Nonparametric Identification and Estimation of the Extended Roy Model

Byoung G. Park[‡]
SUNY Albany

October 2013

Abstract

This paper proposes a new identification and estimation method for the extended Roy model, in which the agents maximize their utility rather than just outcome. The identification results substantially relax conventional functional form restrictions. No functional form restriction is imposed on the distribution of the potential outcomes. Hence it allows for nonseparable functional forms and/or disturbances of an arbitrary dimension. The utility functions are allowed to be nonlinear so that it can accommodate important features of utility functions such as the concavity of the utility functions. The identification method does not require the instrument to have a large support. I show that (i) when the instrument is continuous, possibly having a bounded support, the model is point-identified on a certain identification region, and that (ii) when only discrete instruments are available, sharp bounds for the model are obtained. The key assumption of the method is the monotonicity of the selection with respect to the instrument. Based on the identification result, I propose a nonparametric estimation procedure that builds upon a simulation-based method proposed by Dette et al. (2006). The estimator is easy to implement in practice because it only uses a closed form formula and straightforward simulations. I show that the estimator possesses a standard nonparametric rate of convergence, and examine its efficacy in finite samples by Monte Carlo simulations.

Keywords Roy model, treatment effect, nonparametric identification, nonparametric estimation

*Department of Economics, SUNY Albany, Email : bpark2@albany.edu

†This paper is a chapter of my Ph.D. dissertation at Yale University. I thank my advisors, Yuichi Kitamura, Edward Vytlacil, and Donald Andrews for their guidance and support. I also benefited from discussions with Joseph Altonji, Xiaohong Chen, Sukjin Han, James Heckman, Stefan Holderlein, Hidehiko Ichimura, Sung Jae Jun, Hiroaki Kaido, Jihyung Lee, Sokbae Lee, Charles Manski, Peter Phillips, Christopher Udry, and Yoon-Jae Whang. I also thank Tavneet Suri for suggesting the data set. Finanacial supports from Cowles foundation and Korea Foundation of Advanced Studies are gratefully acknowledged.

1 Introduction

Self-selection has been at the core of econometrics and empirical economics for decades. The Roy model is a structural model of self-selection that has been used to control for self-selection bias in a large class of applications. Examples include the choice of schooling (Willis and Rosen, 1979), immigration (Borjas, 1987), labor union status (Lee, 1978).

Also, the Roy model provides a useful framework to estimate heterogeneous treatment effects. The Roy model shares the potential outcome framework in common with the Rubin causal model (Holland (1986)), which is the standard model for causal inference. Thus it is straightforward to apply the Roy model to causal inference. See Heckman and Vytlacil (2005) and Eisenhauer et al. (2011)for detailed discussion.

In this paper, I propose a new identification and estimation method for the extended Roy model. Consider a binary choice from two states. Let $d \in \{0, 1\}$ index the states. Each agent is endowed with a pair of potential outcomes (Y_0, Y_1) associated with the respective state. However, the researcher observes the realized outcome Y only, which is given by

$$Y = DY_1 + (1 - D)Y_0$$

where D is the selection indicator.

Suppose that there are observed vectors X and Z. I assume

$$(Y_0, Y_1)|X, Z \sim F_{Y_0Y_1|X} \tag{1}$$

where $F_{Y_0Y_1|X}$ is an unknown conditional distribution function, and

$$D = 1\{\mathcal{U}_1(Y_1, X, Z) > \mathcal{U}_0(Y_0, X, Z)\} \tag{2}$$

where $\mathcal{U}_0$ and $\mathcal{U}_1$ are unknown utility functions, and $1\{\cdot\}$ is the indicator function. I will refer to (1) and (2) as the outcome equation and the selection equation, respectively.

The outcome equation specifies the joint distribution of the potential outcomes conditional on observed variables. The only restriction imposed in (1) is that the conditional joint distribution of the potential outcomes given X and Z is independent of Z. That is, I assume that Z is independent of the potential outcomes conditional on X almost surely. It is a standard assumption imposed on the instrumental variables. In this paper, I use Z as the instrumental variables for the endogenous variable D.

The outcome equation (1) relaxes conventional functional form restrictions. The standard formulation of the outcome equation in the literature is an additively separable form. For $d \in \{0, 1\}$,

$$Y_d = m_d(X) + \varepsilon_d \tag{3}$$

where m_d is an unknown function, ε_d is an unobservable disturbance, and $X \perp (\varepsilon_0, \varepsilon_1)$. (E.g., Heckman (1990), Maurel and D'Haultfoeuille (2011), and French and Taber (2011)). In additively separable models, the effect of X on Y_d is assumed to be homogeneous across agents after conditioning on observables, which is hard to justify with economic theory in many applications. Furthermore, due to the independence assumption, the covariance between Y_0 and Y_1 is not affected by X, which can be too restrictive. Equation (1) relaxes these functional form assumptions and enables a flexible specification of the effect of X. Furthermore, (1) does not restrict the dimension of unobserved disturbances so that it allows for general unobserved heterogeneity.

The selection equation (2) is also very general. It extends the Roy model studied in Heckman and

Honore (1990) to accommodate utility functions. In their paper, the agents are assumed to choose the state with the higher potential outcome. By assuming that the agents maximize their utility rather than the outcome, it is possible to consider utility components other than the outcome such as costs of choosing each state or nonpecuniary preferences. In particular, it can accommodate important features of utility functions such as concavity and/or nonadditive utility factors. While there have been many papers allowing for utility components other than the outcome (Bayer et al. (2011), Maurel and D'Haultfoeuille (2011), and French and Taber (2011)), this the first paper that allows the potential outcome to enter the utility function nonlinearly and nonadditively.

However, it should be noted that the selection equation in (2) assumes that unobserved heterogeneity in the selection comes only from the potential outcomes. For instance, it rules out the agents' imperfect expectation of the potential outcome or unobserved heterogeneity in preferences. ? refer to such a model as the extended Roy model to distinguish from the generalized model in which additional unobserved heterogeneity in the selection is allowed. In general, the additional unobserved heterogeneity in the generalized Roy model complicates the identification of the joint distribution of the potential outcomes because there are always two sources of unobserved heterogeneity: one is the unrealized potential outcome, and the other is the unobserved heterogeneity in the selection equation. Even under more assumptions such as large support assumptions, separating the two unobserved heterogeneities and identifying the joint distribution is not an easy task, which is discussed in a companion paper (Park (2012)). Since this paper aims to identify the joint distribution of (Y_0, Y_1) under few functional form assumptions, it assumes the extended Roy model.

The parameters of interest are the joint distribution of the potential outcomes, $F_{Y_0Y_1|X}$, and the utility functions. The two parameters are nonparametrically identified under mild conditions. One of the main contributions of this paper is to identify the joint distribution of the potential outcomes. The joint distribution, as opposed to the marginal distribution, of the potential outcomes in the extended or generalized Roy model is not identified in the literature.[1] But the identification of the joint distribution of the potential outcomes is of great important in treatment effect analysis. If we think of choosing state 1 as participating in a treatment and state 0 as not, $Y_1 - Y_0$ can be thought of as the causal effect of the treatment on the outcome. Idnetification of two marginal distributions of Y_0 and Y_1 is not enough to identify the distribution of $Y_1 - Y_0$. So the treatment analysis has to detour this problem. An example of this problem is the quantile treatment effects (QTE). The QTE is widely used to capture the heterogeneity of the treatment effects. A natural way to define the QTE is $Q_{Y_1-Y_0|X}(\tau|x)$, where $\tau \in (0,1)$ and $Q_{A|B}(\tau|b)$ represents the τ-th quantile of A conditional on B being b for generic random variables A and B. However, due to the lack of the identification of the joint distribution, the QTE is defined as $Q_{Y_1|X}(\tau|x) - Q_{Y_0|X}(\tau|x)$ in Chernozhukov and Hansen (2005). In general, the two parameters are different. Hence, in a rigorous sense, the commonly used QTE can be misleading. The identification of the joint distribution presented in this paper enables researchers to use the proper QTE.

Identification of utility functions without linearity assumption is another important contribution of this paper to the literature. In many empirical applications, the causal effect of a treatment on the welfare is as important as the treatment effect on the outcome. (Eisenhauer et al. (2011)) Using the utility functions in (2), we can define the causal effect of a treatment on the welfare by $\mathcal{U}_1(Y_1, X, Z) - \mathcal{U}_0(Y_0, X, Z)$. Without linearity of utility functions in the potential outcomes, the welfare effect can not be expressed in terms of the usual treatment effect. Nonlinearity of utliity functions is important to understand welfare implication

[1] In Heckman and Honore (1990), the joint distribution is identified in the original Roy model. But it does not extend to the extended or generalized Roy model. Vijverberg (1993) tries to infer the correlation coefficient between Y_0 and Y_1 using the positive semi-definiteness of the covariance matrix.

of a treatment on heterogeneous agents. For example, consider the choice of attending college. Those who are from rich families may have little marginal utility from the wage increment by college degree, and care more about other non-pecuniary benefits such as pride, building social network, etc. In such cases, they may attend college even though the wage increase by college attendance is negative. That is, the treatment effect (on the outcome) can be negative even when the welfare effect is positive. On the other hand, it is also possible that the poor do not attend college due to financial constraints even though the wage increase is substantially positive. This example shows that the welfare effect of a treatment can be totally different from the treatment effect, and also the direction of the difference can be heterogeneous.

The key assumption of this paper is the uniformity assumption introduced in Heckman and Vytlacil (2005); Heckman et al. (2006) or the monotonicity assumption in Imbens and Angrist (1994) and Vytlacil (2003). These assumptions imply that the response to a change in the instrumental variable, Z, should be in the same direction for all the agents. It rules out, for example, the case in which the change in Z makes state 0 more attractive relative to state 1 for some agents, while it makes state 1 more attractive relative to state 0 for the others. A change in Z should be either in favor of state 0 or state 1 for all the agents. It will be discussed in more detail in the body of this paper. To explain the intuition of the identification method, suppose that when Z increases, state 0 becomes more attractive and some agents will change their selection from state 1 to state 0. The researcher observes the changes in the distribution of observed outcomes using the variation in Z. I exploit the idea that the inflow into state 0 and the outflow out of state 1 sum to zero. Matching the inflows and outflows of two states gives the key equation used for the identification.

My identification strategy is totally different from the prevalent identification method using the so-called "identification at infinity" argument. The identification at infinity argument relies on the idea that the selection problem disappears when a state is chosen with probability one. To make the choice probability arbitrarily close to one, a sufficiently large variation in instrumental variables should be assumed. It bears two potential problem due to its nature of utilizing events in the limit. First, it requires a large support assumption on instruments, which often fails to hold in practice. Second, the estimation based on the identification-at-infinity is "irregular" in the sense that it relies on observations near infinity. (Andrews and Schafgans (1998); Khan and Tamer (2010))

This paper addresses the two central problems of the identification-at-infinity method. First, the identification method proposed in this paper does not require any support condition on the instrumental variable. Instead, it gives different identification results for different types of instruments. When the instrument is continuous, possibly having a bounded support, the model is point-identified. When only discrete instruments are available, bounds for the model are obtained and shown to be sharp. Second, the estimation method based on the new identification strategy possesses a standard nonparametric rate of convergence. The estimation procedure using the identification-at-infinity argument is irregular, potentially resulting in a slower rate of convergence.

I propose a simulation-based estimator for the model with a continuous instrument. The method proceeds in two steps. In the first stage, the distribution function of observed outcomes and its derivatives with respect to the instrument are estimated by a local linear estimator. In the second stage, the parameters of interest are estimated by solving equations obtained from the first stage estimates. I adopt and extend the simulation-based method proposed by Dette et al. (2006); Dette and Scheder (2006); Dette and Volgushev (2008) to solve the equations under the monotonicity assumption. The estimator is easy to implement because it only uses a closed-form formula and straightforward simulations. I derive its rate of convergence and conduct a set of Monte Carlo simulations to examine its efficacy in finite samples.

The remainder of the paper is organized as follows. Section 2 presents the identification results. Subsections 2.1 and 2.2 provide a point-identification result for continuous instruments and a partial identification result for discrete instruments, respectively. In Section 3, I develop a nonparametric estimator based on the identification results and derive the convergence rate of the estimator. Section 4 presents Monte Carlo simulation results to check the performance of the estimator in finite samples and compares it with benchmark estimators. Section 6 summarizes the results and provides some future research directions. Appendix A proves an asymptotic theory for the simulation-based estimator used in Section 3 under high-level assumptions. Appendix B provides a high-level conditions for stochastic equicontinuity. Tables are deferred to the end of the paper.

2 Identification

Suppose that $Y_d \in \mathbb{R}$ for each $d \in \{0,1\}$ and let $X \in \mathcal{X} \subset \mathbb{R}^{m_X}$ and $Z \in \mathcal{Z} \subset \mathbb{R}^{m_Z}$. I call X and Z covariates and instruments, respectively. Throughout this paper, I assume that $m_Z = 1$ for simplicity of notation. Generalization to the case of $m_Z > 1$ is straightforward.

We need a normalization because utility functions are identified only up to a monotone transformation. Assume that the utility function $\mathcal{U}_1$ in equation (2) is strictly increasing in Y_1 for any X and Z almost surely. Then there exists the inverse of $\mathcal{U}_1$ with respect to Y_1 with X and Z being fixed, denoted by $\mathcal{U}_1^{-1}$. Let $h = \mathcal{U}_1^{-1} \circ \mathcal{U}_0$. Then, the selection equation (2) can be written as follows:

$$D = 1\{Y_1 > h(Y_0, X, Z)\} \tag{4}$$

Now h can be regarded as a utility function normalized in units of Y_1.

The model can be summarized by the following three equations.

$$(Y_0, Y_1)|X, Z \sim F_{Y_0 Y_1|X}$$

$$D = 1\{Y_1 > h(Y_0, X, Z)\}$$

$$Y = DY_1 + (1 - D)Y_0$$

The parameters of interest are the joint distribution of the potential outcomes, $F_{Y_0 Y_1|X}$, and the utility function, h. For identification, I assume that the researcher observes the joint distribution of the state choice D, realized outcome Y, covariates X, and instrument Z.

Now I fix notation. Throughout this paper, random variables are denoted by upper case letters and their realizations by lower case letters. Let G_d for $d \in \{0,1\}$ be the conditional distribution function of observed outcomes in state d:

$$G_d(y|x,z) = \Pr(Y \leq y, D = d|X = x, Z = z)$$

Note that G_d is identified by definition because it is a distribution function of observed variables. When it is obvious, the conditioning variables are abbreviated. For example, $\Pr(\cdot|x,z)$ abbreviates $\Pr(\cdot|X = x, Z = z)$. Let $supp(\cdot)$ represent the support of a random variable. Also, $supp(\cdot|\cdot)$ signifies the conditional support of a random variable conditional on another random variable.

In the following subsections, I present two identification results. One is a point-identification result when the instrument is continuous, and the other is a partial identification result when the instrument is discrete.

2.1 Continuous Instrumental Variable

First I consider a continuous instrument case. Throughout this section, I fix X at x and the results are conditional on x. The following assumptions are used:

Assumption 2.1. *The following assumptions hold conditional on $X = x$ with probability one:*

(a) (Y_0, Y_1) is independent of the instrument Z.

(b) The distribution of (Y_0, Y_1) has a density $f_{Y_0Y_1|X}$ and a support equal to $\mathbb{R}^2$.

(c) h is strictly increasing in Y_0 for any Z.

(d) h is strictly increasing in Z for any Y_0.

(e) The distribution of Z has a density $f_{Z|X}$.

(f) h is differentiable with respect to Z for any Y_0.

Assumption 2.1(a) is the standard assumption on the instrumental variables. Assumption 2.1(b) guarantees that observed distributions have well-defined densities and allows one to ignore the indecisive case of $Y_1 = h(Y_0, x, z)$. This assumption can be weakened, for example, to incorporate mixed continuous and discrete distribution by specifying a tie breaking rule. The support is mild. To identify parameters locally at a certain point, the support assumption can be relaxed. In the case of bounded support, Bayer et al. (2011) presents an identification strategy that utilizes the boundedness of the support. Assumption 2.1(c) is an axiom of utility and naturally satisfied in most applications.

Assumption 2.1(d) imposes a monotonicity of selection in the instrument, which is the key for the identification. It is the Roy model's version of the monotonicity assumption in Imbens and Angrist (1994) and Vytlacil (2003). Intuitively speaking, it implies that when the instrument increases, the utility for state 0 increases more or decreases less than that for state 1. A more general assumption is the so-called uniformity assumption. The uniformity assumption in this model is stated as follows.

Assumption (Uniformity)**.** *Let x and distinct z, z' be given. Suppose that either of the following statements is true*

(i) $h(Y_0, x, z) \geq h(Y_0, x, z')$ for any Y_0 almost surely.

(ii) $h(Y_0, x, z) \leq h(Y_0, x, z')$ for any Y_0 almost surely.

Note that under the uniformity assumption, it is allowed that h is increasing in Z at one value of Z, but decreasing in Z at a different value of Z. The direction is testable by testing the direction of the propensity score, $\Pr(D = 1|X, Z)$, with respect to Z. Though the uniformity assumption is more general, I will assume the monotonicity assumption in the rest of the paper for the convenience of proofs. It is easy to replace the monotonicity assumption with the uniformity assumption.

Assumption 2.1(f) guarantees the differentiability of G_d with respect to z. I denote the derivative of G_d with respect to z by

$$G_d^{(1)}(y_d|x, z) = \frac{\partial}{\partial z} G_d(y_d|x, z)$$

Let $h^{-1}(y_1, x, z)$ be the inverse function of $h(y_0, x, z)$ with respect to y_0 at (x, z). That is, $y_1 = h(y_0, x, z)$ is equivalent to $y_0 = h^{-1}(y_1, x, z)$. Assumption 2.1(c) guarantees that h^{-1} is uniquely defined.

The following is an intermediate result, which is used for the identification of underlying parameters.

Lemma 2.1. *Under Assumption 2.1, if $y_1 = h(y_0, x, z)$, then*

$$G_0(y_0|x, z) + G_1(y_1|x, z) = F_{Y_0Y_1|X}(y_0, y_1|x)$$

and

$$G_0^{(1)}(y_0|x,z) + G_1^{(1)}(y_1|x,z) = 0$$

.

Proof. Using the selection equation and Assumption 2.1(a), the observed distribution functions, G_0 and G_1, can be written as

$$\begin{aligned} G_0(y_0|x,z) &= \Pr\left(Y \leq y_0, D = 0|x,z\right) \\ &= \Pr\left(Y_0 \leq y_0, Y_1 \leq h(Y_0,x,z)|x\right) \end{aligned} \tag{5}$$

and

$$\begin{aligned} G_1(y_1|x,z) &= \Pr\left(Y \leq y_1, D = 1|x,z\right) \\ &= \Pr\left(Y_1 \leq y_1, Y_1 > h(Y_0,x,z)|x\right) \end{aligned} \tag{6}$$

Note that conditioning on Z disappears because Z affects G_0 and G_1 only via h by the independence assumption between Z and (Y_0, Y_1) conditional on X.

By the monotonicity of h in Y_0, $Y_0 \leq y_0$ and $Y_1 \leq h(Y_0,x,z)$ imply

$$\begin{aligned} Y_1 &\leq h(Y_0,x,z) \\ &\leq h(y_0,x,z) \\ &= y_1 \end{aligned}$$

Thus,

$$\begin{aligned} G_0(y_0|x,z) &= \Pr\left(Y_0 \leq y_0, Y_1 \leq h(Y_0,x,z)|x,z\right) \\ &= \Pr\left(Y_0 \leq y_0, Y_1 \leq y_1, Y_1 \leq h(Y_0,x,z)|x\right) \end{aligned}$$

Similarly,

$$\begin{aligned} G_1(y_1|x,z) &= \Pr\left(Y_1 \leq y_1, Y_1 > h(Y_0,x,z)|x\right) \\ &= \Pr\left(Y_0 \leq y_0, Y_1 \leq y_1, Y_1 > h(Y_0,x,z)|x\right) \end{aligned}$$

Therefore,

$$\begin{aligned} G_0(y_0|x,z) + G_1(y_1|x,z) &= \Pr\left(Y_0 \leq y_0, Y_1 \leq y_1, Y_1 \leq h(Y_0,x,z)|x\right) \\ &\quad + \Pr\left(Y_0 \leq y_0, Y_1 \leq y_1, Y_1 > h(Y_0,x,z)|x\right) \\ &= \Pr(Y_0 \leq y_0, Y_1 \leq y_1|x) \\ &= F_{Y_0Y_1|X}(y_0,y_1|x) \end{aligned}$$

which proves the first statement of the lemma.

The probability (5) can be expressed as an integral as follows:

$$G_0(y_0|x,z) = \int_{-\infty}^{y_0} \int_{-\infty}^{h(v_0,x,z)} f_{Y_0Y_1|X}(v_0,v_1|x)dv_1dv_0$$

Differentiating this with respect to z yields

$$G_0^{(1)}(y_0|x,z) = \int_{-\infty}^{y_0} \left(\frac{\partial h(v,x,z)}{\partial z}\right) f_{Y_0Y_1|X}(v,h(v,x,z)|x)dv \tag{7}$$

Similarly, we can write $G_1(y_1|x,z)$ as

$$G_1(y_1|x,z) = \int_{-\infty}^{h^{-1}(y_1,x,z)} \int_{h(v_0,x,z)}^{y_1} f_{Y_0Y_1|X}(v_0,v_1|x)dv_0dv_1$$

Now differentiate this with respect to z to obtain

$$\begin{aligned} G_1^{(1)}(y_1|x,z) &= \left(\frac{\partial h^{-1}(y_1,x,z)}{\partial z}\right) \int_{h(h^{-1}(y_1,x,z),x,z)}^{y_1} f_{Y_0Y_1|X}(h^{-1}(y_1,x,z),v_1|x)dv_1 \\ &\quad + \int_{-\infty}^{h^{-1}(y_1,x,z)} \left(-\frac{\partial h(v_0,x,z)}{\partial z}\right) f_{Y_0Y_1|X}(v_0,h(v_0,x,z)|x)dv_0 \end{aligned}$$

Observe that the range of the first integral is simply a point $\{Y_1 = y_1\}$ because $h(h^{-1}(y_1,x,z),x,z) = y_1$. Since the set $\{Y_1 = y_1\}$ is of measure zero with respect to the Lebesgue measure, any integral over this set with respect to a measure that is absolutely continuous with respect to the Lebesgue measure yields zero under Assumption 2.1(b). Hence,

$$G_1^{(1)}(y_1|x,z) = -\int_{-\infty}^{h^{-1}(y_1,x,z)} \left(\frac{\partial h(v,x,z)}{\partial z}\right) f_{Y_0Y_1|X}(v,h(v,x,z)|x)dv \tag{8}$$

Combining equations (7) and (8), we have

$$G_0^{(1)}(y_0|x,z) + G_1^{(1)}(y_1|x,z) = \int_{h^{-1}(y_1,x,z)}^{y_0} \left(\frac{\partial h(v,x,z)}{\partial z}\right) f_{Y_0Y_1|X}(v,h(v,x,z)|x)dv \tag{9}$$

It is now clear that if $y_1 = h(y_0,x,z)$,

$$G_0^{(1)}(y_0|x,z) + G_1^{(1)}(y_1|x,z) = 0$$

□

Figure I illustrates the first result of the lemma. Let x and z be given and put Y_0 and Y_1 on the vertical and horizontal axes, respectively. The utility function is a upward sloping curve on the plane. Any point below the curve satisfies $Y_1 < h(Y_0,x,z)$. Hence the agents whose potential outcomes lie below the curve will choose state 0. Furthermore, Y_0 will be observed for the agents below the curve. For an arbitrary y_0, the probability measure on the shaded area represents $G_0(y_0|x,z)$. By the same argument, the probability measure on the shaded area above the curve gives G_1. In particular, when $y_1 = h(y_0,x,z)$, the two areas jointly make a rectangle. Therefore, the probability measure on the rectangle equals to the sum of the two observed probability measures as in the lemma.

Figure I

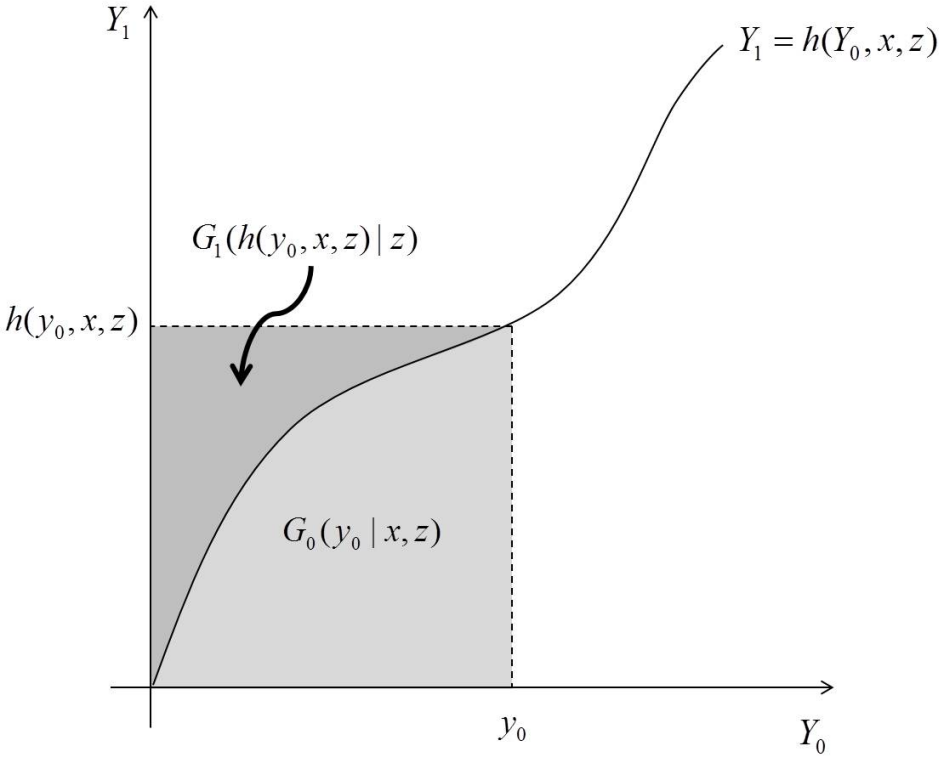


Figure II

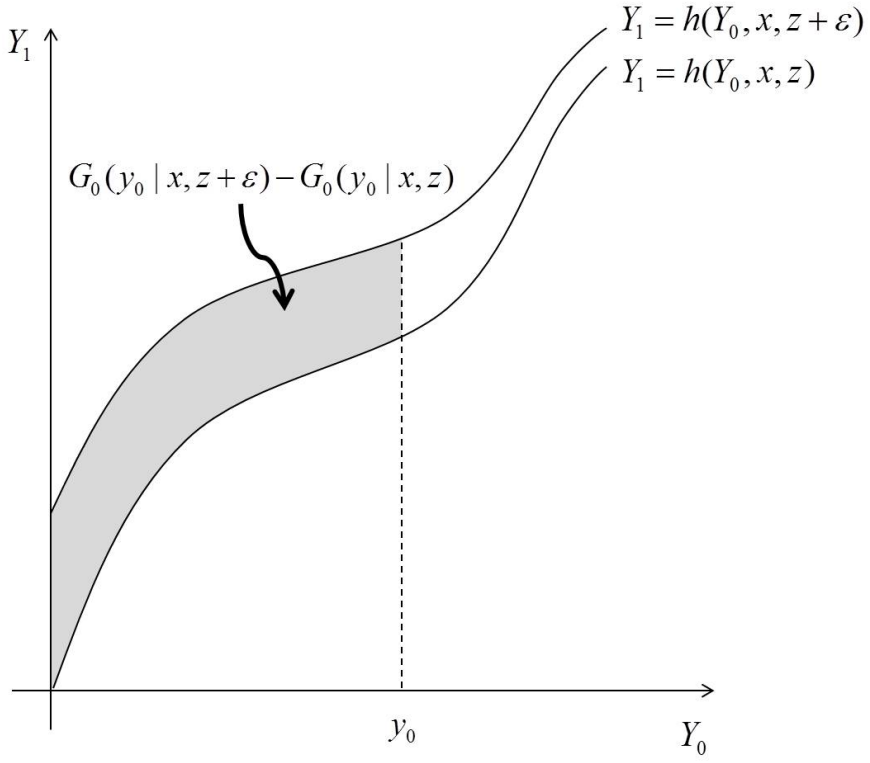
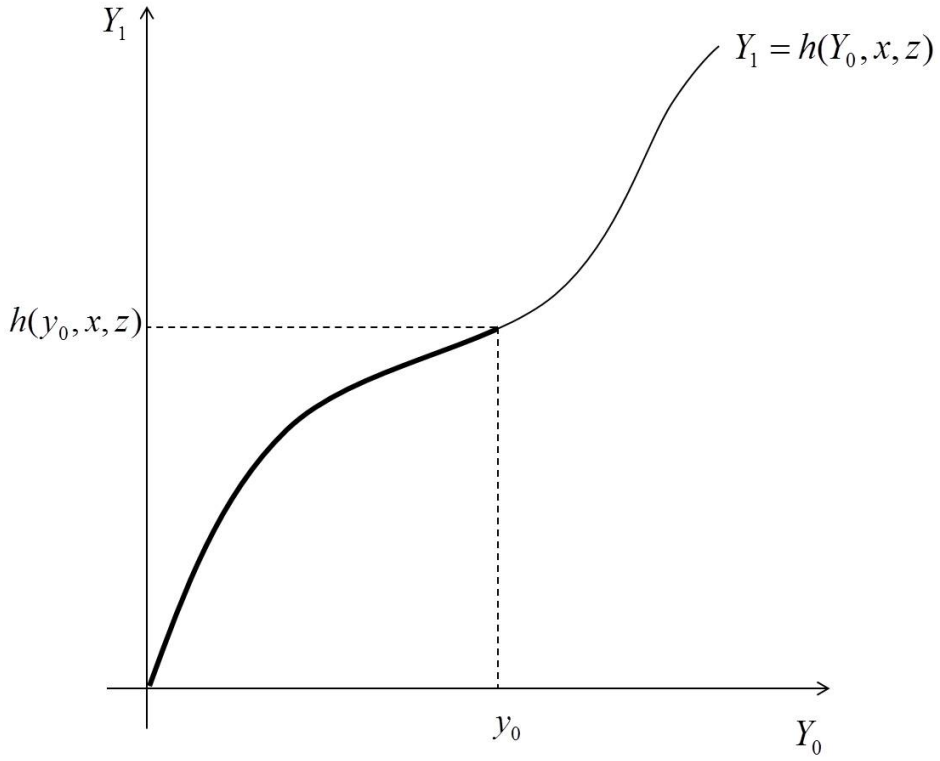


Figure III



Now consider an increase in Z by ε for an arbitrary positive ε. Since h is strictly increasing in z, the utility function shifts upward as in Figure II. An increase in Z will increase the utility of choosing state 0 and so more agents will choose state 0. With y_0 being fixed, $G_0(y_0|x, z+\varepsilon) - G_0(y_0|x, z)$ captures the increase in the probability measure of the agents who choose state 0 and receive an outcome less than or equal to y_0. The shaded area in Figure II represent the subpopulation who changed the choice from state 1 to state 0 due to the increase in Z. If the increase in Z is infinitesimal, $G_0^{(1)}(y_0|x, z)$ captures the infinitesimal increase in the distribution function G_0. The shaded area in Figure II will shrink to a curve that is depicted as a bold line in Figure III. Note that any point on the bold line satisfies $Y_1 = h(Y_0, x, z)$, which implies that the utilities for both states are the same. Therefore $G_0^{(1)}(y_0|x, z)$ can be viewed as the density of the event that Y_0 is less than or equal to y_0 and two states are indifferent.

Similarly, $G_1^{(1)}(y_1|x, z)$ is the density of the event that Y_1 is less than or equal to y_1 and two states are indifferent. Note that $G_1^{(1)}$ is negative because G_1 decreases as Z increases. As Figure III shows, $G_0^{(1)}(y_0|x, z)$ and $G_1^{(1)}(h(y_0, x, z)|x, z)$ capture the same density, though the signs are different. Based on this idea, the following theorem proves the identification of h.

Theorem 2.1. *Under Assumption 2.1, h is identified for any $(y_0, x, z) \in supp(Y_0, X, Z)$.*

Proof. Let $(y_0, x, z) \in supp(Y_0, X, Z)$ be given. By Lemma 2.1, we know that if $y_1 = h(y_0, x, z)$, then

$$G_0^{(1)}(y_0|x, z) + G_1^{(1)}(y_1|x, z) = 0 \tag{10}$$

I prove that (10) holds only if $y_1 = h(y_0, x, z)$. Pick any $y_1' > y_1$. From (8), we have

$$G_1^{(1)}(y_1'|x,z) - G_1^{(1)}(y_1|x,z) = -\int_{h^{-1}(y_1,x,z)}^{h^{-1}(y_1',x,z)} \left(\frac{\partial h(v,x,z)}{\partial z}\right) f_{Y_0Y_1|X}(v, h(v,x,z)|x)dv$$

By Assumption 2.1, $\frac{\partial h}{\partial z} > 0$ and $f_{Y_0Y_1|X} > 0$. Thus, the integrand is strictly positive on a neighborhood of $y_1 = h(y_0, x, z)$. Since $h^{-1}(y_1', x, z) > h^{-1}(y_1, x, z)$, we have

$$G_1^{(1)}(y_1'|x,z) < G_1^{(1)}(y_1|x,z)$$

By exactly the same reasoning, for $y_1'' < y_1$,

$$G_1^{(1)}(y_1''|x,z) > G_1^{(1)}(y_1|x,z)$$

Therefore equation (10) has a unique solution, and we already know that the solution is $h(y_0, x, z)$. Since $G_0^{(1)}$ and $G_1^{(1)}$ are identified, $h(y_0, x, z)$ is also identified. □

Equation (10) is the key equation for identification of $h(y_0, x, z)$. As I mentioned previously, $G_0^{(1)}$ captures the density of inflow toward state 0 caused by an increase in Z and $G_1^{(1)}$ the outflow out of state 1. Since there are only two choice options, the inflow and the outflow must sum to zero. Figure III shows that the zero-sum condition holds when $y_1 = h(y_0, x, z)$. The theorem states that for each y_0 there is unique y_1 that satisfies the zero-sum condition in (10) and the solution is $h(y_0, x, z)$.

Now I consider the identification of $F_{Y_0Y_1|X}$. Though I do not require a large support assumption on Z, the support of Z does affect the identification region. Let $\mathcal{Z}(x)$ be the support of Z conditional on $X = x$. Also define

$$\mathcal{H}(x) = \{(y_0, y_1) \in supp(Y_0, Y_1|x) : y_1 = h(y_0, x, z) \text{ for some } z \in \mathcal{Z}(x)\}$$

Intuitively speaking, an agent whose potential outcomes lie in $\mathcal{H}(x)$ is indifferent between the two states at a certain value of Z in $\mathcal{Z}(x)$. In other words, an agent outside $\mathcal{H}(x)$ chooses one state no matter what the value of Z is. Since identification relies on the changes in the choice caused by Z, such agents outside $\mathcal{H}(x)$ do not provide any information for identification. Hence, $\mathcal{H}(x)$ is the identification region for $F_{Y_0Y_1|X}$.

Theorem 2.2. *Under Assumption 2.1, $F_{Y_0Y_1|X}(y_0, y_1|x)$ is identified for any $x \in \mathcal{X}$ and $(y_0, y_1) \in \mathcal{H}(x)$.*

Proof. Suppose that $(y_0, y_1) \in \mathcal{H}(x)$ is given. Then, there exists $\zeta \in \mathcal{Z}$ such that $y_1 = h(y_0, x, \zeta)$. By Lemma 2.1,

$$F_{Y_0Y_1|X}(y_0, y_1|x) = G_0(y_0|x, \zeta) + G_1(y_1|x, \zeta)$$

Since G_0 and G_1 are observed, only unknown is ζ. By Theorem 2.1, h is identified at (y_0, x, z) for any $z \in \mathcal{Z}(x)$. Hence ζ is also identified. Therefore $F_{Y_0Y_1|X}$ is identified. □

Consider a large support assumption that $\mathcal{H}(x)$ includes the support of (Y_0, Y_1) conditional on $X = x$. Under the large support assumption, $F_{Y_0Y_1|X}$ is point identified everywhere on its support.

Even if the large support assumption fails, useful treatment effect parameters can still be identified. Let $Q_{Y_1-Y_0|X}(\tau|x)$ be the τth quantile of $Y_1 - Y_0$ conditional on $X = x$ for $\tau \in [0, 1]$. The following result gives the conditions under which $Q_{Y_1-Y_0|X}(\tau|x)$ is identified with a bounded support of the instrument.

Assumption 2.2. *Suppose that $\mathcal{Z}(x) = [\underline{z}, \bar{z}]$ where $-\infty < \underline{z} < \bar{z} < \infty$. Assume that the following inequality holds almost surely.*

$$h(Y_0, x, \underline{z}) \leq Y_0 + Q_{Y_1 - Y_0|X}(\tau|x) \leq h(Y_0, x, \bar{z})$$

Corollary 2.1. *Under Assumption 2.1 and 2.2, $Q_{Y_1-Y_0|X}(\tau|x)$ is identified.*

Proof. It suffices to show that $\Pr(Y_1 - Y_0 \leq Q_{Y_1-Y_0|X}(\tau|x)|x)$ is identified. Note that

$$\begin{aligned} \Pr(Y_1 - Y_0 \leq Q_{Y_1-Y_0|X}(\tau|x)|x) &= \Pr(Y_1 - Y_0 \leq Q_{Y_1-Y_0|X}(\tau|x), Y_1 \leq h(Y_0, x, \underline{z})|x) \\ &\quad + \Pr(Y_1 - Y_0 \leq Q_{Y_1-Y_0|X}(\tau|x), Y_1 > h(Y_0, x, \underline{z})|x) \end{aligned} \tag{11}$$

By Assumption 2.2, $Y_1 \leq h(Y_0, x, \underline{z})$ implies $Y_1 - Y_0 \leq Q_{Y_1-Y_0|X}(\tau|x)$. Thus, the first term in (11) can be written as

$$\begin{aligned} \Pr(Y_1 - Y_0 \leq Q_{Y_1-Y_0|X}(\tau|x), Y_1 \leq h(Y_0, x, \underline{z})|x) &= \Pr(Y_1 \leq h(Y_0, x, \underline{z})|x) \\ &= \Pr(D = 0|x, \underline{z}) \end{aligned}$$

Note that $\Pr(D = 0|x, \underline{z})$ is an observed probability, so it is identified.

Now I prove that the second term in (11) is also identified. Note that $Y_1 - Y_0 \leq Q_{Y_1-Y_0|X}(\tau|x)$ implies that $Y_1 \leq h(Y_0, x, \bar{z})$ almost surely by Assumption 2.2. For any $y_1 \in [h(y_0, x, \underline{z}), h(y_0, x, \bar{z})]$, by the intermediate value theorem, there exists $\zeta \in \mathcal{Z}(x)$ such that $y_1 = h(y_0, x, \zeta)$. Hence a point (y_0, y_1) such that $Y_1 - Y_0 \leq Q_{Y_1-Y_0|X}(\tau|x)$ and $Y_1 > h(Y_0, x, \underline{z})$ is an element of $\mathcal{H}(x)$ almost surely. Since the conditional joint distribution function of (Y_0, Y_1) conditional on $X = x$ is identified on $\mathcal{H}(x)$ by Theorem 2.2, the second term in (11) is identified. □

Remark 2.1. The relationship $Y_1 = Y_0 + Q_{Y_1-Y_0|X}(\tau|x)$ can be depicted as a diagonal line on two dimensional plane of (Y_0, Y_1). Assumption 2.2 states that the diagonal line lies in $\mathcal{H}(x)$. For heuristic explanation, suppose that the utility function h is additively separable in Y_0 and nonpecuniary preference: $h(Y_0, X, Z) = Y_0 + \varphi(X, Z)$. In this case, Assumption 2.2 is equivalent to a simple assumption that $Q_{Y_1-Y_0|X}(\tau|x) = \varphi(x, z)$ for some $z \in \mathcal{Z}(x)$. That is, the support of $\varphi(x, Z)$ is the identification region for $Q_{Y_1-Y_0|X}(\tau|x)$.

2.2 Discrete Instrumental Variable

In this subsection, I consider the case in which the instrumental variable is discrete. In this case, point-identification of the model is impossible. Instead, bounds are obtained. Suppose that Z takes a value from $\mathcal{Z} = \{0, 1, \cdots, K\}$.

Assumption 2.3. *Conditional on $X = x$, the following assumptions hold with probability one:*

(a) The distribution of (Y_0, Y_1) is absolutely continuous with respect to the Lebesgue measure with a density $f_{Y_0Y_1|X}$.

(b) h is strictly increasing in Y_0 for any Z.

(c) h is strictly increasing in Z for any Y_0.

Assumption 2.3 is the same as Assumption 2.1 except for the support condition on Z and the differentiability condition. As the derivatives of the observed distributions in the continuous instrument case play the key role, the differentials with respect to Z in the observed distributions are important in discrete case.

For $z \in \{1, 2, \cdots, K\}$ and $d \in \{0, 1\}$, define

$$\Delta G_d(y_0|x, z) = G_d(y_0|x, z) - G_d(y_0|x, z-1)$$

using expression (5), this can be written as

$$\Delta G_0(y_0|x, z) = \Pr\left(Y_0 \leq y_0, h(Y_0, x, z-1) \leq Y_1 < h(Y_0, x, z)|x\right) \tag{12}$$

It is obvious that $\Delta G_0(y_0|x, z)$ is non-negative and non-decreasing in y_0. Similarly, we have

$$\begin{aligned}
\Delta G_1(y_1|x, z) &= G_1(y_1|x, z) - G_1(y_0|x, z-1) \\
&= \Pr\left(Y_1 \leq y_1, Y_1 \geq h(Y_0, x, z)|x\right\} - \Pr\{Y_1 \leq y_1, Y_1 \geq h(Y_0, x, z-1)|x) \\
&= -\Pr\left(Y_1 \leq y_1, h(Y_0, x, z-1) \leq Y_1 < h(Y_0, x, z)|x\right)
\end{aligned} \tag{13}$$

Note that it takes a non-positive value and is non-increasing in y_1.

Theorem 2.3. *Under Assumption 2.3, for $z \in \{1, 2, \cdots, K\}$, the upper bound for $h(y_0, x, z)$ is given by*

$$h^U(y_0, x, z) = \inf\{y_1 : \Delta G_0(y_0|x, z) + \Delta G_1(y_1|x, z) \leq 0\} \tag{14}$$

and the lower bound for $h(y_0, x, z-1)$ by

$$h_L(y_0, x, z-1) = \sup\{y_1 : \Delta G_0(y_0|x, z) + \Delta G_1(y_1|x, z) \geq 0\} \tag{15}$$

Furthermore, the bounds are sharp.

Proof. I prove the theorem for the upper bound only. The lower bound can be shown similarly. I will establish that for $y_1 \geq h(y_0, x, z)$,

$$\Delta G_0(y_0|x, z) + \Delta G_1(y_1|x, z) \leq 0$$

If the claim is true, $h^U(y_0, x, z) \geq h(y_0, x, z)$. Suppose that y_0, y_1, x and z are fixed. If $y_1 \geq h(y_0, x, z)$, $Y_0 \leq y_0$ and $Y_1 < h(Y_0, x, z)$ imply $Y_1 \leq y_1$. Thus, we have

$$\begin{aligned}
\Delta G_0(y_0|x, z) &= \Pr\left(Y_0 \leq y_0, h(Y_0, x, z-1) \leq Y_1 < h(Y_0, x, z)|x\right) \\
&\leq \Pr\left(Y_1 \leq y_1, h(Y_0, x, z-1) \leq Y_1 < h(Y_0, x, z)|x\right) \\
&= -\Delta G_1(y_1|x, z)
\end{aligned}$$

Note that by Assumption 2.3(a) a strict inequality can be replaced with a weak one and vice versa.

To show that the bound is sharp, I will illustrate that any value smaller than $h^U(y_0, x, z)$ can be also smaller than $h(y_0, x, z)$ under certain circumstances. Let $\bar{y}_1$ be an arbitrary value smaller than $h^U(y_0, x, z)$. By definition, $\Delta G_0(y_0|x, z) + \Delta G_1(\bar{y}_1|x, z) > 0$. Fix a small $\varepsilon > 0$ and let $\mathcal{A} = \{(Y_0, Y_1) : h(Y_0, x, z-1) \leq Y_1 < h(Y_0, x, z)\}$. Also let

$$\begin{aligned}
\mathcal{A}_0(\varepsilon) &= \mathcal{A} \cap \{Y_0 \leq y_0, Y_1 > h(y_0, x, z) - \varepsilon\} \\
\mathcal{A}_1(\varepsilon) &= \mathcal{A} \cap \{Y_0 > y_0, Y_1 \leq h(y_0, x, z) - \varepsilon\}
\end{aligned}$$

Suppose that $\Pr(\mathcal{A}_0(\varepsilon)|x) = \Pr(\mathcal{A}_1(\varepsilon)|x) + \Delta G_0(y_0|x,z) + \Delta G_1(\bar{y}_1|x,z)$. Since $\mathcal{A}_0(\varepsilon)$ and $\mathcal{A}_1(\varepsilon)$ are disjoint regions having a positive Lebesgue measure, their probabilities can be manipulated easily without hurting any assumptions imposed for the theorem. A simple but tedious algebra gives

$$\Delta G_0(y_0|x,z) + \Delta G_1(h(y_0,x,z) - \varepsilon|x,z) = \Pr(\mathcal{A}_0(\varepsilon)|x) - \Pr(\mathcal{A}_1(\varepsilon)|x)$$

Thus,

$$\Delta G_1(h(y_0,x,z) - \varepsilon|x,z) = \Delta G_1(\bar{y}_1|x,z)$$

Since $\Delta G_1(y_1|x,z)$ is strictly monotone, $\bar{y}_1 = h(y_0,x,z) - \varepsilon < h(y_0,x,z)$. Therefore, h^U is sharp. □

Note that there is no bound for the boundary values of Z. That is, $h(y_0,x,0)$ has no lower bound and $h(y_0,x,K)$ has no upper bound. This is related to the identification region $\mathcal{H}(x)$ as discussed in the previous section.

The following theorem gives bounds for the joint distribution function. The bounds extend the bounds proposed by Peterson (1976) to a general censoring scheme that includes an instrument.

Theorem 2.4. *For fixed (y_0, y_1, x), let*

$$\begin{aligned} F^U_{Y_o Y_1|X}(y_0,y_1|x) &= \inf_{z\in\mathcal{Z}} \{G_0(y_0|x,z) + G_1(y_1|x,z)\} \\ F^L_{Y_0 Y_1|X}(y_0,y_1|x) &= \sup_{z\in\mathcal{Z}} \left\{G_0(y_0^L(z)|x,z) + G_1(y_1^L(z)|x,z)\right\} \end{aligned}$$

where

$$\begin{aligned} y_0^L(z) &= \min\{y_0, \inf\{t : \Delta G_0(t|x,z) + \Delta G_1(y_1|x,z) \le 0\}\} \\ y_1^L(z) &= \min\{y_1, \inf\{t : \Delta G_0(y_0|x,z) + \Delta G_1(t|x,z) \le 0\}\} \end{aligned}$$

Under the assumptions of Theorem 2.3,

$$F^L_{Y_0Y_1|X}(y_0,y_1|x) \le F_{Y_0Y_1|X}(y_0,y_1|x) \le F^U_{Y_0Y_1|X}(y_0,y_1|x)$$

and the bounds are sharp.

Proof. For any $z \in \mathcal{Z}$,

$$\begin{aligned} G_0(y_0|x,z) &= \Pr\{Y_0 \le y_0, Y_1 < h(Y_0,x,z)|x\} \\ &\ge \Pr\{Y_0 \le y_0, Y_1 \le y_1, Y_1 < h(Y_0,x,z)|x\} \end{aligned} \tag{16}$$

and

$$\begin{aligned} G_1(y_1|x,z) &= \Pr\{Y_1 \le y_1, Y_1 \ge h(Y_0,x,z)|x\} \\ &\ge \Pr\{Y_0 \le y_0, Y_1 \le y_1, Y_1 \ge h(Y_0,x,z)|x\} \end{aligned} \tag{17}$$

Since

$$\begin{aligned} & \Pr\{Y_0 \le y_0, Y_1 \le y_1, Y_1 < h(Y_0, x, z)|x\} \\ & + \Pr\{Y_0 \le y_0, Y_1 \le y_1, Y_1 \ge h(Y_0, x, z)|x\} \\ = \; & \Pr\{Y_0 \le y_0, Y_1 \le y_1|x\} \end{aligned}$$

we have

$$G_0(y_0|x,z) + G_1(y_1|x,z) \ge \Pr\{Y_0 \le y_0, Y_1 \le y_1|x\}$$

for any $z \in \mathcal{Z}$.

Note that $y_1^L(z)$ is the minimum of y_1 and the lower bound for $h(y_0, x, z)$ given in Theorem 2.3. Thus $Y_1 \le y_1^L(z)$ implies $Y_1 \le h(y_0, x, z)$ as well as $Y_1 \le y_1$.

$$\begin{aligned} G_1(y_1^L(z)|x,z) &= \Pr\{Y_1 \le y_1^L(z), Y_1 \ge h(Y_0, x, z)|x\} \\ &\le \Pr\{Y_1 \le h(y_0, x, z), Y_1 \le y_1, Y_1 \ge h(Y_0, x, z)|x\} \\ &\le \Pr\{Y_0 \le y_0, Y_1 \le y_1, Y_1 \ge h(Y_0, x, z)|x\} \end{aligned}$$

The last inequality holds because $Y_1 \le h(y_0, x, z)$ implies $Y_0 \le y_0$ when $Y_1 \ge h(Y_0, x, z)$ holds due to the monotonicity of h. By switching the role of Y_0 and Y_1, it can be shown that

$$G_0(y_0^L(z)|x,z) \le \Pr\{Y_0 \le y_0, Y_1 \le y_1, Y_1 < h(Y_0, x, z)|x\}$$

Therefore, for any $z \in \mathcal{Z}$,

$$G_0(y_0^L(z)|x,z) + G_1(y_1^L(z)|x,z) \le \Pr\{Y_0 \le y_0, Y_1 \le y_1|x\}$$

Now I show that the bounds are sharp. I will provide an example in which the bounds are actually the same as the object. Suppose that there exists $\zeta \in \{0, 1, 2, \cdots, K\}$ such that $y_1 = h(y_0, x, \zeta)$. By the monotonicity of h, $Y_0 \le y_0$ and $Y_1 < h(Y_0, x, \zeta)$ imply $Y_1 \le h(y_0, x, \zeta) = y_1$. Hence, the inequality in (16) holds as an equality. Similarly, (17) becomes an equality for $z = \zeta$. Therefore,

$$G_0(y_0|x,\zeta) + G_1(y_1|x,\zeta) = \Pr\{Y_0 \le y_0, Y_1 \le y_1|x\} = F_{Y_0Y_1|X}(y_0, y_1|x) \tag{18}$$

which implies

$$F^U_{Y_0Y_1|X}(y_0, y_1|x) = F_{Y_0Y_1|X}(y_0, y_1|x)$$

Note that

$$y_1^L(z) = \min\{y_1, h^U(y_0, x, z)\}$$

where h^U is the bound for h defined in Theorem 2.3. Hence, if $y_1 = h(y_0, x, \zeta)$,

$$h^U(y_0, x, \zeta) \ge h(y_0, x, \zeta) = y_1$$

and so

$$y_1^L(\zeta) = y_1$$

Similarly, it can be shown that $y_0^L(\zeta) = y_0$. Then,

$$\begin{aligned} F^L_{Y_0Y_1|X}(y_0, y_1|x) &\geq G_0(y_0|x,\zeta) + G_1(y_1|x,\zeta) \\ &= F_{Y_0Y_1|X}(y_0, y_1|x) \end{aligned}$$

Hence, the bounds are sharp. □

3 Estimation

In this section, I consider the estimation of the model with a continuous instrument. Suppose that we observe a random sample of $\{(D_i, Y_i, Z_i, X_i) : i = 1, \ldots, n\}$ where $D_i \in \{0, 1\}$, $Y_i \in \mathbb{R}$, $Z_i \in \mathcal{Z} \subset \mathbb{R}$ and $X_i \in \mathcal{X} \subset \mathbb{R}^{m_X}$. The identification results imply that $h(y_0, x, z)$ is the unique solution to

$$G_0^{(1)}(y_0|x,z) + G_1^{(1)}(y_1|x,z) = 0$$

for fixed (y_0, x, z). Meanwhile, for given (y_0, y_1, x), $F_{Y_0Y_1|X}(y_0, y_1|x)$ is the solution to

$$F_{Y_0Y_1|X}(y_0, y_1|x) = G_0(y_0|x,\zeta) + G_1(y_1|x,\zeta)$$

for some ζ, where ζ solves $y_1 = h(y_0, x, z)$ for the given (y_0, y_1, x).

I propose a nonparametric estimator based on the identification results. The estimation method proceeds in two steps. In the first stage, the conditional distribution function of the observed outcome and the choice, $G_d(y_d|x,z)$, and its derivative with respect to the instrument are estimated by a local linear estimator. Local linear estimators are a standard nonparametric method extensively studied in statistics and econometrics. For example, see Fan and Gijbels (1996), Ruppert and Wand (1994) and the references therein. In particular, the local linear estimator of a conditional distribution function is considered in Yu and Jones (1998), Hall et al. (1999).

In the second stage, I solve equations that are obtained from the first stage estimation. I adopt and extend the simulation-based method proposed in Dette et al. (2006), Dette and Scheder (2006), and Dette and Volgushev (2008). The estimator is computationally attractive: it only requires straightforward simulations and calculations of closed-form expressions. It does not require any numerical optimization. Furthermore, it is applicable to a more general estimation problem of numerically solving equations. Thus, the results obtained here might be of theoretical and practical interest in their own right.

I start by fixing some notation that is used throughout the section. Let $W_i \equiv (X_i', Z_i)'$ and $\mathcal{W} \subset \mathbb{R}^m$ be the support of W. For a positive integer l, $G_d^{(l)}(y|w)$ stands for $\frac{\partial^l}{\partial z^l} G_d(y|w)$ for $w = (x', z)'$. Similarly, $f_W^{(l)}(w)$ is the lth derivative of f_W with respect to z, where f_W is the probability density of W.

Let $\iota = (\iota_1, \cdots, \iota_m)$ be a m-dimensional vector of non-negative integers. Let $|\iota| = \sum_{j=1}^m \iota_p$. For $v = (v_1, \cdots, v_m) \in \mathbb{R}^m$ and a function $\mu : \mathbb{R}^m \to \mathbb{R}$, define

$$v^\iota = v_1^{\iota_1} \times \cdots \times v_m^{\iota_m}$$

and

$$\frac{\partial^\iota \mu}{\partial v^\iota} = \frac{\partial^{|\iota|}\mu}{\partial v_1^{\iota_1} \cdots \partial v_m^{\iota_m}}$$

3.1 First Stage Estimation

The first stage estimation is to estimate the conditional distribution function of (Y, D) conditional on W. Let (y, d, w) be given. We know that

$$E\left[1\{Y_i \leq y, D_i = d\}|W_i\right] = G_d(y|W_i)$$

If we write $\beta_0 = G_d(y|w)$ and $\beta_1 = \frac{\partial G_d(y|w)}{\partial w}$, by Taylor's theorem, we can approximate the right hand side by

$$G_d(y|W_i) \quad \approx \quad \beta_0 + \beta_1'(W_i - w) \tag{19}$$

for $W_i \approx w$.[2]

I consider a product kernel with equal bandwidth across variables. Let k be a kernel function defined on $\mathbb{R}$ and b be a sequence of positive numbers converging to zero. For $v = (v_1, \cdots, v_m) \in \mathbb{R}^m$, define

$$K_b(v) = \frac{1}{b^m} \prod_{j=1}^{m} k\left(\frac{v_j}{b}\right) \tag{20}$$

Based upon the approximation (19), one can estimate the parameters by solving the following weighted least squares problem

$$\hat{\beta} = \arg\min_{(a_0, a_1)} \sum_{i=1}^{n} \left[1\{D_i = d, Y_i \leq y\} - a_0 - a_1'(W_i - w)\right]^2 K_b(W_i - w) \tag{21}$$

where $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1'\right)'$ is an estimator of $\beta = (\beta_0, \beta_1')'$.

The solution to the problem (21) has a closed-form solution. It is convenient to work with matrix notation. Define matrices $\mathbb{W}$ and $\mathbb{Y}$ as follow:

$$\mathbb{W} = \begin{pmatrix} 1 & (W_1 - w)' \\ \vdots & \vdots \\ 1 & (W_n - w)' \end{pmatrix} \text{ and } \mathbb{Y} = \begin{pmatrix} 1\{D_1 = d, Y_1 \leq y\} \\ \vdots \\ 1\{D_n = d, Y_n \leq y\} \end{pmatrix}$$

Further, define $\mathbb{K} = diag\{K_b(W_1 - w), \ldots, K_b(W_n - w)\}$. Then the solution to the weighted least squares problem (21) can be written as

$$\hat{\beta} = (\mathbb{W}'\mathbb{K}\mathbb{W})^{-1}\mathbb{W}'\mathbb{K}\mathbb{Y} \tag{22}$$

I make the following assumptions.

Assumption 3.1. *Let $y \in \mathbb{R}$, $d \in \{0, 1\}$ and $w \in \mathcal{W}$ be given. Suppose that $G_d(y|w) \in (0, 1)$. At (y, d, w), the following assumptions hold*

(a) $\frac{\partial^\iota G_d}{\partial w^\iota}$ is bounded and continuous over $\mathcal{W}$ for any $|\iota| \leq 3$.

(b) $\frac{\partial^\iota f_W}{\partial w^\iota}$ is bounded and continuous on $\mathcal{W}$ for $|\iota| = 0, 1$.

[2] Here I consider a first-order expansion because the higher-order derivatives are not of interest and would make the notation and calculation complicated if included. However, it is possible to approximate up to a higher order. The estimator has smaller bias at a cost of larger variance. This extension is straightforward and well-studied. For example, see Ruppert and Wand (1994) and Fan and Gijbels (1996) among others.

(c) k is a symmetric, positive, second order kernel function. The following constants are finite for $j \leq 8$.

$$\begin{aligned} \kappa_j &= \int u^j k(u) du \\ \nu_j &= \int u^j k^2(u) du \end{aligned}$$

(e) As $n \to \infty$, $b \to 0$ but $nb^{m+2} \to \infty$.

Theorem 3.1. *Under Assumption 3.1, for $d \in \{0, 1\}$,*

$$\begin{aligned} E\left[\hat{G}_d(y|w) - G_d(y|w)|\mathbb{W}\right] &= b^2 B_0(y, d, w) + o_p(b^2) \\ E\left[\hat{G}_d^{(1)}(y|w) - G_d(y|w)|\mathbb{W}\right] &= b^2 B_1(y, d, w) + o_p(b^2) \end{aligned}$$

where

$$B_0(y, d, w) = \frac{\kappa_2}{2} Trace\left(\frac{\partial^2}{\partial w \partial w'} G_d(y|w)\right)$$

and

$$\begin{aligned} B_1(y, d, w) &= \frac{1}{2}\left\{\kappa_4 G_d^{(2)}(y|w) f_W^{(1)}(w) + \kappa_2^2 \frac{\partial G_d^{(1)}(y|w)}{\partial x'} \frac{\partial f_W(w)}{\partial x}\right\} \\ &\quad + \frac{1}{6}\left\{\kappa_4 G_d^{(3)}(y|w) f_W(w) + \kappa_2^2 Trace\left(\frac{\partial^2 G_d^{(1)}(y|w)}{\partial x \partial x'}\right) f_W(w)\right\} \end{aligned}$$

Moreover,

$$\begin{aligned} var\left[\hat{G}_d(y|w)|\mathbb{W}\right] &= \frac{1}{nb^m} V_0(y, d, w) + o_p\left(\frac{1}{nb^m}\right) \\ var\left[\hat{G}_d^{(1)}(y|w)|\mathbb{W}\right] &= \frac{1}{nb^{m+2}} V_1(y, d, w) + o_p\left(\frac{1}{nb^{m+2}}\right) \end{aligned}$$

where

$$V_0(y, d, w) = \frac{\nu_0^m}{f_W(w)} G_d(y|w)\left[1 - G_d(y|w)\right]$$

$$V_1(y, d, w) = \frac{\nu_0^{m-1}}{\nu_2 \kappa_2^2 f_W(w)} G_d(y|w)\left[1 - G_d(y|w)\right]$$

Proof. The conditional mean squared error for a local linear estimator is derived in, e.g., Ruppert and Wand (1994) and Fan et al. (1997). Particularly for the estimates of the derivatives, see Lu (1996). □

The theorem shows the rate of convergence of $\hat{G}_d$ and $\hat{G}_d^{(1)}$. Let r_n represent the convergence rate of $\hat{G}_d^{(1)}$. That is,

$$r_n = \max\left\{b^2, \frac{1}{\sqrt{nb^{m+2}}}\right\}$$

3.2 Estimation of the utility function

In this section, the estimation of $h(y_0, w)$ is discussed. For given (y_0, w), define

$$\mu_h(y_1) = G_0^{(1)}(y_0|w) + G_1^{(1)}(y_1|w)$$

Since (y_0, w) is fixed, μ_h is a function of y_1. By Theorem 2.1, we know that $h(y_0, w)$ is the unique solution to

$$\mu_h(\cdot) = 0 \tag{23}$$

Also, it is shown that μ_h is strictly decreasing in y_1 by the proof of Theorem 2.1. Using the first stage estimates of $G_0^{(1)}$ and $G_1^{(1)}$, a natural estimate of μ_h is

$$\hat{\mu}_h(y_1) = \hat{G}_0^{(1)}(y_0|w) + \hat{G}_1^{(1)}(y_1|w)$$

Note that $\hat{\mu}_h$ might not be decreasing in y_1 because the first stage estimate is not shape-restricted. A natural way to estimate $h(y_0, w)$ is to find y_1 such that

$$\hat{\mu}_h(y_1) = 0 \tag{24}$$

To obtain the numerical solution to equation (24), I use a simulation-based method, which is explained in detail in Appendix A below.

Let F_h^* be a distribution function chosen by the researcher and T_h^* be a random variable with the distribution function F_h^*. Using the monotonicity assumption that μ_h is strictly decreasing,

$$\begin{aligned} \Pr\left(\mu_h(T_h^*) \geq 0\right) &= \Pr(T_h^* \leq \mu_h^{-1}(0)) \\ &= F_h^*(\mu_h^{-1}(0)) \end{aligned}$$

It leads to

$$\mu_h^{-1}(0) = (F_h^*)^{-1}\left(\Pr(\mu_h(T_h^*) \geq 0)\right) \tag{25}$$

Note that F_h^* is known. Let $\{t_{hj}^* : j = 1, \ldots, n\}$ be a generated random sample from F_h^*. Replace μ_h with $\hat{\mu}_h$ and approximate the probability of the event $\{\mu_h(T_h^*) \geq 0\}$ by

$$\frac{1}{n^*} \sum_{j=1}^{n^*} 1\left\{\hat{\mu}_h(t_{hj}^*) \geq 0\right\} \tag{26}$$

Since $h(y_0, w) = \mu_h^{-1}(0)$, using (25) and (26), define an estimator for $h(y_0, w)$ by

$$\hat{h}(y_0, w) = (F_h^*)^{-1}\left(\frac{1}{n^*} \sum_{j=1}^{n^*} 1\left\{\hat{\mu}_h(t_{hj}^*) \geq 0\right\}\right)$$

Note that it only requires drawing a random sample from a known distribution and calculations of closed-form expressions in (26). Hence it is computationally very accessible.

The choice of F_h^* does not affect the first-order asymptotics as long as it satisfies the assumptions below, which are mild. Hence one may choose a distribution function that is tractable. In addition, since μ_h is evaluated at each point of the generated sample, it is desirable that the support of Y_1. One possible choice is a uniform distribution over the range of the observations of Y_1, for example $[Q_{Y_1|W}(0.01|w), Q_{Y_1|W}(0.99|w)]$.

I make the following assumptions.

Assumption 3.2. *Suppose that the following assumptions hold:*

(a) $\frac{\partial^\iota G_d(y|w)}{\partial w^\iota}$ *is bounded and continuous over* $(y, w) \in \mathbb{R} \times \mathcal{W}$ *for any* $|\iota| \leq 3$.

(b) $\frac{\partial G_d(y|w)}{\partial y}$ *is bounded and continuous over* $(y,w) \in \mathbb{R} \times \mathcal{W}$.
(c) $\sup_{u \in \mathbb{R}^m} |u'uK(u)| < \infty$
(d) There exists a positive sequence δ_n *such that* $\delta_n \log(n) = o(1)$.
(e) F_h^* *has a continuous and bounded density* f_h^* *and* $f_h^* > 0$ *on a neighborhood of* $h(y_0, w)$.
(f) $\mu_h(T_h^*)$ *has a bounded density* f_μ^*.
(g) $\frac{1}{\sqrt{n^*}} = o(r_n)$

Remark 3.1. Assumptions 3.2 (a) and (b) are a stronger version of the boundedness condition in Assumption 3.1. Since $\hat{G}_d^{(1)}(y|w)$ is evaluated at different values for the estimation of h, the boundedness condition has to be strengthened accordingly. While the boundedness condition in Assumption 3.1 is with respect to $w \in \mathcal{W}$, Assumptions 3.2 (a) and (b) are with respect to both y and w.

Theorem 3.2. *Let* $y_0 \in \mathbb{R}$ *and* $w \in \mathcal{W}$ *be given. Under Assumptions 3.1 and 3.2,*

$$\hat{h}(y_0, w) - h(y_0, w) = -\left(\frac{\partial G_1^{(1)}(h(y_0, w)|w)}{\partial y_1}\right)^{-1} [\hat{\mu}_h(h(y_0, w))] + o_p(r_n)$$

Proof. It is a direct result of Lemma A.1, which is given in Appendix. Assumption A.1 (a) and (b) are satisfied because $\mu_h(y_1)$ is strictly monotone and differentiable in y_1. Assumption A.1 (c) and (d) follow from Theorem 3.1 and Lemma B.2, respectively. Finally, (e), (f) and (g) are assumed in Assumption 3.2. □

Remark 3.2. Theorem 3.2 implies that the convergence rate of $\hat{h}(y_0, w)$ is the same as that of $\hat{\mu}_h$. Since

$$\hat{\mu}_h(y_1) = \hat{G}_0^{(1)}(y_0|w) + \hat{G}_1^{(1)}(y_1|w)$$

it follows from Theorem 3.1 that the bias and the standard error of $\hat{\mu}_h$ have an order of b^2 and $\frac{1}{\sqrt{nb^{m+2}}}$, respectively. Therefore, the convergence rate of $\hat{h}(y_0, w)$ is $r_n = \max\left\{b^2, \frac{1}{\sqrt{nb^{m+2}}}\right\}$, which is the standard rate of convergence for nonparametric estimators of first-order derivatives.

3.3 Estimation of the joint distribution

In this subsection, I consider the estimation of $F_{Y_0Y_1|X}(y_0, y_1|x)$ for fixed (y_0, y_1, x). By Theorem 2.2, we know that

$$F_{Y_0Y_1|X}(y_0, y_1|x) = G_0(y_0|x, \zeta) + G_1(y_1|x, \zeta)$$

where ζ satisfies $y_1 = h(y_0, x, \zeta)$. So a natural estimator of $F_{Y_0Y_1|X}(y_0, y_1|x)$ is

$$\tilde{F}_{Y_0Y_1|X}(y_0, y_1|x) = \hat{G}_0(y_0|x, \zeta) + \hat{G}_1(y_1|x, \zeta) \tag{27}$$

However, it is infeasible because ζ is unknown. To estimate ζ for fixed (y_0, y_1, x), define

$$\mu_\zeta(z) = G_0^{(1)}(y_0|x, z) + G_1^{(1)}(y_1|x, z)$$

It is clear that ζ satisfies $\mu_F(\zeta) = 0$. We can use the simulation-based estimator that is used to estimate $h(y_0, w)$. Let F_ζ^* be a distribution function chosen by the researcher and T_ζ^* be a random variable with the distribution function F_ζ^*. Draw a random sample $\{t_{\zeta j}^* : j = 1, \ldots, n^*\}$ from F_ζ^*. Define the simulation-based

estimator of ζ by

$$\hat{\zeta} = \left(F_\zeta^*\right)^{-1}\left(\frac{1}{n^*}\sum_{j=1}^{n^*} 1\left\{\hat{\mu}_\zeta(t_{\zeta j}^*) \geq 0\right\}\right)$$

Plug $\hat{\zeta}$ into (27) to obtain

$$\hat{F}_{Y_0Y_1|X}(y_0, y_1|x) = \hat{G}_0(y_0|x, \hat{\zeta}) + \hat{G}_1(y_1|x, \hat{\zeta})$$

The following assumptions are required.

Assumption 3.3. *Suppose that the assumptions of Theorem 3.1 hold. Further assume that*

(a) On a neighborhood of w, f_W is bounded away from zero.

(b) The kernel k is uniformly continuous.

(c) There exists a sequence of positive numbers δ_n such that $\delta_n = o(b)$, $\delta_n \log(n) = o(1)$ and $r_n = o(\delta_n)$.

(d) F_ζ^* has a continuous and bounded density f_ζ^* and $f_\zeta^* > 0$ on a neighborhood of ζ.

(e) $\mu_\zeta(T_\zeta^*)$ has a bounded density f_μ^*.

(f) $\frac{1}{\sqrt{n^*}} = o\left(r_n\right)$

Theorem 3.3. *Under Assumptions 3.1 and 3.3,*

$$\begin{aligned}\hat{F}_{Y_0Y_1|X}(y_0, y_1|x) - F_{Y_0Y_1|X}(y_0, y_1|x) &= \hat{G}_0(y_0|x, \zeta) - G_0(y_0|x, \zeta) + \hat{G}_1(y_1|x, \zeta) - G_1(y_1|x, \zeta) \\ &\quad + \left(\frac{G_0^{(1)}(y_0|x, \zeta) + G_1^{(1)}(y_1|x, \zeta)}{G_0^{(2)}(y_0|x, \zeta) + G_1^{(2)}(y_1|x, \zeta)}\right)\hat{\mu}_\zeta(\zeta) + o_p(r_n)\end{aligned}$$

Proof. Note that

$$\begin{aligned}\hat{F}_{Y_0Y_1|X}(y_0, y_1|x) - F_{Y_0Y_1|X}(y_0, y_1|x) &= \hat{G}_0(y_0|x, \hat{\zeta}) + \hat{G}_1(y_1|x, \hat{\zeta}) - G_0(y_0|x, \hat{\zeta}) - G_1(y_1|x, \hat{\zeta}) \\ &\quad + G_0(y_0|x, \hat{\zeta}) + G_1(y_1|x, \hat{\zeta}) - G_0(y_0|x, \zeta) - G_1(y_1|x, \zeta) \quad (28)\end{aligned}$$

By Lemma A.1, along the same line as the proof of Theorem 3.2, we can show that

$$\hat{\zeta} - \zeta = \left[G_0^{(2)}(y_0|x, \zeta) + G_1^{(2)}(y_1|x, \zeta)\right]^{-1}\hat{\mu}_\zeta(\zeta) + o_p(r_n) \quad (29)$$

Since $\hat{\zeta} - \zeta = O_p(r_n)$ and $r_n = o(\delta_n)$, $\hat{\zeta} \in \mathcal{N}(\zeta, \delta_n)$ with probability approaching to one. By Lemma B.3, we have

$$\hat{G}_d(y_d|x, \hat{\zeta}) - G_d(y_d|x, \hat{\zeta}) = \hat{G}_d(y_d|x, \zeta) - G_d(y_d|x, \zeta) + o_p(r_n) \quad (30)$$

By the delta-method,

$$G_d(y_d|x, \hat{\zeta}) - G_d(y_d|x, \zeta) = G_d^{(1)}(y_d|x, \zeta) \cdot \left[\hat{\zeta} - \zeta\right] + o_p\left(\left|\hat{\zeta} - \zeta\right|\right)$$

Hence,

$$G_0(y_0|x, \hat{\zeta}) + G_1(y_1|x, \hat{\zeta}) - G_0(y_0|x, \zeta) - G_1(y_1|x, \zeta) = \left(\frac{G_0^{(1)}(y_0|x, \zeta) + G_1^{(1)}(y_1|x, \zeta)}{G_0^{(2)}(y_0|x, \zeta) + G_1^{(2)}(y_1|x, \zeta)}\right)\hat{\mu}_\zeta(\zeta) + o_p(r_n) \quad (31)$$

From (28), (30) and (31), we have

$$\begin{aligned}\hat{F}_{Y_0Y_1|X}(y_0,y_1|x) - F_{Y_0Y_1|X}(y_0,y_1|x) &= \hat{G}_0(y_0|x,\zeta) - G_0(y_0|x,\zeta) + \hat{G}_1(y_1|x,\zeta) - G_1(y_1|x,\zeta) \\ &\quad + \left(\frac{G_0^{(1)}(y_0|x,\zeta) + G_1^{(1)}(y_1|x,\zeta)}{G_0^{(2)}(y_0|x,\zeta) + G_1^{(2)}(y_1|x,\zeta)}\right)\hat{\mu}_\zeta(\zeta) + o_p(r_n)\end{aligned}$$

□

Remark 3.3. It follows from (29) that the convergence rate of $\hat{\zeta}$ is the same as that of $\hat{G}_0^{(1)} + \hat{G}_1^{(1)}$. By Theorem 3.1, we know that the convergence rate of $\hat{\zeta}$ is r_n. Since $\hat{F}_{Y_0Y_1|X}$ depends not only on $\hat{\zeta}$ but also on $\hat{G}_0 + \hat{G}_1$, we have to compare the relative orders of $\hat{\zeta}$ and $\hat{G}_d$. Theorem 3.1 gives that the orders of the biases of $\hat{G}_d$ and $\hat{G}_d^{(1)}$ are the same, but the orders of their standard errors are different: $\hat{G}_d^{(1)}$ is slower. Hence the convergence rate of $\hat{F}_{Y_0Y_1|X}$ is at least as slow as $\hat{G}_d^{(1)}$. It is also possible that the estimation error in $\hat{G}_d$ is of negligible order relative to $\hat{\zeta}$ if $b^2 = o(\frac{1}{\sqrt{nb^{m+2}}})$, which requires that the bandwidth is small enough.

4 Monte Carlo experiment

This section presents some Monte Carlo experiment results to demonstrate the finite-sample performance of the nonparametric estimator proposed in Section 3. Using the nonparametric estimation procedure, I estimate two parameters, the utility function and the treatment effect, which are of principal interest in practice. I consider six different designs of experiments. In the first three designs, the assumptions needed for the nonparametric estimator are satisfied. In the other designs, the selection equation has another random component so that the extended Roy model assumption is violated. I also estimate the model using two benchmark methods : one is Heckman's two-step estimator (Heckman (1979)) for a parametric model with a normal distribution assumption, and the other is a semiparametric estimator based on the identification-at-infinity method proposed in Heckman (1990) and Andrews and Schafgans (1998).

4.1 Data Generating Processes

I consider six designs. Design (A) is the standard parametric model given by

$$Y_d = \alpha_d + X\beta_d + \varepsilon_d \tag{32}$$

$$D = 1\{Y_1 > Y_0 + Z\delta\} \tag{33}$$

where X and Z are independent standard normal random variables and $(\varepsilon_0, \varepsilon_1)$ follows a bivariate normal distribution independently of X and Z. Parameter values are set as $\alpha_0 = \alpha_1 = 0$, $\beta_1 = 1$, $\beta_0 = 0.5$, $\delta = 1$ and

$$\begin{pmatrix}\varepsilon_0 \\ \varepsilon_1\end{pmatrix} \sim N\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}1 & 0.5 \\ 0.5 & 1\end{pmatrix}\right) \tag{34}$$

In Design (B), skewed disturbance terms are considered. The selection equation is given by (33) and the outcome equation is similar to the specifications above except that the disturbance term ε_d is replaced by ε_d^{skew}. The new disturbance term ε_d^{skew} is generated from the following formula

$$\varepsilon_d^{skew} = \frac{\exp(\varepsilon_d) - E\left[\exp(\varepsilon_d)\right]}{\sqrt{var[\exp(\varepsilon_d)]}}$$

where ε_d is generated according to (34). That is, ε_d^{skew} is a bivariate log-normal distribution with its mean and variance normalized to zero and one, respectively. All the parameter values are the same as Design (A).

Finally, Design (C) is to accommodate nonlinear utility functions. The outcome equation is given by (32) and (34), while the selection equation takes the form of

$$D = 1\{\mathcal{U}(Y_1) > \mathcal{U}(Y_0) + Z\delta\}$$

where $\mathcal{U}(y) = -\exp(-y)$. Note that this design incorporates a quasi-linear *constant absolute risk aversion* (CARA) utility function $\mathcal{U}$. All the parameter values are the same as the base design.

For the robustness check, I consider three misspecified cases, which are deviations from the preceeding designs. For Designs (D) and (E), I generate data from the selection equation with an additional random error ξ of the following form:

$$D = 1\{Y_1 > Y_0 + Z\delta + \xi\}$$

where ξ is an standard normal random variable independent of all variables. For Design (F), I consider

$$D = 1\{\mathcal{U}(Y_1) > \mathcal{U}(Y_0) + Z\delta + \xi\}$$

The outcome equations for Designs (D), (E), (F) are the same as that of Design (A), (B), (C), respectively.

For each design, a hypothetical data is generated for different sample sizes of $n \in \{200, 500, 1000\}$. The estimated parameters are the utility function $h(y_0, x, z)$ at $(y_0, x, z) = (0, 0, 0)$ and the median treatment effect, $median(Y_1 - Y_0|X = 0)$.

4.2 Estimation Procedure

The preliminary tuning parameters used in the estimation are as follows. In the first stage, the conditional distribution function and its derivatives are calculated by the local linear estimator with the Gaussian kernel. I estimate the model for different bandwidths of $\{0.5, 1.0, 1.5, 2.0, 2.5\}$.

The size of the generated sample for numerical inversion method, n^*, is set equal to 1000. In general, the larger n^*, the better as long as the computing time and cost permits. In practice, I suggest to set n^* no lower than the sample size of the data. It is generated from a uniform distribution over the interval between upper and lower 0.5 percentile values.

$h(y_0, x, z)$ is estimated at $y_0 = 0$, $x = 0$ and $z = 0$. Following are the procedures that I use to estimate $h(y_0, x, z)$ at a fixed (y_0, x, z).

1. F_Y^* is set as the uniform distribution function over $[Q_{Y_1}(0.05), Q_{Y_1}(0.95)]$, where Q_{Y_1} is the sample quantile of Y conditional on $D = 1$.

2. Draw a random sample $\{y_j^* : j = 1, 2, \cdots, n_Y^*\}$ from F_Y^*, where $n_Y^* = 1000$. I just used an equally-spaced grid.

3. Calculate $\hat{\mu}_h(y_j^*) = \hat{G}_0^{(1)}(y_0|x, z) + \hat{G}_1^{(1)}(y_j^*|x, z)$ for each $j = 1, 2, \cdots, n_Y^*$. Using the closed-form formula for local linear estimators, the calculation is easy to implement.

4. Let $\bar{n}_Y$ be the number of $\hat{\mu}_h$'s such that $\hat{\mu}_h(y_j^*) \leq 0$.

5. $F_Y^{*^{-1}}(\bar{n}_Y / n_Y^*)$ is the estimate of $h(y_0, x, z)$.

Table 1 exhibits the results of estimation of h.

The estimation of $median(Y_1 - Y_0|x)$ requires a different procedure. Since $F_{Y_0Y_1|X}(y_0, y_1|x)$ is the estimated object, an extra step is needed to calculate $median(Y_1 - Y_0|x)$ from $F_{Y_0Y_1|X}$. Let me first explain how to estimate $F_{Y_0Y_1|X}(y_0, y_1|x)$ for a fixed (y_0, y_1, x).

1. Let F_Z^* be a uniform distribution over $[Q_Z(0.05), Q_Z(0.95)]$.
2. Draw a random sample $\{z_j^* : j = 1, 2, \cdots, n_Z^*\}$ from F_Z^*, where $n_Z^* = 1000$.
3. Calculate $\hat{\mu}_F(z_j^*) = \hat{G}_0^{(1)}(y_0|x, z_j^*) + \hat{G}_1^{(1)}(y_1|x, z_j^*)$ for $j = 1, 2, \cdots, n_Z^*$ using the formula for the local linear estimators.
4. Let $\bar{n}_Z$ be the number of non-positive $\hat{G}_2^{(1)}(y_0, y_1|x, z_j^*)$'s among $\{\hat{G}_2^{(1)}(y_0, y_1|x, z_j^*) : j = 1, 2, \cdots, n_Z^*\}$ and $\bar{z}^*$ be the $\bar{n}_z$th smallest element of $\{z_j^* : j = 1, 2, \cdots, n_Z^*\}$.
5. Using the estimates calculated in the second step, $\hat{G}_0(y_0|x, \bar{z}^*) + \hat{G}_1(y_1|x, \bar{z}^*)$ is the final estimate of $F_{Y_0Y_1|X}(y_0, y_1|x)$.

The resulting $F_{Y_0Y_1|X}$ may be outside $[0, 1]$. One may adjust the estimates to satisfy the restriction, but I don't in the simulations.

To calculate $median(Y_1 - Y_0|x)$, I estimate $F_{Y_0Y_1|X}(y_0, y_1|x)$ at 200×200 points of (y_0, y_1). The points divides $\mathbb{R}^2$ into 201×201 squares and the probability measure on each square can be calculated. Then I treat (Y_0, Y_1) as a discrete random variable distributed on 201×201points, which represent the squares. Under the presumption that (Y_0, Y_1) is discrete, we know the whole distribution of the random variable. Hence we can calculate the distribution of $Y_1 - Y_0$ and so the median of $Y_1 - Y_0$.

4.3 Benchmark and alternative methods

As a benchmark, I also estimate the model using Heckman's two-step method (Heckman (1979)). It is a widely used method to adjust selection bias in many empirical studies. Let me briefly explain how Heckman's two step estimator works. In the first step, the selection equation is modeled as follows:

$$D_i = 1\{\pi_0 + X_i\pi_1 + Z_i\pi_2 \geq \xi_i\} \tag{35}$$

with

$$\xi_i | X_i, Z_i \sim N(0, 1)$$

and it is estimated by the probit model. In the second stage, using the first stage estimates and the normal distribution assumption, the selection bias in the outcome equation is adjusted. For each $d \in \{0, 1\}$,

$$E[Y_i | X_i, Z_i, D_i = d] = \alpha_d + X_i\beta_d + \sigma\lambda(\hat{\pi}_0 + X_i\hat{\pi}_1 + Z_i\hat{\pi}_2)$$

where λ is the inverse Mills ratio. By regressing Y on X, Z and $\lambda(\hat{\pi}_0 + X\hat{\pi}_1 + Z\hat{\pi}_2)$, we can estimate α_d and β_d accounting for the selection bias.

Note that the first step is correctly specified in Design (A). Meanwhile, in Design (B), it is misspecified because the disturbance ξ is not normally distributed, which results in misspecification bias in the first stage. In Design (C), it is also misspecified not only because the disturbance is not normal, but also because the selection equation is not linear in X and Z.

I also estimate the model with a semiparametric method based on the identification-at-infinity argument proposed by Andrews and Schafgans (1998). The basic idea is that state 1 will be chosen with a probability close to 1 if Z_i is low enough. Hence, using the observations for which Z_i is smaller than a certain threshold, we can estimate the model without selection bias. For the subsample such that $D_i = 1$, I compute $\hat{Y}_i = Y_i - X_i\hat{\beta}_1$ using the estimate of β_d obtained by Heckman's two-step estimator. The estimate of α_1, $\tilde{\alpha}_1$, is defined by

$$\tilde{\alpha}_1 = \frac{\sum_{i=1}^n \hat{Y}_i 1\{D_i = 1, Z_i \leq \gamma_n\}}{\sum_{i=1}^n 1\{D_i = 1, Z_i \leq \gamma_n\}}$$

where γ_n is a threshold. That is, $\tilde{\alpha}_1$ is the sample mean of $\hat{Y}_i$ conditional on $D_i = 1$ and Z_i being smaller than the threshold. The threshold is defined by a sample quantile of Z and I consider three different quantiles: 0.02, 0.05 and 0.1. Similarly, $\tilde{\alpha}_0$ can be computed using the observations for which $D_i = 0$ and Z_i is larger than a threshold, which is set as either 0.98, 0.95 or 0.9 sample quantile of Z. In theory, the threshold should approach toward 0 or 1 as the sample size increases. However there is no theory how to choose the threshold. I also tried 1% threshold level, but I encountered some cases in the simulations in which I could not calculate the estimator because $\sum_{i=1}^n 1\{D_i = 1, Z_i \leq \gamma_n\} = 0$.

It should be noted that the estimator requires a consistent preliminary estimator for β_d, which is usually taken from Heckman's two-step estimator. In Designs (B), (C), (E), (F), Heckman's two-step estimator is not consistent. To prevent possible misspecification error in computing $\hat{Y}_i$, I only estimate Designs (A) and (D). Also note that the large support assumption holds because the support of Z is the whole real line. Even though all assumptions for the estimator are satisfied, it is known that the estimator has a slower rate of convergence because the effective sample size used to estimate $\tilde{\alpha}_d$ is smaller than the whole sample size. For example, if the sample size is 1000 and the extreme 5% observations are used, the effective sample size used to calculate $\tilde{\alpha}_d$ is only 50.

One obstacle of the comparison is that these estimators do not estimate $median(Y_1 - Y_0|x)$ and $\mathcal{U}(Y_d)$ directly. Though it is possible to use $\hat{\alpha}_1 - \hat{\alpha}_0$ or $\tilde{\alpha}_1 - \tilde{\alpha}_0$ as estimators for $median(Y_1 - Y_0|x)$ under additional distributional assumptions, it makes the comparison even unclearer because the bias of $\hat{\alpha}_1$ (or $\tilde{\alpha}_1$) can be canceled by that of $\hat{\alpha}_0$ (or $\tilde{\alpha}_0$). Hence, I report the root mean squared error (RMSE), bias, and standard error of the respective intercept estimator rather than those of the difference of the two intercept estimators. I could not find any legitimate comparison benchmark for the utility function as a function of Y_0. The first-stage probit model in (35) treats Y_0 as a part of disturbance term.

4.4 Results

I calculate the root mean squared error (RMSE), the bias and the standard error (SE) of the estimators, based on 1,000 times of repetitions. Table 1 presents the results for the estimates of the utility function, h. In all the deisgns, the bias is quite small regardless of the sample size and the bandwidth. The standard error heavily depends on the sample size and the bandwidth, which is consistent with the theory. It is hard to compare the results across different designs because the shape of the underlying parameters affects the results.

It should be addressed that the biases in the latter three designs are not severe even though the model is misspecified. On the other hand, the standard errors in the latter three designs are larger than those in the correctly-specified first three designs. The increase in the standard error is rather natural, since the models include additional randomness in the selection equation. This shows that the estimator does not suffer from

severe misspecification error up to an additional independent randomness in the selection equation.

Table 2 shows the performance of the median treatment effect estimates.The performance is remarkably good and consistent with theory in all the settings. Even in the misspecified designs, the estimator shows nice performances.

Table 3 summarizes the performance of Heckman's two-step estimator. As expected, in Designs (A) and (D), the estimator performs very well as the parametric assumptions are all satisfied. However, in the other designs, the estimates suffer from misspecification errors, having large biases. Notably, the estimates of α_d have larger biases than those of β_d. This implies that a misspecification error might be more critical when estimating treatment effects that are associated with the intercepts rather than the slope coefficients.

Table 4 presents the results of the semiparametric estimator proposed by Andrews and Schafgans (1998). I only estimate Designs (A) and (D) because the preliminary estimates, $\hat{\beta}_d$, are not consistent in the other designs, which violates an assumption required for the consistency of the estimator. The results are sensitive to the threshold level. As the threshold level increases, the effective sample size used in the estimation decreases and thus the standard errors become larger. But the bias gets smaller as the effective sample size becomes smaller. This is because the identification-at-infinity argument may not hold for non-extreme values of Z.

Table 5 allows us to compare the RMSE's of the three estimators. For comparison, I use the results of my nonparametric estimator for the bandwidth of the smallest RMSE and Andrews-Schafgans' semiparametric estimator for the effective sample size being 10% of the whole sample size. Direct comparison is somewhat difficult because the targetting parameters are different. Furthermore, my estimator is nonparametric while the others are parametric and semi-parametric, respectively. In general, nonparametric estimators tend to show larger standard errors compared to parametric or semiparametric estimators. This explains why the RMSE's of the nonparametric estimator are larger than those of the other estimators in Designs (A) and (D). However, in Designs (B) and (C) in which Heckman's two-step estimator shows high RMSE due to the misspecification error, the nonparametric estimator has smaller RMSE's than those of the parametric estimator. In Designs (E) and (F), where both of estimators are misspecified, still the nonparametric estimator works fine.

For more detailed analysis, the biases of the estimators are reported in Table 6. As expected by the theory, the nonparametric estimator has quite small biases in correctly-specified settings. In the misspecified setting in Designs (D), (E) and (F), the biases are still very small. While, Heckman's two-step estimator is very accurate in Designs (A) and (D) in which the parametric assumptions are satisfied. But, in the other designs, the bias is not ignorable, which is a consequence of the misspecification. Andrews-Schafgans' estimator performs well, but its biases tend to be larger than those of the nonparametric estimator. Though the bias would diminish as the threshold gets higher, in practice the bias can not be eleminated completely.

Overall, the nonparametric estimator developed in this paper performs well in all simulation designs, while the parametric estimator can be biased when the parametric assumptions fail. Most importantly, the nonparametric estimator outperforms Andrews-Schafgans' semiparametric estimator even when the model is correctly specified.

Finally, it is notable that the computational cost of the estimation is fairy low. For the sample size of 1000, one iteration for the estimation of h takes less than one second and one iteration for the estimation of the median treatment effect takes about 6 seconds. The code is written and run in MATLAB on a desktop PC with 3GHz CPU.

5 Application to Hybrid Maize Adoption

Agriculture is the most important sector and maize is a major food crop in Malawi. Hybrid maize is a variety of maize that has been developed to improve agricultural yield. To resolve the food security problem and foster economic growth, the government has implemented a large-scale agricultural input subsidy programs since the 1980s. Even though one of the principal goals of the government's programs is to encourage farmers to plan hybrid maize, the adoption rate is still around 60%.

In this section, I investigate the farmers' adoption decision of hybrid maize in the framework of the extended Roy model. I attempt to answer two policy-relevant questions : First, does hybrid maize really improve yield? Second, how much subsidy is needed to maximize yield? Regarding the first question, I estimate the treatment effect of hybrid maize on yield. I also estimate the relationship between the effectiveness of hybrid maize and other agricultural inputs such as soil quality and the amount of fertilizer use. The estimation results suggest that hybrid maize is effective in improving yields in general. Then, why do farmers not plant hybrid maize? There must be a cost of planting hybrid maize, which dominates the yield increment. It naturally raises the second question. Since the utility function is estimated, the subjective cost of planting hybrid maize directly follows. I calculate the optimal subsidy level that depends on observable characteristics of farmers.

5.1 Decision to Adopt Hybrid Maize

This subsection describes how farmers' decisions to adopt hybrid maize can be analyzed within the framework of the Roy model. Suppose that farmers face a discrete choice between planting traditional maize or hybrid maize. Let $D = 1$ represent planting hybrid maize and $D = 0$ traditional maize. Both types of seeds yield the same type of crop. Only the amount of crop differs. For $d \in \{0, 1\}$, Y_d signifies the potential yield of maize for each type of maize seed. The potential outcomes are heterogeneous across farmers. Thus, the returns to planting hybrid maize is also heterogeneous. I do not impose any functional form restrictions on the distribution of the potential yields:

$$(Y_0, Y_1)|X \sim F_{Y_0 Y_1|X}$$

where X includes the amount of fertilizer, the area of plot, years of schooling of the farmer, and soil quality dummy.

Knowing their own potential yields, the farmers take more components into account. They have to buy hybrid maize seeds from the market if they decide to plant hybrid maize, while they can just use the previous year's crop as seeds if plant traditional maize. I assume that the farmer's decision is made according to the following selection equation:

$$D = 1\{Y_1 > h(Y_0, X, Z)\}$$

h is the utility function measured in terms of yield of hybrid maize. It can be interpreted as the yield from hybrid maize that is equivalent to Y_0 from traditional maize. If $h(Y_0, X, Z)$ is higher than Y_0, there are some obstacles that hinder the farmers from planting hybrid maize. Let $\phi(Y_0, X, Z) = h(Y_0, X, Z) - Y_0$ and rewrite the selection equation as

$$D = 1\{Y_1 > Y_0 + \phi(Y_0, X, Z)\}$$

Note that it nests the usual extended Roy model specification

$$D = 1\{Y_1 > Y_0 + \phi(X, Z)\}$$

where $\phi(X, Z)$ is the cost of planting hybrid maize, which does not allow interaction between (X, Z) and Y_d. $\phi(Y_0, X, Z)$ captures the cost of planting hybrid maize.

On the other hand, the cost $\phi(Y_0, X, Z)$ can be viewed as the optimal subsidy level to maximize the yield. Consider a policy such that subsidy S, measured in terms of Y_1, is offered to farmers who plant hybrid maize. This policy affects the selection, so the selection equation becomes

$$D = 1\{Y_1 + S > h(Y_0, X, Z)\} \tag{36}$$

If the policy maker's objective is to maximize the yield, the optimal subsidy S^* must induce

$$D = 1\{Y_1 > Y_0\} \tag{37}$$

When $S = \phi(Y_0, X, Z)$, the selection equation (36) becomes (37). Therefore, we can calculate the optimal subsidy level that follows the estimation of h.

I use distance to road as an instrumental variable. In Malawi, the maize market is monopolized by a government-owned corporation, the Agricultural Development Marketing Corporation (ADMARC). ADMARC monopolizes the distribution of agricultural inputs and sets crop prices. There are local private buyers/sellers that intermediates the farmers and ADMARC. To plant hybrid maize, the farmers must either buy the seeds from a local private seller or travel to an outlet post of ADMARC. Hence, the cost of planting hybrid maize is correlated with the distance to road. Its relevance has been confirmed in various studies on hybrid maize adoption. Among papers on the adoption of hybrid maize in Malawi, Zeller et al. (1998) find that the traveling cost to the market is one of the significant determinants of hybrid maize adoption and Chirwa (2005) obtains a similar result with the distance to the nearest market. Suri (2011) examined the effect of hybrid seed using Kenyan data and used the distance to the nearest fertilizer seller as an instrument. Also, since covariates include geographic variables such as the area of plot and soil quality, the potential indirect effect of the instrument on the yield is controlled.

5.2 Data

The data comes from the Integrated Household Survey (IHS) in Malawi conducted by the Government of Malawi through the National Statistical Office jointly with the World Bank. The survey is conducted roughly every 5 years and the latest one is third round in 2010 and 2011. As the sample households in each round have no link, I only use the latest survey as a cross-sectional data. The sample households are nationally representative, drawn from all regions over Malawi. The survey provides a detailed information on a huge number of households : it covers 12,271 households and provides various demographic, agricultural and geographic variables for each household.[3]

Here are some sample exclusion criteria. I excluded the plots if a crop other than traditional local maize or hybrid maize is planted; or if multiple crops are planted. Further, recycled hybrid maize is excluded because it is known that recycled hybrid maize has little difference from the traditional one. I excluded

[3]For more detailed information, visit http://go.worldbank.org/6A7GUDQ1Q0

the observations whose yield is not measured in a metric unit - kilogram or ton. All observations with any missing items are removed.

My sample includes the farmers who planted maize - either traditional or hybrid - in the rainy season in 2009-2010. The unit of observation is a plot. There are several reasons that I use a plot as the unit. First, the information on agricultural activities - for example, the amount of fertilizer use, the amount of yield - are recorded for each plot. Second, a farmer may have multiple plots and may decide to plant different types of maize seeds on different plots. Therefore, a plot is the unit in which the decision to use hybrid seeds is made. Therefore, it is natural to consider a plot as the unit.

The descriptions of the variables are given in Table 7. Table 8 shows some descriptive statistics of related variables. Note that it is clear that hybrid maize yields more than traditional maize, and the hybrid maize users live closer to road on average than the traditional maize users.

Preliminarily, some frequently used models are estimated. First, I estimate a probit model to see determinants of the adoption of hybrid maize under parametric assumptions and the results are reported in Table 9. The coefficient to the distance to road is significantly negative. This result supports the relationship between the cost of hybrid maize and the distance to road. Next, I regress yield on the adoption of hybrid maize and other variables and the results are reported in Table 10. OLS results suggest that hybrid maize has a positive effect on yield. I also estimate an instrumental variable regression model treating hybrid maize adoption as an endogenous variable and the distance to road as an excluded variable. Still the results indicate hybrid maize increases yields. Though, the regression results can be limited in a sense that the effect of hybrid maize on yields is assumed to be merely a constant shift in the intercept. Heckman's two-step estimator allows for the coefficients differ across two types of maize. Table 11 reports the estimated coefficients using Heckman's two-step estimator. The implied effect of hybrid maize on yields is calculated and reported in Table 12.

5.3 Estimation

I estimate the full nonparametric model. In the first stage, I estimate the distribution of observed harvest conditional on covariates and the instrument, $G_d(y|x,z)$, and its derivative with respect to the instrument using a local linear estimator. I use a product kernel using Gaussian kernel. The bandwidth parameters are chosen by cross validation method as follows. First, since different bandwidths should be applied to different variables according to their degrees of dispersion, the bandwidth of a variable is fixed to be proportional to the standard deviation of the variable up to a coefficient. Then I compute the coefficient that applies to all variables by a cross-validation method. Another problem is that in the estimation of $G_d(y|w)$, the theoretically optimal bandwidth may be different for different values of y. Since it is hard to differentiate the bandwidths according to y, the bandwidth parameter is chosen to minimize the average squared error over different values of y and applied to all y. The average squared error is calculated on empirical deciles of observed Y and then summed over the nine points except for 0th decile and 10th decile. I denote the deciles as $\{y_{(1)}, y_{(2)}, \cdots, y_{(9)}\}$.

Then, the standard leave-one-out cross-validation method is used. Without using the ith observation, estimate $G_d(y|w)$ at $y \in \{y_{(1)}, \cdots, y_{(9)}\}$ and $w = W_i$ by a local linear estimator. Denote it as $\hat{G}_{d,-i}(y_{(j)}|W_i)$. The bandwidth parameter is chosen to minimize

$$\sum_{i=1}^{n} \sum_{j=1}^{9} \sum_{d \in \{0,1\}} \left(1\{Y_i \leq y_{(j)}, D_i = d\} - \hat{G}_{d,-i}(y_{(j)}|W_i) \right)^2$$

among candidate values of $\{0.3, 0.4, \cdots, 3.0\}$. In noncomputable cases due to matrix singularity, I impute the squared error as 1. When including discrete variables, e.g., dummy for soil quality, the bandwidth parameter is cross-validated for the subsample of each type of soil quality. The cross validation results are given in Table 13.

Then I estimate two parameters: one is the median treatment effect of hybrid maize, $median(Y_1 - Y_0|X)$, and the other is the optimal subsidy level, $h(Y_0, X, Z) - Y_0$. The median treatment effect is estimated at five different levels of fertilizer use: the median of fertilizer use, 54.95 Kg/Acre, and deviations from the median, 34.95, 44.95, 64.95 and 74.95 Kg/Acre. After estimating the median effect using the whole sample, I use a subsample to incorporate the soil quality. Soil quality is given as a binary code - good or poor. Using the subsample for each type of soil quality, the median treatment effect is estimated at five different levels of fertilizer use. Meanwhile, Y_0 is fixed at its median.

The detailed estimation procedure is the same as described in the previous section. 90% confidence interval for each estimate is calculated based on the bootstrap of 1000 repetitions. The program is available for Matlab upon request. A program for Stata is in progress.

5.4 Results

The estimation results for the median treatment effect are given in Table 14. It suggests that the median treatment effect is positive in the whole sample estimates and the subsample estimates conditional on good soil quality. The confidence intervals support the positive treatment effect of hybrid maize. However, when the soil quality is poor, the estimates are negative. The confidence intervals are too wide to tell the sign of the estimates. This implies that hybrid maize does not always increase the yield.

The effects of fertilizer and soil quality are positive from the estimates. In each panel, as the amount of fertilizer use increases, the estimated treatment effect strictly increases. Though the confidence intervals overlap with each other, they are also increasing as a whole. The effect of soil quality is also clearly positive. The estimates for good soil are strictly larger than those for poor soil.

All estimates of the optimal subsidy level are positive. They show that the farmers do not plant hybrid maize even though it yields more. In each panel, the optimal subsidy level increases as the distance to road increases, which is consistent with intuition. Since the farmers living far from road are likely to face higher transaction cost of buying hybrid maize, a higher level of subsidy is needed for them to plant hybrid maize relative to the farmers close to road.

There is a pattern that the optimal subsidy level is higher for the farmers with less years of schooling. There can be two possible reasons: first, highly educated farmers are more likely to be from a rich family. Hence, the borrowing constraint upon buying hybrid maize might be weaker for highly educated farmers. Another explanation is that highly educated farmers are more willing to adopt a new technology, which results in a lower psychological cost of adopting hybrid maize.

Finally, I make a remark on the confidence intervals. The confidence intervals are wider than desirable. This is a common problem with a fully nonparametric approach when multiple conditioning variables are used. In general, functional form restrictions help construct tighter intervals and allow one to use more conditioning variables. A semiparametric estimation method is an interesting topic to pursue.

6 Conclusion

In this paper I have developed a new identification method for the extended Roy model that identifies the joint distribution of the potential outcomes and the utility function. The key assumptions are that the utility function is deterministic conditional on the potential outcomes and observables and that the effect of the instrument on the selection is monotone. In a companion paper, I study the identification of the Roy model with unobserved heterogeneity in preferences.

An advantage of the identification is that it allows one to relax conventional functional form restrictions on the outcome equation and the utility function such as linearity and additive separability. Another advantage of the identification method is that no support assumption is required so that instruments with a bounded or finite support can be used. A point-identification result is given if the instrument is continuous, possibly having a bounded support. Sharp bounds for the model are obtained if the instrument is discrete.

In addition, I have developed a nonparametric estimator based on the identification method using the simulation-based estimator proposed in Dette et al. (2006). It is computationally attractive. I showed that the simulation-based estimator has a standard nonparametric rate of convergence and examined the efficacy of the estimator in finite sample shown by Monte Carlo simulations.

Appendix

A A simulation-based estimator

In this section, I introduce a simulation-based estimator for the solution to a monotone equation. Consider a strictly increasing function $\mu : \mathcal{T} \to \mathbb{R}$ where $\mathcal{T} \subset \mathbb{R}$. Suppose that the parameter of interest, $\theta \in \mathcal{T}$, is the unique solution to

$$\mu(\cdot) = s$$

for a given $s \in \mathbb{R}$. Let $\hat{\mu}$ be a preliminary estimator of μ, which is not necessarily increasing. Many applications of such a setup can be found in economics. A typical example is estimation of quantile functions (e.g., Dette and Volgushev (2008),Chernozhukov et al. (2010)). Economic models with monotonicity are common. For example, in Berry and Haile (2012), market share is a monotone function of price.

In general, estimating θ using $\hat{\mu}$ is numerically formidable. One possible way is to impose a shape restriction when estimating μ. It requires a constrained optimization. Another alternative is the minimum distance estimation to find θ that minimizes the distance between $\hat{\mu}$ and s. For example, the estimator minimizing the distance defined by the squared deviation is given by

$$\hat{\theta} = \min_{t \in \mathcal{T}} \left(\hat{\mu}(t) - s\right)^2$$

It also requires numerical optimization procedure. Furthermore, frequently used methods, such as Newton-Rhapson method, require estimation of the derivative of μ. If $\hat{\mu}$ is not monotone, the criterion function is nonconvex, which makes the optimization more complicated. Another simple way to estimate θ using the monotonicity is to find the smallest or largest value that $\hat{\mu}$ intersects s, e.g.,

$$\hat{\theta} = \inf\{t : \hat{\mu}(t) \geq 0\}$$

It is an ad hoc way to resolve the problem of multiple solutions. Furthermore, it does not give any guideline how to find such a value. For example, in the case of $\hat{\mu}$ being the Nadaraya-Watson estimator, it is very difficult, if not impossible, to evaluate $\hat{\mu}$ at all point.

The method I introduce in this section provides a simple way to solve monotone equations. I extend the estimator proposed by Dette et al. (2006) in several directions. First, in their papers, the first stage estimator is limited to kernel-type regression estimators. In contrast, I consider a general class of first-stage estimators under primitive conditions. Second, I relax the differentiability assumption in Dette et al. (2006) that μ is twice differentiable at θ. I replace the differentiability condition with a mild condition that μ is not 'flat' on a neighborhood of θ. Third, I relax the compactness of $\mathcal{T}$ and accommodate possibly unbounded $\mathcal{T}$. Lastly, I drop redundant smoothing parameters so that the effect of the smoothing parameter is eliminated in finite samples.

Let F^* be a distribution function chosen by the researcher and T^* be a random variable with the distribution function F^* and a support $\mathcal{T}^*$. Suppose that $\mathcal{T}^*$ is included in $\mathcal{T}$ so that $\mu(T^*)$ is well-defined almost surely. For a fixed s, define

$$\pi(s) = \Pr\left(\mu(T^*) \leq s\right) \tag{38}$$

Since μ is increasing and $s = \mu(\theta)$, $\mu(T^*) \leq s$ is equivalent to $T^* \leq \theta$. Thus we have

$$\begin{aligned} \pi(s) &= \Pr\left(T^* \leq \theta\right) \\ &= F^*(\theta) \end{aligned}$$

or

$$\theta = {F^*}^{-1}(\pi(s)) \tag{39}$$

Since F^* is chosen by the researcher, ${F^*}^{-1}$ is known. Hence, in (39), $\pi(s)$ is the only unknown object. Let $\tilde{\pi}(s)$ be an estimator of $\pi(s)$ obtained by replacing μ in (38) with $\hat{\mu}$. That is,

$$\tilde{\pi}(s) = \Pr\left(\hat{\mu}(T^*) \leq s\right)$$

Even though the distribution of T^* is known, computing $\tilde{\pi}$ is not an easy task. Instead of integrating $\hat{\mu}$ with respect to the true F^*, we approximate the integral by a Monte Carlo integration. Let $\{t_j^* : j = 1, \ldots, n^*\}$ be a generated random sample from F^*. Define a simple estimator of $\pi(s)$ by

$$\hat{\pi}(s) = \frac{1}{n^*} \sum_{j=1}^{n^*} 1\{\hat{\mu}(t_j^*) \leq s\}$$

Relative to $\tilde{\pi}$, $\hat{\pi}$ is affected by the additional randomness from the Monte Carlo approximation. However, one can take n^* to be arbitrarily large and $\hat{\pi}$ can be arbitrarily close to $\tilde{\pi}$. Note that still the randomness in $\hat{\mu}$ remains in $\tilde{\pi}$.

The choice of F^* is important for practical reasons. First, F^* should be selected so that it is easy to simulate a random sample from F^* and calculate the inverse ${F^*}^{-1}$. Second, F^* is a solution selection mechanism. For an extreme example, consider two distribution functions F_1^* and F_2^* that have disjoint supports. Let $\mathcal{T}_1^*$ and $\mathcal{T}_2^*$ be the support of F_1^* and F_2^*, respectively. Suppose that $t_1 \leq t_2$ for any $t_1 \in \mathcal{T}_1^*$ and $t_2 \in \mathcal{T}_2^*$. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the estimator of θ using the same first-stage estimate, but different distribution

functions, F_1^* and F_2^*, respectively. Then it is always true that $\hat{\theta}_1 \leq \hat{\theta}_2$ because $F_1^{*^{-1}}(\pi_1) \leq F_2^{*^{-1}}(\pi_2)$ for any π_1 and π_2 on $[0,1]$. However, under the assumptions state below, F^* has no effect on the first-order asymptotic result that follows.

In Dette et al. (2006), the limit theory is shown by the delta method based on a Taylor expansion. This approach can not be applied to the estimator because I do not require differentiability and not smooth the estimator. Thus the proof of the limit theory in this paper is based on stochastic equicontinuity arguments, which is different from Dette et al. (2006).

To simplify notation, define

$$\mathcal{N}(t, \varepsilon) = \{\tilde{t} \in \mathcal{T} : |t - \tilde{t}| \leq \varepsilon\}$$

The following assumptions are used to derive the limit theory for $\hat{\theta}$.

Assumption A.1. *Let $s \in \mathbb{R}$ be given. Define $\Delta_n(t) = \hat{\mu}(t) - \mu(t)$.*

(a) $\mu(t) < s$ for $t < \theta$ and $\mu(t) > s$ for $t > \theta$. Furthermore,

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \inf_{t \notin \mathcal{N}(\theta, \varepsilon)} |\mu(t) - \mu(\theta)| > 0$$

(b) There exists a positive constant C_μ such that

$$C_\mu = \lim_{\varepsilon \to 0} \inf_{t \in \mathcal{N}(\theta, \varepsilon)} \frac{|\mu(t) - \mu(\theta)|}{|t - \theta|}$$

(c) $E[\Delta_n^2(t)]^{1/2} = O(r_n)$ uniformly over $\mathcal{T}$.

(d) For a positive sequence δ_n such that $\delta_n = o(1)$ and $r_n = o(\delta_n)$,

$$\sup_{t \in \mathcal{N}(\theta, \delta_n)} |\Delta_n(t) - \Delta_n(\theta)| = o(r_n)$$

(e) F^* has a support $\mathcal{T}^* \subset \mathcal{T}$ and a density f^* with respect to the Lebesgue measure; and $f^*(t) > 0$ on a neighborhood of θ.

(f) $\mu(T^*)$ has a bounded density f_μ^*.

(g) $\frac{1}{\sqrt{n^*}} = o(r_n)$

Remark A.1. Assumptions A.1(a) and (b) are actually weaker than the monotonicity and differentiability assumptions, respectively. The second part of Assumptions A.1(a) guarantees that θ is the unique solution to $\mu(\cdot) = s$ and that $\mu(t)$ is reasonably distinguishable from $\mu(\theta)$ if t is away from θ. Assumptions A.1(b) requires that μ is not 'flat' on a neighborhood of θ. If μ is flat, we can not estimate $\hat{\theta}$ accurately no matter how accurate $\hat{\mu}$ is. Conversely, if $C_\mu = \infty$, the estimator is very accurate relative to $\hat{\mu}$. $C_\mu = \infty$ occurs if $\mu(\theta) = 0$ and $\lim_{t \to \theta} \mu(t) \neq 0$, in which there is a jump at θ so it is very easy to distinguish θ from its neighborhood.

Remark A.2. Uniformity of the convergence rate in the mean squared error is much weaker than the uniform convergence of $\hat{\mu}$ itself, e.g., see Assumption 2 in Chernozhukov et al. (2010). For most standard nonparametric methods, Assumptions A.1(c) is easy to verify. Assumption A.1(d) is the so-called stochastic equicontinuity. It regulates that $\hat{\mu}$ around the true value θ be smooth enough.

Remark A.3. Without Assumption A.1(e) that the estimator converges to the true parameter θ. If F^* puts no probability measure on a neighborhood of θ, the estimator does not converge to θ. Assumption A.1(f) implies that $\pi(s)$ is uniformly continuous in s.

Lemma A.1. *Under Assumption A.1,*

$$\hat{\theta} - \theta = C_\mu^{-1} \left[-\Delta(\theta) \right] + o_p(r_n)$$

Proof. By Taylor's theorem and the inverse function theorem, we have

$$\begin{aligned} \hat{\theta} - \theta &= {F^*}^{-1}(\hat{\pi}(s)) - {F^*}^{-1}(\pi(s)) \\ &= \left(f^*(\theta)\right)^{-1} \left(\hat{\pi}(s) - \pi(s)\right) + o\left(|\hat{\pi}(s) - \pi(s)|\right) \end{aligned}$$

Hence, it suffices to show

$$\hat{\pi}(s) - \pi(s) = f^*(\theta) C_\mu^{-1} \left[\Delta(\theta) \right] + o_p(r_n)$$

By the central limit theorem, it is easy to show that

$$\hat{\pi}(s) - \tilde{\pi}(s) = \frac{1}{n^*} \sum_{j=1}^{n^*} 1\left\{ \hat{\mu}(t_j^*) \leq s \right\} - \Pr\left(\hat{\mu}(t_j^*) \leq s \right) = O_p\left(\frac{1}{\sqrt{n^*}} \right)$$

and by AssumptionA.1(e), $\hat{\pi}(s) - \tilde{\pi}(s) = o_p(r_n)$.

Now it remains to show that $\tilde{\pi}(s) - \pi(s) = f^*(\theta) C_\mu^{-1} \left[\Delta(\theta) \right] + o(r_n)$. For the sequence δ_n in Assumption A.1(c), let $\mathcal{N}_n = \mathcal{N}(\theta, \delta_n) \cap \mathcal{T}$. Observe that

$$\begin{aligned} \tilde{\pi}(s) - \pi(s) &= \int 1\left\{ \hat{\mu}(T^*) \leq s \right\} dF^* - \int 1\left\{ \mu(T^*) \leq s \right\} dF^* \\ &= \int_{\mathcal{N}_n} 1\left\{ \hat{\mu}(T^*) \leq s \right\} dF^* - \int_{\mathcal{N}_n} 1\left\{ \mu(T^*) \leq s \right\} dF^* \\ &\quad + \int_{\mathcal{N}_n^c} 1\left\{ \hat{\mu}(T^*) \leq s \right\} dF^* - \int_{\mathcal{N}_n^c} 1\left\{ \mu(T^*) \leq s \right\} dF^* \\ &\equiv \Pi_1 + \Pi_2 \end{aligned}$$

where

$$\begin{aligned} \Pi_1 &= \int_{\mathcal{N}_n} 1\left\{ \hat{\mu}(T^*) \leq s \right\} dF^* - \int_{\mathcal{N}_n} 1\left\{ \mu(T^*) \leq s \right\} dF^* \\ \Pi_2 &= \int_{\mathcal{N}_n^c} 1\left\{ \hat{\mu}(T^*) \leq s \right\} dF^* - \int_{\mathcal{N}_n^c} 1\left\{ \mu(T^*) \leq s \right\} dF^* \end{aligned}$$

I will establish that

$$\Pi_1 = f^*(\theta) C_\mu^{-1} \left[-\Delta_n(\theta) \right] + o_p(r_n) \tag{40}$$

and

$$\Pi_2 = o_p(r_n) \tag{41}$$

Note that, for any $t \in \mathbb{R}$,

$$\begin{aligned} & 1\{\hat{\mu}(t) \leq s\} - \int 1\{\mu(t) \leq s\} \\ = \; & 1\{\hat{\mu}(t) \leq s < \mu(t)\} - 1\{\mu(t) \leq s < \hat{\mu}(t)\} \\ = \; & 1\{\Delta(t) \leq s - \mu(t) < 0\} - 1\{0 \leq s - \mu(t) < \Delta(t)\} \end{aligned} \tag{42}$$

I first show equation (40). Let $C_\Delta = \sup_{t \in \mathcal{N}_n} |\Delta(t) - \Delta(\theta)|$. If $t \in \mathcal{N}_n$, we have $\Delta(\theta) - C_\Delta \leq \Delta(t) \leq \Delta(\theta) + C_\Delta$. Moreover, by AssumptionA.1(a), if $t \in \mathcal{N}_n$,

$$\begin{aligned} |s - \mu(t)| &= |\mu(\theta) - \mu(t)| \\ &\geq C_\mu |\theta - t| \end{aligned}$$

Hence for $t \in \mathcal{N}_n$ and $t \geq \theta$, $\Delta(t) \leq s - \mu(t)$ implies $\Delta(\theta) - C_\Delta \leq C_\mu(\theta - t)$. For $t \in \mathcal{N}_n$ and $t < \theta$, $s - \mu(t) < \Delta(t)$ implies $C_\mu(\theta - t) < \Delta(\theta) + C_\Delta$. Therefore we have

$$\begin{aligned} & -1\{0 \leq C_\mu(\theta - t) < \Delta(\theta) + C_\Delta\} && (43) \\ \leq\ & 1\{\Delta(t) \leq s - \mu(t) < 0\} - 1\{0 \leq s - \mu(t) < \Delta(t)\} \\ \leq\ & 1\{\Delta(\theta) - C_\Delta \leq C_\mu(\theta - t) < 0\} && (44) \end{aligned}$$

Using these bounds and equation (42), bounds for Π_1 are given by

$$\begin{aligned} & -\int 1\{0 \leq C_\mu(\theta - T^*) < \Delta(\theta) + C_\Delta\} dF^* && (45) \\ \leq\ & \Pi_1 \\ \leq\ & \int 1\{\Delta(\theta) - C_\Delta \leq C_\mu(\theta - T^*) < 0\} dF^* && (46) \end{aligned}$$

I will prove that both (45) and (46) have the same limit. First, note that equation (46) can be written as

$$\Pr\left(\theta < T^* \leq \theta + C_\mu^{-1}(-\Delta(\theta) + C_\Delta)\right) = F^*\left(\theta + C_\mu^{-1}(-\Delta(\theta) + C_\Delta)\right) - F^*(\theta)$$

By the Taylor theorem, it becomes

$$f^*(\theta) C_\mu^{-1} \left[-\Delta_n(\theta) + C_\Delta\right] + o_p\left(|-\Delta_n(\theta) + C_\Delta|\right)$$

Since $\Delta(\theta) = O(r_n)$ and $C_\Delta = o(r_n)$ by Assumption A.1(b) and (c), we have that the upper bound converges to

$$f^*(\theta) C_\mu^{-1} \left[-\Delta_n(\theta)\right] + o_p\left(r_n\right)$$

Similarly, the limit of the lower bound is also given by

$$f^*(\theta) C_\mu^{-1} \left[-\Delta_n(\theta)\right] + o_p\left(r_n\right)$$

By the sandwich principle, we have

$$\Pi_1 = f^*(\theta) C_\mu^{-1} \left[-\Delta_n(\theta)\right] + o_p\left(r_n\right)$$

as in (40).

Now I show that $\Pi_2 = o(r_n)$. Since $|\Pi_2| \leq 1$, it suffices to show $E[|\Pi_2|] = o(r_n)$. If $t \in \mathcal{N}_n^c$, by Assumption A.1(a),

$$\begin{aligned} |\mu(\theta) - \mu(t)| &\geq |\mu(\theta) - \mu(\theta \pm \delta_n)| \\ &\geq C_\mu \delta_n \end{aligned}$$

Note that equation (42) implies that

$$\begin{aligned} |1\{\hat{\mu}(t) \leq s\} - 1\{\mu(t) \leq s\}| &\leq 1\{\Delta(t) \leq s - \mu(t) < 0\} + 1\{0 \leq s - \mu(t) < \Delta(t)\} \\ &= 1\{|\Delta(t)| > |s - \mu(t)|\} \end{aligned}$$

Using the fact that $1\{a \geq b\} \leq \frac{a^2}{b^2}$ for any positive real numbers a and b,

$$1\{|\Delta(t)| > |s - \mu(t)|\} \leq \frac{\Delta^2(t)}{(s - \mu(t))^2}$$

Using these inequalities, we obtain

$$\begin{aligned} E\left[|\Pi_2|\right] &\leq E\left[\int_{\mathcal{N}_n^c} |1\{\hat{\mu}(T^*) \leq s\} - 1\{\mu(T^*) \leq s\}| dF^*\right] \\ &\leq \int_{\mathcal{N}_n^c} \frac{E\left[\Delta^2(T^*)\right]}{(s - \mu(T^*))^2} dF^* \end{aligned}$$

Since $E\left[\Delta^2(T^*)\right]^{1/2}$ is uniformly $O(r_n)$ by Assumption A.1(b), the nominator is uniformly $O(r_n^2)$. Let $U = s - \mu(T^*)$ and F_U^* be the implied distribution function. For $T^* \in \mathcal{N}_n^c$, $|U| \geq C_\mu \delta_n$. Hence, by a change-of-variable method,

$$\int_{\mathcal{N}_n^c} \frac{1}{(s - \mu(T^*))^2} dF^* = \int_{C_\mu \delta_n}^{\infty} \frac{1}{U^2} dF_U^* + \int_{-\infty}^{-C_\mu \delta_n} \frac{1}{U^2} dF_U^*$$

By Assumption, the density of F_U^* is bounded and let $C_{f^*} \geq f_U^*$. So we can bound

$$\begin{aligned} \int_{C_\mu \delta_n}^{\infty} \frac{1}{U^2} dF_U^* &\leq \int_{C_\mu \delta_n}^{\infty} \frac{C_{f^*}}{U^2} dU \\ &= \frac{C_{f^*}}{C_\mu \delta_n} \end{aligned}$$

and similarly, $\int_{-\infty}^{-C_\mu \delta_n} \frac{1}{U^2} dF_U^* \leq \frac{C_{f^*}}{C_\mu \delta_n}$. Hence

$$E\left[|\Pi_2|\right] \leq O\left(\frac{r_n^2}{\delta_n}\right) = o(r_n)$$

Thus we get the desired result as in equation (41). □

A.1 Algorithm for the method

In practice, the estimator proceeds as follows:

1. Pick a distribution function F^* to satisfy the required assumptions. Draw a random sample $\{t_j^* : j = 1, 2, \cdots, n^*\}$ from F^*.

2. Compute $\hat{\mu}(t_j^*)$ for $j = 1, 2, \cdots, n^*$. If $\hat{\mu}$ has a closed form formula, it is easy to compute.

3. Compute

$$\hat{\theta} = {F^*}^{-1}\left(\frac{1}{n^*}\sum_{j=1}^{n^*} 1\{\hat{\mu}(t_j^*) \leq s\}\right)$$

4. To solve $\mu(\cdot) = s'$ for a different $s' \neq s$, reuse $\{\hat{\mu}(t_j^*) : j = 1, 2, \cdots, n^*\}$ and just repeat Step 3 unless there is any reason for F^* to fail any of the assumptions.

When choosing F^*, there are several desirable attributes of F^*. For the sake of computational ease, it should be easy to compute the inverse of F^*. Note that ${F^*}^{-1}$ is used twice in the algorithm. First, to draw a random sample from F^*, it is usually carried out by take ${F^*}^{-1}$on a uniform random sample. Also, in Step 3, ${F^*}^{-1}$ is needed. If there is any information such as θ lies in a certain region, one can utilize the information by restricting the support of F^* to the region. On the other hand, if the support of F^* is too large that some points are far from observations, there can be numerical singularity condition when computing $\hat{\mu}$. In general, when there is no additional information, a uniform distribution on the range of observations is a standard choice. Anyhow, the effect of F^* disappears in the limit.

B Lemmas for Stochastic Equicontinuity

In this section, I verify the stochastic equicontinuity conditions using high-level assumptions. To accommodate an abstract random element, I introduce triangular arrays of random processes: let $\xi \in \Xi$ be a generic random element. For each $n = 1, 2, \ldots$, let Ψ_n be a class of measurable functions from Ξ to $\mathbb{R}$. Also assume that Ψ_n can be indexed by a set T. That is,

$$\Psi_n = \{\psi_n(\cdot, t) : t \in T\}$$

for some function $\psi_n : \Xi \times T \to \mathbb{R}$.

For a random sample $\{\xi_1, \ldots, \xi_n\}$, I simplify the notation by writing $\psi_{ni}(t)$ instead of $\psi_n(\xi_i, t)$. The triangular array of random processes $\{\psi_{ni}(t) : t \in T, i = 1, \ldots, n\}$ is said to be i.i.d. within each row if and only if $\{\xi_1, \cdots, \xi_n\}$ is i.i.d. for each n. Call a measurable mapping $\bar{\psi}_n : \Xi \to \mathbb{R}$ an envelope for Ψ_n if $|\psi_{ni}(\xi, t)| \leq \bar{\psi}_n(\xi)$ for all $t \in T$. Define L_p norm between two processes as follows:

$$\|\psi_n(\cdot, t)\|_p = \left[\int |\psi_n(\xi, t)|^p \, d\mathbb{P}(\xi)\right]^{1/p}$$

for $p \in [1, \infty)$ and $\|\psi_n(\cdot, t)\|_\infty$ refers to the supremum norm using the notion of essential supremum.

To deal with the abstract space, I define a measure for the complexity of Ψ_n's.

Definition B.1. For each $\varepsilon > 0$, the L_1-covering number, $N_1(\varepsilon, \Psi_n)$, is the smallest value of J for which there exists $\psi_n^{(1)}, \cdots, \psi_n^{(J)} \in \Psi_n$ such that for any $\psi_n \in \Psi_n$, there always exists j^* that satisfies $\left\|\psi_n - \psi_n^{(j^*)}\right\|_1 \leq \varepsilon$.

Following are the conditions imposed.

Assumption B.1. *Let Ψ_n be a class of real-valued measurable functions defined on Ξ. Suppose that there is an envelope function $\bar{\psi}_n : \Xi \to \mathbb{R}$ for Ψ_n with $\left\|\bar{\psi}_n\right\|_\infty < \infty$. Let$\{\psi_{ni}(t) \in \Psi_n : t \in T, i = 1, \cdots, n\}$ be a triangular array of random processes that are i.i.d. within each row. Suppose that the following are true.*

(a) There exist positive constants C_1 and C_2 such that

$$N_1(\varepsilon \left\|\bar{\psi}_n\right\|_\infty, \Psi_n) \leq C_1 \varepsilon^{-C_2}$$

(b) Letρ_n be a sequence such that $\|\psi_n\|_2 \sqrt{\frac{\log(n)}{n}} = o(\rho_n)$.

Lemma B.1. *Under Assumption B.1,*

$$\sup_{t \in T} \left| \frac{1}{n} \sum_{i=1}^{n} \psi_{ni}(t) - E\left[\psi_{ni}(t)\right] \right| = o_p(\rho_n) \tag{47}$$

Proof. It can be proved using Theorem 2.37 in Pollard (1984). Using his notation, the stochastic process f_{ni} is taken to be $\frac{1}{\left\|\bar{\psi}_n\right\|_\infty} \psi_{ni}$ and let $\mathcal{F}_n = \left\{ \frac{1}{\left\|\bar{\psi}_n\right\|_\infty} \psi_{ni} : \psi_n \in \Psi_n \right\}$. I prove that

$$\sup_{t \in T} \left| \frac{1}{n} \sum_{i=1}^{n} f_{ni}(t) - E\left[f_{ni}(t)\right] \right| = o_p(\left\|\bar{\psi}_n\right\|_\infty \rho_n)$$

To check the covering number condition, let $J_n = N_1(\varepsilon \left\|\bar{\psi}_n\right\|_\infty, \Psi_n)$ and $\{\psi_n^{(j)} : j = 1, \ldots, J_n\}$ be the approximating functions as in Definition B.1. It implies that there exists a subset $\{\psi_n^{(1)}, \cdots, \psi_n^{(J_n)}\}$ of Ψ_n such that for any $\psi_n \in \Psi_n$ there always exists $j \in \{1, \ldots, J_n\}$ satisfying $\left\|\psi_n - \psi_n^{(j)}\right\|_1 \leq \varepsilon \left\|\bar{\psi}_n\right\|_\infty$. By dividing every element of $\{\psi_n^{(1)}, \cdots, \psi_n^{(J_n)}\}$ by $\left\|\bar{\psi}_n\right\|_\infty$, we can construct a subset $\{f_n^{(1)}, \cdots, f_n^{(J_n)}\}$ of $\mathcal{F}_n$. Now consider an arbitrary $f_n \in \mathcal{F}_n$. By hypothesis, for $\psi_n = \left\|\bar{\psi}_n\right\|_\infty f_n$, there exists $j \in \{1, \ldots, J_n\}$ such that $\left\|\psi_n - \psi_n^{(j)}\right\|_1 \leq \varepsilon \left\|\bar{\psi}_n\right\|_\infty$. It is straightforward to show that there must exist

$$\left\|f_n - f_n^{(j)}\right\|_1 = \frac{1}{\left\|\bar{\psi}_n\right\|_\infty} \left\|\psi_n - \psi_n^{(j)}\right\| \leq \varepsilon$$

Thus, we have $N_1(\varepsilon \left\|\bar{\psi}_n\right\|_\infty, \Psi_n) = N_1(\varepsilon, \mathcal{F}_n)$. Therefore, Assumption B.1(a) implies the covering number condition in Pollard (1984).

By construction, it is obvious that $|f_{ni}| \leq 1$ almost surely for all n. Furthermore, $\|f_{ni}\|_2 \leq \frac{\sigma_n}{\left\|\bar{\psi}_n\right\|_\infty}$ for all n. Using his notation, let $\delta_n = \frac{\sigma_n}{\left\|\bar{\psi}_n\right\|_\infty}$ and $\alpha_n = \frac{\left\|\bar{\psi}_n\right\|_\infty \rho_n}{\delta_n^2}$. Observe that

$$n\delta_n^2\alpha_n^2 \quad = \quad n\frac{\rho_n^2}{\sigma_n^2}$$

and by Assumption B.1(b), we have$\log(n) = o\left(n\frac{\rho_n^2}{\sigma_n^2}\right)$. Hence, all the conditions for Theorem 2.37 in Pollard (1984) are satisfied. As a direct result of the theorem, we have

$$\sup_{t \in T} \left| \frac{1}{n} \sum_{i=1}^{n} f_{ni}(t) - E\left[f_{ni}(t)\right] \right| = o_p\left(\frac{\rho_n}{\left\|\bar{\psi}_n\right\|_\infty} \right)$$

By multiplying $\left\|\bar{\psi}_n\right\|_\infty$ on both sides, we have the desired result:

$$\sup_{t \in T} \left| \frac{1}{n} \sum_{i=1}^{n} \psi_{ni}(t) - E\left[\psi_{ni}(t)\right] \right| = o_p\left(\rho_n\right)$$

□

The following lemma is used to verify Assumption A.1 (d) to prove Theorem 3.2.

Lemma B.2. *Let (y_0, w) be given and $\theta = h(y_0, w)$. Under Assumption 3.1 and 3.2,*

$$\sup_{y \in \mathcal{N}(\theta, \delta_n)} \left| \hat{G}_d^{(1)}(y|w) - \hat{G}_d^{(1)}(\theta|w) - G_d^{(1)}(y|w) + G_d^{(1)}(\theta|w) \right| = o_p(r_n)$$

The proof is long and provided in Subsection B.1.

For the estimation of the joint distribution, the following lemma is used to prove Theorem 3.3.

Lemma B.3. *Let $(y_0, y_1) \in \mathbb{R}^2$ and $x \in \mathcal{X}$ be given. Under Assumptions 3.1 and 3.3, we have*

$$\sup_{z \in \mathcal{N}(\zeta, \delta_n)} \left| \hat{G}_d(y|x, z) - \hat{G}_d(y|x, \zeta) - G_d(y|x, z) + G_d(y|x, \zeta) \right| = o_p(r_n) \tag{48}$$

and

$$\sup_{z \in \mathcal{N}(\zeta, \delta_n)} \left| \hat{G}_d^{(1)}(y|x, z) - \hat{G}_d^{(1)}(y|x, \zeta) - G_d^{(1)}(y|x, z) + G_d^{(1)}(y|x, \zeta) \right| = o_p(r_n) \tag{49}$$

B.1 Proof of Lemma B.2

Let

$$\beta = \left(G_d(y|w), \frac{\partial G_d(y|w)}{\partial w'} \right)'$$

and recall that the weighted-least-square estimate of β by solving the weigh problem (21) is given by

$$\hat{\beta} = (\mathbb{W}' \mathbb{K} \mathbb{W})^{-1} \mathbb{W}' \mathbb{K} \mathbb{Y} \tag{50}$$

Note that $\mathbb{Y}$ is a vector of $1\{Y_i \leq y, D_i = 1\}$. Define a random variable $R_i(y)$ by

$$R_i(y) = 1\{Y_i \leq y, D_i = 1\} - G_d(y|W_i)$$

By construction, $E[R_i(y)|\mathbb{W}] = 0$ and $|R_i(y)| \leq 1$.

By the Taylor theorem,

$$G_d(y|W_i) = G_d(y|w) + \frac{\partial G_d(y|w)}{\partial w'}(W_i - w) + Q_i(y, w)$$

where $Q_i(y, w)$ is the remainder term given by

$$Q_i(y, w) = \frac{1}{2}(W_i - w)' \frac{\partial^2 G_d(y|\bar{w}_i)}{\partial w \partial w'}(W_i - w)$$

for some intermediate value $\bar{w}_i$ that lies between W_i and w. By Assumption 3.2(a),

$$|Q_i(y, w)| \leq \frac{C}{2} \|W_i - w\|^2$$

for some constant C.

Now we can write

$$1\{Y_i \leq y, D_i = d\} = G_d(y|w) + \frac{\partial G_d(y|w)}{\partial w'}(W_i - w) + Q_i(y,w) + R_i(y)$$

or

$$\mathbb{Y} = \mathbb{W}\beta + \mathbb{Q} + \mathbb{R} \tag{51}$$

where $\mathbb{Q}$ and $\mathbb{R}$ are the vectors of Q_i's and R_i's, respectively. Plug the expression (51) into the formula (50) to obtain

$$\hat{\beta} - \beta = (\mathbb{W}'\mathbb{K}\mathbb{W})^{-1}\mathbb{W}'\mathbb{K}(\mathbb{Q} + \mathbb{R}) \tag{52}$$

To study the asymptotic behaviors of $\hat{\beta} - \beta$, it is useful to adopt the following normalization. Let $\mathbb{B}$ be a $(m+1) \times (m+1)$ diagonal matrix defined by

$$\mathbb{B} = \begin{pmatrix} 1 & 0 \\ 0 & bI_m \end{pmatrix}$$

where I_m is an $m \times m$ identity matrix and b is the bandwidth. Rewrite equation (52) as

$$\hat{\beta} - \beta = \mathbb{B}^{-1}\left(\frac{1}{n}\mathbb{W}_b'\mathbb{K}\mathbb{W}_b\right)^{-1}\frac{1}{n}\mathbb{W}_b'\mathbb{K}(\mathbb{Q} + \mathbb{R}) \tag{53}$$

where $\mathbb{W}_b = \mathbb{W}\mathbb{B}^{-1}$, the vector of deviations of W_i's from w normalized by the bandwidth, i.e.,

$$\mathbb{W}_b = \begin{pmatrix} 1 & \left(\frac{W_i - w}{b}\right)' \\ \vdots & \vdots \\ 1 & \left(\frac{W_n - w}{b}\right)' \end{pmatrix}$$

We are interested only in $\hat{G}_d^{(1)}(y|w) - G_d^{(1)}(y|w)$. Since Z_i is the mth element of W_i, $\hat{G}_d^{(1)}(y|w) - G_d^{(1)}(y|w)$ is the $(m+1)$th element of $\hat{\beta} - \beta$. Hence, I particularly let $\omega_n(w)$ be the $(m+1)$th diagonal element in $\left(\frac{1}{n}\mathbb{W}_b'\mathbb{K}\mathbb{W}_b\right)^{-1}$. Further, the $(m+1)$th element of $\frac{1}{n}\mathbb{W}_b'\mathbb{K}(\mathbb{Q} + \mathbb{R})$ can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i - z}{b}\right)K_b(W_i - w)\left[R_i(y) + Q_i(y,w)\right]$$

Therefore, the formula for $\hat{G}_d^{(1)}(y|w) - G_d^{(1)}(y|w)$ is

$$\hat{G}_d^{(1)}(y|w) - G_d^{(1)}(y|w) = \frac{\omega_n(w)}{bn}\sum_{i=1}^{n}\left(\frac{Z_i - z}{b}\right)K_b(W_i - w)\left[R_i(y) + Q_i(y,w)\right] \tag{54}$$

For each $t \in [-1, 1]$, define

$$\psi_{ni}(t) = \frac{1}{b}\left(\frac{Z_i - z}{b}\right)K_b(W_i - w)\left[R_i(\theta + \delta_n t) + Q_i(\theta + \delta_n t, w) - R_i(\theta) - Q_i(\theta, w)\right] \tag{55}$$

Then, for any $y \in \mathcal{N}(\theta, \delta_n)$,

$$\hat{G}_d^{(1)}(y|w) - \hat{G}_d^{(1)}(\theta|w) - G_d^{(1)}(y|w) + G_d^{(1)}(\theta|w) = \frac{\omega_n(w)}{n} \sum_{i=1}^{n} \psi_{ni}(t)$$

for some $t \in [-1, 1]$ such that $y = \theta + \delta_n t$.

Therefore, what we want to show becomes

$$\sup_{t \in [-1,1]} \left| \frac{\omega_n(w)}{n} \sum_{i=1}^{n} \psi_{ni}(t) \right| = o_p(r_n)$$

By the results in Ruppert and Wand (1994) and Lu (1996), we know that

$$\omega_n(w) = \frac{1}{\kappa_2 f_W(w)} + o_p(1)$$

and

$$E[\psi_{ni}(t)] = b^2 \kappa_2 f_W(w) \left[B_1(\theta + \delta_n t, d, w) - B_1(\theta, d, w) \right] + o(r_n)$$

where B_1 is the leading term of the bias as defined in Theorem 3.1. By Assumption 3.1, B_1 is bounded and continuous in y. Hence, as $\delta_n = o(1)$,

$$B_1(\theta + \delta_n t, d, w) - B_1(\theta, d, w) = o(1)$$

uniformly and so $E[\psi_{ni}(t)]$ is $o(r_n)$ uniformly over $t \in [-1, 1]$.

Note that

$$\begin{aligned} \sup_{t \in [-1,1]} \left| \frac{\omega_n(w)}{n} \sum_{i=1}^{n} \psi_{ni}(t) \right| &= |\omega_n(w)| \sup_{t \in [-1,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \psi_{ni}(t) \right| \\ &\leq |\omega_n(w)| \sup_{t \in [-1,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \psi_{ni}(t) - E[\psi_{ni}(t)] \right| \\ &\quad + |\omega_n(w)| \sup_{t \in [-1,1]} |E[\psi_{ni}(t)]| \end{aligned}$$

Now it remains to show that

$$\sup_{t \in [-1,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \psi_{ni}(t) - E[\psi_{ni}(t)] \right| = o_p(r_n)$$

The proof closely follows Lemma B.1. Let $\Psi_n = \{\psi_n(t) : t \in [-1, 1]\}$.

Lemma B.4. *Under the assumptions of Lemma B.2,*

$$\sup_{t \in [-1,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \psi_{ni}(t) - E[\psi_{ni}(t)] \right| = o_p(r_n)$$

Proof. I verify each condition for Lemma B.1. C will denote a generic positive constant with different values in different places. First, I claim that $Q_i(y, w)$ is of negligible order compared to the term arising from $R_i(y)$.

By Assumption 3.2(a), there exists a positive constant C such that

$$(W_i - w)' \frac{\partial^2 G_d(y|\bar{w}_i)}{\partial w \partial w'} (W_i - w) \leq C(W_i - w)'(W_i - w)$$

Using the standard change-of-variable method for kernel estimators, let $W_i = w + bU$ and $\bar{w}_i = w + b\bar{u}$. Then,

$$\begin{aligned} Q_i(y, w) &= \frac{b^2}{2} U' \frac{\partial^2 G_d(y|w + b\bar{u})}{\partial w \partial w'} U \\ &\leq Cb^2 U'U \end{aligned}$$

for some positive constant C. By Assumption 3.2 (c), $U'U$ is bounded when multiplied by the kernel function. While $R_i(y)$ does not shrink, the order of $Q_i(y, w)$ is b^2, which converges to zero. Therefore, the leading term of $\psi_{ni}(t)$ is associated with $R_i(y)$.

STEP 1. We know that $|R_i(y)| \leq 1$ for any y. Also note that

$$\left| \left(\frac{Z_i - z}{b} \right) K_b(W_i - w) \right| \leq \frac{1}{b^m} \sup_{u \in \mathbb{R}^m} |uK(u)|$$

Therefore,

$$\left| \frac{1}{b} \left(\frac{Z_i - z}{b} \right) K_b(W_i - w) \left\{ R_i(\theta + \delta_n t) - R_i(\theta) \right\} \right| \leq \frac{C}{b^{m+1}}$$

for a positive constant C. Since $Q_i(y, w)$ of negligible order, $\left\| \bar{\psi}_n \right\|_\infty$ has an order of $\frac{1}{b^{m+1}}$.

STEP 2. In this step, I calculate the covering number. To calculate the L_1-distance between two elements of Ψ_n, take arbitrary t and t' from $[-1, 1]$. Without loss of generality, suppose that $t < t'$. Then

$$\begin{aligned} R_i(\theta + \delta_n t) - R_i(\theta + \delta_n t') &= 1\{\theta + \delta_n t < Y_i \leq \theta + \delta_n t', D_i = d\} \\ &\quad - G_d(\theta + \delta_n t'|W_i) + G_d(\theta + \delta_n t|W_i) \end{aligned}$$

Thus,

$$\begin{aligned} E\left[|R_i(\theta + \delta_n t) - R_i(\theta + \delta_n t')| \mid W_i \right] &\leq 2 \left| G_d(\theta + \delta_n t'|W_i) - G_d(\theta + \delta_n t|W_i) \right| \\ &\leq 2 \sup_{(y,w) \in \mathbb{R} \times \mathcal{W}} \left| \frac{\partial G_d(y|w)}{\partial y} \right| \delta_n |t - t'| \end{aligned}$$

and by the same algebra as in STEP 1, we can show that

$$E \left| \frac{1}{b} \left(\frac{Z_i - z}{b} \right) K_b(W_i - w) \left\{ R_i(\theta + \delta_n t) - R_i(\theta + \delta_n t') \right\} \right| \leq \frac{C}{b^{m+1}} |t - t'|$$

Also, it follows from STEP 1 that the term containing Q_i is ignorable. Hence, we have

$$\left\| \psi_{ni}(t) - \psi_{ni}(t') \right\|_1 \leq \frac{C}{b^{m+1}} |t - t'|$$

For any $\varepsilon \in (0, 1)$, let $\varepsilon_n = b^{m+1} \left\| \bar{\psi}_n \right\|_\infty \varepsilon$ and let J be the greatest integer smaller than or equal to $\frac{2C}{\varepsilon_n}$. Always we can take a set $\{t_j \in [-1, 1] : j = 1, 2, \cdots, J\}$ such that for any $t \in [-1, 1]$ there always exists $j^* \in \{1, 2, \cdots, J\}$ satisfying $|t - t_{j^*}| \leq \frac{\varepsilon_n}{C}$. Then, for any $\psi_{ni}(t) \in \Psi_n$, there always exists $\psi_{ni}(t_{j^*}) \in \Psi_n$

such that

$$\begin{aligned}\|\psi_{ni}(t) - \psi_{ni}(t_{j^*})\|_1 &\leq \frac{C}{b^{m+1}}|t - t_{j^*}| \\ &\leq \frac{\varepsilon_n}{b^{m+1}} = \left\|\bar{\psi}_n\right\|_\infty \varepsilon\end{aligned}$$

Hence $\{\psi_{ni}^{(j)} \in \Psi_n : \psi_{ni}^{(j)} = \psi_n(\cdot, t_j), j = 1, 2, \cdots, J\}$ satisfies the condition in Definition B.1. Since $J \leq \frac{2C}{\varepsilon_n} = \frac{2C}{\bar{\kappa}_1 \varepsilon}$, the assumption is verified.

STEP 3. Now I compute the order of $E\left[\psi_{ni}^2(t)\right]^{1/2}$. The algebra is exactly the same as the algebra used to calculate the variance in Theorem 3.1. Note that

$$\begin{aligned}E[R_i^2(\theta + \delta_n t)|W_i] &= var[1\{Y_i \leq \theta + \delta_n t, D_i = d\} - 1\{Y_i \leq \theta, D_i = d\}|W_i] \\ &\leq |G_d(\theta + \delta_n|W_i) - G_d(\theta|W_i)| \\ &\leq C\delta_n\end{aligned}$$

for some constant C. Hence, the leading term of $E[\psi_{ni}^2(t)]$ is

$$\begin{aligned}E\left[\frac{1}{b^2}\left(\frac{Z_i - z}{b}\right)^2 K_b^2(W_i - w) R_i^2(\theta + \delta_n t)\right] &= \frac{C\delta_n}{b^2} E\left[\left(\frac{Z_i - z}{b}\right)^2 K_b^2(W_i - w)\right] \\ &\leq \frac{C\delta_n}{b^{m+2}}\end{aligned}$$

Still Q_i is of smaller order than R_i by the same reasoning as in the previous steps. Therefore, $\|\psi_{ni}\|_2$ has an order of $\sqrt{\frac{\delta_n}{b^{m+2}}}$.

STEP 4. Finally I verify Assumption B.1(b). In our case, ρ_n is replaced with r_n. We want to show that

$$\|\psi_{ni}\|_2 \sqrt{\frac{\log(n)}{n}} = o(r_n)$$

Since $r_n \geq \frac{1}{\sqrt{nb^{m+2}}}$, by Assumption 3.2(d),

$$\|\psi_{ni}\|_2 \sqrt{\frac{\log(n)}{n}} = \frac{\sqrt{\delta_n \log(n)}}{\sqrt{nb^{m+2}}} = o(r_n)$$

as desired. □

B.2 Proof of Lemma B.3

Equation (48) and (49) can be proved along the same lines. Since the proof for equation (49) is more involved, I only prove the equation. For ease of notation, let $w_\zeta = (x', z)'$. From the expression (53), we have

$$\hat{G}_d^{(1)}(y|w) - G_d^{(1)}(y|w) = \frac{\omega_n(w)}{n} \sum_{i=1}^n \frac{1}{b}\left(\frac{Z_i - z}{b}\right) K_b(W_i - w)\left\{R_i(y) + Q_i(y, w)\right\} \tag{56}$$

Corollary 2 in Masry (1996) implies that

$$\sup_{z\in\mathcal{N}(\zeta,\delta_n)}\left|\omega_n(w)-\frac{1}{\kappa_2 f_W(w)}\right|=o_p(1)$$

Since f_W is bounded and continuous and $\delta_n=o(1)$,

$$\sup_{z\in\mathcal{N}(\zeta,\delta_n)}\left|\frac{1}{\kappa_2 f_W(w)}-\frac{1}{\kappa_2 f_W(w_\zeta)}\right|=o(1)$$

Thus,

$$\sup_{z\in\mathcal{N}(\zeta,\delta_n)}\left|\omega_n(w)-\frac{1}{\kappa_2 f_W(w_\zeta)}\right|=o_p(1)$$

For each $t\in[-1,1]$, define

$$\begin{aligned}\phi_{ni}(t) &= \frac{1}{b}\left\{\left(\frac{Z_i-z_t}{b}\right)K_b(W_i-w_t)-\left(\frac{Z_i-\zeta}{b}\right)K_b(W_i-w_\zeta)\right\}R_i(y)\\ &\quad+\frac{1}{b}\left(\frac{Z_i-z_t}{b}\right)K_b(W_i-w_t)Q_i(y,w_t)\\ &\quad-\frac{1}{b}\left(\frac{Z_i-\zeta}{b}\right)K_b(W_i-w_\zeta)Q_i(y,w_\zeta)\end{aligned}$$

where $z_t=\zeta+\delta_n t$ and $w_t=(x',z_t)'$. Let Φ_n be the family of ϕ_{ni}'s, i.e.,

$$\Phi_n=\{\phi_{ni}(t):t\in[-1,1]\}$$

Since

$$\begin{aligned}&\sup_{z\in\mathcal{N}(\zeta,\delta_n)}\left|\hat{G}_d^{(1)}(y|w)-\hat{G}_d^{(1)}(y|w_\zeta)-G_d^{(1)}(y|w)+G_d^{(1)}(y|x,w_\zeta)\right|\\ =\;&\left(\frac{1}{\kappa_2 f_W(w_\zeta)}+o_p(1)\right)\sup_{t\in[-1,1]}\left|\frac{1}{n}\sum_{i=1}^n\phi_{ni}(t)\right|\\ \le\;&\left(\frac{1}{\kappa_2 f_W(w_\zeta)}+o_p(1)\right)\sup_{t\in[-1,1]}\left|\frac{1}{n}\sum_{i=1}^n\phi_{ni}(t)-E[\phi_{ni}(t)]\right|\\ &+\left(\frac{1}{\kappa_2 f_W(w_\zeta)}+o_p(1)\right)\sup_{t\in[-1,1]}|E[\phi_{ni}(t)]|\end{aligned}$$

it suffices to show

$$\sup_{t\in[-1,1]}\left|\frac{1}{n}\sum_{i=1}^n\phi_{ni}(t)-E[\phi_{ni}(t)]\right|=o_p(r_n) \tag{57}$$

and

$$\sup_{t\in[-1,1]}|E[\phi_{ni}(t)]|=o(r_n) \tag{58}$$

(57) is shown in Lemma below. To see (58), note that $\frac{1}{\kappa_2 f_W(w_\zeta)}E[\phi_{ni}(t)]$ is the bias of $\hat{G}_d^{(1)}(y|w_t)-\hat{G}_d^{(1)}(y|w_\zeta)$. By Theorem 3.1, we know

$$E[\phi_{ni}(t)]=\kappa_2 f_W(w_\zeta)\left[b^2\left(B_1(y,w_t)-B_1(y,w_\zeta)\right)+o(b^2)\right]$$

Since B_1 is continuous in w and $|w_t - w_\zeta| \leq \delta_n$, it follows that

$$\sup_{t\in[-1,1]} |E[\phi_{ni}(t)]| = o(r_n)$$

Lemma B.5. *Under the assumptions of Theorem 3.1 and Lemma B.3,*

$$\sup_{t\in[-1,1]} \left| \frac{1}{n}\sum_{i=1}^{n} \phi_{ni}(t) - E\left[\phi_{ni}(t)\right] \right| = o_p(r_n)$$

Proof. It is a result of Lemma B.1. As in Lemma B.4, $Q_i(y,w)$ is of smaller order than $R_i(y)$, so will be ignored when calculating convergence rates. Also, C will denote a generic positive constant, which may represent different values in different places.

STEP 1. For $z \in \mathcal{N}(\zeta, \delta_n)$,

$$\begin{aligned} \left| \left(\frac{Z_i - z}{b}\right) K_b(W_i - w) - \left(\frac{Z_i - \zeta}{b}\right) K_b(W_i - w_\zeta) \right| &\leq \frac{C}{b^m} |\frac{z-\zeta}{b}| \\ &\leq \frac{C\delta_n}{b^{m+1}} \end{aligned}$$

for some positive constant C. Since $|R_i(y)| \leq 1$,

$$\left| \frac{1}{b} \left\{ \left(\frac{Z_i - z}{b}\right) K_b(W_i - w) - \left(\frac{Z_i - \zeta}{b}\right) K_b(W_i - w_\zeta) \right\} R_i(y) \right| \leq \frac{C\delta_n}{b^{m+2}}$$

I will set the envelope function $\bar{\phi}_n = \frac{\bar{C}\delta_n}{b^{m+2}}$ for a sufficiently large positive constant $\bar{C}$.

STEP 2. I calculate the covering number. For $t, t' \in [-1,1]$, by the same algebra used in STEP 1,

$$|\phi_{ni}(t) - \phi_{ni}(t')| \leq \frac{C\delta_n}{b^{m+2}} |t - t'|$$

Thus, $\|\phi_{ni}(t) - \phi_{ni}(t')\|_1 \leq \frac{C_1\delta_n}{b^{m+2}}|t-t'|$ for a positive constant C_1.

Let $\varepsilon > 0$ be given. And let J be the largest integer such that $J \leq \frac{2C_1\varepsilon}{\bar{C}}$. We can take a partition $\{t_1, t_2, \cdots, t_J\}$ on $[-1,1]$ such that $\{\mathcal{N}(t_j, \frac{\bar{C}\varepsilon}{C_1}) : j = 1, 2, \cdots, J\}$ covers $[-1,1]$. Consider a finite subset of Φ_n, $\{\phi_{ni}(t_1), \phi_{ni}(t_2), \cdots, \phi_{ni}(t_J)\}$. For any $\phi_{ni}(t) \in \Phi_n$, there exists $j^* \leq J$ such that $|t - t_{j^*}| \leq \frac{\bar{C}\varepsilon}{C_1}$. Then,

$$\begin{aligned} \|\phi_{ni}(t) - \phi_{ni}(t_{j^*})\|_1 &\leq \frac{C_1\delta_n}{b^{m+2}}|t - t_{j^*}| \\ &\leq \frac{\bar{C}\delta_n}{b^{m+2}}\varepsilon \\ &= \left\|\bar{\phi}_n\right\|_\infty \varepsilon \end{aligned}$$

Thus, Assumption B.1 (a) is satisfied.

STEP 3. I calculate the order of $\|\phi_{ni}\|_2$. Since the terms associated with $Q_i(y,w)$ are of negligible order,

the leading term of $|\phi_{ni}(t)|^2$ is

$$\begin{aligned}
& \frac{1}{b^2}\left\{\left(\frac{Z_i - z_t}{b}\right) K_b(W_i - w_t) - \left(\frac{Z_i - \zeta}{b}\right) K_b(W_i - w_\zeta)\right\}^2 R_i^2(y) \\
= \; & \frac{1}{b^2}\left(\frac{Z_i - z_t}{b}\right)^2 K_b^2(W_i - w_t) R_i^2(y) + \frac{1}{b^2}\left(\frac{Z_i - \zeta}{b}\right)^2 K_b^2(W_i - w_\zeta) R_i^2(y) \\
& -\frac{2}{b^2}\left(\frac{Z_i - z_t}{b}\right)\left(\frac{Z_i - \zeta}{b}\right) K_b(W_i - w_t) K_b(W_i - w_\zeta) R_i^2(y)
\end{aligned}$$

The same algebra used to calculate the variances in Theorem 3.1 yields

$$E\left[\frac{1}{b^2}\left(\frac{Z_i - z_t}{b}\right)^2 K_b^2(W_i - w_t) R_i^2(y)\right] = \frac{\kappa_2^2 f_W^2(w_t)}{b^{m+2}} V_1(y, w_t) + o\left(\frac{1}{b^{m+2}}\right)$$

and

$$E\left[\frac{1}{b^2}\left(\frac{Z_i - \zeta}{b}\right)^2 K_b^2(W_i - w_\zeta) R_i^2(y)\right] = \frac{\kappa_2^2 f_W^2(w_\zeta)}{b^{m+2}} V_1(y, w_\zeta) + o\left(\frac{1}{b^{m+2}}\right)$$

where V_1 is the leading term of variance defined in Theorem 3.1. By Assumption3.1, V_1 is continuous in w. Since $|w_t - w_\zeta| \leq \delta_n = o(1)$,

$$f_W^2(w_t) V_1(y, w_t) = f_W^2(w_\zeta) V_1(y, w_\zeta) + o(1)$$

Assumption 3.3 implies that there exists a positive constant C such that

$$\begin{aligned}
|K_b(W_i - w_t) - K_b(W_i - w_\zeta)| &\leq \frac{C}{b^m}|w_t - w_\zeta| \\
&= \frac{C}{b^m}\delta_n
\end{aligned}$$

for any $t \in [-1, 1]$. By Assumption 3.3, we have $\frac{\delta_n}{b} = o(1)$. Thus it follows that

$$\left(\frac{Z_i - z_t}{b}\right)\left(\frac{Z_i - \zeta}{b}\right) K_b(W_i - w_t) K_b(W_i - w_\zeta) = \left(\frac{Z_i - \zeta}{b}\right)^2 K_b^2(W_i - w_\zeta) + o\left(\frac{1}{b^m}\right)$$

and a straightforward but tedious calculus yields

$$\begin{aligned}
& E\left[\frac{1}{b^2}\left(\frac{Z_i - z_t}{b}\right)\left(\frac{Z_i - \zeta}{b}\right) K_b(W_i - w_t) K_b(W_i - w_\zeta) R_i^2(y)\right] \\
= \; & \frac{\kappa_2^2 f_W^2(w_\zeta)}{b^{m+2}} V_1(y, w_\zeta) + o\left(\frac{1}{b^{m+2}}\right)
\end{aligned}$$

Therefore, the order of $\|\phi_{ni}\|_2$ is $\sqrt{\frac{\delta_n}{b^{m+2}}}$. By Assumption 3.3,

$$\begin{aligned}
\|\phi_{ni}\|_2 \sqrt{\frac{\log(n)}{n}} &= \frac{1}{\sqrt{n b^{m+2}}}\sqrt{\delta_n \log(n)} \\
&= o(1)
\end{aligned}$$

Hence, Assumption B.1(b) is satisfied. □

Table 1: Estimation of the utility function

	$n = 200$			$n = 500$			$n = 1000$		
b	RMSE	bias	SE	RMSE	bias	SE	RMSE	bias	SE
				Design (A)					
0.5	0.768	0.073	0.765	0.409	0.036	0.408	0.281	0.004	0.281
1	0.398	0.001	0.398	0.232	0.010	0.232	0.172	0.002	0.172
1.5	0.368	-0.003	0.368	0.222	0.003	0.222	0.161	0.004	0.161
2	0.367	-0.005	0.367	0.223	0.001	0.223	0.161	0.004	0.161
2.5	0.369	-0.006	0.369	0.226	0.000	0.226	0.162	0.004	0.162
				Design (B)					
0.5	0.387	0.126	0.366	0.172	0.086	0.148	0.128	0.074	0.104
1	0.220	0.079	0.205	0.145	0.081	0.121	0.115	0.073	0.089
1.5	0.219	0.067	0.209	0.144	0.072	0.125	0.113	0.067	0.091
2	0.224	0.061	0.215	0.147	0.067	0.130	0.114	0.063	0.095
2.5	0.228	0.058	0.220	0.149	0.065	0.135	0.115	0.061	0.097
				Design (C)					
0.5	0.419	-0.015	0.419	0.247	0.017	0.246	0.182	0.006	0.182
1	0.303	0.008	0.303	0.182	0.027	0.180	0.136	0.022	0.135
1.5	0.299	0.014	0.299	0.183	0.031	0.181	0.136	0.030	0.132
2	0.303	0.018	0.302	0.187	0.034	0.184	0.139	0.033	0.135
2.5	0.308	0.021	0.307	0.191	0.034	0.187	0.142	0.035	0.137
				Design (D)					
0.5	1.056	0.147	1.045	0.643	0.044	0.642	0.448	0.032	0.447
1	0.550	0.032	0.549	0.331	0.023	0.330	0.225	0.007	0.225
1.5	0.489	0.017	0.489	0.295	0.016	0.294	0.200	0.004	0.200
2	0.478	0.008	0.478	0.288	0.014	0.288	0.196	0.002	0.196
2.5	0.480	0.006	0.480	0.288	0.014	0.288	0.195	0.002	0.195
				Design (E)					
0.5	0.840	0.268	0.797	0.394	0.113	0.377	0.238	0.087	0.221
1	0.356	0.088	0.345	0.210	0.074	0.197	0.148	0.061	0.135
1.5	0.321	0.065	0.314	0.197	0.062	0.187	0.137	0.050	0.127
2	0.315	0.055	0.310	0.198	0.056	0.190	0.136	0.046	0.128
2.5	0.317	0.050	0.313	0.200	0.053	0.192	0.137	0.043	0.130
				Design (F)					
0.5	0.853	0.079	0.850	0.529	0.013	0.528	0.383	0.025	0.382
1	0.465	0.030	0.465	0.293	0.031	0.291	0.210	0.022	0.209
1.5	0.425	0.028	0.424	0.266	0.032	0.264	0.187	0.024	0.186
2	0.422	0.027	0.421	0.263	0.033	0.261	0.184	0.028	0.182
2.5	0.421	0.026	0.420	0.265	0.033	0.263	0.184	0.031	0.182

Table 2: Simulation results for the median treatment effect

	$n = 200$			$n = 500$			$n = 1000$		
b	RMSE	bias	SE	RMSE	bias	SE	RMSE	bias	SE
				Design (A)					
0.5	0.376	-0.012	0.375	0.236	0.015	0.236	0.185	-0.007	0.185
1	0.275	-0.008	0.275	0.168	0.007	0.168	0.113	0.005	0.113
1.5	0.253	-0.008	0.252	0.155	0.005	0.155	0.104	0.008	0.103
2	0.249	-0.007	0.248	0.157	0.004	0.157	0.107	0.006	0.107
2.5	0.250	-0.006	0.250	0.158	0.004	0.158	0.110	0.005	0.110
				Design (B)					
0.5	0.182	-0.002	0.182	0.119	0.000	0.119	0.090	-0.012	0.090
1	0.175	-0.006	0.175	0.118	-0.005	0.118	0.084	-0.014	0.083
1.5	0.168	0.002	0.168	0.116	0.006	0.115	0.084	0.001	0.084
2	0.168	0.017	0.167	0.119	0.028	0.115	0.089	0.032	0.083
2.5	0.173	0.027	0.170	0.125	0.043	0.117	0.096	0.049	0.083
				Design (C)					
0.5	0.277	0.007	0.277	0.167	0.032	0.164	0.125	0.024	0.122
1	0.217	0.039	0.213	0.137	0.056	0.125	0.104	0.056	0.087
1.5	0.206	0.053	0.199	0.138	0.068	0.120	0.113	0.073	0.086
2	0.205	0.061	0.195	0.140	0.072	0.120	0.117	0.078	0.087
2.5	0.205	0.063	0.195	0.141	0.072	0.121	0.116	0.075	0.089
				Design (D)					
0.5	0.450	0.011	0.450	0.286	0.008	0.286	0.194	-0.005	0.194
1	0.334	-0.001	0.334	0.213	0.004	0.213	0.143	-0.001	0.143
1.5	0.312	-0.001	0.312	0.199	0.002	0.199	0.134	0.000	0.134
2	0.306	-0.002	0.306	0.196	0.003	0.196	0.132	0.000	0.132
2.5	0.305	-0.001	0.305	0.196	0.004	0.196	0.133	0.000	0.133
				Design (E)					
0.5	0.260	-0.010	0.260	0.154	-0.009	0.154	0.110	-0.015	0.109
1	0.226	-0.009	0.226	0.148	-0.007	0.148	0.108	-0.007	0.108
1.5	0.223	0.003	0.223	0.146	0.013	0.145	0.105	0.016	0.104
2	0.222	0.014	0.222	0.147	0.027	0.144	0.108	0.032	0.103
2.5	0.224	0.019	0.223	0.150	0.035	0.146	0.112	0.042	0.104
				Design (F)					
0.5	0.372	0.012	0.372	0.231	0.019	0.230	0.165	0.011	0.164
1	0.282	0.020	0.281	0.182	0.037	0.179	0.132	0.037	0.127
1.5	0.268	0.030	0.267	0.176	0.046	0.170	0.129	0.048	0.120
2	0.267	0.036	0.265	0.174	0.049	0.167	0.128	0.049	0.118
2.5	0.266	0.038	0.263	0.173	0.050	0.166	0.127	0.048	0.117

Table 3: Simulation results for Heckman's two-step estimator

	$n = 200$			$n = 500$			$n = 1000$		
	RMSE	bias	SE	RMSE	bias	SE	RMSE	bias	SE
				Design (A)					
$\hat{\alpha}_0$	0.180	0.009	0.180	0.110	0.002	0.110	0.079	0.001	0.079
$\hat{\alpha}_1$	0.178	-0.005	0.178	0.114	0.002	0.114	0.078	-0.002	0.078
$\hat{\beta}_0$	0.109	0.000	0.109	0.071	-0.001	0.071	0.048	-0.001	0.048
$\hat{\beta}_1$	0.107	0.006	0.107	0.071	0.003	0.071	0.049	0.001	0.049
				Design (B)					
$\hat{\alpha}_0$	0.334	-0.257	0.214	0.300	-0.267	0.137	0.281	-0.265	0.094
$\hat{\alpha}_1$	0.339	-0.264	0.213	0.298	-0.265	0.136	0.290	-0.273	0.100
$\hat{\beta}_0$	0.137	-0.025	0.134	0.096	-0.028	0.092	0.068	-0.028	0.062
$\hat{\beta}_1$	0.146	0.031	0.142	0.094	0.030	0.088	0.071	0.030	0.064
				Design (C)					
$\hat{\alpha}_0$	0.239	0.197	0.136	0.208	0.190	0.084	0.198	0.187	0.063
$\hat{\alpha}_1$	0.300	0.260	0.149	0.283	0.267	0.093	0.273	0.265	0.064
$\hat{\beta}_0$	0.091	0.029	0.087	0.063	0.027	0.057	0.047	0.027	0.039
$\hat{\beta}_1$	0.216	-0.186	0.111	0.199	-0.186	0.072	0.195	-0.188	0.050
				Design (D)					
$\hat{\alpha}_0$	0.225	0.000	0.225	0.142	0.001	0.142	0.098	0.004	0.097
$\hat{\alpha}_1$	0.223	-0.008	0.223	0.139	-0.001	0.139	0.099	-0.003	0.099
$\hat{\beta}_0$	0.114	-0.003	0.114	0.071	-0.001	0.071	0.048	0.001	0.048
$\hat{\beta}_1$	0.107	-0.003	0.107	0.070	-0.002	0.070	0.050	0.000	0.050
				Design (E)					
$\hat{\alpha}_0$	0.340	-0.176	0.291	0.250	-0.181	0.173	0.217	-0.175	0.128
$\hat{\alpha}_1$	0.343	-0.178	0.293	0.261	-0.190	0.179	0.225	-0.183	0.131
$\hat{\beta}_0$	0.139	-0.010	0.138	0.092	-0.008	0.091	0.063	-0.009	0.062
$\hat{\beta}_1$	0.144	0.004	0.144	0.089	0.012	0.088	0.066	0.011	0.065
				Design (F)					
$\hat{\alpha}_0$	0.229	0.125	0.192	0.176	0.126	0.122	0.148	0.124	0.082
$\hat{\alpha}_1$	0.274	0.177	0.209	0.226	0.187	0.128	0.201	0.179	0.092
$\hat{\beta}_0$	0.093	0.014	0.092	0.059	0.015	0.057	0.043	0.015	0.040
$\hat{\beta}_1$	0.195	-0.161	0.109	0.174	-0.160	0.069	0.166	-0.159	0.050

Table 4: Simulation results for Andrews-Schafgans' semiparametric estimator

		$n = 200$			$n = 500$			$n = 1000$		
	γ_n	RMSE	bias	SE	RMSE	bias	SE	RMSE	bias	SE
					Design (A)					
	2%	0.482	0.039	0.480	0.315	0.040	0.313	0.218	0.012	0.218
$\tilde{\alpha}_0$	5%	0.316	0.041	0.314	0.205	0.042	0.200	0.148	0.036	0.144
	10%	0.236	0.065	0.227	0.155	0.066	0.140	0.123	0.063	0.105
	2%	0.494	0.019	0.493	0.318	0.021	0.317	0.219	0.021	0.218
$\tilde{\alpha}_1$	5%	0.322	0.035	0.320	0.209	0.038	0.205	0.149	0.043	0.143
	10%	0.245	0.062	0.237	0.159	0.062	0.146	0.121	0.063	0.104
					Design (D)					
	2%	0.512	0.064	0.508	0.331	0.045	0.328	0.230	0.040	0.226
$\tilde{\alpha}_0$	5%	0.342	0.080	0.333	0.219	0.061	0.210	0.160	0.059	0.149
	10%	0.253	0.092	0.236	0.172	0.078	0.154	0.135	0.082	0.107
	2%	0.512	0.032	0.511	0.325	0.028	0.324	0.237	0.049	0.232
$\tilde{\alpha}_1$	5%	0.333	0.050	0.329	0.218	0.051	0.212	0.156	0.058	0.145
	10%	0.251	0.074	0.240	0.171	0.075	0.153	0.133	0.079	0.107

Table 5: Comparison of RMSE

	Nonparametric	Hekcman's 2-step		Andrews-Schafgans	
n	$med(Y_1 - Y_0 \mid X = 0)$	α_0	α_1	α_0	α_1
		Design (A)			
200	0.249	0.181	0.179	0.236	0.245
500	0.155	0.110	0.114	0.155	0.159
1000	0.104	0.079	0.078	0.123	0.121
		Design (B)			
200	0.168	0.332	0.335		
500	0.116	0.299	0.295	N.A.	
1000	0.084	0.280	0.289		
		Design (C)			
200	0.205	0.238	0.301		
500	0.137	0.209	0.283	N.A.	
1000	0.104	0.199	0.272		
		Design (D)			
200	0.305	0.226	0.224	0.253	0.251
500	0.196	0.143	0.139	0.172	0.171
1000	0.132	0.097	0.100	0.135	0.133
		Design (E)			
200	0.222	0.338	0.342		
500	0.146	0.249	0.261	N.A.	
1000	0.105	0.216	0.225		
		Design (F)			
200	0.266	0.224	0.278		
500	0.173	0.174	0.228	N.A.	
1000	0.127	0.148	0.202		

The table reports the root mean squared error of three estimators. The bandwidth for the nonparametric estimator is chosen to minimize RMSE among five values considered in the simulations and the threshold level for Andrews-Schafgans' estimator is fixed at 10% which gives the smallest RMSE among the levels considered in the simulations.

Table 6: Comparison bias

	Nonparametric	Heckman's 2-step		Andrews-Schafgans	
n	$med(Y_1 - Y_0 \mid X = 0)$	α_0	α_1	α_0	α_1
		Design (A)			
200	-0.007	0.011	-0.004	0.039	0.019
500	0.005	0.002	0.003	0.040	0.021
1000	0.008	0.001	-0.002	0.012	0.021
		Design (B)			
200	0.002	-0.253	-0.260		
500	0.006	-0.266	-0.263	N.A.	
1000	-0.014	-0.264	-0.271		
		Design (C)			
200	0.061	0.198	0.259		
500	0.056	0.192	0.266	N.A.	
1000	0.056	0.190	0.264		
		Design (D)			
200	-0.001	0.000	-0.007	0.064	0.032
500	0.003	0.002	0.000	0.045	0.028
1000	0.000	0.004	-0.002	0.040	0.049
		Design (E)			
200	0.014	-0.174	-0.176		
500	0.013	-0.180	-0.189	N.A.	
1000	0.016	-0.174	-0.183		
		Design (F)			
200	0.038	0.125	0.176		
500	0.050	0.128	0.186	N.A.	
1000	0.048	0.125	0.178		

The table reports the bias of three estimators. The bandwidth for the nonparametric estimator is choosen to minimize RMSE among $\{0.5, 1.1.5, 2.0, 2.5\}$ and the threshold level for Andrews-Schafgan's estimator is fixed at 2% which gives the smallest bias among the levels considered in the simulations.

Table 7: Description of Variables

label	description	Unit
yield	Log of maize harvest per acre	Log(Kg/Acre)
hybrid	Indicator variable for planting hybrid maize	binary
fertil	Amount of fertilizer used per 10 acres	Kg/(10*Acre)
labor	Days of labor worked on the plot per 10 acres	Days/(10*Acre)
area	Area of plot	Acre
*rain*09	Rainfall over July 2009 to June 2010	mm/100
avgrain	Rainfall over the last decade	mm/100
road	Distance to secondary road network	Km
$soil - good$	Dummy for soil quality. 1 if good	binary
$soil - poor$	Dummy for soil quality. 1 if poor	binary
edu	Years of schooling of head of household, with top code 15.	years
sex	Dummy for sex of head of household. 1 if male	binary

Table 8: Summary Statistics

	All			Non-hybrid			Hybrid		
	mean	median	std. dev.	mean	median	std. dev.	mean	median	std. dev.
hybrid	0.5561	1	0.4970	0	0	0	1	1	0
yield	5.9386	5.9915	0.9534	5.7092	5.7978	0.8915	6.1218	6.1598	0.9619
fertil	8.0522	5.4945	21.0635	5.2878	3.3333	8.7661	10.2591	7.1429	26.9417
labor	7.3329	5.2941	12.1202	7.0105	5.2381	13.6382	7.5902	5.3333	10.7544
area	1.0217	0.8700	0.8332	1.0697	0.9500	0.8121	0.9833	0.8100	0.8480
*rain*09	9.2221	8.6000	1.8383	9.0004	8.3700	1.6594	9.3992	8.8100	1.9521
avgrain	8.5536	8.3100	0.8601	8.4358	8.1600	0.7973	8.6477	8.4000	0.8964
edu	5.4910	6.0000	4.4037	4.5153	4.0000	4.1352	6.2699	7.0000	4.4576
road	1.2247	1.6572	1.7528	1.4944	1.9657	1.6631	1.0094	1.3913	1.7930
$soil - good$	0.4607	0	0.4986	0.4674	0	0.4992	0.4553	0	0.4982
$soil - poor$	0.0958	0	0.2944	0.0917	0	0.2887	0.0992	0	0.2990
sex	0.7744	1	0.4180	0.7464	1	0.4353	0.7967	1	0.4026
sample size		2212			982			1230	

Table 9: Probit regression

	coefficient	t-stat
$fertil$	0.0250	7.18 ***
$labor$	-0.0047	-1.78 *
$area$	-0.0459	-1.36
$rain09$	-0.0214	-0.59
$soil - good$	-0.0568	-0.97
$soil - poor$	0.1003	1.02
edu	0.0390	5.57
sex	0.0791	1.16
$avgrain$	0.1469	1.89 *
$road$	-0.0605	-3.60 ***
constant	-0.0605	-3.60 ***
sample size	2212	

Dependent variable is *hybrid*.

Table 10: OLS and IV regression results

	OLS		IV regression	
	coefficient	t-stat	coeffcient	t-stat
$hybrid$	0.2683	7.16 ***	1.0172	2.39 **
$fertil$	0.0114	11.38 ***	0.0097	6.59 ***
$labor$	0.0045	2.56 **	0.0052	2.67 ***
$area$	-0.1808	-8.12 ***	-0.1600	-5.96 ***
$rain09$	0.0394	3.92 ***	0.0234	1.65 **
$soil - good$	0.0665	1.73 **	0.0774	1.84 **
$soil - poor$	-0.2136	-3.30 ***	-0.2416	-3.36 ***
edu	0.0294	6.62 ***	0.0150	1.57
sex	0.0815	1.82 **	0.0693	1.41
constant	5.2503	49.22 ***	5.0562	31.72
sample size	2212		2212	
adj. R^2	0.2036		0.0632	

Dependent variable is *yield*. For instruments, *road* and *avgrain* are used.

Table 11: Heckman's two-step estimator

	Non-hybrid		Hybrid	
	coefficient	t-stat	coefficient	t-stat
$fertil$	0.0022	3.88 ***	0.0040	1.22
$labor$	-0.0006	-0.60	0.0077	1.11
$area$	-0.0260	-2.07 **	-0.0862	-1.20
$rain09$	0.0766	3.94 ***	-0.0623	-1.83
$soil-good$	-0.0293	-0.52	0.1671	1.40
$soil-poor$	-0.3490	-3.62 ***	-0.2516	-1.26
edu	0.0135	1.13	-0.0325	-1.68 *
sex	0.0463	0.73	0.0808	0.57
constant	4.8974	24.06 ***	8.2951	12.91 ***
sample size	982		1230	

Dependent variable is *yield*.

Table 12: Average treatment effect using Heckman's two-step estimator

	Average treatment effect of hybrid maize
Other regressors fixed at their mean	1.7760
Other regressors fixed at their median	1.7879

The average treatment effects are calculated using the results in Table 11. They imply the expected difference in yields between hybrid maize and non-hybrid maize when other regressors are fixed at their mean or median.

Table 13: Cross-validation results

b	Sum of squared errors		
	whole sample	poor soil quality	good soil quality
0.3	7614.2	5233.1	4040.4
0.4	6782.1	3970.5	3614.2
0.5	6570.1	4678.1	3422.2
0.6	6504.5	4771.0	3440.8
0.7	6446.3	4805.2	3327.1
0.8	6404.6	4849.3	3213.5
0.9	6376.5	4869.8	3201.8
1.0	6340.7	4830.8	3105.9
1.1	6332.2	4723.0	3025.4
1.2	6332.7	4441.2	2980.6
1.3	6337.6	3991.5	2958.6
1.4	6341.5	3652.0	2945.1
1.5	6342.2	3523.4	2933.7
1.6	6340.0	3486.3	2922.2
1.7	6335.9	**3474.7**	2911.1
1.8	6331.1	3520.0	2901.6
1.9	6332.7	3518.1	2894.5
2.0	6326.4	3518.7	2889.6
2.1	**6325.1**	3518.4	2886.4
2.2	6326.4	3519.0	2884.3
2.3	6329.6	3519.8	2882.8
2.4	6334.0	3520.4	2881.7
2.5	6338.9	3520.5	2881.0
2.6	6343.9	3520.1	**2880.7**
2.7	6348.7	3519.2	2880.8
2.8	6353.2	3518.0	2881.4
2.9	6357.5	3516.9	2882.6
3.0	6361.6	3516.4	2884.1

The minimum values are reported in bold.

Table 14: Median treatment effect of hybrid maize

fertil	median treatment effect	90% C.I.	
	whole sample		
34.95 Kg/acre	0.5123	-0.0904	2.1093
44.95 Kg/acre	0.5725	0.0603	2.4709
54.95 Kg/acre	0.9944	0.2109	2.7421
64.95 Kg/acre	1.4162	0.3013	2.8626
74.95 Kg/acre	1.9888	0.4219	2.9832
	good soil quality		
34.95 Kg/acre	0.8533	-0.2297	2.1660
44.95 Kg/acre	1.0830	0.1969	2.7896
54.95 Kg/acre	1.3784	0.3938	3.0850
64.95 Kg/acre	2.1004	0.5907	3.2819
74.95 Kg/acre	2.7896	0.7876	3.4131
	poor soil quality		
34.95 Kg/acre	-0.3892	-1.0009	1.2511
44.95 Kg/acre	-0.3058	-0.8897	1.6125
54.95 Kg/acre	-0.1112	-0.6950	2.0017
64.95 Kg/acre	0.1390	-0.5560	2.2520
74.95 Kg/acre	0.3892	-0.4170	2.4188

The results are conditional on other variables being fixed at their respective median. The final two columns report the lower bound and the upper bound of 90% confidence interval calculated from bootstrap of 1000 repetitions.

Table 15: Optimal Subsidy Level

distance	optimal subsidy level	90% C.I.	
	no schooling		
1km	0.3808	-0.1367	1.7904
5km	0.3928	-0.1367	1.8931
10km	0.4887	-0.1534	1.9348
	6 years of schooling		
1km	0.2909	-0.1367	1.1406
5km	0.3328	-0.1256	1.3572
10km	0.3808	-0.1256	1.7043
	9 years of schooling		
1km	0.2249	-0.1367	1.0434
5km	0.2789	-0.1339	1.1684
10km	0.3508	-0.1145	1.4655

The results are conditional on other variables being fixed at their respective median. The final two columns report the lower bound and the upper bound of 90% confidence interval calculated from bootstrap of 1000 repetitions.

References

Andrews, D., Schafgans, M., 1998. Semiparametric estimation of the intercept of a sample selection model. The Review of Economic Studies 65 (3), 497.

Bayer, P., Khan, S., Timmins, C., 2011. Nonparametric identification and estimation in a roy model with common nonpecuniary returns. Journal of Business and Economic Statistics 29 (2), 201–215.

Berry, S., Haile, P., 2012. Identification in differentiated products markets using market level data.

Borjas, G., 1987. Self-selection and the earnings of immigrants. The American Economic Review, 531–553.

Chernozhukov, V., Fernández-Val, I., Galichon, A., 2010. Quantile and probability curves without crossing. Econometrica 78 (3), 1093–1125.

Chernozhukov, V., Hansen, C., 2005. An iv model of quantile treatment effects. Econometrica 73 (1), 245–261.

Chirwa, E., 2005. Adoption of fertiliser and hybrid seeds by smallholder maize farmers in southern malawi. Development Southern Africa 22 (1), 1–12.

Dette, H., Neumeyer, N., Pilz, K., 2006. A simple nonparametric estimator of a strictly monotone regression function. Bernoulli 12 (3), 469–490.

Dette, H., Scheder, R., 2006. Strictly monotone and smooth nonparametric regression for two or more variables. Canadian Journal of Statistics 34 (4), 535–561.

Dette, H., Volgushev, S., 2008. Non-crossing non-parametric estimates of quantile curves. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (3), 609–627.

Eisenhauer, P., Heckman, J., Vytlacil, E., 2011. The generalized roy model and the cost-benefit analysis of social programs.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., Engel, J., 1997. Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. Annals of the Institute of Statistical Mathematics 49 (1), 79–99.

Fan, J., Gijbels, I., 1996. Local polynomial modelling and its applications. Vol. 66. Chapman & Hall/CRC.

French, E., Taber, C., 2011. Identification of models of the labor market. Handbook of Labor Economics 4, 537–617.

Hall, P., Wolff, R., Yao, Q., 1999. Methods for estimating a conditional distribution function. Journal of the American Statistical Association, 154–163.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.

Heckman, J., 1990. Varieties of selection bias. The American Economic Review 80 (2), 313–318.

Heckman, J., Honore, B., 1990. The empirical content of the roy model. Econometrica: Journal of the Econometric Society, 1121–1149.

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

Heckman, J. J., Vytlacil, E., 2005. Structural equations, treatment effects, and econometric policy evaluation. Econometrica, 669–738.

Holland, P. W., 1986. Statistics and causal inference. Journal of the American statistical Association 81 (396), 945–960.

Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica: Journal of the Econometric Society, 467–475.

Khan, S., Tamer, E., 2010. Irregular identification, support conditions, and inverse weight estimation. Econometrica 78 (6), 2021–2042.

Lee, L., 1978. Unionism and wage rates: a simultaneous equations model with qualitative and limited dependent variables. International economic review 19 (2), 415–433.

Lu, Z., 1996. Multivariate locally weighted polynomial fitting and partial derivative estimation. Journal of Multivariate Analysis 59 (2), 187–205.

Masry, E., 1996. Multivariate local polynomial regression for time series: uniform strong consistency and rates. Journal of Time Series Analysis 17 (6), 571–599.

Maurel, A., D'Haultfoeuille, X., 2011. Inference on an extended roy model, with an application to schooling decisions in france. Economic Research Initiatives at Duke (ERID) Working Paper No. 101.

Park, B., 2012. Identification of the generalized roy model with an application to panel data models, unpublished manuscript, Department of Economics, Yale University.

Peterson, A., 1976. Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. Proceedings of the National Academy of Sciences of the United States of America, 11–13.

Pollard, D., 1984. Convergence of stochastic processes. Springer.

Ruppert, D., Wand, M., 1994. Multivariate locally weighted least squares regression. The annals of statistics, 1346–1370.

Suri, T., 2011. Selection and comparative advantage in technology adoption. Econometrica 79 (1), 159–209.

Vijverberg, W. P., 1993. Measuring the unidentified parameter of the extended roy model of selectivity. Journal of Econometrics 57 (1), 69–89.

Vytlacil, E., 2003. Independence, monotonicity, and latent index models: An equivalence result. Econometrica 70 (1), 331–341.

Willis, R., Rosen, S., 1979. Education and self-selection. The Journal of Political Economy, 7–36.

Yu, K., Jones, M., 1998. Local linear quantile regression. Journal of the American Statistical Association, 228–237.

Zeller, M., Diagne, A., Mataya, C., 1998. Market access by smallholder farmers in malawi: Implications for technology adoption, agricultural productivity and crop income. Agricultural Economics 19 (1), 219–229.