Identifying the Average Treatment Effect in a Two Threshold Model

Arthur Lewbel and Thomas Tao Yang[*][†]
Boston College

July 2013

Abstract

Assume individuals are treated if a latent variable, containing a continuous instrument, lies between two thresholds. We place no functional form restrictions on the latent errors. Here unconfoundedness does not hold and identification at infinity is not possible. Yet we still show nonparametric point identification of the average treatment effect. We provide an associated root-n consistent estimator. We apply our model to reinvestigate the inverted-U relationship between competition and innovation, estimating the impact of moderate competitiveness on innovation without the distributional assumptions required by previous analyses. We find no evidence of an inverted-U in US data.

JEL Codes: C14, C21, C26

Keywords: Average treatment effect, Ordered choice model, Special regressor, Semiparametric, Competition and innovation, Identification.

[*] Corresponding Author: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, https://www2.bc.edu/~lewbel/. Thomas Tao Yang, Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467, USA. yangta@bc.edu.

[†] We would like to thank Aamir Hashmi for providing his data. All errors are our own.

1 Introduction

Suppose an outcome Y is given by

$$Y = Y_0 + (Y_1 - Y_0) D \tag{1}$$

where Y_0 and Y_1 are potential outcomes as in Rubin (1974) and Angrist, Imbens, and Rubin (1996), and D is a binary treatment indicator. Generally, point identification of the average treatment effect (ATE) $E(Y_1 - Y_0)$ requires either (conditional or unconditional) unconfoundedness, or an instrument for D that can drive D to zero and to one (with probability one), or functional restrictions on the joint distribution of Y_0, Y_1 and D. In contrast, we provide a novel point identification result (and an associated estimator) for the ATE in a model where none of these conditions hold.

Let V be a continuous instrument that affects the probability of treatment but not the outcome, and let X denote a vector of other covariates. In our model, D is given by a structure that is identical to one of the middle choices in an ordered choice model, that is,

$$D = I\left[\alpha_0(X) \leq V + U \leq \alpha_1(X)\right] \tag{2}$$

where $I(\cdot)$ is the indicator function that equals one if $\cdot$ is true and zero otherwise, U is a latent error term, and $\alpha_0(X)$ and $\alpha_1(X)$ are unknown functions. The joint distribution of $(U, Y_0, Y_1 \mid X)$ is assumed to be unknown.

In the special case of this model where $\alpha_0(X)$ is linear and $\alpha_1(X) - \alpha_0(X)$ is constant, treatment is given by the more standard looking ordered choice specification

$$D = I\left(\delta_0 \leq X'\beta_1 + \beta_2 V + U \leq \delta_1\right)$$

for constants δ_0, δ_1, β_1, and β_2. However, we don't impose these linearity restrictions. In addition, unlike standard ordered choice models, we allow the distribution of U to depend on X in completely unknown ways. Equivalently, the covariates X can all be endogenous regressors, with no available associated instruments. The only covariate we require to be exogenous is V.

The proposed model is confounded, because the unobservable U that affects D can be correlated with Y_0 and Y_1, with or without conditioning on X. No parametric or semiparametric restrictions are placed on the distribution of $(U, Y_0, Y_1 \mid X)$, so treatment effects are not identified by functional form restrictions on the distributions of unobservables. We assume V has large support, but the model is not identified at infinity. This is because both very large and very small values of V drive the probability of treatment close to zero, but no value of V (or of other covariates) drives the probability of treatment close to one. So in this framework none of the conditions that are known to permit point identification of the ATE hold. Even a local ATE (LATE) is not identified in the usual way, because monotonicity of treatment with respect to the instrument is not satisfied. Nevertheless, we show that the ATE is identified in our model, using a special regressor argument as in Lewbel (1998, 2000, 2007). We also provide conditions under which the corresponding estimate of the ATE converges at rate root n.

To illustrate the model and foreshadow our later empirical application, suppose the outcome Y is a measure of innovation in an industry and $D = 1$ when a latent measure of competitiveness in the industry lies between two estimated thresholds, otherwise $D = 0$. According to the "Inverted-U" theory in Aghion, Bloom, Blundell, Griffith, and Howitt (2005) (hereafter ABBGH), industries with intermediate levels of competitiveness have more innovation than those with low levels or high levels of competition. As in Revenga (1990, 1992), Bertrand (2004), and Hashmi (2012), we use a source-weighted average of industry exchange rates as an instrumental variable for competition, which we take to be our special regressor V. This instrument is computed from the weighted average of the US dollar exchange rate with the currencies of its trading partners. When V is low, products from the U.S. becomes relatively cheaper, thereby reducing competition by driving out competitors. The treatment effect we estimate is therefore the gains in innovation that result from facing moderate (rather than very low or very high) levels of competition.

Existing methods for point identifying ATE's are discussed in surveys such as Heckman and Vytlacil (2007a, 2007b) and Imbens and Wooldridge (2009). The early treatment effects literature achieves identification by assuming unconfoundedness, see, e.g., Cochran and Rubin (1973), Rubin (1977), Barnow, Cain, and Goldberger (1980), Rosenbaum and Rubin (1983), Heckman and Robb (1984), and Rosenbaum (1995). As noted by ABBGH, competition and innovation are mutually endogenous. Much of what determines both is difficult to observe or even define, making it very unlikely that unconfoundedness would hold, regardless of what observable covariates one conditions upon.

Without unconfoundedness, instrumental variables have been used in a variety of ways to identify treatment effects. Instead of estimating the ATE, Imbens and Angrist (1994) show identification of a local average treatment effect (LATE), which is the ATE for a subpopulation called compliers (the definition of who compliers are, and hence the LATE, depends on the choice of instrument). An assumption for identifying the LATE is that the probability of treatment increase monotonically with the instrument. This assumption does not hold in our application, since both increasing or decreasing V sufficiently causes the probability of treatment to decrease. Kitagawa (2009) shows that, if possible, point identification of the ATE without identification at infinity, based only on an exogenous instrument, would require instrument nonmonotonicity. Our model possesses this necessary nonmonotonicity (another example of such nonmonotonicity is Gautier and Hoderlein 2011).

Building on Björklund and Moffitt (1987), in a series of papers Heckman and Vytlacil (1999, 2005, 2007a) describe identification of a marginal treatment effect (MTE) as a basis for program evaluation. The MTE is based on having a continuous instrument, as we do. However, identification of the ATE using the MTE requires the assumption that variation in V can drive the probability of treatment to either zero or one, and hence depends on an identification at infinity argument. As we have already noted, identification at infinity is not possible in our model, since no value of V can drive the probability of treatment to one.

A few other papers consider identification of treatment effects in ordered choice models, such as Angrist and Imbens (1995) and Heckman, Urzua, and Vytlacil (2006). However, these papers deal with models having more information that ours, i.e, observing extreme as well as middle choices, and they consider identification of LATE and MTE, respectively, not ATE.

The way we achieve identification here is based on special regressor methods, particularly Lewbel (2007), which exploits a related result to identify a class of semiparametric selection models. The instrumental variable V needs to be continuous, conditionally independent of other variables and have a large support, which are all standard assumptions for special regressor based estimators. See, e.g., Dong and Lewbel (2012), Lewbel, Dong, and Yang (2012), and Lewbel (2012). Some of the previously discussed papers also implicitly assume a special regressor, notably, Heckman, Urzua, and Vytlacil (2006).

In addition to the ATE, our methods can be immediately extended to estimate quantile treatment effects as in Abadie, Angrist, and Imbens (2002), Chernozhukov and Hansen (2005). Bitler, Gelbach, and Hoynes (2006), or Firpo (2006). This is done by replacing Y with $I(Y \leq y)$ in our estimator.

Our empirical application uses panel data. Our identification method extends to the following panel data case

$$D_{it} = I(\alpha_0(x_{it}) \leq a_i + b_t + V_{it} + U_{it} \leq \alpha_1(x_{it})), \tag{3}$$

$$Y_{it} = \widetilde{a}_i + \widetilde{b}_t + Y_{0it} + (Y_{1it} - Y_{0it})D_{it}, \tag{4}$$

where $a_i, \widetilde{a}_i, b_t, \widetilde{b}_t$ are individual and time dummies in selection and outcome equations. Now Y_{0it} and Y_{1it} are potential outcomes that are purged of cross section and time fixed effects, so the interpretation of the ATE $\mathrm{E}(Y_{1it} - Y_{0it})$ is analogous to the interpretation of treatment effects in difference-in-difference models, except that here D_{it} can be endogenous and hence correlated with the potential outcomes.

Analogous to Honore and Lewbel (2002), our identification and estimation strategy overcomes the incidental parameters problem, attaining a rate root n estimate for the ATE, even when the treatment equation contains fixed effects a_i and when the number of time periods is fixed. We also provide asymptotics allowing for both a_i and b_t fixed effects, by letting both n and the number time periods go to infinity.

In the panel context, if unconfoundedness held so that $Y \perp D|X$ and if in addition a_i and b_t were absent from the selection equation, then one could achieve identification via the difference-in-difference (DID) method. See, e.g., Ashenfelter (1978), Ashenfelter and Card (1985), Cook and Tauchen (1982, 1984), Card (1990), Meyer, Viscusi, and Durbin (1995), Card and Krueger (1993, 1994) and many others for applications of DID methods. In contrast, we obtain identification without unconfoundedness, while still allowing for a_i and b_t fixed effects.

In the next section we introduce the model and establish the consistency and asymptotic normality of our cross section and panel estimators. In section three, we apply our estimators to investigate the relationship between competition and innovation. We also implement simulation experiments to evaluate small sample properties of our estimators, using a Monte Carlo design that replicates features of our empirical data. This is followed by conclusions, and by an appendix containing proofs. Another Appendix provides an evaluation of how the robustness of our approach compares to more structural models (of the type others have used to evaluate competitiveness and innovation) in the presence of measurement errors, because competitiveness is likely to not be measured very precisely. Finally, in a supplemental appendix separate from the main paper, we provide details regarding application of relatively standard semiparametric methods for deriving the limiting distribution of our estimators.

2 The Model

In this section we first prove identification of the unconditional ATE, and of the ATE conditioning on X, in our model. The proof we provide is constructive, and we next describe a corresponding estimator. This is followed by some extensions, in particular, a panel data estimator with fixed effects. The remaining parts of this section then provide root n limiting distribution theory for the estimators.

2.1 Identification and Estimation

Let Ω and f denote supports and density functions, e.g., Ω_x and f_x are the support and density function for the random variable X. Let $\widehat{\mathrm{E}}(.)$ denote the sample mean of the argument inside, and let $\widehat{f}(.)$ and $\widehat{\mathrm{E}}(.|.)$ denote nonparametric Nadayara-Watson kernel density and kernel regression estimators, with bandwidth denoted h. For notational convenience, h is assumed the same for all covariates. We use R to denote any set of residual terms that are proven to be asymptotically irrelevant for our derived limiting distributions.

Assumption 2.1 *We observe realizations of an outcome Y, binary treatment indicator D, a covariate V, and a $k \times 1$ covariate vector X. Assume the outcome Y and treatment indicator D are given by equations (1) and (2) respectively, where $\alpha_0(X)$ and $\alpha_1(X)$ are unknown threshold functions, U is an unobserved latent random error, and Y_0 and Y_1 are unobserved random untreated and treated potential outcomes. The joint distribution of (U, Y_0, Y_1), either unconditional or conditional on X, is unknown.*

Assumption 2.2 *Assume $E(Y_j|X,V,U) = E(Y_j|X,U)$ for $j = 0,1$, and $V \perp U \mid X$. Assume $V \mid X$ is continuously distributed with probability density function $f(V \mid X)$. For all $x \in supp(X)$, the $supp(V \mid X = x)$ is an interval on the real line, and the interval $[\inf supp(\alpha_0(X) - U \mid X = x), \sup supp(\alpha_1(X) - U \mid X = x)]$ is contained in $supp(V \mid X = x)$.*

Assumption 2.1 defines the model, while Assumption 2.2 says that V is an instrument, in that V affects the probability of treatment but not outcomes (after conditioning on X). The instrument V is also continuously distributed, and has a large enough support so that, for any values U and X may take on, V can be small enough to make $D = 0$ or large enough to make $D = 0$. But no value of V and X will force $D = 1$, so identification at infinity is not possible.[1]

In this model, obtaining identification by imposing unconfoundedness would be equivalent to assuming that U was independent of $Y_1 - Y_0$, possibly after conditioning on covariates X. However, we do not make any assumption like this, so unconfoundedness does not hold. Alternatively, one might parametrically model the dependence of $Y_1 - Y_0$ on U to identify the model. In contrast we place no restrictions on the joint distribution of (U, Y_0, Y_1), either unconditional or conditioning upon X.

Assumption 2.3 *For some positive constant τ, define the trimming function $I_\tau(v, x) = I[\inf \mathit{supp}(V|X = x) + \tau \leq v \leq \sup \mathit{supp}(V|X = x) - \tau]$. Assume that the interval $[\inf \mathit{supp}\,(\alpha_0(X) - U \mid X = x),$ $\sup \mathit{supp}\,(\alpha_1(X) - U \mid X = x)]$ is contained in $\{v : I_\tau(v, x) = 1\}$.*

Assumption 2.4 *Assume there exists a positive constant $\widetilde{\tau} < \tau$ such that, for all v,x having $I_{\widetilde{\tau}}(v, x) = 1$, the density $f(v|x)$ is bounded away from zero (except possibly on a set of measure zero) and is bounded. $f_x(X)$ and Y are also bounded, and $f_x(X)$ is bounded away from zero.*

Assumption 2.3 is not necessary for identification, but is convenient for simplifying the limiting distribution theory for the estimator we construct based on the identification. In particular, this assumption permits fixed trimming that avoids boundary bias in our kernel estimators. This assumption could be relaxed using asymptotic trimming arguments. Some components of Assumption 2.4 might also be relaxed via asymptotic trimming. Define the function $\psi(X)$ by

$$\psi(X) = \frac{\mathrm{E}\left[I_\tau DY/f(V \mid X) \mid X\right]}{\mathrm{E}\left[I_\tau D/f(V \mid X) \mid X\right]} - \frac{\mathrm{E}\left[I_\tau (1 - D) Y/f(V \mid X) \mid X\right]}{\mathrm{E}\left[I_\tau (1 - D)/f(V \mid X) \mid X\right]} \tag{5}$$

Theorem 2.1 *Let Assumptions 2.1, 2.2, 2.3 and 2.4 hold. Then*

$$\psi(X) = E(Y_1 - Y_0 \mid X)$$

[1] If instead of the ordered choice $D = I\left[\alpha_0(X) \leq V + U \leq \alpha_1(X)\right]$ we had a threshold crossing binary choice $D = I\left(\alpha_0(X) \leq V + U\right)$, then Assumption 2.2 would suffice to use "identification at infinity" to identify the treatment effect, by using data where V was arbitrarily low to estimate $E(Y_0 \mid X)$ and data where V was arbitrarily high to estimate $E(Y_1 \mid X)$. However, in our ordered choice model identification at infinity is not possible, since no value of V guarantees with high probability that Y will equal Y_1.

Proof. See Appendix. ■

Theorem 2.1 shows identification of the conditional ATE since $\psi(X)$ is defined in terms of moments and densities of observed variables. Equation (5) would still hold without Assumption 2.3 and without the trimming terms I_τ, but their inclusion simplifies the later asymptotics.

It follows immediately from Theorem 2.1 that $\Psi =$E$[\psi(X)]$ equals the ATE, which is therefore identified and can be consistently estimated by $\widehat{\Psi} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\psi}(x_i)$ where

$$\widehat{\psi}(x) = \frac{\widehat{\mathrm{E}}\left[I_\tau DY/\widehat{f}(V \mid X) \mid X = x\right]}{\widehat{\mathrm{E}}\left[I_\tau D/\widehat{f}(V \mid X) \mid X = x\right]} - \frac{\widehat{\mathrm{E}}\left[I_\tau (1-D) Y/\widehat{f}(V \mid X) \mid X = x\right]}{\widehat{\mathrm{E}}\left[I_\tau (1-D)/\widehat{f}(V \mid X) \mid X = x\right]},$$

with uniformly consistent kernel estimators $\widehat{f}$ and $\widehat{\mathrm{E}}$. Later we apply standard theory on two step estimators with a nonparametric first step to obtain the root n limiting distribution of $\widehat{\Psi}$.

If V had a uniform distribution, thereby making f a constant, then equation (5) would simplify to standard propensity score weighting which is consistent when there is no confounding. So in our model, weighting by $f(V \mid X)$, i.e., weighting by the density of the instrument, essentially fixes the problem of confounding. This density weighting is a feature of some special regressor estimators like Lewbel (2000), (2007), and indeed V has the properties of a special regressor, including appearing additively to unobservables in the model, a continuous distribution, large support, and conditional independence.

2.2 Extensions

The above identification and associated estimator can be extended to handle independent random thresholds, that is, all the results go through if $\alpha_1(X)$ and $\alpha_0(X)$ are replaced with random variables α_1 and α_0, provided that $(\alpha_0, \alpha_1) \perp (U, Y_1, Y_0) \mid X$.

Our results also immediately extend to permit estimation of quantile treatment effects. The proof of Theorem 2.1 shows that the first term in Equation (5) equals E$(Y_1 \mid X)$ and the second term equals E$(Y_0 \mid X)$. Suppose we strengthen the assumption that E$(Y_j \mid X, V, U) =$E$(Y_j \mid X, U)$ for $j = 0, 1$ to say that $F_j(Y_j \mid X, V, U) = F_j(Y_j \mid X, U)$, where F_j is the distribution function of Y_j for $j = 0, 1$. Then one can apply Theorem 2.1 replacing Y with $I(Y \leq y)$ for any y, and thereby estimate E$(I(Y_j \leq y) \mid X) = F_j(y \mid X)$. Given this identification and associated estimators for the distributions $F_j(y \mid X)$ of the counterfactuals Y_j, we could then immediately recover quantile treatment effects.

Now consider panel data. Define the treatment indicator equation as

$$D_{it} = I(\alpha_0(x_{it}) \leq a_i + b_t + V_{it} + U_{it} \leq \alpha_1(x_{it})), \tag{6}$$

and the outcome equation as

$$Y_{it} = \widetilde{a}_i + \widetilde{b}_t + Y_{0it} + (Y_{1it} - Y_{0it})D_{it}, \tag{7}$$

where $a_i, b_t, \widetilde{a}_i, \widetilde{b}_t$ are equivalent to coefficients times individual and time dummies in these two equations. For example, b_t equals some unknown scalar times a dummy variable that equals one for all observations in time period t and zero otherwise.

As before, the observables in the model are the outcome Y, treatment D, instrument V, and covariate vector X. We interpret $(a_i, b_t, \widetilde{a}_i, \widetilde{b}_t)$ as fixed effects, in that they will not be estimated and they can correlate with unobservables and with X in unknown ways. However, we do require $(a_i, b_t, \widetilde{a}_i, \widetilde{b}_t)$ to be random variables, because we make an independence assumption involving them and the instrument V in Assumption 2.5 below. However, the joint distribution of $(a_i, b_t, \widetilde{a}_i, \widetilde{b}_t, U_{it}, Y_{0it}, Y_{1it})$ conditional or unconditional on X_{it}, is unknown. A similar assumption regarding fixed effects in discrete choice appears in Honore and Lewbel (2002).

Assumption 2.5 *For individuals i and time periods t, $a_i, b_t, \widetilde{a}_i, \widetilde{b}_t$ are random variables.*

$$E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{jit} | X_{it}, V_{it}, a_i, b_t, U_{it}\right) = E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{jit} \middle| X_{it}, a_i, b_t, U_{it}\right),$$

for $j = 0, 1$. $V_{it} \perp a_i, b_t, U_{it} | X_{it}$.

Assumption 2.6 *Assumption 2.3 holds after replacing supp[$\alpha_0(x) - u, \alpha_1(x) - u$] with supp[$\alpha_0(x_{it}) - \widetilde{a}_i - \widetilde{b}_t - u_{it}, \alpha_1(x_{it}) - \widetilde{a}_i - \widetilde{b}_t - u_{it}$].*

Assumption 2.5 is essentially a panel data version of Assumption 2.2.

Theorem 2.2 *Let Assumption 2.1, 2.4, 2.5, and 2.6 hold for each individual i in each time period t. Let f_{v_t} denote the density of V in time t. Then*

$$\frac{E[I_{\tau it} D_{it} Y_{it} / f_{v_t}(V_{it}|X_{it}) | X_{it}]}{E[I_{\tau it} D_{it} / f_{v_t}(V_{it}|X_{it}) | X_{it}]} - \frac{E[I_{\tau it}(1 - D_{it}) Y_{it} / f_{v_t}(V_{it}|X_{it}) | X_{it}]}{E[I_{\tau it}(1 - D_{it}) / f_{v_t}(V_{it}|X_{it}) | X_{it}]} = E(Y_{1it} - Y_{0it} | X_{it}). \tag{8}$$

Note that the expectations in Theorem 2.2 are both over time and over individuals. Equation (8) is essentially identical to $\psi(X)$ in Equation (5), and therefore the corresponding estimator for panel data is essentially identical to the cross section case from Theorem 2.1, despite the addition of fixed effects. If the distribution of V varies by time then the density of V must be estimated separately in each time period, but other than that estimation is the same, just averaging across all individuals in all time periods. Estimation does not require first differencing or other similar techniques to remove the fixed effects. As one can see from the proof of Theorem 2.2, this convenience comes from the fact that fixed effects in the treatment equation are eliminated by taking expectations after density weighting, and fixed effects in the outcome equation are canceled out when differencing the average effect for the untreated from the average effect for the treated.

2.3 Consistency and Asymptotic Normality

To ease the notational burden, for the purpose of deriving asymptotic theory we will write $I_{\tau i}D_i$ as just D_i and $I_{\tau i}(1-D_i)$ as just $(1-D_i)$, so the fixed trimming term $I_{\tau i}$ will be implicit. Our identification theorem permits fixed trimming, which then allows our limiting distribution derivation to follow standard arguments like those in Newey and McFadden (1994), avoiding the complications associated with boundary bias. These assumptions also avoid rates of convergence issues, yielding estimation of the ATE at rate root n. As noted briefly in Lewbel (2000b) and discussed more thoroughly in Khan and Tamer (2010), without fixed trimming obtaining root n rates with inverse density weighted estimators like ours would generally require V to have very thick tails (such as having infinite variance). Otherwise, attainable rates of convergence may be considerably slower than root n. Our fixed trimming avoids this thick tails requirement.

For this section, proofs and the standard assumptions regarding kernels, bandwidths and smoothness are provided in an Appendix, while assumptions that require some discussion are kept in the main text.

2.3.1 Cross Section Case

We first derive properties for the version of our estimator in which the density function $f\left(V|X\right)$ is known, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\widehat{\mathrm{E}}\left(\left.\frac{D_iY_i}{f(v_i|x_i)}\right| x_i\right)}{\widehat{\mathrm{E}}\left(\left.\frac{D_i}{f(v_i|x_i)}\right| x_i\right)}-\frac{\widehat{\mathrm{E}}\left(\left.\frac{(1-D_i)Y_i}{f(v_i|x_i)}\right| x_i\right)}{\widehat{\mathrm{E}}\left(\left.\frac{1-D_i}{f(v_i|x_i)}\right| x_i\right)}\right]. \tag{9}$$

This version of our estimator might be used in cases when V is determined by experimental design, but will otherwise typically be infeasible. We then extend these results to handle the usual case where f is estimated instead of known. Denote $h_{1i} = \frac{D_i Y_i}{f(v_i|x_i)}$, $g_{1i} = \frac{D_i}{f(v_i|x_i)}$, $h_{2i} = \frac{(1-D_i)Y_i}{f(v_i|x_i)}$, $g_{2i} = \frac{1-D_i}{f(v_i|x_i)}$, $\psi_1(x_i) = \frac{\mathrm{E}(h_{1i}|x_i)}{\mathrm{E}(g_{1i}|x_i)}$, and $\psi_2(x_i) = \frac{\mathrm{E}(h_{2i}|x_i)}{\mathrm{E}(g_{2i}|x_i)}$. From Theorem 2.1,

$$\psi_1(x_i) = \mathrm{E}(Y_1|x_i), \quad \psi_2(x_i) = \mathrm{E}(Y_0|x_i).$$

The sample counterpart estimate for $\psi_1(x_i)$ is then

$$\widehat{\psi}_1(x_i) = \frac{\widehat{\mathrm{E}}\left(h_{1i}|\, x_i\right)}{\widehat{\mathrm{E}}\left(g_{1i}|\, x_i\right)}. \tag{10}$$

Define

$$\mathrm{E}\left(\widetilde{h}_{1i}\middle|\, x_i\right) = \mathrm{E}\left(h_{1i} f_x(x_i)|\, x_i\right), \tag{11}$$

$$\mathrm{E}\left(\widetilde{g}_{1i}|\, x_i\right) = \mathrm{E}\left(g_{1i} f_x(x_i)|\, x_i\right) \tag{12}$$

which can be estimated using leave-one-out kernel estimators

$$\widehat{\mathrm{E}}\left(\widetilde{h}_{1i}\middle|\, x_i\right) = \frac{1}{nh^k}\sum_{j=1, j\neq i}^{n} h_{1j} K\left(\frac{x_j - x_i}{h}\right), \tag{13}$$

$$\widehat{\mathrm{E}}\left(\widetilde{g}_{1i}|\, x_i\right) = \frac{1}{nh^k}\sum_{j=1, j\neq i}^{n} g_{1j} K\left(\frac{x_j - x_i}{h}\right), \tag{14}$$

so we can write

$$\widehat{\psi}_1(x_i) = \frac{\widehat{\mathrm{E}}\left(\widetilde{h}_{1i}\middle|\, x_i\right)}{\widehat{\mathrm{E}}\left(\widetilde{g}_{1i}|\, x_i\right)}. \tag{15}$$

Replacing the subscript 1 with 2, similarly define $\widehat{\psi}_2(x_i), \widehat{\mathrm{E}}\left(h_{2i}|\, x_i\right), \widehat{\mathrm{E}}\left(g_{2i}|\, x_i\right), \mathrm{E}\left(\widetilde{h}_{2i}\middle|\, x_i\right)$, $\mathrm{E}(\widetilde{g}_{2i}|\, x_i)$, $\widehat{\mathrm{E}}\left(\widetilde{h}_{2i}\middle|\, x_i\right)$, and $\widehat{\mathrm{E}}\left(\widetilde{g}_{2i}|\, x_i\right)$. The resulting estimator (9) is then

$$\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i)\right].$$

Assumptions 4.1 4.2, and 4.3, provided in the Appendix, are all standard. Given these assumptions, the consistency of estimator (9) is established as follows.

Theorem 2.3 *Let the Assumptions in Theorem 2.1 and Assumptions 4.1, 4.2, 4.3, hold, and let* $h \to 0$ *and* $nh^k \to \infty$, *as* $n \to \infty$. *Then Equation (9) is a consistent estimator of* $E(Y_1 - Y_0)$.

The proof is in the Appendix. For asymptotic normality, we make additional standard assumptions of boundedness, smoothness, and Lipschitz conditions, used to control bias and the size of residuals, given by Assumption 4.4 in the Appendix and the folowing.

Assumption 2.7 *Define*

$$q_{1i} = \left(\frac{h_{1i}}{E(g_{1i}|x_i)} - \frac{E(h_{1i}|x_i)g_{1i}}{[E(g_{1i}|x_i)]^2}\right) - \left(\frac{h_{2i}}{E(g_{2i}|x_i)} - \frac{E(h_{2i}|x_i)g_{2i}}{[E(g_{2i}|x_i)]^2}\right) + E(Y_1 - Y_0|x_i) - E(Y_1 - Y_0).$$

and assume the second moment of q_{1i} *exists.*

The following theorem gives the asymptotics for estimator (9), based on showing that q_{1i} defined above is the influence function for $\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i)$−E$(Y_1 - Y_0)$.

Theorem 2.4 *Let Assumptions 4.4, 2.7, and all of the Assumptions in Theorem 2.3 hold. Let* $nh^{2p} \to 0$ *and* $n^{1-\varepsilon}h^{2k+2} \to \infty$ *for any small positive* ε *as* $n \to \infty$. *Then*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i) - E(Y_1 - Y_0)\right] = \frac{1}{n}\sum_{i=1}^{n} q_{1i} + o_p\left(\frac{1}{\sqrt{n}}\right)$$

and, since observations are i.i.d. across i,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i) - E(Y_1 - Y_0)\right] \xrightarrow{d} N(0, var(q_{1i})).$$

The proof of this theorem is straightforward but lengthy, and so is provided in a supplemental appendix. The influence function q_{1i} can be decomposed into two pieces. One piece, E$(Y_1 - Y_0|x_i)$− E$(Y_1 - Y_0)$, comes from estimating expectations as sample averages, while the remaining terms in q_{1i} embody the effects of the first stage nonparametric estimates. These additional terms in q_{1i} correspond to $\delta(z_i)$ in Theorem 8.1 of Newey and McFadden (1994).

We now consider the case where the conditional density $f(V|X)$ is estimated (nonparametrically) instead of being known. The resulting estimator is now

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\widehat{\mathrm{E}}\left(\frac{D_iY_i}{\widehat{f}(v_i|x_i)}\middle| x_i\right)}{\widehat{\mathrm{E}}\left(\frac{D_i}{\widehat{f}(v_i|x_i)}\middle| x_i\right)} - \frac{\widehat{\mathrm{E}}\left(\frac{(1-D_i)Y_i}{\widehat{f}(v_i|x_i)}\middle| x_i\right)}{\widehat{\mathrm{E}}\left(\frac{1-D_i}{\widehat{f}(v_i|x_i)}\middle| x_i\right)}\right]. \tag{16}$$

Redefine the sample counterpart estimates for $\psi_1(x_i)$ as

$$\widehat{\psi}_1(x_i) = \frac{\widehat{\mathrm{E}}\left(\widehat{h}_{1i}\middle|\, x_i\right)}{\widehat{\mathrm{E}}\left(\widehat{g}_{1i}\middle|\, x_i\right)}, \tag{17}$$

where

$$\widehat{h}_{1i} = \frac{D_i Y_i}{\widehat{f}(v_i|x_i)},\ \widehat{g}_{1i} = \frac{D_i}{\widehat{f}(v_i|x_i)}.$$

Similarly, the sample counterpart estimates for $\mathrm{E}\Big(\widetilde{h}_{1i}\Big|\, x_i\Big)$ and $\mathrm{E}(\widetilde{g}_{1i}|\, x_i)$ for estimator (16) are

$$\widehat{\mathrm{E}}\left(\widehat{\widetilde{h}}_{1i}\middle|\, x_i\right) = \frac{1}{nh^k}\sum_{j=1,j\neq i}^{n} \widehat{h}_{1j} K\left(\frac{x_j - x_i}{h}\right), \tag{18}$$

$$\widehat{\mathrm{E}}\left(\widehat{\widetilde{g}}_{1i}\middle|\, x_i\right) = \frac{1}{nh^k}\sum_{j=1,j\neq i}^{n} \widehat{g}_{1j} K\left(\frac{x_j - x_i}{h}\right). \tag{19}$$

Let $\widehat{\psi}_2(x_i)$, $\widehat{\mathrm{E}}\left(\widehat{\widetilde{h}}_{2i}\middle|\, x_i\right)$, and $\widehat{\mathrm{E}}\left(\widehat{\widetilde{g}}_{2i}\middle|\, x_i\right)$ be defined analogously, with subscript 1 replaced by 2. Estimator (16) can then be written as

$$\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i)\right].$$

Theorem 2.5 *Let all the Assumptions in Theorem 2.3 hold, except that* $nh^{k+1} \to \infty$, *as* $n \to \infty$. *Then Equation (16) is a consistent estimator of* $E(Y_1 - Y_0)$.

For asymptotic normality, similar to what was imposed for estimator (9), make Assumption 4.5 in the Appendix to control the bias and the size of residuals, and the following.

Assumption 2.8 *Define*

$$q_{2i} = \left[\frac{h_{1i}}{E(g_{1i}|x_i)} - \frac{E(h_{1i}|x_i, v_i)}{E(g_{1i}|x_i)} - \frac{E(h_{1i}|x_i)\, g_{1i}}{[E(g_{1i}|x_i)]^2} + \frac{E(h_{1i}|x_i)\, E(g_{1i}|x_i, v_i)}{[E(g_{1i}|x_i)]^2}\right] - \left[\frac{h_{2i}}{E(g_{2i}|x_i)}\right.$$
$$\left. - \frac{E(h_{2i}|x_i, v_i)}{E(g_{2i}|x_i)} - \frac{E(h_{2i}|x_i)\, g_{2i}}{[E(g_{2i}|x_i)]^2} + \frac{E(h_{2i}|x_i)\, E(g_{2i}|x_i, v_i)}{[E(g_{2i}|x_i)]^2}\right] + E(Y_1 - Y_0|x_i) - E(Y_1 - Y_0),$$

and assume the second moment of q_{2i} *exists.*

The next theorem provides asymptotics for estimator (16), by showing that q_{2i} is the influence function for $\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i) - \mathrm{E}(Y_1 - Y_0)$. The additional terms in q_{2i} compared with q_{1i}, and the expectation terms that condition on both X and V instead of just on X, are due to nonparametric estimation of $f(V|X)$. This makes $\widehat{f}_{xv}$ appear in the estimator, where before we only had $\widehat{f}_x$.

Theorem 2.6 *Let Assumptions 4.5, 2.8 and all Assumptions in Theorem 2.5 hold. Let* $nh^{2p} \to 0$, $n^{1-\varepsilon}h^{4k+4} \to \infty$ *for a very small positive* ε, *as* $n \to \infty$. *Then*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i) - E(Y_1 - Y_0)\right] = \frac{1}{n}\sum_{i=1}^{n} q_{2i} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

and since observations are i.i.d. across i

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i) - E(Y_1 - Y_0)\right] \xrightarrow{d} N(0, var(q_{2i})). \tag{20}$$

This theorem is proved in the supplemental appendix. Compared to Theorem 2.4, in this theorem stronger rate restrictions are imposed on n and h, because estimation of $f(V|X)$ requires another summation inside estimator (16).

2.3.2 Panel Data Case

The panel version of our estimator is essentially identical to averaging our cross section estimator across multiple time periods, because, as noted in the proof of Theorem 2.2, the estimator automatically accounts for fixed effects. Deriving the asymptotic properties of the panel estimator is therefore relatively straightforward but tedious. The main differences from the cross section case come from allowing the distribution of V to vary over time, and accounting for the impact of fixed effects. To simplify this analysis and to match our empirical application, we take X_{it} to be constant in equations (6) and (7), yielding the model

$$Y_{it} = a_i + b_t + Y_{0it} + (Y_{1it} - Y_{0it})\, D_{it}, \tag{21}$$

$$D_{it} = I\left[0 \le \widetilde{a}_i + \widetilde{b}_t + V_{it} + U_{it} \le \alpha\right], \tag{22}$$

where $i = 1, 2, ..., n$, $t = 1, 2, ..., T$, and α is an unknown constant. The sample counterpart we estimate is then

$$\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})}{\widehat{f}_{v_t}(v_{it})}}. \tag{23}$$

The presence of fixed effects affects rates of convergence of the estimator. We consider asymptotics where T is small relative to n. We show $\sqrt{T}$ convergence rates in the presence of both time and cross section dummies (i.e., a_i, b_t, $\widetilde{a}_i$, and $\widetilde{b}_t$) and $\sqrt{n}$ convergence rates with fixed T, without time dummies (keeping a_i and $\widetilde{a}_i$ but dropping b_t and $\widetilde{b}_t$).

Assumption 2.9 *$n \to \infty, T \to \infty$, and $\frac{T}{n} \to 0$.*

Assumption 2.10 *$a_i, \widetilde{a}_i$ are i.i.d. across i. $b_t, \widetilde{b}_t$ are i.i.d. across t. (Y_{0it}, Y_{1it}) are identically distributed across i, t. $(U_{it}, Y_{0it}, Y_{1it}) \perp (U_{i't'}, Y_{0i't'}, Y_{1i't'})$ for any $i \neq i'$, $t \neq t'$. $(U_{it}, Y_{0it}, Y_{1it}) \perp (U_{it'}, Y_{0it'}, Y_{1it'})\, |a_i, \widetilde{a}_i$ for any i, $t \neq t'$. $(U_{it}, Y_{0it}, Y_{1it})|\, b_t, \widetilde{b}_t$ are i.i.d. across i for any given t.*

The assumption that (Y_{0it}, Y_{1it}) is identically distributed over t as well as over i for each t is made for convenience, and could be relaxed at the expense of additional notation that would include redefining the estimand to be the average value over time of $\mathrm{E}(Y_1 - Y_0|t)$. We could allow heterogeneity (non-identical distributions) over the time dimension for other variables as well, but we do rely on the i.i.d. assumption across i, conditional on t. Variables with the same i or the same t subscript are correlated with each other through individual or time dummies.

In Assumption 2.10, we define $a_i, \widetilde{a}_i, b_t, \widetilde{b}_t$ as random variables, but we estimate the model treating them as one would handle fixed effects, without imposing the assumptions that would be required for random effects estimators. For example, a_i and b_t are allowed to be correlated with U_{it} and Y_{it} in arbitrary unknown ways. Also, based on our identification theorem, a_i and b_t are not estimated, but are instead eliminated from the model when taking expectations, analogous to the elimination of fixed effects by differencing in linear panel models.

Assumption 2.11 *$E\Big(\widetilde{a}_i + \widetilde{b}_t + Y_{jit}\Big|\, V_{it}, a_i, b_t, U_{it}\Big) = E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{jit}\Big|\, a_i, b_t, U_{it}\right)$, for $j = 0, 1$. $V_{it} \perp a_i, b_t, U_{it}$. V_{it} are independent across i and t. V_{it} are identically distributed across i given t, with distribution $f_{v_t}(V_{it})$.*

For each time period t, Assumption 2.11 is equivalent to the cross section special regressor assumption without X. In addition it is assumed that special regressor observations are independent over time, but the distribution of V_{it} is allowed to vary with t. This independence assumption could be relaxed, and it would even be possible to let V_{it} not vary with t, though this would entail dropping the cross section fixed effects from the model.

As in the cross section case, we first show consistency of our panel estimator (23), and then give its limiting distribution.

Assumption 2.12 *$f_{v_t}(v_{it})$ is three times continuous differentiable in v_{it}. Second moments of $\frac{D_{it}Y_{it}}{f_{v_t}(v_{it})}$, $\frac{D_{it}}{f_{v_t}(v_{it})}$, $\frac{(1-D_{it})Y_{it}}{f_{v_t}(v_{it})}$, and $\frac{(1-D_{it})Y_{it}}{f_{v_t}(v_{it})}$ are uniformly bounded by a positive number M.*

Smoothness of f_{v_t} is imposed to ensure the consistency of $\widehat{f}_{v_t}$. The uniform boundedness in Assumption 2.12 is stronger than necessary in that it suffices to prove convergence of the estimator in mean square. We use this stronger norm to simplify coping with the fixed effects dummies in equations (21) and (22).

Theorem 2.7 *Let all Assumptions in Theorem 2.2 hold, taking X to be a constant. Also let Assumption 4.3, 2.9, 2.10, 2.11, 2.12 hold and $h \to 0$, $nh \to \infty$, as $n \to \infty$. Then Equation (23) is a consistent estimator of $E(Y_1 - Y_0)$.*

The asymptotics of estimator (23) uses the additional assumption 4.6 in the Appendix, which plays the same role as Assumption 4.4, 4.5.

Define

$$\Lambda_{1it} = \frac{\left(Y_{it} - \mathrm{E}(\widetilde{a}_i + \widetilde{b}_t + Y_1)\right) D_{it} - \mathrm{E}\left[\left(Y_{it} - \mathrm{E}(\widetilde{a}_i + \widetilde{b}_t + Y_1)\right) D_{it} \middle| v_{it}\right]}{f_{v_t}(v_{it})},$$

$$\Lambda_{2it} = \frac{\left(Y_{it} - \mathrm{E}(\widetilde{a}_i + \widetilde{b}_t + Y_0)\right) (1 - D_{it}) - \mathrm{E}\left[\left(Y_{it} - \mathrm{E}(\widetilde{a}_i + \widetilde{b}_t + Y_0)\right) (1 - D_{it}) \middle| v_{it}\right]}{f_{v_t}(v_{it})},$$

$$\Pi_{1it} = \frac{D_{it}}{\widehat{f}_{v_t}(v_{it})},\ \overline{\Pi}_1 = \mathrm{E}\left(\frac{D_{it}}{f_{v_t}(v_{it})}\right),\ \Pi_{2it} = \frac{1 - D_{it}}{\widehat{f}_{v_t}(v_{it})},\ \overline{\Pi}_2 = \mathrm{E}\left(\frac{(1 - D_{it})}{f_{v_t}(v_{it})}\right).$$

Using these definitions, the following theorem provides asymptotics for estimator (23).

Theorem 2.8 *Let all Assumptions in Theorem 2.7 and Assumption 4.6 hold. Assume that* $nh^{2p} \to 0$, $n^{1-\varepsilon}h^4 \to \infty$ *for a very small positive* ε, *as* $n \to \infty$. *Then*

$$\sqrt{T}\left[\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})}{\widehat{f}_{v_t}(v_{it})}} - E(\widetilde{a}_i + \widetilde{b}_t + Y_1) + E(\widetilde{a}_i + \widetilde{b}_t + Y_0)\right]$$
$$\xrightarrow{d} N\left(0, \frac{var\left(E\left[\Lambda_{1it}\middle| b_t, \widetilde{b}_t\right]\right)}{\overline{\Pi}_1^2} - \frac{2cov\left(E\left[\Lambda_{1it}\middle| b_t, \widetilde{b}_t\right], E\left[\Lambda_{2it}\middle| b_t, \widetilde{b}_t\right]\right)}{\overline{\Pi}_1\overline{\Pi}_2} + \frac{var\left(E\left[\Lambda_{2it}\middle| b_t, \widetilde{b}_t\right]\right)}{\overline{\Pi}_2^2}\right).$$

This Theorem is proved in the supplemental appendix. For the case when n and T go to infinity at the same speed, with additional i.i.d. assumptions across t, we could obtain similar results as above. More practically, if time dummies are dropped from the model then we can obtain the faster rate of convergence $\sqrt{n}$ instead of $\sqrt{T}$, and we can obtain this rate with fixed T. We conclude this section by deriving this latter result.

Modify the model and assumption as follows. Let

$$Y_{it} = a_i + Y_{0it} + (Y_{1it} - Y_{0it})\,D_{it}, \tag{24}$$

$$D_{it} = I\left[0 \le \widetilde{a}_i + V_{it} + U_{it} \le \alpha\right]. \tag{25}$$

Assumption 2.13 T *is finite,* $n \to \infty$.

Assumption 2.14 $a_i, \widetilde{a}_i$ *are i.i.d. across* i. (Y_{0it}, Y_{1it}) *are identically distributed across* i, *and* t. *Observations are i.i.d. across* i.

Since T is now fixed, we can treat terms inside summation over t as a single term and we do not need to impose assumptions on the structure along t dimension.

Assumption 2.15 *The special regressor* V_{it} *satisfies* $E(\widetilde{a}_i + Y_{jit}|\, V_{it}, a_i, U_{it}) = E\left(\widetilde{a}_i + Y_{jit}|\, a_i, U_{it}\right)$, *for* $j = 0, 1$. $V_{it} \perp a_i, U_{it}$. V_{it} *are independent across* i *and* t. V_{it} *are identically distributed across* i *given* t, *with distribution* $f_{v_t}(V_{it})$.

Assumption 2.14 and 2.15 are simplified versions of Assumption 2.10 and 2.11 respectively. Define

$$\widetilde{\Lambda}_{1it} = \frac{(Y_{it} - \mathrm{E}(\widetilde{c}_i + Y_1))\,D_{it} - \mathrm{E}\left[(Y_{it} - \mathrm{E}(\widetilde{a}_i + Y_1))\,D_{it}|\, v_{it}\right]}{f_{v_t}(v_{it})},$$

$$\widetilde{\Lambda}_{2it} = \frac{\left(Y_{it} - \mathrm{E}(\widetilde{c}_i + Y_0)\right)(1 - D_{it}) - \mathrm{E}\left[\left(Y_{it} - \mathrm{E}(\widetilde{a}_i + Y_0)\right)(1 - D_{it}) \middle| v_{it}\right]}{f_{v_t}(v_{it})},$$

$$\widetilde{\Pi}_{1it} = \frac{D_{it}}{\widehat{f}_{v_t}(v_{it})},\ \widetilde{\overline{\Pi}}_1 = \mathrm{E}\left(\frac{D_{it}}{f_{v_t}(v_{it})}\right),\ \widetilde{\Pi}_{2it} = \frac{1 - D_{it}}{\widehat{f}_{v_t}(v_{it})},\ \widetilde{\overline{\Pi}}_2 = \mathrm{E}\left(\frac{1 - D_{it}}{f_{v_t}(v_{it})}\right).$$

Corollary 2.9 *Let all Assumptions in Theorem 2.2 hold, taking X to be a constant. Also let Assumption 4.3, 2.12, 4.6, 2.13, 2.14, 2.15 hold and* $h \to 0,\ nh \to \infty$, *as* $n \to \infty$. *Then*

$$\sqrt{n}\left[\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})}{\widehat{f}_{v_t}(v_{it})}} - E(\widetilde{a}_i + \widetilde{b}_t + Y_1) + E(\widetilde{a}_i + \widetilde{b}_t + Y_0)\right]$$
$$\xrightarrow{d} N\left(0, \frac{var\left(\frac{1}{T}\sum_{t=1}^{T}\widetilde{\Lambda}_{1it}\right)}{\widetilde{\overline{\Pi}}_1^2} - \frac{2cov\left(\frac{1}{T}\sum_{t=1}^{T}\widetilde{\Lambda}_{1it}, \frac{1}{T}\sum_{t=1}^{T}\widetilde{\Lambda}_{2it}\right)}{\widetilde{\overline{\Pi}}_1\widetilde{\overline{\Pi}}_2} + \frac{var\left(\frac{1}{T}\sum_{t=1}^{T}\widetilde{\Lambda}_{2it}\right)}{\widetilde{\overline{\Pi}}_2^2}\right).$$

The proof of this Corollary follows directly from the proof of Theorem 2.8.

3 Competition and Innovation

We apply our model to test the the "Inverted-U" theory of ABBGH (Aghion, Bloom, Blundell, Griffith, and Howitt 2005) relating innovation investments to competitiveness in an industry. ABBGH consider two types of oligopoly industries, called Neck-and-Neck (NN) industries, in which firms are technologically close to equal, and Leader-Laggard industries, where one firm is technologically ahead of others. For these industries there are two opposing effects of competition on innovation. One is the *Schumpeterian effect*, where increased competition reduces profits and thus reduces the incentive to innovate. The second is the *escape-competition effect,* where firms innovate to increase the profits associated with being a leader. For these latter firms, increased competition increases the incentive to innovate. ABBGH argue that the escape-competition effect dominates in NN industries while the Schumpeterian effect dominates in LL industries. This theory results in an inverted-U relationship, because low levels of competition are associated with NN industries and hence with low innovation, by the escape-competition effect, and high levels of competition are associated with LL industries, again leading to low innovation but now by the

Schumpeterian effect. In contrast, with an intermediate level of competition, both NN and LL industries innovate to some extent, yielding a higher overall level of innovation in steady state than in either the low or high competition industries.

ABBGH find empirical support for the inverted-U based mainly on UK data. Hashmi (2012) revisits the relationship using a richer dataset from the US, and finds no inverted-U. Hashmi notes that his finding can be reconciled with the ABBGH model by the assumption that US industries are dominated by NN industries.

For identification and estimation, both the ABBGH and Hashmi empirical results depend heavily on functional form assumptions, by fully parameterizing both the relationship of competitiveness to innovation and the functional form of error distributions. In contrast, we apply our model to test for an inverted-U relationship with minimal restrictions on functional forms and error distributions.

3.1 Data

Our sample, from Hashmi (2012), consists of US three-digit level industry annual data from 1976 to 2001. There are 116 industries, resulting in 2716 industry-year observations. Our analysis is based on three key variables: a measure of industry competitiveness, a measure of industry innovation, and a source-weighted average of industry exchange rates that serves as an instrument, and hence as our special regressor. Summary statistics for this data are reported in Table 1. We only applied our estimator to Hashmi's data and not to ABBGH's data, because the latter does not contain a continuous instrumental variable that can be used as a special regressor.

The measure of the level of competition for industry i at time t, denoted c_{it}, is defined by

$$c_{it} = 1 - \frac{1}{n_{it}} \sum\nolimits_{i=1}^{n_{it}} l_{it}, \tag{26}$$

where i indexes firms, l_{it} is the Lerner index of the price the cost margin of firm i in year t, and n_{it} is the number of firms in industry i in year t. The higher c_{it} is, the higher is the level of competition. The innovation index, denoted y_{it}, is a measure of citation-weighted patent counts, constructed using data from the NBER Patent Data Project. Details regarding the construction of this data can be found in Hashmi (2012).

To deal with the mutual endogeneity of competition and innovation, Hashmi uses a source-weighted average of industry exchange rates as instrument variable for competition. This instrument, V_{it}, is a weighted average of the US dollar exchange rate with the currencies of trading partners, with weights that vary by industry according to the share of each country in the imports to the US. This instrument has been used in other similar applications, including Revenga (1990, 1992) and Bertrand (2004).

3.2 Model Specifications

Hashmi (2012) adopts a control function approach to deal with endogeneity. In a first stage, c_{it} is regressed on V_{it}, industry dummies and time dummies, so

$$c_{it} = V_{it}\beta + a_i + b_t + w_{it}, \tag{27}$$

where a_i and b_t are fixed effects (coefficients of industry and time dummies) and w_{it} is the error from the first stage regression. The fitted residuals $\widehat{w}_{it}$ from this regression are then included as additional regressors in an outcome equation of the form

$$\ln(y_{it}) = \widetilde{a}_i + \widetilde{b}_t + \theta_0 + \theta_1 c_{it} + \theta_2 c_{it}^2 + \delta \widehat{w}_{it} + \varepsilon_{it}, \tag{28}$$

where $\widetilde{a}_i$ and $\widetilde{b}_t$ are outcome equation fixed effects (coefficients of industry and time dummies). Hashmi estimates the coefficients in equation (28) by maximum likelihood, where the distribution of errors ε_{it} is determined by assuming that $\ln(y_{it})$ has a negative binomial distribution, conditional on c_{it}, industry, and year dummies. This model assumes the relationship of $\ln(y)$ to c is quadratic, with an inverted-U shape if θ_1 is positive and θ_2 negative. The industry and time dummies cannot be differenced out in this model, and so are estimated along with the other parameters.

In addition to the possibility that this quadratic is misspecified, or that the endogeneity takes a form that is not completely eliminated by the control function addition of $\widehat{w}$ as a regressor, or that the distribution is not negative binomial, Hashmi's estimates could also suffer the from the problem of incidental parameters (Neyman and Scott 1948). This problem is that the need to estimate industry and time fixed effects results in inconsistent parameter estimates unless both T and n go to infinity. In this application neither T nor n is particularly small, but the presence of the fixed effects still results in over 100 nuisance parameters to estimate, which can lead to imprecision. Our intention is not to criticize Hashmi's or ABBGH's model, but only to point out that there are many reasons why it is desirable to provide a less parametric alternative, to verify that their results are not due to potential model specification or estimation problems.

To apply our estimator, let the treatment indicator D_{it} equal one for industries i that have neither very low nor very high levels of competition in period t, and otherwise let $D_{it} = 0$. We then let innovation y_{it} be determined by

$$y_{it} = \widetilde{a}_i + \widetilde{b}_t + Y_{0it} + (Y_{1it} - Y_{0it})D_{it}. \tag{29}$$

where $\widetilde{a}_i$, $\widetilde{b}_t$ are the industry and time dummies respectively, and Y_{0it} are Y_{1it} are unobserved potential outcomes for industry i in time t, after controlling for time and industry fixed effects. Unlike the specific functional form imposed by equation (28), equation (29) is almost completely unrestricted. Both Y_{1it} and Y_{0it} are random variables with completely unknown distributions that can be correlated with each other, and with the error term in the D_{it} equation, in completely unknown ways. We will then estimate the ATE $E(Y_{1it} - Y_{0it})$, which equals the average difference in outcomes y (after controlling for fixed effects), between industries with moderate levels of competitiveness, versus industries that have very low or very high levels of competitiveness.

What our model assumes about the treatment indicator D_{it} is

$$D_{it} = I(\alpha_0 \leq a_i + b_t + V_{it} + U_{it} \leq \alpha_1), \tag{30}$$

where a_i and b_t are industry and time dummies, U_{it} are unobserved, unknown factors that affect competition, and α_0 and α_1 are unknown constants. The way to interpret equation (30) is that the latent variable c^*_{it} given by

$$c^*_{it} = a_i + b_t + V_{it} + U_{it} \tag{31}$$

is some unobserved true level of competitiveness of industry i in time t. Our model does not require the observed competitiveness measure c_{it} to equal the true measure c^*_{it}, but if they do happen to be equal then our model implies that Hashmi's equation (27) is correctly specified. Note when comparing the models for c^*_{it} and c_{it} to each other that replacing c^*_{it} with βc^*_{it} to make equation (31) line up with equation (30) is a free scale normalization that can be made without loss of generality, because the definition of D_{it} is unaffected by rescaling c^*_{it}.

As in Hashmi's model, our estimator assumes that V_{it} is a valid instrument, affecting competitiveness c_{it}^* and hence the treatment indicator D_{it}, but not directly affecting the outcome y_{it}. We also require that V_{it} has a large support. This appears to be the case in our data, e.g., the exchange rate measure sometimes as much as doubles or halves over time even within a single industry, and varies substantially across industries as well.

3.3 Measurement Errors in Competitiveness

In our empirical application, we define D_{it} to be one when the observed c_{it} lies between the .25 and .75 quantiles of the empirical c_{it} distribution (we also experiment with other quantiles). This is therefore consistent with equation (27) if c_{it} is linear in c_{it}^*. However, our model remains consistent even if c_{it} differs greatly from c_{it}^*, as long as the middle 50% of industry and time periods in the c_{it} distribution corresponds to the middle 50% of industry and time periods in the c_{it}^* distribution.

More generally, suppose c_{it} equals c_{it}^* plus some measurement error. Then the Hashmi model, even if correctly specified, will be consistent only if this measurement error satisfies the conditions necessary for validity of their control function estimator. Some control function estimators remain consistent in models containing measurement errors that are classical, i.e., independent of the true c_{it}^* and of the true model. However, the Hashmi control function estimator would not be consistent even with classical measurement errors, because equation (28) is nonlinear in the potentially mismeasured variable c_{it} (this is not intended as a criticism of Hashmi's empirical application, since that work uses control functions only to deal with endogeneity and never made any claims regarding measurement errors).

In contrast, our estimator can remain consistent in theory even with measurement errors that are large and nonclassical, as long as c_{it} correctly sorts industries into moderate versus nonmoderate levels of competitiveness. However, in practice, measurement error in c_{it} will likely cause some industries to be misclassified, so D_{it} is likely to be mismeasured for some industries (particularly for some that are near the .25 and .75 quantile cutoffs). Also, in practice we should expect Hashmi's control function specification to at least partly correct for potential measurement error.

To summarize: competitiveness is difficult to precisely define and measure, and as a result the impact of measurement errors on this analysis could be large. One advantage of our methodology is that it only depends on sorting industries into two groups (that is, moderate versus extreme levels of competitiveness as indicated by D_{it}). While this sorting discards some information and may therefore cost some efficiency, it will also mitigate measurement error biases, because only a small number of observations of D_{it} are likely to be mismeasured even if most or all of the c_{it} observations are mismeasured to some extent. To check whether this intuition is correct, in an appendix we do a monte carlo analysis that compares the accuracy of our estimator with that of Hashmi's in the presence of measurement errors.

3.4 Estimation

Our estimator is quite easy to implement, in part because it does not entail any numerical searches or maximizations. We first estimate the density of V_{it} separately for each year, using a standard kernel density estimator

$$\widehat{f}_{v_t}(v_{it}) = \frac{1}{n-1} \sum_{j\neq i,,j=1}^{n} \frac{1}{h} K\left(\frac{v_{it}-v_{jt}}{h}\right).$$

Note that the density is estimated at each of the data points v_{it}. We employ a Gaussian kernel function K, and choose the bandwidth h using Silverman's rule of thumb. Our estimator involves dividing by these nonparametric density estimates, which can result in outlier observations when $\widehat{f}$ is close to zero. As suggested in Lewbel (2000) and Dong and Lewbel (2012) for other special regressor based estimators, we trim out (i.e., discard from the sample) the 2% of observations with the smallest values of $\widehat{f}_{v_t}$. This defines the trimming function $I_\tau(v)$ from our asymptotic theory.

Given the density estimates $\widehat{f}_{v_t}(v_{it})$, our resulting estimate of the ATE $E(Y_{1it} - Y_{0it})$ is then given by

$$\text{Trim-ATE} = \frac{\sum_i \sum_t I_\tau(v_{it}) D_{it} Y_{it} / \widehat{f}_{v_t}(V_{it})}{\sum_i \sum_t I_\tau(v_{it}) D_{it} / \widehat{f}_{v_t}(V_{it})} - \frac{\sum_i \sum_t I_\tau(v_{it})(1-D_{it}) Y_{it} / \widehat{f}_{v_t}(V_{it})}{\sum_i \sum_t I_\tau(v_{it})(1-D_{it}) / \widehat{f}_{v_t}(V_{it})} \tag{32}$$

where the i and t sums are over the 98% of observations that were not trimmed out. This model corresponds to the estimator (23), which has standard errors that we calculate based on the asymptotic distribution provided in Theorem 2.8. To assess the effect of the trimming on this estimator, we construct a corresponding estimate of ATE that is not trimmed, given by

$$\text{No-Trim-ATE} = \frac{\sum_i \sum_t D_{it} Y_{it} / \widehat{f}_{v_t}(V_{it})}{\sum_i \sum_t D_{it} / \widehat{f}_{v_t}(V_{it})} - \frac{\sum_i \sum_t (1 - D_{it}) Y_{it} / \widehat{f}_{v_t}(V_{it})}{\sum_i \sum_t (1 - D_{it}) / \widehat{f}_{v_t}(V_{it})}. \tag{33}$$

For comparison, in addition we calculate a Naive-ATE estimator given by

$$\text{Naive-ATE} = \frac{\sum_i \sum_t D_{it} Y_{it}}{\sum_i \sum_t D_{it}} - \frac{\sum_i \sum_t (1 - D_{it}) Y_{it}}{\sum_i \sum_t (1 - D_{it})}. \tag{34}$$

This Naive-ATE just subtracts the average value of Y_{it} when $D_{it} = 0$ from the average value of Y_{it} when $D_{it} = 1$. This would be a consistent estimator of the ATE if treatment were unconfounded, that is, if low or high competitiveness as indicated by D_{it} was randomly assigned over firms and time periods. One could also consider a LATE estimator such as an instrumental variables regression of Y on D using V as an instrument. However, as noted in the introduction, LATE requires that the probability of treatment increase monotonically with the instrument. This requirement does not hold in our application, since both increasing or decreasing V sufficiently causes the probability of treatment to decrease.

We also compare our results to a parametric maximum likelihood estimate of the ATE (denoted ML-ATE) assuming a Heckman (1979) type selection model for treatment. This model assumes equations (29) and (30) hold and that U, Y_0, Y_1 are jointly normally distributed. Let Φ denote the standard normal cumulative distribution function, $\theta_0 = E(Y_0)$, $\theta_1 = E(Y_1)$, and $\sigma = cov[U, Y_0, Y_1]$ be the three by three covariance matrix of elements σ_{kl} for $k = 1, 2, 3$ and $l = 1, 2, 3$. Then the ML-ATE is defined by

$$\begin{aligned}
&\text{ML-ATE } = \widehat{\theta}_1 - \widehat{\theta}_0 \quad \text{where} \quad \left[\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\alpha}_0, \widehat{\alpha}_1, [\widehat{\sigma}_{kl}]_{3\times 3}\right] = \arg\max \sum_i \sum_t \\
&\left\{ (1 - D_{it}) \log \left(\Phi\left(\frac{Y_{it} - \theta_0}{\sigma_{22}}\right) \left[\Phi\left(\frac{\alpha_0 - V_{it} - \frac{\sigma_{12}}{\sigma_{22}}(Y_{it} - \theta_0)}{\sqrt{\sigma_{11} - \sigma_{12}^2/\sigma_{22}}} \right) + 1 - \Phi\left(\frac{\alpha_1 - V_{it} - \frac{\sigma_{12}}{\sigma_{22}}(Y_{it} - \theta_0)}{\sqrt{\sigma_{11} - \sigma_{12}^2/\sigma_{22}}} \right) \right] \right) \right. \\
&\left. + D_{it} \log \left(\Phi\left(\frac{Y_{it} - \theta_1}{\sigma_{33}}\right) \left[\Phi\left(\frac{\alpha_1 - V_{it} - \frac{\sigma_{13}}{\sigma_{33}}(Y_{it} - \theta_1)}{\sqrt{\sigma_{11} - \sigma_{13}^2/\sigma_{33}}} \right) - \Phi\left(\frac{\alpha_0 - V_{it} - \frac{\sigma_{13}}{\sigma_{33}}(Y_{it} - \theta_1)}{\sqrt{\sigma_{11} - \sigma_{13}^2/\sigma_{33}}} \right) \right] \right) \right\}.
\end{aligned}$$

3.5 Empirical Results

Figure 1 shows our kernel density estimates $\widehat{f}_{v_t}$ for each year t. The estimates can be seen to vary quite a bit over time, so we use separate density estimates for each year instead of assuming a constant distribution across years. Figure 2 shows a scatterplot of our competitiveness and innovation data. It is difficult to discern any clear relationship between the two by eye.

Table 2 shows our main empirical results. The first row of Table 2 provides estimates where D_{it} is defined to equal one for the middle half of the data, that is, D_{it} equals one for firms and years that lie between the 25th and 75th percent quantiles of the observed measure of competition, making half the observations treated and the other half untreated. Other rows of Table 2 report results using different quantiles to define D_{it}. In each row of Table 2 we report four estimates of ATE, as described in the previous section. Standard errors for all the estimates are provided in parentheses.

An inverted-U would imply a positive ATE, but all of our estimates are negative, confirming Hashmi's finding that the inverted-U is not present in US data, perhaps because the US is dominated by NN type industries. For example, our main estimate from the first row of Table 2 is that the Trim-ATE equals -3.9, and is strongly statistically significant. We conclude that Hashmi's main result regarding signs of the effect is likely genuine and not due to possible model specification errors.

We also find that failure to appropriately control for error correlations between competitiveness and innovation substantially biases the magnitudes of estimated treatment effects. Our semiparametric estimates of the ATE are 50% to 100% larger than both the naive estimates that ignore these correlations, and the maximum likelihood estimates that allow for correlations but requires the errors to be jointly normally distributed.

Attempts to find a positive ATE by experimenting with more unusual quantiles for defining D_{it} were for the most part fruitless. An exception, based on examination of Figure 2, was to define the left and right thresholds by 0.62 (10%) and 0.68 (20%) respectively. This implies a heavily skewed inverted U where 80% of firms are in the upper tail. This yields a positive ATE of 8.66, but this model is implausible, since it treats a very narrow spike in Figure 2 as the set of all moderately sized firms. We also experimented with varying the degree of trimming, but we only report results without trimming and with 2% percent trimming because the impacts of other changes in trimming were very small.

3.6 Monte Carlo designed for the empirical example

To assess how the estimator works in small samples, we provide two sets of Monte Carlo experiments. We designed these experiments to closely match moments and other features of our empirical data, to see how likely our estimator is to perform well in a controlled setting that mimics our actual application. The number of observations is set to 2716, the same as the number of observations in our empirical dataset. The same four estimators we applied on the actual data, Trim-ATE, No-Trim-ATE, Naive-ATE and ML-ATE, are analyzed in each set of Monte Carlo simulations

Let e_{1i}, e_{2i}, e_{3i}, and V_i be random variables that are drawn independently of each other. We consider a few different distributions for these variables as described below. The counterfactual outcomes in our simulation are defined by

$$Y_{0i} = \theta_0 + \theta_{01} e_{1i} + \theta_{02} e_{3i} \text{ and } \ Y_{1i} = \theta_1 + \theta_{11} e_{2i} + \theta_{12} e_{3i}.$$

True competitiveness is constructed to equal $V_i + \theta_2 e_{3i}$, and treatment D_i is defined to equal one for observations i that lie between the 25th and 75th quantile of the distribution of $V_i + \theta_2 e_{3i}$. The observed outcome is then constructed as

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) D_i.$$

For simplicity, fixed effect type dummies are omitted from the model. Note that e_{3i} appears in D_i, Y_{0i}, and Y_{1i}, and so is the source of confounding in this model. By construction, the unobserved U_i in our theoretical model is given by $U_i = \theta_2 e_{3i}$. Let θ denote the vector of parameters $(\theta_0, \theta_1, \theta_2, \theta_{01}, \theta_{02}, \theta_{11}, \theta_{12})$. In each Monte Carlo experiment the parameter vector θ is set to match moments and outcomes of our actual data, specifically, they are set to make the ATE $\theta_1 - \theta_0$ equal our estimate -3.90, and to make the mean and variance of Y_i and D_i, and the covariance between Y_i and D_i, equal the values observed in our data. The variance of V_i is freely normalized (inside the binomial response indicator) to equal one.

The ML-ATE estimator is asymptotically efficient when e_{1i}, e_{2i}, and e_{3i} are normally distributed. In our first experiment we let e_{1i}, e_{2i}, e_{3i}, and V_i each have a standard normal distribution, so the resulting ML-ATE estimates can then serve as an efficient benchmark.

As noted by Khan and Tamer (2010), single threshold crossing model special regressor estimators converge at slow rates when f_v has thin tails, as in the previous design. Although their results are not directly applicable to this paper's two threshold model, it is still sensible to see if our estimator works better with thicker tails, so our second experiment gives e_{1i}, e_{2i}, e_{3i}, and V_i each a uniform distribution on $[-0.5, 0.5]$. Note this is still likely not the best case for our estimator, since Khan and Tamer (2010) note that special regressor methods converge fastest when V has a thick tail and all other variables have thin tails.

Both the normal and uniform designs have symmetric errors, which favors the ML alternative over our estimator. However, with symmetric errors it is impossible to define a vector θ that matches all the moments of the empirical data, because symmetry prevents matching the empirical covariance between Y and D. Therefore, in both designs we choose values for θ that match all the other moments and come as close as possible to matching this covariance (the required values for θ are given in the footnote of Table 3).

To match the empirical correlation between Y and D along with other moments, we next consider designs that introduce asymmetry into the confounder e_{3i}. In our third experiment, we let e_{1i}, e_{2i}, and V_i be standard normal and let e_{3i} be a modified normal, equaling a standard normal with probability one half when $e_{3i} < 0$ and equaling θ_3 times a standard normal with probability one half when $e_{3i} \geq 0$. When then choose θ_3 along with the other elements of θ to match the moments of the empirical data including the covariance of Y with D. This required setting $\theta_3 = 2.65$. Similarly, in a fourth experiment we let e_{1i}, e_{2i}, and V_i be uniform on $[-0.5, 0.5]$ and take e_{3i} to equal a (demeaned) mixed uniform distribution. This mixture was uniform on $[-2, 0]$ with probability one half and uniform on $[0, 5]$ with probability one half, before demeaning.

Each of these four Monte Carlo experiments was replicated 10,000 times, and the results are summarized in Table 3. Panel A in Table 3 is the symmetric normal design. Because of symmetry, all of the estimators in this design are unbiased. ML, being efficient here, has the lowest root mean squared error (RMSE), and the naive estimator is almost as efficient as ML in this case, since it just involves differencing simple covariance estimates. Our Trim-ATE estimator performs reasonably well compared to the efficient estimator, being unbiased and having a RMSE of .43 versus the efficient .30. Trimming improves the RMSE enormously here, as expected because f_v has thin tails, which produces outliers in the denominator of averages weighted by f_v.

Panel B of Table 3 shows that, in the symmetric uniform design, all four estimators are almost identical. The happens because, with V is uniform, $\widehat{f}_v$ is close to a constant, and the estimators for the average effects of the treated and the untreated are close to their sample means.

In the asymmetric designs, given in panels C and D of Table 3, the ML-ATE and Naive-ATE are no longer consistent, and both become substantially downward biased, with an average value of about one half the true value of -3.90. In contrast, our trimmed and untrimmed ATE estimates had far smaller downward biases, resulting in much smaller RMSE, particularly for the Trim-ATE.

The differences in biases between the inconsistent estimators (ML-ATE and NAIVE-ATE) and our proposed estimator in these asymmetric Monte Carlos closely match the observed differences in our empirical application estimates. Specifically, in case 1 of Table 2 the estimated ATE using the ML and Naive estimators is about one half the estimate of -3.90 we obtained using Trim-ATE. This provides evidence that the Monte Carlo results in panels C and D of Table 3 are relevant for assessing the empirical performance of our proposed estimator.

In addition to assessing the quality of estimators we also assess the quality of associated standard error estimates, by providing, in the last column of Table 3, the percentage of times the true ATE fell in the estimated 95% confidence interval (defined as the estimated ATE plus or minus two estimated standard errors). In the symmetric designs all the estimated standard errors for all the estimators were too large, yielding overly conservative inference. In the asymmetric designs the estimated 95% confidence intervals of the inconsistent estimators ML-ATE and NAIVE-ATE were very poor, containing the true value less than 25% of the time. The No-Trim-ATE did much better, but our preferred estimator, Trim-ATE, was by far the best, giving correct 95% coverage in panel C, and conservative 99% coverage in panel D.

4 Conclusions

In this article, we propose a new method to estimate the average treatment effect in a two threshold model, where the treated group is a middle choice. In our application, treatment is defined as facing an intermediate level of competition, versus a low or high level of competition.

The proposed model is confounded, because the unobservables that affect the treatment indicator D can be correlated in unknown ways with potential outcomes Y_0 and Y_1, with or without conditioning on other covariates. No parametric or semiparametric restrictions are placed on distributions of treatment and potential outcomes, so treatment effects are not identified by functional form. Our model assumes a continuous instrument V with large support, but treatment effects are not identified at infinity, because both very large and very small values of V drive the probability of treatment close to zero, while no value of V (or of other covariates) drives the probability of treatment close to one. So in this framework none of the conditions that are known to permit point identification of the ATE hold. Even the monotonicity conditions usually required for identifying LATE are not satisfied. Nevertheless, we show that the ATE is identified, using a special regressor argument, and we provide conditions under which the corresponding estimate of the ATE is consistent, and root-n normal. Root-n consistency is even obtained in a panel context with fixed effects, despite nonlinearities that would usually induce an incidental parameters problem in the equaiton determining probability of treatment. We provide Monte Carlo results that show that our estimator works well in small samples (comparable to the data in our empirical application) we show in an Appendix that our estimator is relatively robust to measurement error and misspecification.

We use our method to investigate the relationship between competition and innovation. Our estimates using a dataset from Hashmi (2012) confirm Hashmi's findings that an inverted-U is not present in US data. We also find that standard parametric model and naive treatment effect estimators substantially underestimate the magnitude of the treatment effect in this context.

References

[1] Abadie, A., J. Angrist, and G. Imbens, (2002), "Instrumental Variables Estimation of Quantile Treatment Effects,"Econometrica. Vol. 70, No. 1, 91-117.

[2] Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt, (2005) "Competition and innovation: an inverted-U relationship," Quarterly Journal of Economics, 120(2):701-28

[3] 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 90:430, 431–442.

[4] Angrist, J. D., G. W. Imbens, and D. Rubin, (1996) "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association 91:434, 444–455.

[5] Ashenfelter, O. (1978), “Estimating the Effect of Training Programs on Earnings,”Review of Economics and Statistics, 60, 47-57.

[6] Ashenfelter, O., and D. Card, (1985), “Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs,” Review of Economics and Statistics, 67, 648-660.

[7] Barnow, B. S., G. G. Cain, and A. S. Goldberger (1980): “Issues in the Analysis of Selectivity Bias,” in Evaluation Studies, Vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage, 43–59.

[8] Bertrand, M. (2004), "From the invisible handshake to the invisible hand? How import competition changes the employment relationship," Journal of Labor Economics, 22(4):722-65.

[9] Bitler, M., J. Gelbach, and H. Hoynes (2002) “What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments,” unpublished paper, Department of Economics, University of Maryland.

[10] Björklund, A. and R. Moffitt, (1987), “The Estimation of Wage Gains and Welfare Gains in Self–Selection Models,”Review of Economics and Statistics, Vol. LXIX, 42–49.

[11] Card, D., (1990), “The Impact of the Mariel Boatlift on the Miami Labor Market,” Industrial and Labor Relations Review 43, 245-257.

[12] Card, D., and A. Krueger, (1993), “Trends in Relative Black-White Earnings Revisited,” American Economic Review, vol. 83, no. 2, 85-91.

[13] Card, D., and A. Krueger, (1994): “MinimumWages and Employment: A Case Study of the Fast-food Industry in New Jersey and Pennsylvania,” American Economic Review, 84 (4), 772-784.

[14] Chernozhukov, V., and C. Hansen, (2005), “An IV Model of Quantile Treatment Effects,” Econometrica, 73(1), 245-261.

[15] Cook, P.J., and G. Tauchen, (1982), "The effect of Liquor Taxes on Heavy Drinking," Bell Journal of Economics, 13(2): 379-90.

[16] Cook, P.J., and G. Tauchen, (1982), "The effect of Minimum Drinking age Legislation on Youthful Auto Fatalities, 1970-1977," Journal of Legal Studies, 13(1): 169-90.

[17] Cochran, W., AND D. Rubin (1973): "Controlling Bias in Observational Studies: A Review," Sankhyā, 35, 417–446.

[18] Dong Y., and A. Lewbel (2012), "A Simple Estimator for Binary Choice Models with Endogenous Regressors," forthcoming, Econometric Reviews.

[19] Firpo, S. (2006), "Efficient Semiparametric Estimation of Quantile Treatment Effects," Econometrica, 75(1), 259-276.

[20] Gautier, E. and S. Hoderlein (2011), "A Triangular Treatment Effect Model with Random Coefficients in the Selection Equation," unpublished manuscript.

[21] Hashmi, A.R. (2012), "Competition and innovation: the inverted-U relationship revisited," forthcoming in Review of Economic Statistics.

[22] Heckman, J., (1979), "Sample Selection Bias as a Specification Error," Econometrica, 47(1), 153-162.

[23] Heckman, J., and R. Robb (1984): "Alternative Methods for Evaluating the Impact of Interventions," Longitudinal Analysis of Labor Market Data, ed. by J. Heckman and B. Singer. Cambridge, U.K.: Cambridge University Press, 156–245.

[24] Heckman, J.J., S. Urzua, and E. Vytlacil (2006), "Understanding instrumental variables in models with essential heterogeneity," Review of Economics and Statistics, 88(3): 389–432.

[25] Heckman, J. J., and E. Vytlacil (1999) "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," Proceedings of the National Academy of Science, USA, 96, 4730–4734.

[26] Heckman, J. J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," Econometrica, 73, 669–738.

[27] Heckman, J. J., and E. Vytlacil (2007a) "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Evaluation of Public Policies," Handbook of Econometrics, J.J. Heckman and E.E. Leamer (eds.), Vol. 6, North Holland, Chapter 70.

[28] Heckman, J. J., and E. Vytlacil (2007b) “Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments,” Handbook of Econometrics, J.J. Heckman and E.E. Leamer (eds.), Vol. 6, North Holland, Chapter 71.

[29] Honore, B. and A. Lewbel, (2002) "Semiparametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors," Econometrica, 70, 2053-2063.

[30] Imbens, G., and J. Angrist (1994), “Identification and Estimation of Local Average Treatment Effects,” Econometrica, Vol. 61, No. 2, 467-476.

[31] Imbens, G., and J. Wooldridge (2009), "Recent Development in the Econometrics of Program Evaluation," Journal of Economic Literature, 47:1, 5-86.

[32] Khan, S, and E. Tamer (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," Econometrica, 6, 2021-2042.

[33] Kitagawa, T. (2009), "Identification Region of the Potential Outcome under Instrument Independence," unpublished manuscript.

[34] Lewbel, A. (1998), "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," Econometrica, 66, 105-122.

[35] Lewbel, A. (2000), "Semiparametric qualitative response model estimation with unknown heteroscedasticity and instrumental variables," Journal of Econometrics, 97, 145-177.

[36] Lewbel, A. (2007), "Endogenous selection or treatment model estimation," Journal of Econometrics, 141, 777-806.

[37] Lewbel, A. (2012), "An overview of the special regressor method," a chapter in Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, Oxford University Press.

[38] Lewbel, A., Y. Dong, and T.T. Yang (2012), "Comparing Features of Convenient Estimators for Binary Choice Models With Endogenous Regressors," Canadian Journal of Economics, 45, 809-829.

[39] Meyer, B., K. Viscusi and D. Durbin, (1995), “Workers’ Compensation and Injury Duration: Evidence from a Natural Experiment,” American Economic Review, Vol. 85, No. 3, 322-340.

[40] Newey, W. K. and D. McFadden (1994), ”Large Sample Estimation and Hypothesis Testing,” in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[41] Neyman, J., and E.L. Scott, (1948), "Consistent estimation from partially consistent observations," Econometrica 16, 1-32.

[42] Revenga, A. (1990), "Essays on labor market adjustment and open economics." PhD diss., Harvard University, Economics Department.

[43] Revenga, A. (1992), "Exporting jobs? The impact of import competition on employment and wages in U.S. manufacturing," Quarterly Journal of Economics, 107, no. 1:255–84.

[44] Rosenbaum, P., and D. Rubin (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” Biometrika, 70, 41–55.

[45] Rubin, D. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," Journal of Educational Psychology, 66, 688–701.

Appendix A: Robustness to Measurement Errors

Observable indices of competitiveness of an industry, like the average Lerner index in equation (26), may be relatively crude measures of true competitiveness. In this section we therefore assess the robustness of our estimator, relative to a parametric model estimator like Hashmi’s, to measurement error in the index of competitiveness. we first show that both models, as one would expect, become inconsistent if competitiveness is mismeasured, even when the models are otherwise correctly specified. However, we also show that the bias in our estimator resulting from measurement error is quite small relative to alternative estimators.

First consider the case where competitiveness is mismeasured, but a parametric model like Hashmi’s (dropping fixed effects for simplicity) is the correct specification in terms of true competitiveness. This model assumes

$$\ln Y = \theta_0 + \theta_1 c^* + \theta_2 c^{*2} + \widetilde{e}, \tag{35}$$

where $\ln Y$ is logged innovation, c^* is the true level of competitiveness, and $\widetilde{e}$ is an error term. For simplicity we ignore discreteness in $\ln Y$, and we assume c^* can be linearly decomposed into the observable instrument V and an unobserved independent component W, so

$$c^* = V + W. \tag{36}$$

Assume validity of Hashmi's control function type assumption that $\widetilde{e} = \lambda W + e$ where e is independent of W and V, so

$$\ln Y = \theta_0 + \theta_1 c^* + \theta_2 c^{*2} + \lambda W + e \tag{37}$$

In this model, if c^* were observed, then control function estimation (first regressing c^* on a constant and V, getting the residuals $\widehat{W}$, and then regressing $\ln Y$ on a constant, c^*, c^{*2}, and $\widehat{W}$) would consistently estimate the θ coefficients and hence any desired treatment effects based on θ.

Now assume the observable competitiveness measure c equals the true measure c^* plus measurement error c_e, so

$$c = c^* + c_e, \tag{38}$$

where c_e is the measurement error and independent of c^* and e. To take the best case scenario for the parametric model, assume that the measurement error c_e has mean zero and is independent of V, W, and e.

Substituting equation (38) into Equation (37) gives

$$\ln Y = \theta_0 + \theta_1 c + \theta_2 c^2 + \lambda W + e^* \tag{39}$$

where

$$e^* = \theta_1 c_e - 2\theta_2 c c_e - \theta_2 c_e^2 + e.$$

The error e^* does not have mean zero and correlates with c and c^2, which makes the control function estimator inconsistent. Unlike the case of linear models with independent mean zero measurement errors, the control function estimator is not consistent because of the nonlinearity in this model.

Now consider applying our nonparametric estimator to this model. The treatment indicator D that we would construct is defined as equaling one for firms in the .25 to .75 quantile of c and zero otherwise, while the corresponding indicator D^* based on the true measure of competitiveness equals one for firms in the .25 to .75 quantile of c^* and zero otherwise. Unless the measurement error c_e is extremely large, for the large majority of firms D will equal D^*. This is part of what makes our estimator more robust to measurement error. Even if all firms have c mismeasured to some extent, most will still be correctly classified in terms of D.

To check the relative robustness of these estimators to measurement error, we perform additional Monte Carlo analysis. As before, we construct simulated data to match moments and the sample size of the empirical data set, and to make what would be the true treatment effect in the model match our empirical estimate of -3.9. We do two simulations, one using normal errors and one based on uniform errors, as before. In both, V and W are scaled to have equal magnitudes, so $V = \delta_0 + \delta_1 \varepsilon_1$ and $W = \delta_0 + \delta_1 \varepsilon_2$. To match data moments, the normal error simulations set $\delta_0 = 0.375$, $\delta_1 = 0.0733$, and $c_e = \kappa_1 \varepsilon_3$ where ε_1, ε_2, and ε_3 are independent standard normals and κ_1 is a constant with values that we vary to obtain different magnitudes of measurement error. The uniform error simulations set $\delta_0 = \delta_1 = 0.25$, and $c_e \sim \kappa_2(\varepsilon_3 - 0.5)$, where now ε_1, ε_2, and ε_3 are independent random variables that are uniformly distributed on $[0, 1]$.

To check for robustness against an alternative specification as well as measurement error, we also generate data replacing the quadratic form in Equation (35) with the step function

$$\ln Y = \theta_0 + (\theta_1 - \theta_0) D^* + \widetilde{e}, \tag{40}$$

where D^*, D, c^*, c, V, W, and e are all defined as above.

The Monte Carlo results, based on 10,000 replications, are reported in Tables 4 and 5. In addition to trying out the four estimators we considered earlier, (Trim-ATE, No-Trim-ATE, Naive-ATE, and ML-ATE) we also apply the control function estimator described above, analogous to Hashmi's estimator.

Our main result is that, with both normal and uniform errors, the greater the magnitude of measurement error is (that is, the larger the κ_1 and κ_2 are), the better our estimator performs relative to other estimators. For the quadratic model without measurement error the control function would be a consistent parametric estimator and so should outperforms our semiparametric estimator. We find this also holds with very small measurement error (e.g., $\kappa_1 = .02$ in the left side block of Table 4), however, both control function and Trim-ATE perform about equally at $\kappa_1 = .03$, and at the still modest measurement error level of $\kappa_2 = .04$, Trim-ATE has smaller RMSE (root mean squared error) than all the other estimators, including control function. Similar results hold for the uniform error model reported in Table 5. Also, in the step function model (shown on the right side of Tables 4 and 5) our Trim-ATE is very close to, or superior to, all the other estimators including control functions at all measurement error levels.

It is worth noting that possible measurement error affects our empirical application only because we defined treatment D in terms of an observed, possibly mismeasured underlying variable, competitiveness. In other applications the treatment indicator may be observed without error even when an underlying latent measure is completely unobserved. For example, suppose an outcome Y is determined in part by an individual's chosen education level, which in turn is determined by an ordered choice specification. The true education level of a student might be unobserved, but a treatment D defined as having graduated high school but not college could still be correctly measured.

Appendix B: Additional Assumptions and Proofs

Proof of Theorem 2.1. To prove this look first at

$$\begin{aligned}
\mathrm{E}\left(\frac{I_\tau DY}{f\left(V \mid X\right)} \mid U, X\right) &= \mathrm{E}\left[\mathrm{E}\left(\frac{I_\tau DY_1}{f\left(V \mid X\right)} \mid V, U, X\right) \mid U, X\right] \\
&= \mathrm{E}\left[\frac{I_\tau I\left[\alpha_0\left(X\right) \leq V + U \leq \alpha_1\left(X\right)\right] E\left(Y_1 \mid V, U, X\right)}{f\left(V \mid X\right)} \mid U, X\right] \\
&= \int_{supp(V|U,X)} \frac{I_\tau I\left[\alpha_0\left(X\right) - U \leq v \leq \alpha_1\left(X\right) - U\right] E\left(Y_1 \mid U, X\right)}{f\left(v \mid X\right)} f\left(v \mid U, X\right) dv \\
&= \int_{\alpha_0(X)-U}^{\alpha_1(X)-U} \frac{E\left(Y_1 \mid U, X\right)}{f\left(v \mid X\right)} f\left(v \mid X\right) dv = E\left(Y_1 \mid U, X\right) \int_{\alpha_0(X)-U}^{\alpha_1(X)-U} 1 dv \\
&= \left[\alpha_1\left(X\right) - \alpha_0\left(X\right)\right] E\left(Y_1 \mid U, X\right),
\end{aligned}$$

the fourth equality holds by Assumption 2.3.

Therefore

$$E\left(\frac{I_\tau DY}{f\left(V \mid X\right)} \mid X\right)=\left[\alpha_1\left(X\right)-\alpha_0\left(X\right)\right]E\left(Y_1 \mid X\right)$$

The same analysis dropping Y gives

$$E\left(\frac{I_\tau D}{f\left(V \mid X\right)} \mid X\right)=\alpha_1\left(X\right)-\alpha_0\left(X\right)$$

so

$$E\left(\frac{I_\tau DY}{f\left(V \mid X\right)} \mid X\right)=E\left(Y_1 \mid X\right)E\left(\frac{I_\tau D}{f\left(V \mid X\right)} \mid X\right)$$

Similarly,

$$\begin{aligned}
E\left(\frac{I_\tau\left(1-D\right)Y}{f\left(V \mid X\right)} \mid X\right) &= E\left(\frac{I_\tau\left(1-D\right)Y_0}{f\left(V \mid X\right)} \mid X\right)\\
&= E\left(\frac{I_\tau Y_0}{f\left(V \mid X\right)} \mid X\right)-E\left(\frac{I_\tau DY_0}{f\left(V \mid X\right)} \mid X\right)\\
&= E\left(Y_0 \mid X\right)E\left(\frac{I_\tau}{f\left(V \mid X\right)} \mid X\right)-\left[\alpha_1\left(X\right)-\alpha_0\left(X\right)\right]E\left(Y_0 \mid X\right)\\
&= E\left(Y_0 \mid X\right)E\left(\frac{I_\tau}{f\left(V \mid X\right)}-\left[\alpha_1\left(X\right)-\alpha_0\left(X\right)\right] \mid X\right)\\
&= E\left(Y_0 \mid X\right)E\left(\frac{I_\tau\left(1-D\right)}{f\left(V \mid X\right)} \mid X\right)
\end{aligned}$$

Together these equations prove the result. ■

Proof of Theorem 2.2. The proof is analogous to that of Theorem 2.1.

$$
\begin{aligned}
& E\left(\frac{I_{\tau it} D_{it} Y_{it}}{f_{v_t}\left(V_{it}|X_{it}\right)} \right| U_{it}, a_i, b_t, X_{it} \Bigg) = E\left[E\left(\frac{I_{\tau it} D_{it}\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it}\right)}{f_{v_t}\left(V_{it}|X_{it}\right)} \mid V_{it}, U_{it}, a_i, b_t, X_{it} \right) \mid U_{it}, a_i, b_t, X_{it} \right] \\
= \; & E\left[\frac{I_{\tau it} I\left(\alpha_0(X_{it}) \leq a_i + b_t + V_{it} + U_{it} \leq \alpha_1(X_{it})\right) E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid V_{it}, u_{it}, a_i, b_t, X_{it}\right)}{f_{v_t}\left(V_{it}|X_{it}\right)} \mid U_{it}, a_i, b_t, X_{it} \right] \\
= \; & \int_{supp(V_{it}|U_{it}, a_i, b_t, X_{it})} \frac{I_{\tau it} I\left(\alpha_0(X_{it}) - a_i - b_t - U_{it} \leq v_{it} \leq \alpha_1(X_{it}) - a_i - b_t - U_{it}\right)}{f_{v_t}\left(v_{it}|X_{it}\right)} \\
& E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid U_{it}, a_i, b_t, X_{it}\right) f_{v_t}\left(v_{it} \mid U_{it}, a_i, b_t, X_{it}\right) dv_{it} \\
= \; & \int_{\alpha_0(X_{it}) - a_i - b_t - U_{it}}^{\alpha_1(X_{it}) - a_i - b_t - U_{it}} \frac{E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid U_{it}, a_i, b_t, X_{it}\right)}{f_{v_t}\left(v_{it}|X_{it}\right)} f_{v_t}\left(v_{it}|X_{it}\right) dv_{it} \\
= \; & E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid U_{it}, a_i, b_t, X_{it}\right) \int_{\alpha_0(X_{it}) - a_i - b_t - U_{it}}^{\alpha_1(X_{it}) - a_i - b_t - U_{it}} 1 dv_{it} \\
= \; & E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid U_{it}, a_i, b_t, X_{it}\right) \left[\alpha_1(X_{it}) - \alpha_0(X_{it})\right]
\end{aligned}
$$

and therefore

$$
\begin{aligned}
& E\left[I_{\tau it} D_{it} Y_{it} / f_{v_t}\left(V_{it}|X_{it}\right) | X_{it}\right] \\
= \; & E\left[E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid U_{it}, a_i, b_t, X_{it}\right) \left[\alpha_1(X_{it}) - \alpha_0(X_{it})\right] | X_{it} \right] \\
= \; & E\left(Y_{1it} + \widetilde{a}_i + \widetilde{b}_t \right| X_{it} \Big) \left[\alpha_1(X_{it}) - \alpha_0(X_{it})\right].
\end{aligned}
$$

Given the above result, the rest of the proof follows similarly as in the proof for Theorem 2.1. ■

Assumption 4.1 *Observations are i.i.d. across i.*

Assumption 4.2 *$E(h_{1i}|x_i)$, $E(h_{2i}|x_i)$, $E(g_{1i}|x_i)$, $E(g_{2i}|x_i)$, $f_x(x_i)$, $f_v(v_i)$, and $f_{xv}(x_i, v_i)$ are three times continuously differentiable in x, v. $E(h_{1i}|x_i)$, $E(h_{2i}|x_i)$, $E(g_{1i}|x_i)$, and $E(g_{2i}|x_i)$ are bounded. $E(h_{2i}|x_i)$ and $E(g_{2i}|x_i)$ are bounded away from zero. Second moments of h_{1i}, g_{1i}, h_{2i}, and g_{2i} exist.*

Assumption 4.3 *The kernel functions $K(v)$, $K(x)$, and $K(x, v)$ have supports that are convex and bounded on $\mathbb{R}^1$, $\mathbb{R}^k$, and $\mathbb{R}^{k+1}$ respectively. Each kernel function integrates to one over its* *support, is symmetric around zero, and has order p, i.e., for $K(x)$,*

$$\int_{\mathbb{R}^k} x_1^{l_1}...x_k^{l_k} K(x)dx = 0 \quad \textit{for } l_1 + ... + l_k < p,$$

$$\int_{\mathbb{R}^k} x_1^{l_1}...x_k^{l_k} K(x)dx \neq 0 \quad for\ some\ l_1 + ... + l_k = p,$$

where $l_1, ..., l_k$ are nonnegative integers and $\int K(x)^2 dx, \int \|x\| K(x)dx$ are finite. This similarly holds for $K(v)$ and $K(x, v)$.

Proof of Theorem 2.3. Under Assumptions 4.1, 4.2, 4.3, and the assumption on n and h, $\widehat{\mathrm{E}}\left(h_{1i}|\, x_i\right)$, $\widehat{\mathrm{E}}\left(h_{2i}|x_i\right)$, $\widehat{\mathrm{E}\left(g_{1i}|x_i\right)}$, $\widehat{\mathrm{E}}\left(g_{2i}|x_i\right)$ are uniformly consistent estimates of $\mathrm{E}(h_{1i}|x_i)$, $\mathrm{E}(h_{2i}|x_i)$, $\mathrm{E}(g_{1i}|x_i)$, $\mathrm{E}(g_{2i}|x_i)$, respectively (see, e.g., Theorem 2.2 in Li and Racine 2007). So Equation (9) is equal to

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathrm{E}\left(\left.\frac{D_iY_i}{f(v_i|x_i)}\right| x_i\right)}{\mathrm{E}\left(\left.\frac{D_i}{f(v_i|x_i)}\right| x_i\right)} - \frac{\mathrm{E}\left(\left.\frac{(1-D_i)Y_i}{f(v_i|x_i)}\right| x_i\right)}{\mathrm{E}\left(\left.\frac{1-D_i}{f(v_i|x_i)}\right| x_i\right)}\right] + o_p(1) = \frac{1}{n}\sum_{i=1}^{n}\psi\left(x_i\right) + o_p(1).$$

By Kolmogorov's law of large numbers, this converges to $\mathrm{E}[\psi\left(X\right)]$, which equals $\mathrm{E}(Y_1 - Y_0)$ by Theorem 2.1. ■

Assumption 4.4 *Let $r_1(x_i) = \frac{1}{E(\widetilde{g}_{1i}|x_i)}$, $s_{1i} = h_{1i}$, $r_2(x_i) = \frac{E(\widetilde{h}_{1i}|x_i)}{E(\widetilde{g}_{1i}|x_i)^2}$, $s_{2i} = g_{1i}$, $r_3(x_i) = \frac{1}{E(\widetilde{g}_{2i}|x_i)}$, $s_{3i} = h_{2i}$, $r_4(x_i) = \frac{E(\widetilde{h}_{2i}|x_i)}{E(\widetilde{g}_{2i}|x_i)^2}$, and $s_{4i} = g_{2i}$. Then for each $r_j(x_i)$ and s_{ji}, $j = 1, 2, 3, 4$, and f_x there exists some positive numbers M_1, M_2, M_3 such that*

$$|E(s_{ji}|x_i + e_x) - E(s_{ji}|x_i)| \leq M_1 \|e_x\|,$$

$$|r_j(x_i + e_x) - r_j(x_i)| \leq M_2 \|e_x\|,$$

$$|f_x(x_i + e_x) - f_x(x_i)| \leq M_3 \|e_x\|.$$

$E(s_{ji}|x_i)$, $r_j(x_i)$, and f_x are bounded and p-th order differentiable in x, and the p-th order derivatives are bounded.

Proof of Theorem 2.5. By $nh^{k+1} \to \infty$, for the same reasons as in Theorem 2.3, $\widehat{f}(v_i|x_i)$ is a uniformly consistent estimator for $f(v_i|x_i)$. By Assumption 4.2, those terms inside the estimate are bounded, so Equation (16) is equal to

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\widehat{\mathrm{E}}\left(\frac{D_iY_i}{f(v_i|x_i)}\middle|\, x_i\right)}{\widehat{\mathrm{E}}\left(\frac{D_i}{f(v_i|x_i)}\middle|\, x_i\right)}-\frac{\widehat{\mathrm{E}}\left(\frac{(1-D_i)Y_i}{f(v_i|x_i)}\middle|\, x_i\right)}{\widehat{\mathrm{E}}\left(\frac{1-D_i}{f(v_i|x_i)}\middle|\, x_i\right)}\right]+o_p(1),$$

where the first term converges to $\mathrm{E}(Y_1 - Y_0)$ by Theorem 2.3. ■

Assumption 4.5 *Let* $r_5(x_i) = \frac{1}{E(\widetilde{g}_{1i}|x_i)}$, $s_{5i} = \frac{D_iY_i}{f_{xv}(x_i,v_i)}$, $r_6(x_i) = r_5(x_i)$, $s_{6i} = \frac{D_iY_if_x(x_i)}{f^2_{xv}(x_i,v_i)}$, $r_7(x_i) = \frac{1}{E(\widetilde{g}_{2i}|x_i)}$, $s_{7i} = \frac{(1-D_i)Y_i}{f_{xv}(x_i,v_i)}$, $r_8(x_i) = r_7(x_i)$, *and* $s_{8i} = \frac{(1-D_i)Y_if_x(x_i)}{f^2_{xv}(x_i,v_i)}$. *Then for each* $r_j(x_i)$ *and* s_{ji}, $j = 5, 6, 7, 8$, f_x, f_{xv} *there exists some positive constants* M_1, M_2, M_3 *and* M_4 *such that*

$$|E(s_{ji}|x_i + e_x, v_i + e_v) - E(s_{ji}|x_i, v_i)| \le M_1 \left\|(e_x, e_v)\right\|,$$

$$|r_j(x_i + e_x) - r_j(x_i)| \le M_2 \left\|e_x\right\|,$$

$$|f_x(x_i + e_x) - f_x(x_i)| \le M_3 \left\|e_x\right\|,$$

$$|f_{xv}(x_i + e_x, v_i + e_v) - f_{xv}(x_i, v_i)| \le M_4 \left\|(e_x, e_v)\right\|.$$

$E(s_{ji}|x_i, v_i)$, $r_j(x_i)$, f_x, and f_{xv} are bounded and p-th order differentiable in x, and the p-th order derivatives are also bounded.

Proof of Theorem 2.7. Similar to the proof of Theorem 2.5, $\widehat{f_{v_t}}$ is a uniformly consistent estimator for f_{v_t}, and as a result equation (23) is equal to

$$\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}Y_{it}}{f_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})}}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})Y_{it}}{f_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})}{f_{v_t}(v_{it})}}+o_p(1). \tag{41}$$

By Assumptions 2.9, 2.10, 2.11, 2.12 from Lemma 4.5, and 4.6 in the Supplemental Appendix, we have that the probability limit of Equation (41) is

$$\frac{\mathrm{E}\left(\frac{D_{it}Y_{it}}{f_{v_t}(v_{it})}\right)}{\mathrm{E}\left(\frac{D_{it}}{f_{v_t}(v_{it})}\right)}-\frac{\mathrm{E}\left(\frac{(1-D_{it})Y_{it}}{f_{v_t}(v_{it})}\right)}{\mathrm{E}\left(\frac{(1-D_{it})}{f_{v_t}(v_{it})}\right)},$$

which equals $\mathrm{E}(Y_1 - Y_0)$ by Theorem 2.2. ■

Assumption 4.6 *Let* $s_{9it} = D_{it}Y_{it}$ *and* $s_{10it} = (1 - D_{it})\,Y_{it}$*. Then for each* s_{jit}*,* $j = 9, 10$ *and* f_{v_t}*, there exists some positive numbers* M_1 *and* M_2 *such that*

$$|E(s_{jit}|v_{it} + e_v) - E(s_{jit}|v_{it})| \leq M_1\,|e_v|\,,$$

$$|f_{v_t}(v_{it} + e_v) - f_{v_t}(v_{it})| \leq M_2\,|e_v|\,.$$

$E\left(s_{jit}|v_{it}\right),\ f_{v_t}$ *are bounded and p-th order differentiable in* v*, and their p-th order derivatives are also bounded.*

Table 1: Summary Statistics of the US Dataset

	MEAN	SD	LQ	MED	UQ
Competition	0.76	0.11	0.70	0.76	0.83
Innovation	5.53	9.98	0.22	1.59	5.77
Source-weighted Interest Rate	0.91	0.23	0.79	0.87	0.99

Note: MEAN = mean. SD = standard errors. LQ = 25% quantile (lower). MED = 50% quantile (median). UQ = 75% quantile (upper).

Table 2: Empirical Estimates in Various Cases

	Right Threshold	Left Threshold	Trim-ATE	No-Trim-ATE	Naive-ATE	ML-ATE
Case 1	25% (0.70)	75% (0.83)	−3.90 (0.78)	−4.25 (0.73)	−1.89 (0.38)	−1.85 (0.39)
Case 2	33% (0.72)	67% (0.80)	−3.27 (0.57)	−3.47 (0.63)	−1.67 (0.36)	−1.69 (0.37)
Case 3	10% (0.63)	90% (0.89)	−2.77 (1.17)	−2.75 (1.07)	−1.95 (0.55)	−4.40 (3.48)
Case 4	20% (0.68)	80% (0.85)	−4.25 (0.84)	−4.62 (0.83)	−2.22 (0.42)	−2.12 (0.43)
Case 5	30% (0.71)	70% (0.82)	−3.54 (0.63)	−3.95 (0.62)	−1.83 (0.36)	−1.81 (0.37)
Case 6	40% (0.74)	60% (0.79)	−2.49 (0.51)	−2.58 (0.59)	−1.18 (0.41)	−1.48 (0.39)

Notes: Right Threshold and Left Threshold are the $\underline{\alpha}$ and $\overline{\alpha}$ in Equation (30) respectively. The first value is the percentage of competition set for the thresholds, with corresponding value of competition in the parenthesis. Four different estimates are reported here, with standard errors in parenthesis. Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|T = 1) - E(Y_0|T = 0)$. ML-ATE is Heckman's selection MLE.

Table 3: Monte Carlo results matching the empirical data

	MEAN(−3.9)	SD	LQ	MED	UQ	RMSE	MAE	MDAE	%2SE
Panel A: Symmetric setting with normal errors									
Trim-ATE	−3.90	0.43	−4.19	−3.90	−3.61	0.43	0.34	0.00	1.00
No-Trim-ATE	−3.90	1.22	−4.67	−3.92	−3.12	1.22	0.95	0.02	1.00
Naive-ATE	−3.90	0.32	−4.11	−3.90	−3.68	0.32	0.25	0.00	1.00
ML-ATE	−3.90	0.30	−4.10	−3.90	−3.70	0.30	0.24	0.00	1.00
Panel B: Symmetric setting with uniform errors									
Trim-ATE	−3.90	0.38	−4.16	−3.90	−3.64	0.38	0.31	0.00	1.00
No-Trim-ATE	−3.90	0.38	−4.16	−3.90	−3.64	0.38	0.31	0.00	1.00
Naive-ATE	−3.90	0.38	−4.16	−3.90	−3.65	0.38	0.30	0.00	1.00
ML-ATE	−3.91	0.38	−4.17	−3.90	−3.65	0.38	0.30	0.00	1.00
Panel C: Asymmetric setting with normal errors									
Trim-ATE	−3.21	0.51	−3.55	−3.21	−2.87	0.86	0.73	0.69	0.95
No-Trim-ATE	−3.65	1.33	−4.50	−3.65	−2.81	1.35	1.06	0.25	0.77
Naive-ATE	−1.99	0.34	−2.21	−2.00	−1.77	1.94	1.91	1.90	0.15
ML-ATE	−1.98	0.35	−2.22	−1.98	−1.75	1.95	1.92	1.92	0.15
Panel D: Asymmetric setting with uniform errors									
Trim-ATE	−3.45	0.48	−3.77	−3.45	−3.12	0.66	0.54	0.45	0.99
No-Trim-ATE	−3.76	1.08	−4.47	−3.76	−3.06	1.09	0.86	0.14	0.85
Naive-ATE	−1.84	0.37	−2.08	−1.84	−1.59	2.10	2.06	2.06	0.09
ML-ATE	−2.07	0.39	−2.34	−2.07	−1.81	1.87	1.83	1.83	0.25

Note: True E(Y_1)−E(Y_0) = −3.9. Parameters set (θ_0, θ_1, θ_{01}, θ_{02}, θ_{11}, θ_{12}, θ_2) for the four MC in order are as follows: (6.94 3.04 5.64 8.44 6.71 4.87 1.06), (6.97 3.07 23.67 −24.30 22.62 25.72 1.07), (6.67 2.77 6.57 −2.91 4.51 −5.43 0.43), (7.41 3.51 8.43 −4.27 5.47 −1.47 0.55). Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|T = 1) - E(Y_0|T = 0)$. ML-ATE is Heckman's selection MLE. All statistics are for the simulation estimates. MEAN = mean. SD = standard errors. LQ = 25% quantile (lower). MED = 50% quantile (median). UQ = 75% quantile (upper). RMSE = root mean square errors. MAE = mean absolute errors. MDAE = median absolute errors. %2SE = percentage of simulations in which the true coefficient was within two estimated standard errors of the estimated coefficient.

Table 4: Robust check: Monte Carlo with normal errors

	Quadratic			Step		
	MEAN (≈ -3.9)	SD	RMSE	MEAN (-3.9)	SD	RMSE
Panel A: $\kappa_1 = 0.02$, Noise Ratio $= 0.19$						
Trim-ATE	-4.23	0.46	0.49	-3.19	0.41	0.82
No-Trim-ATE	-7.79	1.57	4.20	-3.31	1.04	1.20
Naive-ATE	-3.75	0.38	0.39	-3.14	0.34	0.83
ML-ATE	-3.67	0.73	0.76	-3.10	0.66	1.04
Control Function	-3.74	0.24	0.31	-1.38	0.20	2.52
Panel B: $\kappa_1 = 0.03$, Noise Ratio $= 0.28$						
Trim-ATE	-4.08	0.42	0.42	-2.85	0.42	1.11
No-Trim-ATE	-7.68	1.61	4.11	-2.96	1.11	1.46
Naive-ATE	-3.60	0.37	0.47	-2.79	0.34	1.16
ML-ATE	-3.54	0.74	0.82	-2.74	0.64	1.33
Control Function	-3.59	0.23	0.41	-1.33	0.21	2.58
Panel C: $\kappa_1 = 0.04$, Noise Ratio $= 0.36$						
Trim-ATE	-3.93	0.48	0.48	-2.55	0.42	1.41
No-Trim-ATE	-7.63	1.64	4.07	-2.66	1.09	1.65
Naive-ATE	-3.40	0.38	0.62	-2.45	0.34	1.49
ML-ATE	-3.33	0.66	0.87	-2.42	0.59	1.60
Control Function	-3.40	0.26	0.59	-1.26	0.20	2.62

Note: True mean value is -3.9. Noise ratio is defined as the ratio of standard deviation of c_e to the standard deviation of $c*$. The first three and last three columns are the results when the true response forms are quadratic and step function respectively. Five different estimators are reported here. Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|T = 1) - E(Y_0|T = 0)$. ML-ATE is Heckman's selection MLE. Control function approach is defined as in the paper. MEAN = mean. SD = standard errors. RMSE = root mean square errors.

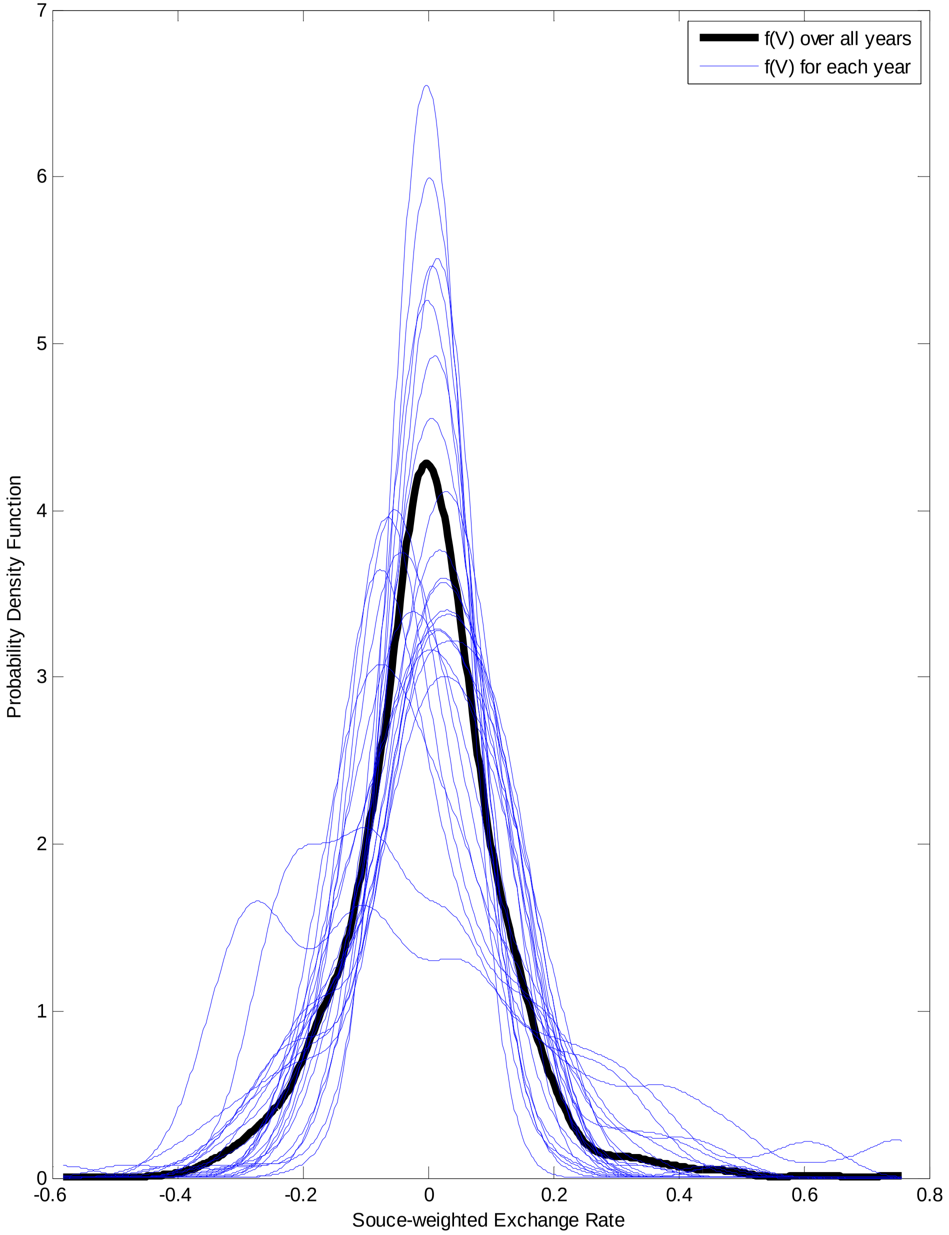

Figure 1: Distribution of V

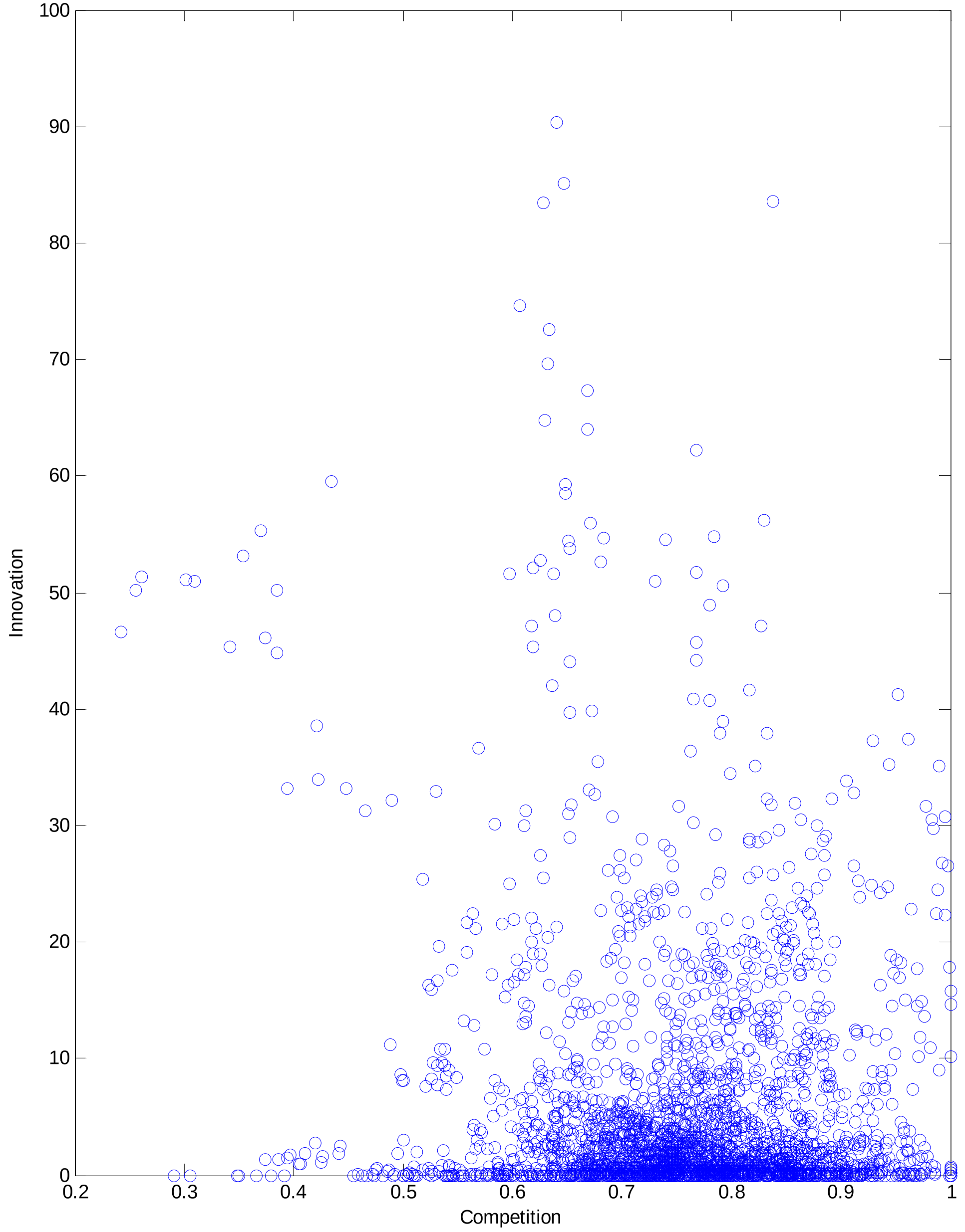

Figure 2: Competition and Innovation