Extrapolation of Instrumental Variables Estimators Towards Externally Valid Parameters

Magne Mogstad* Alexander Torgovitsky†

March 25, 2016

PRELIMINARY DRAFT. PLEASE DO NOT CIRCULATE.

Abstract

We reconsider the empirical content of instrumental variables (IV) estimators under the common instrument monotonicity assumption that there is a weakly separable selection equation. It is well-known that if there is unobserved heterogeneity in the effect of the treatment on the outcome of interest, then the IV estimand can be interpreted as a weighted average of causal effects. It is also well-known that these weights depend on the specific instrument being used and that, as a result, the IV estimand may not be equal to an externally valid parameter that answers an analyst's counterfactual question of interest. We show that, despite this, the IV estimand still contains substantial empirical content for externally valid parameters. Our argument uses the insight of Heckman and Vytlacil (2005) that both externally valid parameters and the IV estimand can be expressed as weighted averages of the same underlying marginal treatment effects (MTEs). Since both sets of weights are known or identified, knowledge of the IV estimand generally places some restrictions on the MTEs, and hence on the logically permissible values of externally valid parameters. We provide a simple computational method for deriving these implied bounds.

*(fill)

†Department of Economics, Northwestern University, a-torgovitsky@northwestern.edu. This research was supported in part by National Science Foundation grant SES-1530538. I am grateful to the Department of Economics at the University of Chicago for supporting me hospitably through a visiting position while I completed this research.

1 Introduction

This paper is about the classic program evaluation problem in which an analyst seeks to measure the causal effect (broadly defined) of a binary treatment $D \in \{0, 1\}$ on a scalar, real-valued outcome Y. Corresponding to the two treatment arms are unobservable potential outcomes, Y_0 and Y_1, that represent the realization of Y that would have been experienced by an agent had their treatment been exogenously set to 0 or 1. The relationship between Y and D is given by the switching regression

$$Y = DY_1 + (1 - D)Y_0. \tag{1}$$

The analyst observes data (Y, D, X, Z) for a random sample of agents, where X are some covariates and Z is an appropriately exogenous instrument. Their goal is to use this observed data to identify an externally valid causal effect, i.e. a summary measure of $Y_1 - Y_0$ that is relevant for answering some specific counterfactual question of interest.

Heckman and Vytlacil (2005) show that a broad array of causal effects for mean outcomes can be represented as weighted averages of marginal treatment effects (MTEs). An implication of their work is that the task of point identifying mean causal effects can be reduced to one of point identifying the MTEs for which these weights are non-zero. This is often difficult, especially when the exogenous variation contained in Z is discrete, which is common in empirical applications, for example in randomized controlled experiments with partial compliance. When Z is continuous, it is generally possible to point identify what Carneiro, Heckman, and Vytlacil (2010) describe as marginal policy relevant treatment effects (PRTEs), or the causal effect on Y of a small change in a variable that affects an agent's realized treatment arm D. However, even in this case, it is still generally not possible for an analyst to point identify infra– or extra–marginal PRTEs without additional assumptions.

In this paper, we develop a procedure for placing bounds on PRTEs and other externally valid causal parameters that is valid regardless of the support of Z. Our procedure utilizes another result derived by Heckman and Vytlacil (2005), which shows that the estimand corresponding to the instrumental variables (IV) estimator can also be written as a weighted average of marginal treatment effects with weights that are identified. We extend this result to show that the same is true for a more general class of IV-like estimators that can be written as cross-moments between the outcome and a vector-valued function of the instruments. This class includes the widely used two-stage least squares (TSLS) estimator, as well as many others. The main insight of

this paper is that even though the IV and TSLS estimators do not, in general, directly identify the causal effect of interest, they still carry information about this causal effect by providing information on the underlying MTEs.

Our primary methodological contribution is to show how to aggregate and extract this information on the MTEs, and then use it to extrapolate and draw formal conclusions about a given, pre-specified, externally valid parameter of interest to the analyst. We show that this can be done by solving linear programming problems. These linear programs determine the largest and smallest value of the externally valid parameter of interest that is consistent with a collection of MTEs that could have also given rise to the observed TSLS estimators. We develop our procedure in the population first. We then modify it to account for statistical variation in a given sample of data and provide a method for constructing valid confidence regions.

(to be added: context in the literature)

2 Model and Assumptions

We augment the switching regression (1) with a weakly separable selection equation

$$D = \mathbb{1}[\nu(Z) - V \geq 0], \tag{2}$$

where V is a continuously distributed unobservable random variable and ν is some unknown function. We maintain the following assumptions on (1) and (2).

A.1 *$\nu(Z)$ is a nondegenerate random variable, conditional on X.*

A.2 *$(Y_0, Y_1, V) \perp\!\!\!\perp Z|X$.*

A.3 *V is continuously distributed.*

Throughout the paper we regard Z as containing X, which simplifies the notation. We refer to the portion of Z that is not contained in X as Z_0. Assumption A.2 requires Z_0 to be exogenous, conditional on X, while A.1 requires that Z_0 is a non-trivial random variable after conditioning on X. Our analysis does not require assumptions about the variation in Z_0 other than A.1. In particular, Z_0 could be a single binary random variable as long as it still has some variation after conditioning on X. In fact, we do not actually need A.1 at all for our analysis to be valid, but the bounds derived when it does not hold will be uninteresting, so we maintain it for emphasis.

Given A.3, one can normalize the selection equation (2) as

$$D = \mathbb{1}[U \leq P(Z)], \tag{3}$$

where $U \equiv F_{V|Z}(V|Z)$ is distributed uniformly on $[0,1]$, independently of Z, and $P(Z) \equiv \mathbb{P}[D=1|Z] = F_{V|Z}(\nu(Z)|Z)$ is the propensity score. Note that, given A.2 and our notational convention that Z contains X, this normalization also implies that $U|X=x$ is distributed uniformly on $[0,1]$ for any supported x. The model defined by (1), (3) and A.1–A.3 is often called the generalized Roy model. Vytlacil (2002) showed that the generalized Roy model is equivalent to the potential outcomes model studied by Imbens and Angrist (1994) under their exogeneity and instrument monotonicity conditions.

Heckman and Vytlacil (2005) argue that analyses of the generalized Roy model can be unified by studying the marginal treatment effect (MTE), which is defined for all $u \in [0,1]$ and $x \in \mathcal{X} \equiv \text{supp}(X)$ as

$$\overline{m}(u,x) \equiv \mathbb{E}[Y_1 - Y_0 | U=u, X=x].$$

In particular, they show that many parameters of interest concerning averages of Y_1 and Y_0 can be expressed as different weighted averages of $\overline{m}$, i.e. as $\int_0^1 \overline{m}(u,x)\omega(u,z)\,du$ for some known or identifiable weighting function ω. In our analysis, it will be more convenient to view the MTE as the difference of two marginal treatment response (MTR) functions $m_1(u,x) \equiv \mathbb{E}[Y_1|U=u,X=x]$ and $m_0(u,x) \equiv \mathbb{E}[Y_0|U=u,X=x]$. In light of the results of Heckman and Vytlacil (2005), we assume that the analyst is interested in a parameter $\beta^\star$ defined as

$$\beta^\star \equiv \int_0^1 \mathbb{E}\left[m_0(u,X)\omega_0^\star(u,Z)\right]\,du, + \int_0^1 \mathbb{E}\left[m_1(u,X)\omega_1^\star(u,Z)\right]\,du, \tag{4}$$

where $\omega_0^\star$ and $\omega_1^\star$ are known or identifiable weighting functions. Our specification of $\beta^\star$ nests parameters that can be written as a weighted average of $\overline{m}$ by taking $\omega_0^\star(u,Z) = -\omega_1^\star(u,Z)$. Some interesting parameters that *can not* be written in form (4) include the quantiles of $Y_1 - Y_0$, or the proportion of agents for whom $Y_1 \geq Y_0$ (Heckman, Smith, and Clements, 1997).

We assume that $m \equiv (m_0, m_1)$ belongs to some class of functions denoted $\mathcal{M}$ that is a subset of a vector space. Later, we will consider placing additional identifying restrictions on these classes and/or on the derived class $\overline{\mathcal{M}} \equiv \{m_1 - m_0 : m \in \mathcal{M}\}$ of MTEs. For now, we only impose the following mild assumption in order to ensure that $\beta^\star$ is well-defined.

A.4 (Existence of moments) *The right-hand side of* (4) *is finite for any* $m \in \mathcal{M}$.

A simple example of weights is $\omega_1^\star(u,z) = 1$ and $\omega_0^\star(u,z) = -1$, in which case $\beta^\star$ is the average treatment effect (ATE). While commonly studied, the ATE may not

be particularly relevant for addressing the effects of a specific policy counterfactual. Heckman and Vytlacil (2005) provide expressions for the weights that correspond to other mean treatment effects, such as the average treatment on the treated (ATT) and the local average treatment effect (LATE). In addition, they introduce a new treatment effect that represents the aggregate mean effect on the population that results from a change in the distribution of the propensity score P. They describe this treatment effect as a policy-relevant treatment effect (PRTE), where the policy refers to the institutional adjustment that affects an agent's treatment decision, as summarized through the propensity score. Our focus on any parameter $\beta^\star$ that can be defined as in (4) is general enough to address both traditional types of causal parameters (ATE, ATT, LATE, etc.) and PRTEs.

One parameter that is of particular interest for the current analysis is the estimand corresponding to the linear instrumental variables estimator that uses some measurable function $J(Z)$ as an instrument for D, when X is either empty or conditioned on nonparametrically. Heckman and Vytlacil (2005) show that this IV estimand, call it $\beta_{\text{IV},J}$, can be written in the form (4). A leading case of interest is when $Z \in \{0,1\}$ is binary, in which case there is only one such non-trivial function J, and $\beta_{\text{IV},J}$ can be interpreted as the local average treatment effect, i.e. the average treatment effect among the "compliers" (Angrist and Imbens, 1995). The next proposition shows that, more generally, any cross-moment of Y with a known or identifiable function of Z can also be expressed as the weighted sum of the two marginal treatment response functions. We refer to such cross-moments as "IV–like" estimands. All proofs are contained in Appendix A.

Proposition 1. *Suppose that s is a known or identifiable, vector-valued function of Z that is measurable and has bounded second moments. We refer to such an s as a* specification *and write $S = s(Z)$ for the random vector it generates. For any s, define $\beta_s \equiv \mathbb{E}[SY]$. Then*

$$\beta_s = \int_0^1 \mathbb{E}\left[m_0(u,X)\omega_{0s}(u,Z)\right]\, du + \int_0^1 \mathbb{E}\left[m_1(u,X)\omega_{1s}(u,Z)\right]\, du,$$
$$\text{where } \omega_{0s}(u,Z) \equiv S\left(1 - \mathbb{1}[P(Z) \geq u]\right)$$
$$\text{and } \omega_{1s}(u,Z) \equiv S\mathbb{1}[P(Z) \geq u],$$

as long as the order of integration with respect to u and expectation can be interchanged.

It is straightforward to verify that Proposition 1 reduces to the weights for $\beta_{\text{IV},J}$ derived by Heckman and Vytlacil (2005) by taking

$$s(Z) = \frac{J(Z) - \mathbb{E}[J(Z)]}{\text{Cov}(J(Z), D)}, \tag{5}$$

which is an identifiable function of Z. As we discuss further in Appendix B, Proposition 1 applies more broadly to cover essentially any well-defined linear IV estimand that uses some function of Z as included and excluded instruments for a set of endogenous variables constructed from D and X. For our analysis, the denominator on the weights for these types of IV estimands (e.g. $\text{Cov}(J(Z), D)$ in (5)) turns out to be irrelevant, since it simply scales the weights by a constant. That is, taking the specification to be $s(Z) = J(Z) - \mathbb{E}[J(Z)]$ will lead to a set of weights with the same empirical content as the choice of s in (5), but which do not necessarily integrate to 1.

The implication of Proposition 1 relevant to this paper is that, while linear IV estimators generally cannot be expected to directly identify an externally valid parameter of direct interest to the analyst, they still carry some information on the underlying MTR functions. If an analyst could extract this information from the IV estimator, then they could use it to draw conclusions about $\beta^\star$. The remainder of the paper is about the development and application of a formal method for performing this task.

3 Extrapolation Towards Externally Valid Parameters

3.1 Bounds Implied by IV–Like Estimands

Throughout this section we assume that the analyst knows the joint distribution of the observed data (Y, D, X, Z) without error. We address issues of statistical inference in the next section.

Let $\mathcal{S}$ denote some finite collection of specifications (functions s), chosen by the analyst, all of which satisfy the conditions set out in Proposition 1. Corresponding with $\mathcal{S}$ is the set of IV–like estimands $\mathcal{B}_\mathcal{S} \equiv \{\beta_s : s \in \mathcal{S}\}$ also defined as in Proposition 1. Our goal is to characterize the range of values of the externally valid parameter $\beta^\star$ that are logically consistent (or "observationally equivalent") with $\mathcal{B}_\mathcal{S}$. To this end, denote the mapping in Proposition 1 as $\Gamma_s : \mathcal{M} \to \mathbb{R}^{d_s}$, defined for any specification s as

$$\Gamma_s(m_0, m_1) = \int_0^1 \mathbb{E}\left[m_0(u, X)\omega_{0s}(u, Z)\right] du + \int_0^1 \mathbb{E}\left[m_1(u, X)\omega_{1s}(u, Z)\right] du,$$

where d_s is the dimension of $S \equiv s(Z)$. We denote the set of all MTR pairs $m \equiv$

(m_0, m_1) that are consistent with the estimands $\mathcal{B}_\mathcal{S}$ as

$$\mathcal{M}^\star \equiv \{m \in \mathcal{M} : \Gamma_s(m) = \beta_s \; \forall s \in \mathcal{S}\}.$$

Similarly, denote the mapping that defines $\beta^\star$ as $\Gamma^\star : \mathcal{M} \to \mathbb{R}$, with

$$\Gamma^\star(m_0, m_1) \equiv \int_0^1 \mathbb{E}[m_0(u, X)\omega_0^\star(u, Z)]\, du + \int_0^1 \mathbb{E}[m_1(u, X)\omega_1^\star(u, Z)]\, du,$$

and define $\mathcal{B}^\star \equiv \{\Gamma^\star(m) : m \in \mathcal{M}^\star\}$. The set $\mathcal{B}^\star$ is the set of values of $\beta^\star$ that are consistent with the analyst's collection of IV–like estimands $\mathcal{B}_\mathcal{S}$. In the following, we will develop a computationally straightforward method for computing $\mathcal{B}^\star$.

The next proposition shows that $\mathcal{B}^\star$ is either empty or an interval as long as $\mathcal{M}$ is assumed to be convex.

Proposition 2. *Suppose that $\mathcal{M}$ is convex. Then either $\mathcal{B}^\star$ is empty or $\mathcal{B}^\star = [\underline{\beta}^\star, \overline{\beta}^\star]$ for some $\underline{\beta}^\star, \overline{\beta}^\star \in \mathbb{R} \cup \{\pm\infty\}$ with $\underline{\beta}^\star \leq \overline{\beta}^\star$.*

The endpoints of $\mathcal{B}^\star$ can be characterized as the solutions to two optimization problems. The lower endpoint is determined by

$$\underline{\beta}^\star = \min_{m \in \mathcal{M}} \Gamma^\star(m) \quad \text{subject to} \quad \Gamma_s(m) = \beta_s \; \forall s \in \mathcal{S}, \tag{6}$$

and the upper endpoint is determined by the analogous maximization problem.

Optimization problem (6) is an infinite-dimensional linear program. In order to solve it, we assume that $\mathcal{M}$ has an exact or approximate representation as a linear mixture of known basis functions. In particular, suppose that every $m_d \in \mathcal{M}_d$ can be represented as

$$m_d(u, x) = \sum_{k=1}^{K_d} \theta_{dk} b_{dk}(u, x) \quad \text{for all } u \in [0, 1] \text{ and } x \in \mathcal{X}, \tag{7}$$

for some $\theta_d \equiv (\theta_{d1}, \ldots, \theta_{dK_d}) \in \mathbb{R}^{K_d}$, where $\{b_{dk}\}_{k=1}^{K_d}$ are some known basis functions that could be different for $d = 0, 1$ if desired. Notationally, we let $\theta \equiv (\theta_0, \theta_1)$, $b_d(u, x) \equiv (b_{d1}(u, x), \ldots, b_{dK_d}(u, x))$ for $d = 0, 1$ and $b(u, x) \equiv (b_0(u, x), b_1(u, x))$, so that (7) can be written more compactly as

$$m(u, x) = \theta' b(u, x) \quad \text{for all } u \in [0, 1] \text{ and } x \in \mathcal{X}. \tag{8}$$

Given (8), the linearity of $\Gamma^\star$ and Γ_s implies that (6) can be written as the following

finite-dimensional linear programming problem

$$\underline{\beta}^\star = \min_\theta \sum_{k=1}^{K_0} \theta_{0k}\Gamma_0^\star(b_{0k}) + \sum_{k=1}^{K_1} \theta_{1k}\Gamma_1^\star(b_{1k}) \tag{9}$$
$$\text{subject to } \sum_{k=1}^{K_0} \theta_{0k}\Gamma_{0s}(b_{0k}) + \sum_{k=1}^{K_1} \theta_{1k}\Gamma_{1s}(b_{1k}) = \beta_s \; \forall s \in \mathcal{S},$$

where we have added the definitions

$$\Gamma_d^\star(b_{dk}) \equiv \int_0^1 \mathbb{E}[b_{dk}(u,X)\omega_d^\star(u,Z)]\,du \text{ for } d = 0,1,\ k = 1,\ldots,K_d$$
$$\text{and } \ \Gamma_{ds}(b_{dk}) \equiv \int_0^1 \mathbb{E}[b_{dk}(u,X)\omega_{ds}(u,Z)]\,du \text{ for } d = 0,1,\ k = 1,\ldots,K_d \text{ and } s \in \mathcal{S}.$$

Since the weights and basis functions are all known or identified, the quantities $\Gamma^\star(b_{dk})$ and $\Gamma_{ds}(b_{dk})$ are also identified for all d, k and s. Hence, in the population we can find $\underline{\beta}^\star$ by solving (9), which is computationally straightforward, even if the dimensions of K_0, K_1 and/or $\mathcal{S}$ are quite large. We find $\overline{\beta}^\star$ by solving the analogous maximization problem. The resulting interval $[\underline{\beta}^\star, \overline{\beta}^\star]$ contains all values of $\beta^\star$ that are consistent with the set of specifications $\mathcal{S}$ considered by the analyst under the assumption that $\mathcal{M}$ only contains functions of the form (7).

In Section 4, we allow K_0 and K_1 to diverge with the sample size as in the method of sieves (Chen, 2007). If the basis functions $\{b_{dk}\}_{k=1}^{K_d}$ are taken to be elements of a flexible approximating class, then this enables us to view our procedure as nonparametrically approximating $\mathcal{M}$. For now, we consider K_0 and K_1 to be fixed and finite. Thus, one should think of (8) as a flexible parametric form for m, with the form becoming more flexible as K_0 and/or K_1 increase. Since linear programs scale well, it is generally feasible to take K_0 and K_1 to be quite large, so that (8) can be a *very* flexible parametric form.

Chen (2007) provides a thorough discussion of several potential linear approximating bases that could be used for the $\{b_{dk}\}_{k=1}^{K_d}$ functions, such as polynomials, splines, wavelets, and B-splines. We refer the reader to that article for a detailed exposition. One factor in choosing an approximating basis is the generality of the space of functions $\mathcal{M}$ that it can approximate well. Another consideration is the rate at which this approximation occurs. As a more practical matter, we will also generally want to be able to impose shape restrictions on elements of $\mathcal{M}$, as discussed in the next section. In the applications in this paper we use either Bernstein polynomials or B-splines as the approximating basis, since for these bases imposing the shape restrictions we consider

turns out to be tantamount to simply adding some linear constraints on θ in (9). See Appendix C for more details.

3.2 Additional Identifying Assumptions

In addition to using (8) as a flexible parametric form (or approximation) for the collection of functions $\mathcal{M}$, one can also impose nonparametric shape restrictions on $\mathcal{M}$. A natural shape restriction to consider is that every $m_d \in \mathcal{M}_d$ has range in $[\underline{y}, \overline{y}]$, where $\underline{y}$ and $\overline{y}$ are known constants. Most outcome variables in economic applications have a logical upper or lower bound (e.g. 0), while some outcome variables, notably binary outcomes, are known to have finite support. If the specification of m in (8) is taken to be so flexible as to be effectively nonparametric, and if the supports of $\omega_1^\star$ and/or $\omega_0^\star$ have some regions that do not overlap with ω_{ds} for any $d = 0, 1$ or $s \in \mathcal{S}$, then such a bounding assumption is necessary to ensure that $\underline{\beta}^\star$ and $\overline{\beta}^\star$ are finite. The intuition for this observation is exactly the same as the worst-case bounds of Manski (1989), but here applied to the regions of $[0, 1]$ for which the data is completely uninformative about $\overline{m}(u, x)$.

Other nonparametric restrictions that can be imposed on m and hence $\overline{m}$ include non-negativity, i.e. $m_1(u, x) \geq m_0(u, x)$ for all $u \in [0, 1]$ and $x \in \mathcal{X}$, if it is reasonable to assume that the treatment does not decrease outcomes, as suggested by Manski (1997). In a similar vein, one could require $\overline{m}(\cdot, x)$ to be weakly decreasing for every x, if it is thought that agents most likely to select into treatment (those with small realizations of U) are also most likely to have the largest gains from the treatment. This is similar to the "monotone treatment selection" assumption of Manski and Pepper (2000).

There is also additional scope for placing parametric assumptions on $\mathcal{M}$, both to tighten the bounds, and to reduce the computational burden of solving (9). The former motivation is more compelling, since the difficulty of solving (9) scales well with K_d. However, if the dimension of X is large, as it often is in applications, then the second motivation may become more important. Regardless of the motivation, it is straightforward to incorporate such a restriction into (7) by assuming, for example, that

$$b_{dk}(u, x) = \overline{b}_{dk}(u) + x'\beta_d \tag{10}$$

for some basis functions $\{\overline{b}_{dk}\}_{k=1}^{K_d}$ of u alone, and a coefficient vector β_d. Such a specification tightens the bounds and reduces the dimensionality of the problem, while still preserving the computationally convenient linear structure in (9). It is straightforward to modify (10) to allow for certain x variables to interact with u if this is considered

important in a given application.

3.3 Numerical Illustration

We illustrate the procedure described in the previous sections with the following data generating process:

$$\begin{aligned} Y_0 &= \mathbb{1}[U \leq V_0] & (\eta_0, \eta_1) &\sim N([-.2, .4], I_2) \\ Y_1 &= \mathbb{1}[U \leq V_1] & U &\sim \text{Unif}[0,1],\ U \perp\!\!\!\perp (\eta_0, \eta_1) \\ D &= \mathbb{1}[U \leq P(Z)] & P(z) &\equiv \Phi(z - .1), \end{aligned} \tag{11}$$

where Φ is the cumulative distribution function of a standard normal distribution and I_2 is the two–dimensional identity matrix. The first distribution of Z that we consider is discrete with support $\mathcal{Z} \equiv \{-.5, -.4, \ldots, .4, .5\}$ having 11 elements. The density (with respect to counting measure) of Z is set to

$$f_Z(z_l) = \frac{\Phi_Z(z_l) - \Phi_Z(z_{l-1})}{\Phi_Z(z_L)}, \tag{12}$$

where z_l is the lth largest element of $\mathcal{Z}$ (with $z_0 \equiv -\infty$ and $z_L \equiv .5$), and Φ_Z is the cumulative distribution function for a $N(0, .3^2)$ distribution. Since $Y \in \{0, 1\}$, we take $\mathcal{M}_0$ and $\mathcal{M}_1$ to consist of only functions that are bounded below by 0 and above by 1.

We consider the task of identifying the average treatment on the treated (ATT) $\beta^\star_{\text{ATT}} \equiv \mathbb{P}[Y_1 - Y_0 | D = 1]$, which can be expressed in form (4) with weights

$$\omega^\star_{1,\text{ATT}}(u, Z) \equiv \frac{\mathbb{1}[P(Z) \geq u]}{\mathbb{P}[D = 1]} \quad \text{and} \quad \omega^\star_{0,\text{ATT}}(u, Z) = -\omega^\star_{1,\text{ATT}}(u, Z).$$

Figure 1 plots these weights against the weights for β_s when s is defined through (5) with $J(Z) = P(Z)$. Heckman and Vytlacil (2005) showed that of all functions J that could be used as instruments in (5), taking $J(Z) = P(Z)$ leads to the weighting function ω_{1s} with the largest support. Evidently, in this DGP this support does not completely overlap with that of $\omega^\star_{1,\text{ATT}}$, which places the most weight on u near 0. Hence, one should expect that, barring any additional assumptions, $\beta^\star_{\text{ATT}}$ will not be point identified. The true value of $\beta^\star_{\text{ATT}}$, i.e. the value implied by (11) is .235.

We model m_0 and m_1 with the Bernstein polynomials and consider different choices of basis length $K \equiv K_0 = K_1$. The $(k-1)$th Bernstein polynomial of degree $K - 1$ is

defined as

$$b_{dk}(u) = \binom{K-1}{k-1} u^{k-1}(1-u)^{(K-k)}$$

for $k = 1, \ldots, K$. The Bernstein polynomials are an example (a historically significant one) of a class of basis functions that can uniformly approximate any continuous function on $[0, 1]$, but there are many others. For our purposes, a convenient property of the Bernstein polynomials is that it is straightforward to require their linear combinations to obey shape restrictions, such as $m_d \in [0, 1]$, by putting linear restrictions on θ. These linear restrictions are added as constraints in (9).

Columns (1)–(4) of Table 1 report the bounds on $\beta^\star_{\text{ATT}}$ derived from solving (9) for increasing values of K when $\mathcal{S}$ contains only the specification s defined through (5) with $J(Z) = P(Z)$. The bounds increase with K, since this increases the number of variables in the program (9). More intuitively, the more flexible the parametric specification is, the fewer assumptions are imposed on the classes $\mathcal{M}_d$, and the larger the bounds. The bounds can be viewed as approaching nonparametric as $K \to \infty$.

Evidently, K has a dramatic effect on the size of the bounds. Some intuition for this effect can be seen in Figure 3, which reports the MTE function obtained at the lower and upper bounds when $K = 5$ (column (2) of Table 1). The lower bound function, displayed in solid blue, tries to become as small as possible on the region near 0 where the ATT weights are large—compare with Figure 1. At the same time, this function needs to still achieve a certain value of β_s and so must be sufficiently large over the region where the weights for β_s (s as in (5) and $J(Z) = P(Z)$) are non-zero. The optimal lower bound function tries to plunge as quickly as possible to 0 while still maintaining this restriction. The larger that K is, the faster it can plunge, since larger values of K allow for more quickly adjusting functions. For comparison, Figure 4 shows the same plot with $K = 15$ (column (4) of Table 1). The dotted blue line in both figures shows an MTE function that delivers a value of $\beta^\star_{\text{ATT}}$ that is smaller than $\underline{\beta}^\star$, but which does not deliver the required value of β_s, and is therefore logically inconsistent with the observed data. Analogous intuition applies to the optimal and infeasible MTE functions for the upper bound, which are shown in red.

Columns (5)–(8) of Table 1 report analogous bounds when $\mathcal{S}$ is increased to also contain the specification of s in (5) with $J(Z) = Z$. As expected, the bounds are uniformly smaller than for columns (1)–(4), since more restrictions are maintained in the optimization problem (9). In fact, with $K = 2$, $\mathcal{B}^\star$ shrinks to a singleton, since there is only one choice of θ that is consistent with both β_s in $\mathcal{B}_\mathcal{S}$. Intuitively, using

a larger set of specifications corresponds to exploiting more of the observed data, and therefore strengthens the observational equivalence requirement. Given the maintained assumptions of the generalized Roy model, the only potential reason to prefer columns (1)–(4) to columns (5)–(8) is statistical. We discuss statistical inference in Section 4.

Columns (12)–(15) of Table 1 push this observation farther by including specifications that correspond to the s defined through (5) for three other additional choices of $J(Z)$. The weights corresponding to these additional specifications are shown in Figure 2. The impact on the bounds is so dramatic that even with $K = 5$ the bounds are empty, i.e. the program in (9) is infeasible. This yields two observations. First, the five specifications included in $\mathcal{S}$ have tremendous identifying content for $\beta^{\star}_{\text{ATT}}$. Second, the true MTE function corresponding to the data generating process, while nearly linear (see Figure 3), is actually the difference of two normal cumulative distribution functions (see (11)) and so does not have an exact representation as a Bernstein polynomial with $K = 5$ terms. Hence, not only do the bounds derived from (9) have substantial identifying content, they can also be used to falsify a given parametric specification. When K is increased to 10 and 15, the parametric form becomes sufficiently flexible as to be able to rationalize all of the specifications in $\mathcal{S}$ simultaneously. However, the resulting MTE function is sufficiently restricted as to be logically inconsistent with a wide range of values of $\beta^{\star}_{\text{ATT}}$, thereby leading to very tight bounds.

Table 2 shows how this observation could be exploited systematically. Here, the specifications are chosen so that their respective instruments are indicator functions, each one carrying some unique information not contained in any of the others. For each additional set of specifications that is added there is a dramatic decrease in the length of the bounds. With the full specification, we need very large polynomial lengths K to find functions that do not lead to empty bounds, and even so the bounds are quite tight. We are developing a formal approach for implementing this type of systematic selection of specifications $\mathcal{S}$.

Table 3 repeats the exercise of Table 1 when the support of Z is reduced to the three element set $\mathcal{Z} \equiv \{-.5, 0, .5\}$. The data generating process remains otherwise unchanged, although this modification of the distribution of Z changes $\beta^{\star}_{\text{ATT}}$ slightly (at the fourth digit) through its effect on which agents select into treatment. Figure 5 provides a counterpart to Figure 1 for this case, and shows that the weights are constant over larger regions. The effect that this has the bounds varies considerably depending on which column one considers. Perhaps surprisingly, in some columns this reduction of the support markedly shrinks the bounds. This is likely due to the fact that as the support is shrunk with the density kept at (12), there will be relatively more agents with $Z = -.5$. The important takeaway is that having an instrument with

more points of support is not necessarily beneficial in terms of empirical content.

In Table 4 we instead modify the support of Z to be $\mathcal{Z} \equiv \{0, .1, \ldots, .9, 1\}$, i.e. we add .5 to each support point in the specification for Table 1. This has the effect of increasing the support of $P(Z)$ and consequently shifts the weights for β_s when s is given by (5) with $J(Z) = P(Z)$ to be farther away from those for $\beta^{\star}_{\text{ATT}}$—see Figure 6. As one would expect, the bounds on $\beta^{\star}_{\text{ATT}}$ become wider, reflecting the fact that the parameter being extrapolated to has become increasingly different from those that are being estimated. Intuitively, the process of extrapolating is easier when the degree of extrapolation required is smaller.

4 Statistical Inference

5 Application

6 Conclusion

A Proofs

Proof of Proposition 1. Using (1), note that

$$\beta_s = \mathbb{E}[SDY_1] + \mathbb{E}[S(1-D)Y_0].$$

Then using (3), observe that

$$\mathbb{E}[SDY_1] = \mathbb{E}\left(S\mathbb{1}[P(Z) \geq U]\,\mathbb{E}\left[Y_1|U,X,Z\right]\right) = \mathbb{E}\left[S\mathbb{1}[P(Z) \geq U]m_1(U,X)\right], \qquad (13)$$

where the second equality uses the implication of A.2 that $Y_1 \perp\!\!\!\perp Z|U,X$. Using the normalization of $U|Z$ as uniformly distributed on $[0,1]$, independently of Z, and our notational convention that Z contains X, it follows from (13) that

$$\begin{aligned}\mathbb{E}[SDY_1] &= \mathbb{E}\left[\mathbb{E}\left(S\mathbb{1}[P(Z) \geq U]m_1(U,X)|Z\right)\right] \\ &= \mathbb{E}\left[\int_0^1 S\mathbb{1}[P(Z) \geq u]m_1(u,X)\,du\right] = \int_0^1 \mathbb{E}\left[S\mathbb{1}[P(Z) \geq u]m_1(u,X)\right]\,du,\end{aligned}$$

where the final equality follows by the assumption that the order of integration and expectation can be interchanged. The claimed result follows after using a symmetric argument for Y_0. *Q.E.D.*

Proof of Proposition 2. Observe that the mapping Γ_s is linear in m. Hence, $\mathcal{M}^\star$ is either convex or empty. Since $\Gamma^\star$ is also linear in m, the image of $\mathcal{M}^\star$ under $\Gamma^\star$ is also convex, if it is nonempty. Hence, either $\mathcal{B}^\star$ is empty, or $\mathcal{B}^\star$ is convex, i.e. there exist $\underline{\beta}^\star, \overline{\beta}^\star \in \overline{\mathbb{R}}$ with $\underline{\beta}^\star \leq \overline{\beta}^\star$ such that $\mathcal{B}^\star = [\underline{\beta}^\star, \overline{\beta}^\star]$. *Q.E.D.*

B MTR Weights for Linear IV Estimands

C Approximating Bases

References

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442. 5

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394. 2

Chen, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 5549–5632. 8

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64, 487–535. 4

Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. 1, 2, 4, 5, 6, 10

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 4

Manski, C. F. (1989): "Anatomy of the Selection Problem," *The Journal of Human Resources*, 24, 343–360. 9

——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334. 9

Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010. 9

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341. 4

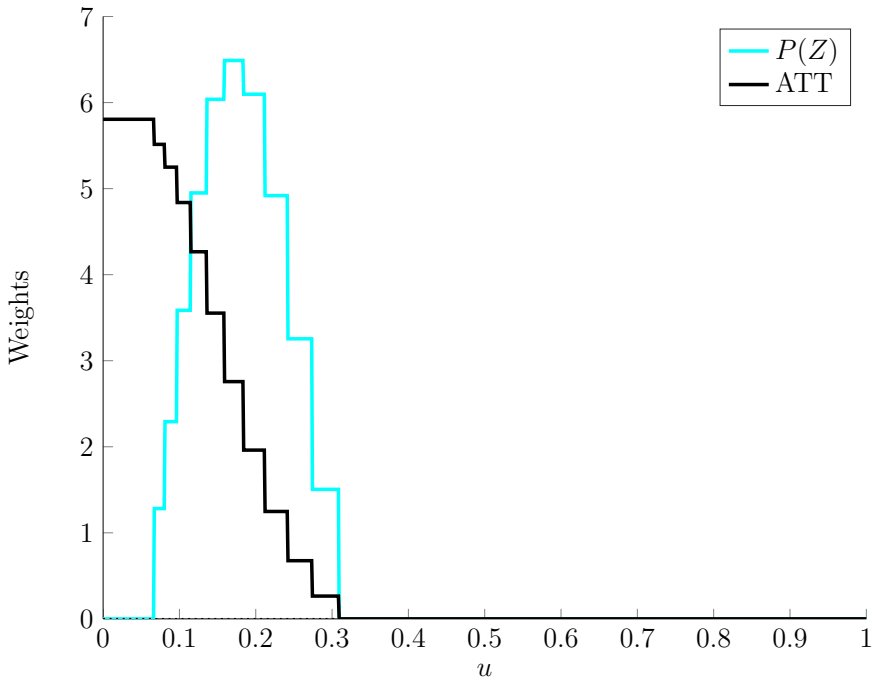


Figure 1: Weights for columns (1)–(4) of Table 1.

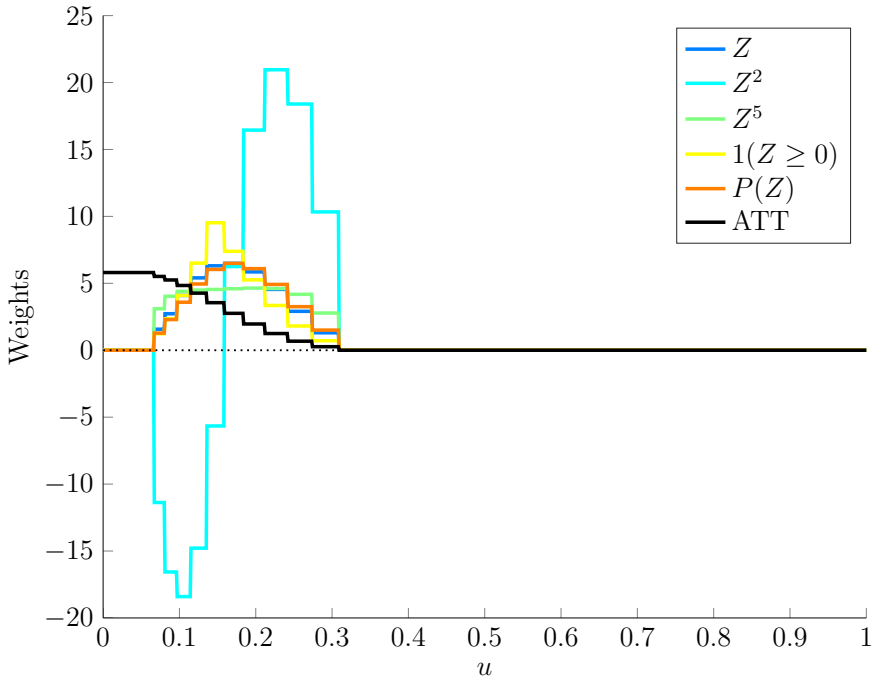


Figure 2: Weights for columns (5)–(8) of Table 2.

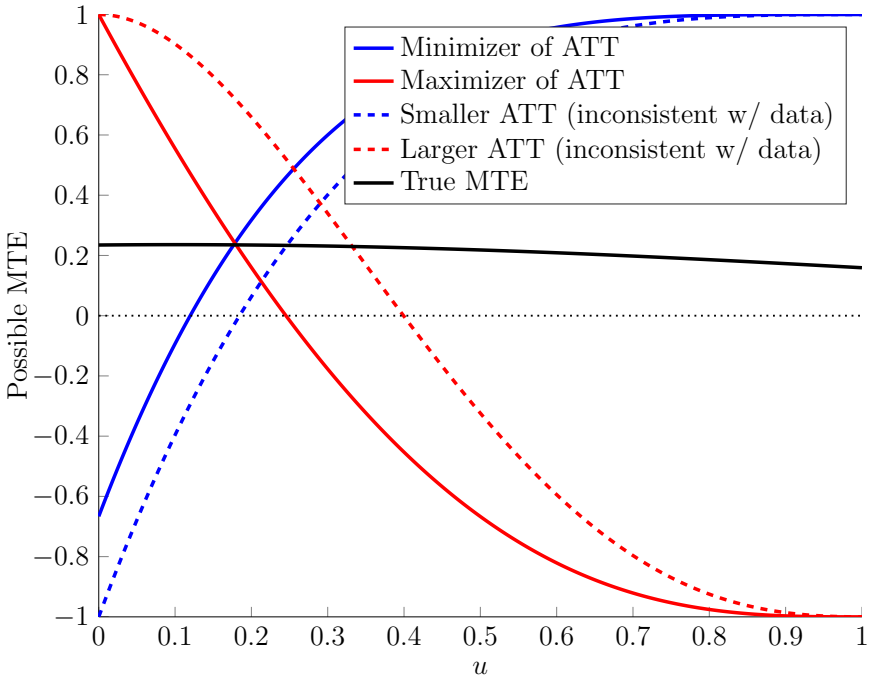


Figure 3: Optimal and infeasible MTEs for column (2) of Table 1.

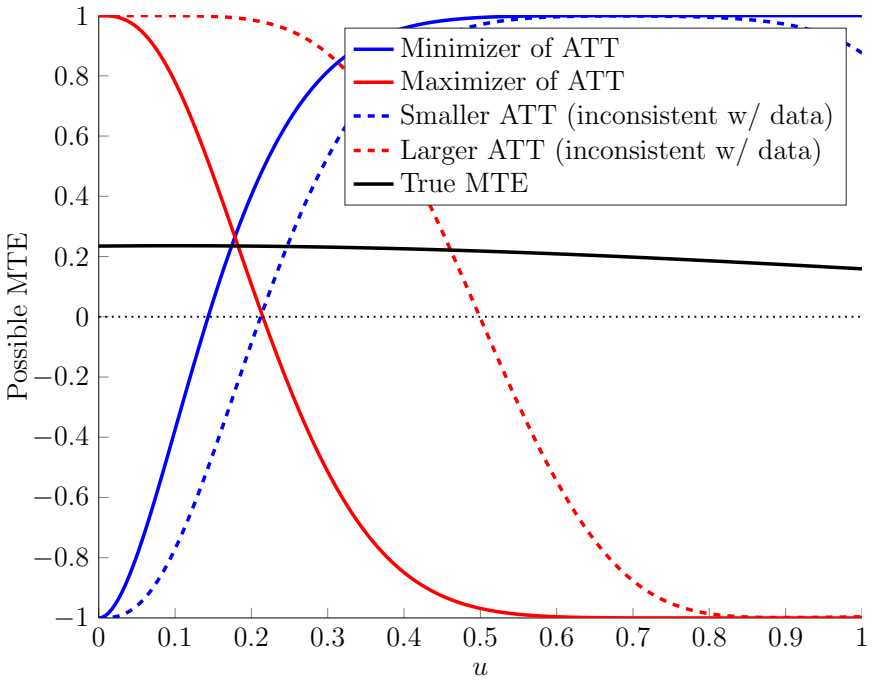


Figure 4: Optimal and infeasible MTEs for column (4) of Table 1.

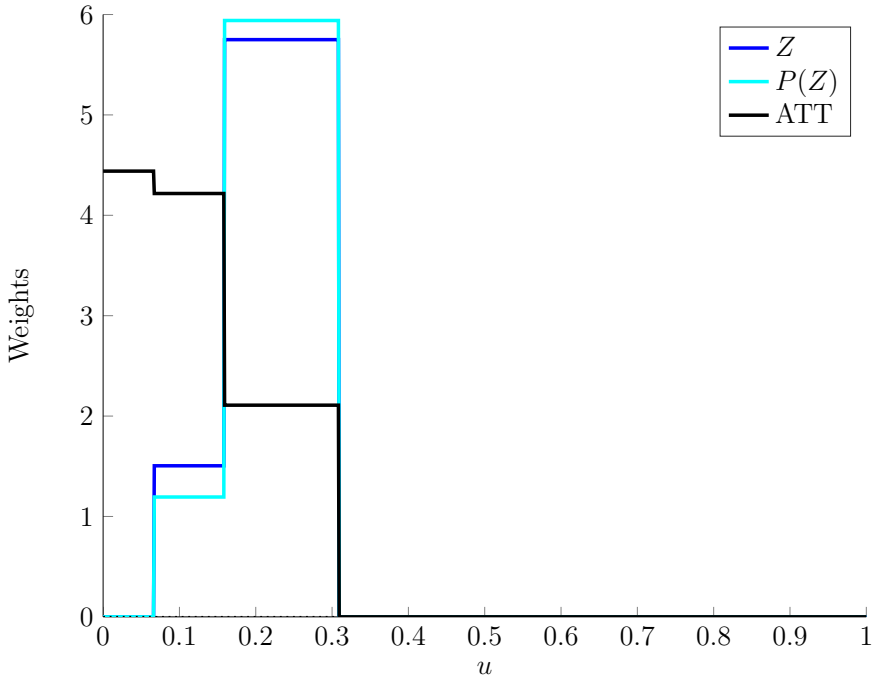


Figure 5: Weights for columns (5)–(8) of Table 3.

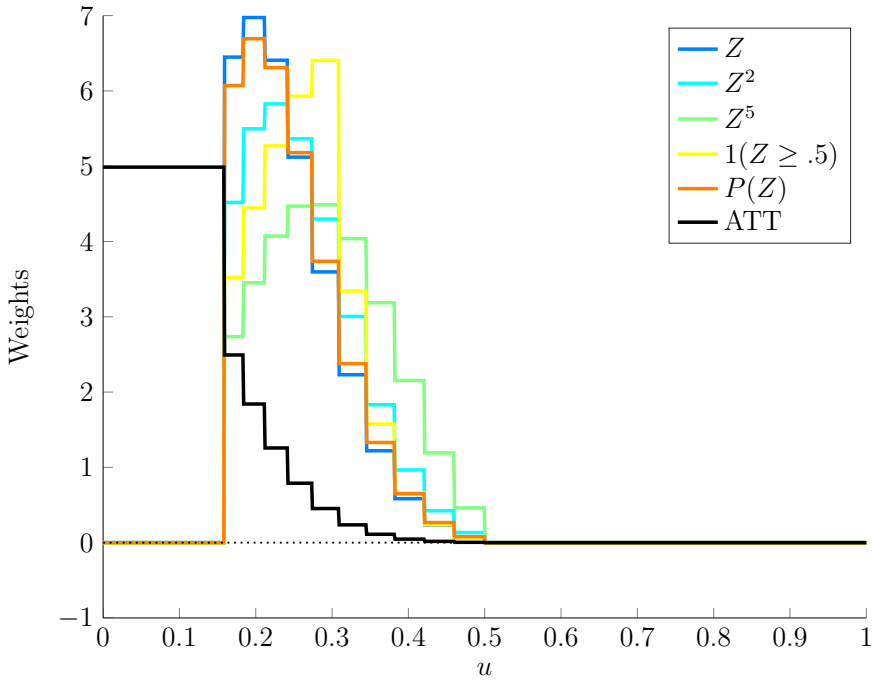


Figure 6: Weights for columns (9)–(12) of Table 4.

$S_{\text{IV}}(Z)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$P(Z)$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Z					✓	✓	✓	✓	✓	✓	✓	✓
Z^2									✓	✓	✓	✓
Z^5									✓	✓	✓	✓
$1(Z \geq 0)$									✓	✓	✓	✓
K	2	5	10	15	2	5	10	15	5	10	15	20
$\underline{\beta}^{\star}_{\text{ATT}}$	.156	-.138	-.300	-.387	.236	.154	-.074	-.113	$\emptyset$	.231	.198	.157
$\overline{\beta}^{\star}_{\text{ATT}}$	.362	.575	.657	.695		.302	.437	.499		.239	.267	.324
$\overline{\beta}^{\star}_{\text{ATT}} - \underline{\beta}^{\star}_{\text{ATT}}$	.206	.713	.957	1.082	0	.148	.511	.612	—	.007	.069	.167

Table 1

$S_{\text{IV}}(Z)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$1(Z \leq 0)$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$1(Z \leq -.1)$				✓	✓	✓	✓	✓	✓	✓	✓	✓
$1(Z \leq .1)$				✓	✓	✓	✓	✓	✓	✓	✓	✓
$1(Z \leq -.2)$							✓	✓	✓	✓	✓	✓
$1(Z \leq .2)$							✓	✓	✓	✓	✓	✓
$1(Z \leq -.3)$										✓	✓	✓
$1(Z \leq .3)$										✓	✓	✓
$1(Z \leq -.4)$										✓	✓	✓
$1(Z \leq .4)$										✓	✓	✓
K	5	10	15	15	30	45	15	30	45	15	30	45
$\underline{\beta}^{\star}_{\text{ATT}}$	-.122	-.287	-.375	-.006	-.128	-.212	.184	.039	-.051	.234	.150	.061
$\overline{\beta}^{\star}_{\text{ATT}}$	.566	.645	.680	.401	.496	.538	.279	.390	.435	.236	.307	.359
$\overline{\beta}^{\star}_{\text{ATT}} - \underline{\beta}^{\star}_{\text{ATT}}$	.688	.932	1.056	.407	.624	.750	.095	.351	.486	.002	.157	.298

Table 2

$S_{\text{IV}}(Z)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$P(Z)$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Z					✓	✓	✓	✓	✓	✓	✓	✓
Z^2									✓	✓	✓	✓
Z^5									✓	✓	✓	✓
$1(Z \geq 0)$									✓	✓	✓	✓
K	2	5	10	15	2	5	10	15	5	10	15	20
$\underline{\beta}^{\star}_{\text{ATT}}$	.142	-.226	-.332	-.391	.235	.186	.051	.012	.186	.051	.012	-.031
$\overline{\beta}^{\star}_{\text{ATT}}$	.381	.559	.639	.675		.277	.368	.403	.277	.368	.403	.416
$\overline{\beta}^{\star}_{\text{ATT}} - \underline{\beta}^{\star}_{\text{ATT}}$	.238	.785	.970	1.066	0	.091	.317	.390	.091	.317	.390	.447

Table 3

$S_{\text{IV}}(Z)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$P(Z)$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Z					✓	✓	✓	✓	✓	✓	✓	✓
Z^2									✓	✓	✓	✓
Z^5									✓	✓	✓	✓
$1(Z \geq .5)$									✓	✓	✓	✓
K	2	5	10	15	2	5	10	15	2	5	10	15
$\underline{\beta}^{\star}_{\text{ATT}}$	.089	-.381	-.536	-.612	.238	-.003	-.293	-.423	$\emptyset$	.231	.107	-.062
$\overline{\beta}^{\star}_{\text{ATT}}$	.463	.685	.799	.828		.513	.625	.688		.241	.378	.467
$\overline{\beta}^{\star}_{\text{ATT}} - \underline{\beta}^{\star}_{\text{ATT}}$	.374	1.066	1.335	1.440	0	.516	.918	1.111	—	.010	.270	.529

Table 4