}}(Z = 1),$ respectively. For illustrative purposes I show what the equation looks like when there are four samples—one with $(Y, D, Z),$ one with $(Y, D),$ one with $(Y, Z),$ one with (D, Z) (The last three samples would clearly be unnecessary in this scenario).

$$\underbrace{\left(\begin{array}{cccccc|c|c}
 1 & 0 & 0 & 0 & 0 & 0 & 0 & 00 \\
 0 & 1 & 0 & 0 & 0 & 0 & 0 & 00 \\
 0 & 0 & 1 & 0 & 0 & 0 & 0 & 00 \\
 -1 & -1 & -1 & 0 & 0 & 0 & q_0 & 00 \\
 0 & 0 & 0 & 1 & 0 & 0 & 0 & 00 \\
 0 & 0 & 0 & 0 & 1 & 0 & 0 & 00 \\
 0 & 0 & 0 & 0 & 0 & 1 & 0 & 00 \\
 0 & 0 & 0 & -1 & -1 & -1 & q_1 & 00 \\
 \hline
 1 & 0 & 0 & 1 & 0 & 0 & 0 & 00 \\
 0 & 1 & 0 & 0 & 1 & 0 & 0 & 00 \\
 0 & 0 & 1 & 0 & 0 & 1 & 0 & 00 \\
 -1 & -1 & -1 & -1 & -1 & -1 & q_0+q_1 & 00 \\
 \hline
 1 & 1 & 0 & 0 & 0 & 0 & 0 & 00 \\
 -1 & -1 & 0 & 0 & 0 & 0 & q_0 & 00 \\
 0 & 0 & 0 & 1 & 1 & 0 & 0 & 00 \\
 0 & 0 & 0 & -1 & -1 & 0 & q_1 & 00 \\
 \hline
 0 & 0 & 0 & 0 & 0 & 0 & 1 & 00 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 01
 \end{array} \right)}_{C_1}
 \underbrace{\left(\begin{array}{cccccccccccc}
 q_0 & 0 & q_0 & 0 & 0 & 0 & 0 & 0 & q_0 & 0 & q_0 & 0 & 0 & 0 & 0 & 0 \\
 0 & q_0 & 0 & q_0 & 0 & q_0 & 0 & q_0 & 0 & 0 & 0 & 0 & q_0 & 0 & q_0 & 0 \\
 0 & 0 & 0 & 0 & q_0 & 0 & q_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 q_1 & q_1 & 0 & 0 & 0 & 0 & 0 & 0 & q_1 & q_1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & q_1 & q_1 & 0 & 0 & q_1 & q_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & q_1 & q_1 & 0 & 0 & 0 & 0 & 0 & 0 & q_1 & q_1 & 0 & 0 \\
 \hline
 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
 \end{array} \right)}_{C_2}
 \underbrace{\left(\begin{array}{l}
 \mathbf{P}(Y_1=0, Y_0=0, D_1=0, D_0=0) \\
 \mathbf{P}(Y_1=0, Y_0=0, D_1=0, D_0=1) \\
 \mathbf{P}(Y_1=0, Y_0=0, D_1=1, D_0=0) \\
 \mathbf{P}(Y_1=0, Y_0=0, D_1=1, D_0=1) \\
 \mathbf{P}(Y_1=0, Y_0=1, D_1=0, D_0=0) \\
 \mathbf{P}(Y_1=0, Y_0=1, D_1=0, D_0=1) \\
 \mathbf{P}(Y_1=0, Y_0=1, D_1=1, D_0=0) \\
 \mathbf{P}(Y_1=0, Y_0=1, D_1=1, D_0=1) \\
 \mathbf{P}(Y_1=1, Y_0=0, D_1=0, D_0=0) \\
 \mathbf{P}(Y_1=1, Y_0=0, D_1=0, D_0=1) \\
 \mathbf{P}(Y_1=1, Y_0=0, D_1=1, D_0=0) \\
 \mathbf{P}(Y_1=1, Y_0=0, D_1=1, D_0=1) \\
 \mathbf{P}(Y_1=1, Y_0=1, D_1=0, D_0=0) \\
 \mathbf{P}(Y_1=1, Y_0=1, D_1=0, D_0=1) \\
 \mathbf{P}(Y_1=1, Y_0=1, D_1=1, D_0=0) \\
 \mathbf{P}(Y_1=1, Y_0=1, D_1=1, D_0=1)
 \end{array} \right)}_p
 =
 \underbrace{\left(\begin{array}{l}
 \mathbf{P}(Y=0, D=0, Z=0) \\
 \mathbf{P}(Y=0, D=1, Z=0) \\
 \mathbf{P}(Y=1, D=0, Z=0) \\
 \mathbf{P}(Y=1, D=1, Z=0) \\
 \mathbf{P}(Y=0, D=0, Z=1) \\
 \mathbf{P}(Y=0, D=1, Z=1) \\
 \mathbf{P}(Y=1, D=0, Z=1) \\
 \mathbf{P}(Y=1, D=1, Z=1) \\
 \hline
 \mathbf{P}(Y=0, D=0) \\
 \mathbf{P}(Y=0, D=1) \\
 \mathbf{P}(Y=1, D=0) \\
 \mathbf{P}(Y=1, D=1) \\
 \hline
 \mathbf{P}(Y=0, Z=0) \\
 \mathbf{P}(Y=1, Z=0) \\
 \mathbf{P}(Y=0, Z=1) \\
 \mathbf{P}(Y=1, Z=1) \\
 \hline
 \mathbf{P}(D=0, Z=0) \\
 \mathbf{P}(D=1, Z=0) \\
 \mathbf{P}(D=0, Z=1) \\
 \mathbf{P}(D=1, Z=1) \\
 \hline
 1 \\
 0 \\
 0
 \end{array} \right)}_{\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}}$$

Figure 1. Equation for Example 8

$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$ has been decomposed into $C_1 C_2$. Let \hat{C}_1 and \hat{C}_2 replace q_0 and q_1 with \hat{q}_0 and \hat{q}_1 , respectively. $\begin{pmatrix} \hat{A}_1 \\ A_2 \end{pmatrix} = \hat{C}_1 \hat{C}_2$. If $q_0 > 0$ and $q_1 > 0$, $\text{rank}(C_2) = \text{rows}(C_2) = \text{rank}(\hat{C}_2) = \text{rows}(\hat{C}_2)$ with probability going to one. Therefore, $\text{rank}\left(\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}\right) = \text{rank}(C_1 C_2) = \text{rank}(C_1)$ and $\text{rank}\left(\begin{pmatrix} \hat{A}_1 \\ A_2 \end{pmatrix}\right) = \text{rank}(\hat{C}_1 \hat{C}_2) = \text{rank}(\hat{C}_1)$. Removing certain rows C_1 and \hat{C}_1 does not reduce the rank and results in the matrix as follows:

$$\left(\begin{array}{cccccc|ccc}
 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
 \hline
 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
 \hline
 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
 \hline
 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
 \hline
 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
 \hline
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
 \end{array} \right)$$

The matrix above has the same rank as C_1 and \hat{C}_1 . Notice that removing rows of C_1 or the equivalent rows of \hat{C}_1 results in this matrix. This can be done similarly with other combinations of variables observed in the samples.