Regression Discontinuity Design: Identification and Estimation of Treatment Effects with Multiple Selection Biases

Muzhe Yang*

Department of Economics, Lehigh University

January 2009

Abstract

Focusing on the "fuzzy" regression discontinuity (RD) design, this paper proposes two easy-to-implement estimators for the average treatment effect in the presence of multiple selection biases—selection on both observables and unobservables. The theoretical results leverage the dual nature of the RD design, both the "borderline experiment" provided near the threshold and the exclusion restriction provided in the selection equation for the choice of treatment. The improvement that the proposed estimators offer in the efficiency-bias trade-off is examined through Monte Carlo experiments and an empirical study of an education program allocated on the basis of test score cutoffs.

*Corresponding author: Department of Economics, Rauch Business Center, Lehigh University, 621 Taylor Street, Bethlehem, PA 18015. Phone: (610) 758-4962. Fax: (610) 758-4677. E-mail: muzheyang@lehigh.edu.

Keywords: fuzzy regression discontinuity design; selection bias; heterogeneous treatment effects; average treatment effect

JEL Classification: C21; C30

1 Introduction

The fundamental problem of causal inference about a treatment effect stems from the impossibility of *observing* the *same* individual simultaneously in treated and untreated states.[1] The identification of treatment effects must rely on comparisons of the outcomes of different individuals with different treatment status, which means differences in outcomes may arise from factors other than the treatment. The most reliable design to deal with this problem is random assignment of the treatment.[2] Unfortunately, for many of the most general questions in the social sciences, random assignment is either too costly to implement or unethical.[3] One ethical procedure that society and governments follow to allocate resources assigns resources on the basis of merit or need, often using eligibility cutoffs for the program in which the odds of qualifying for the intervention change substantially (or "discontinuously") at the cutoffs. For example, researchers note that as long as there is some "noise" or arbitrariness in the eligibility criteria, near the cutoffs, the assignment of resources is "close to random." Although the efficiency of such policy designs can and should be debated, they provide a unique opportunity for evaluating the effect of an intervention while leveraging some of the features of random assignment. The regression discontinuity (RD) design attempts to utilize these discontinuous changes in the prob-

[1] Such a causal effect is defined as the difference between *potential* outcomes in the presence and in the absence of a treatment (Rubin, 1974; Holland, 1986).

[2] In this study, the identification of treatment effects refers to the identification of the effect of a treatment *intervention*, not the effect of a *self-selected* treatment. The former has implications for policy design; the latter can mislead policy making.

[3] In addition to the feasibility issues, random assignment is subject to threats to both its internal and its external validity, such as substitution bias, randomization bias, placebo effects (Malani, 2006), and Hawthorne effects (Winship and Morgan, 1999; Cobb-Clark and Crossley, 2003).

ability of treatment at the eligibility cutoff(s).[4] A "sharp" RD design occurs when the probability of treatment goes from 0 to 1 at the cutoff point. However, this study instead focuses on the so-called "fuzzy" RD design (Trochim, 1984), in which the change in the probability of treatment is less than 1 but still substantial; this design better matches the design of most policy interventions.

Much theoretical research on RD design (Hahn, Todd, and Van der Klaauw, 2001; Lee and Card, 2008) focuses on the effects *at* the eligibility threshold, at which point potential biases disappear while the probability of treatment changes discontinuously.[5] Its similarity to a borderline experiment at the eligibility threshold also has prompted its recognition and exploitation in many empirical studies.[6] However, two concerns arise about a focus on effects at the threshold. First, empirically, the definition of "the limit" can be ad hoc and preclude inference due to a paucity of data. Most applied studies implementing RD design instead use data away from the discontinuity or assume a functional form for the selection bias, due to observables, to obtain estimates and form confidence intervals.[7] In some of these applications, such as Angrist and Lavy (1999) and Chay, McEwan, and Urquiola (2005), as the data away from the discontinuity get trimmed, the confidence intervals grow large enough to disallow the rejection of many hypotheses. Second, in the presence of treatment effect heterogeneity, near the discontinuity, only the average effect for a particular population can be identified in certain conditions (e.g., monotonicity), rather than the average effect for a randomly selected member of the population known as the *average treatment effect* (ATE). The former[8] is useful for policy analysis, because it measures the impact of

[4] For a history and overview of RD design, see Thistlethwaite and Campbell (1960), Goldberger (1972a, 1972b), and Cook (2008).

[5] Lee and Card (2008) further extend the applicability of an RD design to the case in which the selection variable has discrete support.

[6] These studies include Berk and de Leeuw (1999), Black (1999), Black, Galdo, and Smith (2007), Lee (2008), Lemieux and Milligan (2008), Ludwig and Miller (2007), and Van der Klaauw (2002).

[7] These studies include Angrist and Lavy (1999), Chay and Greenstone (2003), Chay, McEwan, and Urquiola (2005), DiNardo and Lee (2004).

[8] This is the local average treatment effect (LATE) (Imbens and Angrist, 1994; Angrist, Imbens,

the eligibility criteria that *were* used by the program; the latter generally is more useful for forecasting the relative benefits of potential policies under consideration. This study attempts to address these concerns by integrating the results from literature on selection biases with RD literature. The derived theoretical results leverage RD's dual nature—that is, that it provides both a borderline experiment near the discontinuity and an instrumental variable (IV) for the actual treatment status. In particular, the treatment probability changes significantly according to the treatment eligibility defined in the RD design, such that if there is no anticipation of a treatment before its implementation, the treatment eligibility should be excluded from the potential outcome in the absence of the treatment. In this sense, the eligibility indicator in RD becomes a strong and valid IV for actual treatment status.

This study proposes the following regression model, which is based on the treatment effect definition (i.e., the difference between two potential outcomes in the presence and in the absence of the treatment):[9]

$$y_j = g_j(z^*) + \eta \cdot j + u_j,\ \mathbb{E}(u_j|z^*) = 0, \tag{1}$$

where $g_j(z^*)$ is continuous in z^*, and j indicates the presence ($j = 1$) or absence ($j = 0$) of a treatment intervention. For the factors that affect y_j, the observable part ($g_j(z^*)$) and the unobservable part (u_j) are assumed to be additively separable. An individual's treatment status (d) indicates either receiving the treatment ($d = 1$) or not ($d = 0$). A random treatment assignment can lead d to j. Here, η is the additive effect, solely due to the treatment. In this model, the ATE is defined as

$$\text{ATE} \equiv \mathbb{E}(y_1 - y_0) = \mathbb{E}\left(\eta + g_1(z^*) - g_0(z^*)\right) \equiv \alpha. \tag{2}$$

and Rubin, 1996).

[9] The following discussion of identification and estimation of treatment effects considers no additional covariates ($\mathbf{x}$), to avoid unnecessary complications. Imbens and Lemieux (2008) discuss adjustments for additional covariates.

The average difference in the *observed* outcome (y, and $y = dy_1 + (1-d)y_0$) between the treatment (T) and the control (C) group can be decomposed into five parts:

$$\mathbb{E}(y|d=1) - \mathbb{E}(y|d=0) = \begin{cases} \mathbb{E}(\eta) & \text{(a)} \\ +\mathbb{E}(g_0(z^*)|d=1) - \mathbb{E}(g_0(z^*)|d=0) & \text{(b)} \\ +\mathbb{E}(g_1(z^*) - g_0(z^*)|d=1) & \text{(c)} \\ +\mathbb{E}(u_0|d=1) - \mathbb{E}(u_0|d=0) & \text{(d)} \\ +\mathbb{E}(u_1 - u_0|d=1) & \text{(e)} \end{cases} . \quad (3)$$

In equation (3), part (a) is the average effect, purely due to the intervention and separate from the observable and the unobservable. Part (b) is the average difference between the T and C groups in the pre-treatment observables. The group difference due to the interaction between observables and the treatment is represented by part (c). Part (d) is the group difference due to pre-treatment unobservables, which causes omitted variable bias (OVB). Finally, part (e) reveals the group difference due to the interaction between the treatment and the unobservables, which causes sorting or selectivity bias. For example, a positive conditional expectation of the potential treatment gain, $u_1 - u_0$, implies cream skimming as happens when a program or treatment is assigned only to individuals who can benefit from it, so as to maximize the program effectiveness. Identifying the treatment effect on a randomly selected individual of a population (ATE) requires the initial identification of part (a) plus $\mathbb{E}\left[g_1(z^*) - g_0(z^*)\right]$. The dual nature of the RD design can take into account several sources of selection bias—selection on the observables (part b), selection on the unobservables (part d), and selection on treatment effect heterogeneity (parts c and e).

For a fuzzy RD design, the selection equation becomes:

$$\begin{aligned} d &= 1\{\pi_0 + \pi_1 z + \pi_2 z^* + v > 0\}, \text{ and} \\ z &= 1\{z^* \leqslant 0\},\ \pi_1 \neq 0,\ v \sim F_v(\cdot), \end{aligned} \tag{4}$$

where v represents the unobservables in the selection process with distribution $F_v(\cdot)$. The variable z^* can be observed and used for the selection eligibility (z), which uses 0 as the cutoff point (selection threshold). The fuzzy RD design forces π_1 to be nonzero. The eligible ($z = 1$) members of the population are more likely to receive the treatment ($d = 1$) if π_1 is positive. Identifying ATE hinges on the relationship between v and u_j ($j = 0, 1$), because if v and u_j are independent ("$\perp\!\!\!\perp$"), *selection on observables* exists, which implies $\mathbb{E}(y_j|d) \neq \mathbb{E}(y_j)$, but $\mathbb{E}(y_j|d, z^*) = \mathbb{E}(y_j|z^*)$ for the observables, z^*. The treatment status (d) becomes exogenous (or "randomly assigned") to potential outcomes (y_j) on average after controlling for the observables (z^*). If v and u_j are not independent, *selection on unobservables* occurs, which implies $\mathbb{E}(y_j|d, z^*) \neq \mathbb{E}(y_j|z^*)$, but $\mathbb{E}(y_j|d, z^*, \epsilon) = \mathbb{E}(y_j|z^*, \epsilon)$ for observables (z^*) and some unobservables (ϵ).[10] In addition, OVB occurs if either u_0 or u_1 is not independent of v. Selectivity or sorting bias arises if the potential gain, $u_1 - u_0$, is not independent of v. Note that an RD's eligibility criterion (z) can change a potential treatment status exogenously if the threshold is unexpected or z^* cannot be manipulated perfectly.[11] The preceding selection equation implies that the eligibility (z) induces a nonsmoothness only in the selection process, not in the potential outcome. It therefore provides a valid exclusion restriction for the outcome equation and serves as a powerful predictor for treatment status. It is worth noting that this article is the first to examine the use of RD's eligibility as an instrument to deal with selection on

[10] See Lee (2005) for formal definitions.

[11] Such a situation may be detectable by checking whether the density of the selection variable is discontinuous at the cutoff point (McCrary, 2008).

the unobservables, namely, part (d) in equation (3), and selection due to unobserved heterogeneities in the treatment effects, or part (e) in equation (3).

Current theoretical RD papers focus instead on treatment effects at the cutoff, assuming selection only on the observables, in which case the bias due to part (e) in equation (3) gets excluded and the bias due to part (d) in equation (3) can be differenced out at the cutoff where $\mathbb{E}(u_0|z^*)$ is *continuous*. The emphasis therefore is on how to control for $g_0(\cdot)$ and $g_1(\cdot)$ (Hahn, Todd, and Van der Klaauw, 2001; Imbens and Lemieux, 2008) to minimize the selection bias due to parts (b) and (c) in equation (3) and estimate the effect at the cutoff. In contrast, this study reformulates RD according to the preceding regression model and focuses on ATE for a predefined population away from the threshold to take into account multiple selection biases, that is, parts (b) to (e) in equation (3). The research question therefore entails the solution to two problems. The first pertains to how to control for part (b) in equation (3) without specifying $g_0(\cdot)$ and deal with part (c) in equation (3) with flexible parametric assumptions on $g_1(\cdot) - g_0(\cdot)$ and thus identify ATE in the absence of parts (d) and (e) in equation (3) for a population near the selection threshold. In this setting, the cutoff selected for an initial program can be reassessed by policymakers. The second problem to solve then becomes how to correct for the impacts of part (d) in equation (3), or the OVB, and part (e) in equation (3), or the sorting bias, on ATE when they are present. This setting allows for forecasting the relative benefits of a policy change when policymakers consider extending the initial program to a larger population, farther away from the threshold.

The response to the first problem proposes a new estimator, an RD robust estimator, for the ATE of a population near the selection threshold. It has greater external validity than existing estimators aimed at measuring the effect only at the threshold. It also achieves good internal validity, in the sense that it controls for part (b) in equation (3) with $g_0(\cdot)$ unspecified, and the parameterization of $g_1(\cdot) - g_0(\cdot)$, which

can be made flexible, is necessary only when treatment effect heterogeneity exists due to the observables (i.e., part (c) in equation (3)). Unlike the alternatives proposed by Hahn, Todd, and Van der Klaauw (2001), this estimator is based on the moment restrictions derived from the conditional mean independence between the unobservables in the potential outcomes and the unobservable in the selection equation, when the selection is based only on the observables. The orthogonality conditions are derived from the residual, that is, the difference between the actual treatment status and the conditional treatment probability (or the propensity score).[12] With selection on observables, the exclusion restriction in the RD design increases the efficiency of the RD robust estimator.

To address the second problem, a second estimator, the correction function estimator, considers the ATE of a population away from the threshold, where selection on unobservables is pertinent. It takes into account the OVB, or part (d) in equation (3), and the sorting bias, which is part (e) in equation (3), using RD's eligibility as the IV for the actual treatment status to correct for (d) in equation (3) and adding correction terms for (e) in equation (3) to the outcome equation. The construction of the correction terms relies on a cubic polynomial moment restriction on the potential treatment gain, $u_1 - u_0$, conditional on v. This restriction allows for a nonlinear sorting pattern, but it is less restrictive than a joint-distributional assumption for (u_0, u_1, v), which is sufficient to account for (d) and (e) in equation (3) simultaneously. This correction function estimator therefore permits the estimation of ATE in the presence of OVB and heterogeneous sorting. It also has greater external validity, in the sense that it bears richer policy implications for a larger population (i.e., farther away from the threshold). However, its internal validity is limited by the parameterization of

[12]Note that the estimators with selection on observables, such as matching (Rosenbaum and Rubin, 1983a, 1983b) or inverse probability weighting (Hogan and Lancaster, 2004; Wooldridge, 2007), are of limited applicability for a fuzzy RD. The overlapping or common-support identification assumption is difficult to meet, and it is completely violated for a sharp RD design.

$g_0(\cdot)$ and $g_1(\cdot)$. This proposed estimator is based on Wooldridge's (2002) work, but it extends existing results to allow for a nonlinear sorting pattern because it adds the cubic polynomial sorting correction terms to the linear specification. This extension is important when both cream skimming and adverse selection may be concerns.

The rest of this article is organized as follows: Section 2 derives identification results and the associated estimators. Section 3 evaluates the estimators' finite sample performances. Section 4 reexamines an empirical study by Chay, McEwan, and Urquiola (2005) to demonstrate the improvement in the efficiency-bias trade-off, using the evaluation of an education program allocated on the basis of test score cutoffs. Both estimators are easy to implement using standard software. Section 5 concludes.

2 Identification and Estimation of Treatment Effects

Removing or correcting for selection biases when identifying ATE hinges on moment restrictions imposed on the unobservables in both the selection and outcome equations. To exploit RD's dual nature, on the one hand, its borderline experiment can establish the RD robust estimator, such that the impacts of (d) and (e) in equation (3) can be plausibly removed close to the threshold; on the other hand, its IV nature, as implied by the selection rule, serves as the correction function estimator to deal with (d) and (e) in equation (3), which also suggests a falsification test for the assumption of selection on observables. For ATE conditional on z^*, we consider two cases: a homogeneous treatment effect that implies a constant difference between $\mathbb{E}(y_1|z^*)$ and $\mathbb{E}(y_0|z^*)$, and a heterogeneous treatment effect that varies with the observable (z^*) and the unobservable (η).

Assumption 1 $\mathbb{E}(y_1 - y_0|z^*) = \eta = \alpha$, *where* $g_1(z^*) = g_0(z^*)$, $u_1 = u_0$, *and* η *is*

constant.

Assumption 2 $\mathbb{E}(y_1 - y_0|z^*) = \mathbb{E}(\eta) + \lambda(z^*)$, *where* $g_1(z^*) - g_0(z^*) = \lambda(z^*)$, *and* $\eta \amalg (d, z^*)$; $\lambda(z^*)$ *is the treatment effect heterogeneity due to observables, and* η *is the treatment effect heterogeneity due to unobservables.*

According to Assumption 2, the observed outcome (y) can be written as

$$y = g_0(z^*) + (\eta + \lambda(z^*))d + e, \text{ where } e \equiv u_0 + d(u_1 - u_0). \tag{5}$$

It then can be rewritten in terms of ATE (α) as

$$y = g_0(z^*) + \alpha d + (\lambda(z^*) - \mathbb{E}(\lambda(z^*)))d + \widetilde{e}, \text{ where } \widetilde{e} \equiv e + d(\eta - \mathbb{E}(\eta)). \tag{6}$$

According to Assumption 1, this model can be simplified to

$$y = g_0(z^*) + \alpha d + u_0. \tag{7}$$

In either case, the observed outcome (y) takes a partially linear form. The main obstacles to identifying α are the presence of $g_0(z^*)$, $\lambda(z^*)$, and the relationship between (u_0, u_1) and v. Identifying ATE in the presence of $\lambda(z^*)$ thus requires an additional parametric assumption.

Assumption 3 $\lambda(z^*) = \mathbf{w}'\gamma$, *where* $\mathbf{w}$ *is a vector including the polynomials of* z^*.

The following identification results reflect the assumption of selection on observables when (u_0, u_1) and v are independent and selection on unobservables when (u_0, u_1) and v are correlated.

2.1 Selection on Observables

According to Assumption 2, the central idea of identifying ATE (α) is to use the conditional moment restrictions derived from selection on observables to generate orthogonality conditions that will remove selection biases, that is, parts (b) and (c) in equation (3), due to $g_0(z^*)$ and $\lambda(z^*)$, respectively. If selection is only on the observables, then $\mathbb{E}(\widetilde{e}|d,z,z^*)=0=\mathbb{E}(\widetilde{e}|z,z^*)$, which implies $\mathbb{E}[(d-\mathbb{E}(d|z,z^*))\widetilde{e}]=0$. Note that $\mathbb{E}[(d-\mathbb{E}(d|z,z^*))|z^*]=0$ as well, which implies $\mathbb{E}[g_0(z^*)(d-\mathbb{E}(d|z,z^*))]=0$. Therefore, the residual in the selection equation, $\widetilde{v}=d-\mathbb{E}(d|z,z^*)$, can orthogonalize both $g_0(z^*)$ and $\widetilde{e}$ in

$$y-\alpha d=g_0(z^*)+d(\lambda(z^*)-\mathbb{E}(\lambda(z^*)))+\widetilde{e}, \tag{8}$$

and also obtain

$$\alpha=\frac{\mathbb{E}[y(d-\mathbb{E}(d|z,z^*))]-\mathbb{E}[(\lambda(z^*)-\mathbb{E}(\lambda(z^*)))\mathbb{V}(d|z,z^*)]}{\mathbb{E}[d(d-\mathbb{E}(d|z,z^*))]}. \tag{9}$$

The identification of α becomes complicated by the presence of $\lambda(z^*)$. However, if there are only innocuous treatment effect heterogeneities, where $\lambda(z^*)=\lambda$ (a constant), then α can be identified as follows:

$$\alpha=\mathbb{E}(\eta+\lambda)=\frac{\mathbb{E}[y(d-\mathbb{E}(d|z,z^*))]}{\mathbb{E}[d(d-\mathbb{E}(d|z,z^*))]}. \tag{10}$$

The following theorem gives the identification results for α in the presence of $\lambda(z^*)$, where $\lambda(z^*)\neq\lambda$ (a constant).

Theorem 1 *With selection on observables, Assumption 2, and Assumption 3,* $\theta\equiv$

$(\alpha, \gamma')'$ *can be identified as follows:*

$$\theta = \mathbb{E}^{-1}(\mathbf{x}\mathbf{x}')\,\mathbb{E}(\mathbf{x}y),$$

where[13]

$$\mathbf{x} \equiv (x_1, \mathbf{x}_2')',\ x_1 \equiv d - \mathbb{E}(d|z, z^*),\ \mathbf{x}_2 \equiv (d - \mathbb{E}(d|z, z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w})).$$

Furthermore, ATE (α) *is given by*

$$\alpha = \frac{\mathbb{E}(x_1 y) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(\mathbf{x}_2 y)}{\mathbb{E}(x_1^2) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)},$$

and the explicit treatment effect heterogeneity (γ) *is given by*

$$\begin{aligned}\gamma \;=\; & \left[\mathbb{E}(\mathbf{x}_2\mathbf{x}_2') - \mathbb{E}(x_1\mathbf{x}_2)\mathbb{E}^{-1}(x_1^2)\mathbb{E}(x_1\mathbf{x}_2')\right]^{-1}\mathbb{E}(\mathbf{x}_2 y) \\ & - \frac{\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)\mathbb{E}(x_1 y)}{\mathbb{E}(x_1^2) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)},\end{aligned}$$

where

$\mathbb{E}(x_1^2) = \mathbb{E}(\mathbb{V}(d|z, z^*))$, $\mathbb{E}(x_1\mathbf{x}_2) = \mathbb{E}[\mathbb{V}(d|z, z^*)\,(\mathbf{w} - \mathbb{E}(\mathbf{w}))]$,
$\mathbb{E}(\mathbf{x}_2\mathbf{x}_2') = \mathbb{E}[\mathbb{V}(d|z, z^*)\,(\mathbf{w} - \mathbb{E}(\mathbf{w}))\,(\mathbf{w} - \mathbb{E}(\mathbf{w}))']$,
$\mathbb{E}(x_1 y) = \mathbb{E}\left[(d - \mathbb{E}(d|z, z^*))\,y\right]$, *and* $\mathbb{E}(\mathbf{x}_2 y) = \mathbb{E}\left[(d - \mathbb{E}(d|z, z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))y\right]$.

Proof. See Appendix A. ■

Note that identifying treatment effects with an RD design and selection on observables is similar to matching (or inverse probability weighting, IPW). Both hinge on a propensity score, $\mathbb{E}(d|z, z^*)$. However, unlike matching (or IPW), the identification strategy proposed here for the fuzzy RD uses randomness (near the cutoff)—the

[13] $\mathbb{E}^{-1}(\mathbf{x}\mathbf{x}') \equiv [\mathbb{E}(\mathbf{x}\mathbf{x}')]^{-1}$.

deviation between the actual treatment status and the associated propensity score—to orthogonalize both observables and unobservables and accredits the difference in observed outcomes (between the treated and the untreated) to the treatment. The resultant two-stage estimator to be constructed uses, in the first stage, a consistent estimator for $\mathbb{E}(d|z,z^*)$, then includes the following regressors for the second stage:

$$\widehat{\mathbf{x}} \equiv \left[\left(d-\widehat{\mathbb{E}}(d|z,z^*)\right),\left(d-\widehat{\mathbb{E}}(d|z,z^*)\right)(\mathbf{w}-\overline{\mathbf{w}})'\right]', \tag{11}$$

where $\mathbf{w}$ is a vector including the polynomials of $\mathbf{z}^*$. We plug $\widehat{\mathbf{x}}$ and y into the RD robust estimator and attain:

$$\widehat{\theta}_{\text{RD-robust}} = \left(\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\left(\sum_{i=1}^{N}\widehat{\mathbf{x}}_i y_i\right). \tag{12}$$

We next present the estimator's large sample properties with its asymptotic variance adjusted in accordance with the generated regressors, which come from the first-stage residuals.

Theorem 2 *With selection on observables, Assumption 2, Assumption 3, and a parametric assumption* $\mathbb{E}(d|z,z^*)=p(z,z^*;\lambda)$*, we have*

$$\sqrt{N}\left(\widehat{\theta}_{RD\text{-}robust}-\theta\right)\xrightarrow{d}N\left(\mathbf{0},A_0^{-1}\Omega A_0^{-1}\right),$$

where

$$\begin{aligned}
&A_0\equiv\mathbb{E}(\mathbf{x}\mathbf{x}'),\ \Omega\equiv\mathbb{V}\left(\mathbf{x}\left(y-\mathbf{x}'\theta\right)-B_0\mathbf{r}(\lambda)\right),\ B_0\equiv\mathbb{E}\left(\left(\theta\otimes\mathbf{x}'\right)'\tfrac{\partial\mathbf{x}}{\partial\lambda'}\right)\\
&\mathbf{x}\equiv\left[\left(d-p(z,z^*;\lambda)\right),\left(d-p(z,z^*;\lambda)\right)\left(\mathbf{w}-\boldsymbol{\mu}\right)'\right]',\ \textit{and}\\
&\sqrt{N}(\widehat{\lambda}-\lambda)=N^{-1/2}\sum_{i=1}^{N}\mathbf{r}_i(\lambda)+o_p(1),\ \mathbb{E}(\mathbf{r}_i(\lambda))=\mathbf{0}.
\end{aligned}$$

Proof. See Appendix B. ■

There is no first stage involving $\mathbb{E}(y|z, z^*)$, which distinguishes this approach from Robinson's (1988) two-stage estimator. The intuition behind this estimator is to use the first stage to control for selection on the observables (z and z^*), then use the "cleaned" residual ($\widehat{\widetilde{v}}$) from the first stage to generate orthogonality conditions, based on the selection-on-observables assumption, to form a moment-based estimation of θ in the second stage. This estimator removes selection bias due to (b) in equation (3) by orthogonalizing $g_0(z^*)$, and $\mathbf{w}$ controls for the bias due to (c) in equation (3). It is robust in the sense that it controls for (b) in equation (3) with $g_0(z^*)$ unspecified, and the parameterization of $g_1(z^*) - g_0(z^*)$ using $\mathbf{w}$ can be flexible if necessary to control for treatment effect heterogeneity due to observables, represented by (c) in equation (3). When selection is only on the observables, the exclusion restriction imposed by a fuzzy RD's eligibility cutoff (z) also brings efficiency gains to this estimator.

2.2 Selection on Unobservables

Identifying treatment effects with an RD design requires that the selection variable (z^*) should not be manipulated perfectly. Otherwise, there will be correlations between the observable (z^*) and the unobservable (u_0, u_1, v), which make the eligibility indicator (z) endogenous and thus invalidate RD's IV nature.[14] Without manipulating the selection variable, identifying ATE with selection on unobservables requires dealing with two additional biases. One is OVB, where $Cov(u_0, v|z^*)$ or $Cov(u_1, v|z^*)$ is not zero, and the other is a sorting (or selectivity) bias, where $Cov(u_1 - u_0, v|z^*)$ is not zero. Possible solutions to these identification problems rely on either the control

[14] In practice, several cases can effectively prevent manipulation. First, the cutoff point is unexpected. If the cutoff point for eligibility comes entirely as a surprise, and no one anticipates such a selection rule, the manipulation of the selection variable is completely avoided. The eligibility criterion is therefore defined by a predetermined selection variable, which is simply not manipulable. Second, individuals only have imperfect control over selection variables. In some cases, manipulation is possible but not perfect. Even for an anticipated threshold, within a small neighborhood, there will be some randomness that prevents perfect manipulation (Lee, 2008).

function or the IV approach. The former, aimed at ATE, follows Heckman's (1979) approach, which obtains $\mathbb{E}(y|d, z^*)$ directly from the distributional assumption for (u_0, u_1, v) by adding two control functions, for both OVB and sorting bias, back to the outcome equation. The latter solution, aimed at LATE instead, replaces the distributional assumption for (u_0, u_1, v) with a restriction that the potential treatment status should respond to the instrument assignment monotonically. The LATE can still be useful for policy analysis, because it measures the impact of the program on "compliers," defined by their potential treatment responses to all possible instrument assignment status. The ATE often appears more useful for forecasting the impact of policies under consideration on a randomly selected individual from the study population. The correction function approach proposed herein provides a middle ground between the two aforementioned methods. Specifically, the proposed approach addresses $\mathbb{E}(y|z, z^*)$ instead of $\mathbb{E}(y|d, z, z^*)$, and therefore, it requires adding $\mathbb{E}(d(u_1 - u_0)|z, z^*)$ back to the outcome equation to remove the sorting bias (which is assumed away in LATE). It identifies ATE through a conditional moment restriction on $\mathbb{E}(u_1 - u_0|v)$, which is assumed to be polynomial in v, instead of a joint distribution for (u_0, u_1, v). This correction function approach derives from a correlated random coefficient model used by Wooldridge (1997, 2002, 2003) but relaxes Wooldridge's (2002) assumption that $\mathbb{E}(u_1 - u_0|v)$ is linear in v to the case that $\mathbb{E}(u_1 - u_0|v)$ can be nonlinear, or at least cubic, in v. The extension to a more flexible correlation structure between the unobservable in the selection (v) and the potential treatment gain ($u_1 - u_0$), which is unobservable to an econometrician, accommodates cases with important economic implications. For example, adverse selection occurs when $\mathbb{E}(u_1 - u_0|v)$ is negative, and cream skimming exists when $\mathbb{E}(u_1 - u_0|v)$ is positive. The following theorem presents the correction terms required for $\mathbb{E}(d(u_1 - u_0)|z, z^*)$, which corresponds to part (e) in equation (3), that is, the sorting bias.

Theorem 3 *With selection on unobservables, Assumption 2, and the following assumptions:*

$$\mathbb{E}(u_1 - u_0|v, z^*) = \mathbb{E}(u_1 - u_0|v) = \xi_1 v + \xi_2 v^2 + \xi_3 v^3, \text{ and } v \sim N(0,1),$$

we have the following three correction terms for $\mathbb{E}(d(u_1 - u_0)|z, z^*)$,[15]

$$\begin{aligned}\mathbb{E}(d(u_1 - u_0)|z, z^*) &= \xi_1 \phi(\pi_0 + \pi_1 z + \pi_2 z^*) + \\ &\quad \xi_2 \left[\Phi(\pi_0 + \pi_1 z + \pi_2 z^*) - (\pi_0 + \pi_1 z + \pi_2 z^*)\phi(\pi_0 + \pi_1 z + \pi_2 z^*)\right] + \\ &\quad \xi_3 \left[2\phi(\pi_0 + \pi_1 z + \pi_2 z^*) + (\pi_0 + \pi_1 z + \pi_2 z^*)^2 \phi(\pi_0 + \pi_1 z + \pi_2 z^*)\right].\end{aligned}$$

Proof. See Appendix C. ■

With three correction terms for $\mathbb{E}(d(u_1 - u_0)|z, z^*)$ added back to the outcome equation:

$$y = g_0(z^*) + \alpha d + d(\lambda(z^*) - \mathbb{E}(\lambda(z^*)) + \mathbb{E}(d(u_1 - u_0)|z, z^*) + \widetilde{e} \quad (13)$$

$$\widetilde{e} \equiv u_0 + d(u_1 - u_0) - \mathbb{E}(d(u_1 - u_0)|z, z^*) + d(\eta - \mathbb{E}(\eta)),\ \mathbb{E}(\widetilde{e}|z, z^*) = 0, \quad (14)$$

ATE (α) can be identified, as suggested by Wooldridge (2002), by using $\mathbb{E}(d|z, z^*)$ or z as the IV for d. We next propose the correction function estimator for α in the presence of multiple selection biases, that is, parts (b) to (e) in equation (3). The OVB, or part (d) in equation (3), can be dealt with according to the IV provided by RD's eligibility cutoff. The sorting bias, part (e) in equation (3), is controlled for by using the correction terms. To control for (b) and (c) in equation (3), we need the parameterization, $g_0(z^*) = \beta_0 + \mathbf{w}'\beta_1$, plus Assumption 3 for the outcome equation. To simplify the notation, we define $\theta \equiv (\beta_0, \beta_1', \alpha, \gamma', \xi_1, \xi_2, \xi_3)'$, $\pi \equiv (\pi_0, \pi_1, \pi_2)'$,

[15] Here, $\phi(\cdot)$ denotes the normal probability density function (pdf), and $\Phi(\cdot)$ denotes the normal cumulative distribution function (cdf).

$\mu \equiv \mathbb{E}(\mathbf{w})$, and $\widetilde{\mathbf{z}} \equiv (1, z, z^*)$. Similar to the RD robust estimator, there are two stages required to implement the correction function estimator. In the first stage, we estimate a probit model for d to obtain $\Phi(\widetilde{\mathbf{z}}'\widehat{\pi})$. In the second stage, the regressors and the instruments in the outcome equation are as follows:

$$\begin{aligned}
\widehat{\mathbf{x}} &\equiv (1, \mathbf{w}', d, d(\mathbf{w} - \widehat{\boldsymbol{\mu}})', \phi(\widetilde{\mathbf{z}}'\widehat{\pi}), \Phi(\widetilde{\mathbf{z}}'\widehat{\pi}) - (\widetilde{\mathbf{z}}'\widehat{\pi})\phi(\widetilde{\mathbf{z}}'\widehat{\pi}), 2\phi(\widetilde{\mathbf{z}}'\widehat{\pi}) + (\widetilde{\mathbf{z}}'\widehat{\pi})^2\phi(\widetilde{\mathbf{z}}'\widehat{\pi}))', \text{ and} \\
\widehat{\mathbf{z}} &\equiv (1, \mathbf{w}', \Phi(\widetilde{\mathbf{z}}'\widehat{\pi}), \Phi(\widetilde{\mathbf{z}}'\widehat{\pi})(\mathbf{w} - \widehat{\boldsymbol{\mu}})', \phi(\widetilde{\mathbf{z}}'\widehat{\pi}), \Phi(\widetilde{\mathbf{z}}'\widehat{\pi}) - (\widetilde{\mathbf{z}}'\widehat{\pi})\phi(\widetilde{\mathbf{z}}'\widehat{\pi}), 2\phi(\widetilde{\mathbf{z}}'\widehat{\pi}) + (\widetilde{\mathbf{z}}'\widehat{\pi})^2\phi(\widetilde{\mathbf{z}}'\widehat{\pi}))'.
\end{aligned}$$

Note that some of the regressors and the instruments are generated from the first stage, due to the estimation of $\widehat{\pi}$ and $\widehat{\boldsymbol{\mu}}$ ($\widehat{\boldsymbol{\mu}} = \overline{\mathbf{w}}$). Therefore, the actual model used for a random sample ($i = 1, 2, \cdots, N$) is:

$$\begin{aligned}
&y_i = \widehat{\mathbf{x}}_i'\theta + \widetilde{\widetilde{e}}_i = \widehat{\mathbf{x}}_i'\theta + (\mathbf{x}_i - \widehat{\mathbf{x}}_i)'\theta + \widetilde{e}_i, \ \mathbb{E}(\widetilde{e}_i|\widetilde{\mathbf{z}}_i) = 0, \text{ and} \\
&d_i = 1\{\widetilde{\mathbf{z}}_i'\pi + v_i > 0\}, \text{ where } v_i \sim \text{i.i.d. } N(0,1).
\end{aligned}$$

An IV-like correction function estimator ($\widehat{\theta}_{\text{crrf}}$) is:

$$\widehat{\theta}_{\text{crrf}} = \left(\sum_{i=1}^{N} \widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\left(\sum_{i=1}^{N} \widehat{\mathbf{z}}_i y_i\right). \tag{15}$$

We next present the estimator's large sample properties, with its asymptotic variance adjusted due to the generated regressors.

Theorem 4 *With selection on unobservables, Assumption 2, Assumption 3, and the following assumptions:*

$\mathbb{E}(u_1 - u_0|v) = \mathbb{E}(u_1 - u_0|v, z^*) = \xi_1 v + \xi_2 v^2 + \xi_3 v^3$, *$v \sim N(0,1)$, and* $g_0(z^*) = \beta_0 + \mathbf{w}'\beta_1$,

we have $\widehat{\theta}_{crrf} \xrightarrow{p} \theta$, *and*

$$\sqrt{N}(\widehat{\theta}_{crrf} - \theta) \xrightarrow{d} N\left(\mathbf{0}, A_0^{-1}\Omega A_0^{\prime -1}\right),$$

where

$$\begin{aligned}
A_0 &\equiv \mathbb{E}(\mathbf{z}\mathbf{x}'),\ \Omega \equiv \mathbb{V}(\mathbf{z}\widetilde{e} - B_0\mathbf{r}(\pi) - B_1(\mathbf{w} - \boldsymbol{\mu})),\ \widetilde{e} \equiv y - \mathbf{x}'\theta, \\
B_0 &\equiv \mathbb{E}\left[\left(2\xi_3 - (\widetilde{\mathbf{z}}'\pi)^2 + \xi_2\widetilde{\mathbf{z}}'\pi - \xi_1\right)(\widetilde{\mathbf{z}}'\pi)\phi(\widetilde{\mathbf{z}}'\pi)\mathbf{z}\widetilde{\mathbf{z}}'\right],\ B_1 \equiv -\mathbb{E}\left(d\mathbf{z}\right)\gamma',\ \textit{and} \\
\mathbf{r}(\pi) &\equiv \mathbb{E}^{-1}\left(\frac{\phi^2(\widetilde{\mathbf{z}}'\pi)\widetilde{\mathbf{z}}\widetilde{\mathbf{z}}'}{\Phi(\widetilde{\mathbf{z}}'\pi)\left(1 - \Phi(\widetilde{\mathbf{z}}'\pi)\right)}\right)\frac{\phi(\widetilde{\mathbf{z}}'\pi)\widetilde{\mathbf{z}}(d - \Phi(\widetilde{\mathbf{z}}'\pi))}{\Phi(\widetilde{\mathbf{z}}'\pi)\left(1 - \Phi(\widetilde{\mathbf{z}}'\pi)\right)}.
\end{aligned}$$

Proof. See Appendix D. ■

The variance in the limit distribution of $\sqrt{N}(\widehat{\theta}_{\text{crrf}} - \theta)$ can be estimated by the sample moments. This correction function estimator deals with the OVB, (d) in equation (3), and the sorting bias, (e) in equation (3), through a two-stage procedure. In the first stage, we use a probit model to obtain $\widehat{\pi}$ and an estimated propensity score $\Phi(\widetilde{\mathbf{z}}'\widehat{\pi})$. In the second stage, there are separate roles for $\widehat{\pi}$ and $\Phi(\widetilde{\mathbf{z}}'\widehat{\pi})$. We use $\widehat{\pi}$ to construct the correction terms for the sorting, based on Theorem 3. Meanwhile, we use either $\Phi(\widetilde{\mathbf{z}}'\widehat{\pi})$ or simply the eligibility indicator (z) of the RD design as the IV for the actual treatment status (d) to correct for OVB. If the selection occurs only on the observables, we can use the deviation ($\widetilde{v}$) between d and $\Phi(\widetilde{\mathbf{z}}'\pi)$, which is exogenous to u_0 and u_1, to orthogonalize the observables ($g_0(z^*)$) and the unobservables ($\widetilde{e}$), as the RD robust estimator does. This approach leaves $g_0(z^*)$ completely unspecified and therefore enhances the study's internal validity. However, if selection also occurs on the unobservables, such orthogonality conditions do not exist, because the deviation $\widetilde{v}$ between d and $\Phi(\widetilde{\mathbf{z}}'\pi)$ will be correlated with u_0, u_1, or $u_1 - u_0$. The correction function estimator requires parameterizing $g_0(z^*)$ to control for selection bias due to (b) in equation (3). Note that both estimators require Assumption 3 to control for

the treatment effect heterogeneity due to the observables, that is, (c) in equation (3). The correction function estimator has greater external validity than the RD robust estimator, because it can deal with selection bias due to (d) and (e) in equation (3), but it has less internal validity due to the parameterization of $g_0(z^*)$. In the presence of multiple selection biases, that is, parts (b) to (e) in equation (3), the choice between the RD robust estimator and the correction function estimator represents a balance between a study's internal and external validity. Such a balance can be guided by the research question or the study population of interest. The RD robust estimator also can be useful for assessing an initial eligibility cutoff if it may need to change in the future. In this situation, evaluating the program's impact for a population close to the initial cutoff point would be necessary. If policymakers want to either extend the initial program to a larger population ("farther away" from the eligibility cutoff point) or make the program mandatory for the entire population, they must identify ATE in the presence of multiple selection biases, that is, parts (b) to (e) in equation (3), and also consider the correction function estimator.

3 Monte Carlo Experiments

In this section, we conduct a series of Monte Carlo experiments to investigate the finite sample performance of the proposed estimators with various sample sizes.[16] We also demonstrate the trade-off between efficiency and bias in estimating the ATE when the effects covary with observables and unobservables.

[16] The series of Monte Carlo experiments uses sample sizes of 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000, 2,000, 3,000, 4,000, 5,000, and 10,000. All simulations are based on 1,000 trials.

3.1 Design and Estimators

We specify the data-generating process (DGP) as follows: For the selection process, we assume that the selection variable, z^*, is uniformly distributed with support $[-1, 1]$. Therefore, $\mathbb{E}(z^*) = 0$, and $\mathbb{V}(z^*) = 1/3$. The eligibility indicator, z, is a binary variable defined (without loss of generality) as $1\{z^* \leqslant 0\}$. The binary treatment status, d, is defined as $1\{\pi_0 + \pi_1 z + \pi_2 z^* + v > 0\}$, where the unobservable in the selection process, v, is assumed to have a standard normal distribution, $N(0, 1)$. An RD design ensures that π_1 will be nonzero, so in the treatment probability, a discontinuity occurs when z^* equals 0. The magnitude of this discontinuity is

$$\delta \equiv \lim_{z^* \downarrow 0} \Pr(d = 1|z^*) - \lim_{z^* \uparrow 0} \Pr(d = 1|z^*). \tag{16}$$

We specify the expectation of the untreated potential outcome, conditional on the selection variable (z^*), $\mathbb{E}(y_0|z^*)$, which corresponds to (b) in equation (3), as follows:

$$\mathbb{E}(y_0|z^*) \equiv g_0(z^*) = \begin{cases} \beta_0 + \beta_1 \cos(h) + \beta_2 h^2 & (-h \leqslant z^* \leqslant h) \\ \beta_0 + \beta_1 \cos(z^*) + \beta_2 z^{*2} & (\text{else}), \end{cases} \tag{17}$$

and the error term, u_0, is given by $u_0 = \xi_0 v + \varepsilon$, where ε has a standard normal distribution, $N(0, 1)$, and is independent of v. For the expectation of the treated potential outcome, conditional on the selection variable (z^*), $\mathbb{E}(y_1|z^*)$, we specify the following:

$$\mathbb{E}(y_1|z^*) \equiv g_1(z^*) = \begin{cases} \beta_0 + \beta_1 \cos(h) + \beta_2 h^2 + \mathbb{E}(\eta) & (-h \leqslant z^* \leqslant h) \\ \beta_0 + \beta_1 \cos(z^*) + \gamma_1 z^* + (\beta_2 + \gamma_2) z^{*2} + \mathbb{E}(\eta) & (\text{else}), \end{cases} \tag{18}$$

and the error term, u_1, is given by $u_1 = (\xi_0 + \xi_1)v + \xi_2 v^2 + \varepsilon$. The η, assumed to be normally distributed as $N(1, 1)$, represents the effect due solely to the treatment,

which is independent of z^*. This point corresponds to part (a) in equation (3). We focus on two models for $g_j(z^*)$ $(j = 0, 1)$ based on the DGP: Model I with h set to be 0 represents the case when $\mathbb{E}(y_j|z^*)$ is a non-constant function of z^*, which accommodates the situation in which treatment effect heterogeneity exists due to z^*, that is, the presence of part (c) in equation (3). Model II with h set to be 1 represents the case when y_j is mean-independent from z^* within the interval $[-1, 1]$, precluding the presence of part (c) in equation (3). The ATE conditional on z^* is therefore

$$\mathbb{E}(y_1 - y_0|z^*) = \begin{cases} \mathbb{E}(\eta) & (-h \leqslant z^* \leqslant h) \\ \gamma_1 z^* + \gamma_2 z^{*2} + \mathbb{E}(\eta) & (\text{else}). \end{cases} \tag{19}$$

Throughout the experiments, we keep the values of the following parameters fixed: $\delta = 0.5$, $\pi_0 = -\Phi^{-1}((\delta+1)/2)$, $\pi_1 = 2\Phi^{-1}((\delta+1)/2)$, $\pi_2 = -1$, $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 1$, $\gamma_1 = 1$, and $\gamma_2 = 3$. Therefore, we have

$$\text{ATE} \equiv \mathbb{E}(y_1 - y_0) = \begin{cases} 2 & (\text{Model I: } h = 0) \\ 1 & (\text{Model II: } h = 1). \end{cases} \tag{20}$$

The treatment effect heterogeneity $(g_1(z^*) - g_0(z^*))$ due to observables can be explained by $\gamma_1 z^* + \gamma_2 z^{*2}$, which corresponds to part (c) in equation (3). The correlation among the three unobservables, u_0, u_1, and v, depends on ξ_0, ξ_1 and ξ_2. Given that $\mathbb{E}(u_1 - u_0|v)$ is equal to $\xi_1 v + \xi_2 v^2$, according to the DGP, the treatment effect heterogeneity $(u_1 - u_0)$ due to unobservables will correlate with v, which corresponds to part (e) in equation (3), unless both ξ_1 and ξ_2 are equal to 0. We vary the values of ξ_0, ξ_1, and ξ_2 to investigate four cases that differ in terms of how the treatment effects covary with observables and unobservables. The first case occurs when treatment selection relies only on observables such that all ξ_0, ξ_1, and ξ_2 are set to 0. In this case, the selection bias is due to observables, that is, part (b) in equation (3), and

the treatment effects covary with the observable (z^*), that is, part (c) in equation (3). Note that the eligibility indicator (z) can be used as an excluded IV in estimating ATE. In the second case, the treatment selection is based on unobservables but only with OVB, or part (d) in equation (3), where ξ_0 is set to 1, but both ξ_1 and ξ_2 are 0. Treatment selection based on unobservables induces both OVB and sorting bias in the third case, and all ξ_0, ξ_1, and ξ_2 are set to 1. In this case, treatment effects covary nonlinearly with the unobservable (v), which corresponds to part (e) in equation (3). The last case is the same as the third case, except that ξ_2 is 0, so the treatment effects covary with the unobservable (v) only linearly. This case also implies that the joint distribution of these three unobservables (u_0, u_1, v) is consistent with a trivariate normal distribution.

In each trial, based on four criteria—mean bias, median bias, root mean squared error (RMSE), and median absolute error—we evaluate the performance of the RD robust estimator and the correction function estimator compared with the other three estimators in Model I and Model II. To simplify the notation, we define π and $\widetilde{\mathbf{z}}$ as $\pi \equiv (\pi_0, \pi_1, \pi_2)'$ and $\widetilde{\mathbf{z}} \equiv (1, z, z^*)$. We also denote $\widehat{\pi}$ as the estimate from a probit model for d. For the proposed RD robust estimator (robust), in Model I, we regress y on $d - \Phi(\widehat{\pi}'\widetilde{\mathbf{z}})$ and $\left(d - \Phi(\widehat{\pi}'\widetilde{\mathbf{z}})\right)(\mathbf{w} - \overline{\mathbf{w}})$ by ordinary least squares (OLS), where $\mathbf{w} = (z^*, z^{*2})$. In Model II, we regress y on d by IV using $d - \Phi(\widehat{\pi}'\widetilde{\mathbf{z}})$ as the instrument for d. To gauge the efficiency loss on the basis of no specification errors, we compare this RD robust estimator with Robinson's (1988) two-stage estimator. In both models, in the first stage, we regress y on $\mathbf{w}$ by OLS, obtaining the residual $\widetilde{y}$; in the second stage, we regress $\widetilde{y}$ on $d - \Phi(\widehat{\pi}'\widetilde{\mathbf{z}})$ and $\left(d - \Phi(\widehat{\pi}'\widetilde{\mathbf{z}})\right)(\mathbf{w} - \overline{\mathbf{w}})$ by OLS, where $\mathbf{w} = (z^*, z^{*2})$. For the proposed correction function estimator (corr func), in Model I, we regress y on 1, d, $\mathbf{w}_1$, $d(\mathbf{w}_2 - \overline{\mathbf{w}}_2)$, $\phi(\widehat{\pi}'\widetilde{\mathbf{z}})$, and $\Phi(\widehat{\pi}'\widetilde{\mathbf{z}}) - (\widehat{\pi}'\widetilde{\mathbf{z}})\phi(\widehat{\pi}'\widetilde{\mathbf{z}})$ by lV using $\Phi(\widehat{\pi}'\widetilde{\mathbf{z}})$ and $\Phi(\widehat{\pi}'\widetilde{\mathbf{z}})(\mathbf{w}_2 - \overline{\mathbf{w}}_2)$ as the (excluded) instruments for d and $d(\mathbf{w}_2 - \overline{\mathbf{w}}_2)$, where $\mathbf{w}_1 = (\cos(z^*), z^*)$, and $\mathbf{w}_2 = (z^*, z^{*2})$. In Model II, we regress y on 1, d, $\phi(\widehat{\pi}'\widetilde{\mathbf{z}})$, and

$\Phi(\widehat{\pi}'\widetilde{\mathbf{z}})-(\widehat{\pi}'\widetilde{\mathbf{z}})\phi(\pi'\widetilde{\mathbf{z}})$ by lV using $\Phi(\widehat{\pi}'\widetilde{\mathbf{z}})$ as the (excluded) instrument for d. To focus on the bias reduction in estimating ATE due to the correction function estimator's ability to deal with OVB and sorting, we preclude its specification errors in $g_0(z^*)$ and $g_1(z^*)$. We also compare this correction function estimator with the control function estimator (ctrl func), on the basis of no specification errors, to gauge the efficiency loss if there is only linear sorting (ξ_2 equal to 0) and the bias reduction if the sorting is nonlinear (ξ_2 not equal to 0). In Model I, we regress y on 1, d, $\mathbf{w}_1$, $d(\mathbf{w}_2-\overline{\mathbf{w}}_2)$, $d(\phi(\widehat{\pi}'\widetilde{\mathbf{z}})/\Phi(\widehat{\pi}'\widetilde{\mathbf{z}}))$, and $(1-d)(\phi(-\widehat{\pi}'\widetilde{\mathbf{z}})/\Phi(-\widehat{\pi}'\widetilde{\mathbf{z}}))$ by OLS, where $\mathbf{w}_1=(\cos(z^*),z^*)$ and $\mathbf{w}_2=(z^*,z^{*2})$. In Model II, we regress y on 1, d, $d(\phi(\widehat{\pi}'\widetilde{\mathbf{z}})/\Phi(\widehat{\pi}'\widetilde{\mathbf{z}}))$, and $(1-d)(\phi(-\widehat{\pi}'\widetilde{\mathbf{z}})/\Phi(-\widehat{\pi}'\widetilde{\mathbf{z}}))$ by OLS. We also consider the OLS estimator with specification errors in $g_0(z^*)$ and $g_1(z^*)$. In Model I, we regress y on 1, d, z^*, and $d(z^*-\overline{z^*})$. In Model II, we regress y on 1 and d. We use the OLS estimator as the baseline for evaluating the efficiency-bias trade-off of the RD robust estimator and the correction function estimator in estimating ATE in the presence of multiple selection biases.

3.2 Results and Discussion

The Monte Carlo experiments show that the proposed estimators perform reasonably well in finite samples. The efficiency-bias trade-off in estimating ATE with various selection biases considered—(b) to (e) in equation (3)—persists in large samples. The results in Table 1 are based on a small sample with 100 observations and a relatively large sample with 1,000 observations. The associated figures demonstrate the estimators' performance relative to the alternative based on three criteria—mean bias, median bias, and RMSE—with a range of sample sizes from 100 to 10,000. All simulations use 1,000 trials.

When the selection bias is only due to observables, in the presence of (b) and (c) in equation (3), the results in Table 1 for Model I reveal that with the fuzzy RD design,

the OLS estimator suffers from severe attenuation bias due to its misspecification in $g_0(z^*)$ and $g_1(z^*)$. The downward bias is approximately 30% in terms of mean bias and 29% in terms of median bias with a small sample of 100 observations. Such a downward bias remains near 28% in terms of both mean bias and median bias when the sample size reaches 1,000 observations. Furthermore, this attenuation bias is approximately 30% for both mean bias and median bias, even with a sample of 10,000 observations, as we show in Figure 1. In contrast, the proposed RD robust estimator, which avoids specifying $g_0(z^*)$ and uses the RD's eligibility indicator as an IV in the d-equation, reduces the attenuation bias to 3% and 4% in terms of mean and median bias, respectively, with 100 observations. The usefulness of this RD robust estimator in removing the attenuation bias due to the misspecification of $g_0(z^*)$ becomes clear with a relatively large sample of 1,000 observations. It results in a slightly upward mean and median bias of 0.4% and 0.1%, respectively. The improvement on the attenuation bias relies on using the first-stage residual, $d - \mathbb{E}(d|z, z^*)$, as an IV to orthogonalize $g_0(z^*)$ and the error terms in the second stage. Those orthogonality conditions largely remove specification errors in $g_0(z^*)$ and the measurement errors. In Model I, the RD's eligibility indicator, which serves as an IV in the d-equation helps reduce the attenuation bias significantly, as Table 1 and Figure 1 show. Furthermore, compared with the Robinson (1988) two-stage estimator and according to the correct specification of $g_0(z^*)$ and $g_1(z^*)$, the efficiency loss that occurs from using the RD robust estimator becomes fairly small when the sample size reaches 200 observations. The kink point in the RMSE panel of Figure 2 demonstrates this loss.

When the OVB and nonlinear sorting bias are both present, the OLS estimator suffers from a severe upward bias in both mean and median bias and in both Model I and II. Figures 3 and 4 show that the magnitude of the upward bias stays relatively constant across the entire range of the sample, from 100 to 10,000 observations. In sharp contrast, in the presence of OVB and nonlinear sorting bias, with a small

sample of 100 observations under Model II, the correction function estimator reduces the upward bias to -24% in mean bias and down to -0.005% in median bias. Thus, the correction function estimator can be sensitive to outliers in a small sample. However, Table 1 shows that in Model II, when the sample size increases to 1,000 observations, the correction function estimator cuts the upward bias down to 4% and 3% in mean and median bias, respectively. Therefore, the correction function estimator appears more reliable in bias reduction as the sample size increases. When OVB and nonlinear sorting exist, Figure 3 shows that the correction function estimator is uniformly more robust, in terms of the reduction in mean and median bias, than is the control function estimator in Model II. Furthermore, Figure 4 shows that if Model II contains only linear sorting and the underlying joint distribution of (u_0, u_1, v) is consistent with a trivariate normal distribution, the efficiency loss of the correction function estimator, traded for its bias reduction, can be greatly reduced when the sample size reaches 400 observations. In Model I, with treatment effect heterogeneity due to z^*, the correction function estimator does not work well in terms of mean and median bias with moderate sample sizes unless the sample size gets very large.

4 Empirical Application

In this section, we reexamine an education program evaluation studied by Chay, McEwan, and Urquiola (2005) based on a fuzzy RD design. We show the improvement associated with using the proposed RD robust estimator on bias reductions and efficiency gains relative to the two-stage least squares (2SLS) estimator used by Chay, McEwan, and Urquiola (2005). When the treatment selection is based on observables only, the RD robust estimator reduces the bias, because it can eliminate specification errors in the untreated potential outcome. Such specification errors may occur when estimating the conditional expectation of the observed outcome using 2SLS. The RD

robust estimator also offers efficiency gains because of the over-identifying restriction imposed by the RD's eligibility criterion, which is exclusive to the selection process, and the propensity score estimated in the first stage (Hirano, Imbens, and Ridder, 2003). This improvement with selection on observables can be falsified by the presence of sorting bias. When heterogeneous treatment effects are induced by sorting, ignoring the sorting will misstate the actual program effects. The correction function estimator suggests a specification test based on the significance of its correction terms.[17] If sorting is detected in a study's chosen population, the assumption of selection on observables should be rejected. Such a specification check is implied by the RD's IV nature; the construction of those correction terms is detailed in Theorem 3. The signs of the correction terms indicate the existence of cream skimming or adverse selection.

To improve school performance, Chile's government initiated the "900 School Program" (P-900, henceforth), a countrywide intervention to target low-performing and publicly funded schools (Chay, McEwan, and Urquiola, 2005) in 1990.[18] Eligibility for this program, based on which approximately 900 schools would be selected, is determined by school-level average test scores of fourth graders in 1988. Specifically, this program's participation was largely determined by whether a school's average test score fell below a cutoff score in its region, chosen by the Ministry of Education. As Chay, McEwan, and Urquiola (2005) emphasize, the schools' 1988 test scores were

[17] Because these correction terms are constructed from the data, we confront the problem of generated regressors. Wooldridge (2002) shows that inferences based on the usual t-statistic, under the null hypothesis, are still valid. However, if the null hypothesis is suspected, a correction should be made for the generated regressor problem. The bootstrap can be used to deal with such problems.

[18] There are four interventions associated with this program: (1) infrastructure improvement, such as building repairs; (2) new instructional materials, including textbooks for students from grades 1 to 4, small classroom libraries, cassette recorders, and copy machines; (3) training workshops (focusing on teaching language and mathematics) for school teachers conducted by local supervisors of the Ministry of Education; and (4) after-school tutoring workshops for third and fourth graders who did not perform well enough relative to their grade level. Each workshop was guided by two trained aides recruited from graduates of local secondary schools. Interventions (1) and (2) were the focus of the first two years (1990 and 1991), and P-900 was expanded to include (3) and (4) in 1992.

collected under a different political regime, at which time there was no anticipation of such an intervention. Therefore, it is plausible that schools had no incentive to manipulate their test performance in 1988 to qualify for the P-900.

The actual P-900 assignment involved two stages. During the first stage in 1988, the Ministry of Education administered countrywide achievement tests to the population of fourth graders. Officials of the Ministry then calculated each school's average test scores in language and mathematics and the average of both averages. These scores were ranked from the highest to the lowest in each of Chile's 13 administrative regions. Separate cutoff scores for each region were determined by the Ministry. Schools for which the overall average fell below the within-region's cutoff score were eligible to participate in the P-900. In the second stage, regional teams of officials added two criteria to filter out some eligible schools. First, to lower program costs, some very small or inaccessible schools were excluded, in part because a parallel program was designed to accommodate them. Second, schools were removed from the preselected list if they had managerial problems, such as misreported enrollment. Using their own discretion, the regional teams also included certain schools that were ineligible according to the first-stage criteria.[19] From a school's perspective, there was no incentive to forgo participation, because the national government covered the full costs. Accordingly, the deviation of schools' P-900 status from their initial eligibility is largely due to unobserved criteria introduced by the program's administers.[20]

To be consistent with Chay, McEwan, and Urquiola (2005), we focus on whether P-900 had significant effects on the test score (mathematics and language) gains of fourth graders over the period 1988-1992. As they point out, average test scores in 1988 may provide a noisy measure of school performances and a misleading ranking of schools. For example, a school's appearance at the bottom of the ranking and its selection

[19] For details, see table 2 in Chay, McEwan, and Urquiola (2005).

[20] For the outcomes of the actual two-stage program assignment and the deviations from the test-score based initial eligibility, see figure 3 in Chay, McEwan, and Urquiola (2005).

into P-900 may be the result of transitory bad luck in the testing year. Because transitory noise can be mean reverting, test scores in this school would rebound in the next period, even in the absence of the P-900 intervention, unless that bad luck is persistent. Thus, ignoring the mean-reversion noise will overstate P-900's actual effect.[21] As Chay, McEwan, and Urquiola (2005) suggest, we can effectively remove the impact of the mean-reversion noise by controlling for a smooth function of the 1988 test score, close to the cutoff point, using RD's borderline experimental nature. Because schools closer to the assignment cutoff are more likely to be randomized into the treatment, mean-reversion noises experienced by these schools are more likely to be similar on average. Thus, the direct impact of a common mean-reversion noise can be absorbed by the intercept term included in the outcome equation. The indirect impact of the mean-reversion noise goes through the 1988 test score, so we can use the 1988 test score as a proxy variable for the transitory noise.[22] The RD design's selection cutoff provides a unique tool to deal with the direct impact of unobservables, such as the mean-reversion noise. However, as we focus on the population near the cutoff, where OVB can be precluded by removing the direct impact of the mean-reversion noise, as argued by Chay, McEwan, and Urquiola (2005), we still need to deal with the sorting bias due to the interaction between the treatment and the unobservables. To detect such sorting biases, we use a t-test for the correction terms, as suggested by Wooldridge (2002).

Table 2 provides estimates of P-900 effects. To deal with selection bias due to observables, we follow Chay, McEwan, and Urquiola (2005), controlling for school-

[21] Chay, McEwan, and Urquiola (2005) find that "transitory noise in average scores, and the resulting mean reversion, lead conventional estimation approaches to overstate greatly the positive impact of P-900. For example, difference-in-differences estimates suggest that P-900 increased 1988-1992 test score gains by 0.4 to 0.7 standard deviations; yet using P-900-type assignment rules, we can generate similar effects during earlier periods in which the program was not yet in operation (1984-88)." Figure 5 in Chay, McEwan, and Urquiola (2005) provides evidence of mean-reversion noises and the program's impact.

[22] In this case, the mean-reversion noise turns into a classical measurement error in the actual 1988 test score.

level socioeconomic status (SES),[23] because P-900 may have encouraged the children of some households to exit or enter the treated schools if parents interpreted program selection as a signal that the institution was not adequately serving their children or if they thought their children could benefit from additional resources. The construction of the correction terms follows Theorem 4.

The 2SLS in Panel A is the one proposed by Chay, McEwan, and Urquiola (2005), which uses the eligibility indicator as the instrument for the P-900 treatment status. We propose the RD robust estimator in Panel A, which uses a probit model for the P-900 treatment status. The regressors included in the probit model in addition to the eligibility indicator (the excluded instrument) are cubic polynomials for the 1988 average test scores, SES in 1990, and the changes in SES between 1990 and 1992. The RD robust estimator uses the first-stage residual, which is the deviation between the P-900 treatment status and the estimated treatment probability, as the instrument for the actual treatment status. The correction function estimator with the cubic polynomial correction terms that we propose appears in Panel B. The correction function estimator with only the linear sorting correction term in Panel B has been proposed by Wooldridge (2002). For cases (1)-(4) in Panels A and B, we include in the outcome equation the P-900 dummy, cubic polynomials for the 1988 average test score, SES in 1990, and the changes in SES between 1990 and 1992, to be consistent with Chay, McEwan, and Urquiola (2005, table 5).

The results depicted in this table highlight the following: First, in Panel A, case (1), the efficiency-bias trade-off emerges when 2SLS is applied. With the full sample, the P-900 effect, which is approximately 0.32 standard deviations in both mathematics and language, is statistically significant. However, in the presence of OVB and the sorting bias, which are likely to exist in the full sample, this estimate is biased

[23] The SES index measures student SES, as reported by the Junta Nacional de Auxilio Escolar y Beca. It is scaled 0-100, with higher values indicating higher SES.

and loses the ATE interpretation, because the sorting bias has been detected by the quadratic correction term in the correction function estimator, as shown in Panel B and the "Full Sample" column. As we focus on the schools close to the selection threshold, the OVB can be removed because the direct impact of the mean-reversion noise is likely to be similar on average between schools just above and those just below the threshold. Without the sorting bias near the selection threshold, the estimates of P-900 effects will regain the ATE interpretation, though at the cost of efficiency. As columns "±5 Points" and "±2 Points" in Panel A show, the 2SLS estimates become statistically insignificant.

Second, in Panel A, case (2), the RD robust estimator improves on the efficiency-bias trade-off. Similar to 2SLS, the RD robust estimator shows that P-900 has a significant effect of roughly 0.31 standard deviations in both mathematics and language with the full sample. However, such effects can be biased in the presence of OVB and the sorting bias, which suggests they have limited internal validity. Considering the schools near the threshold for which the impacts of unobservables on average can be removed and internal validity therefore enhanced, the RD robust estimator detects P-900 effects between 0.22 and 0.34 standard deviations at 1% to 5% significance levels. Compared with the existing RD estimators focusing on the effect only at the threshold, this proposed RD robust estimator can obtain the ATE for a predefined population local to the threshold with greater external validity. Compared with the commonly used 2SLS in empirical RD applications, this estimator also has greater internal validity because it can avoid specification errors that 2SLS incurs when estimating the conditional expectation of the untreated outcomes.

Third, the correction terms in Panel B, case (3), controlling for the sorting, are useful for testing the selection-on-observables assumption on which the 2SLS and the RD robust estimator rely. With the full sample, the sorting bias gets detected by the quadratic sorting correction terms of the correction function estimator (for both

mathematics and language). Thus, the estimated P-900 effect for the full sample, given by either the 2SLS or the RD robust estimator in Panel A (roughly 0.3 standard deviations and statistically significant), is biased and has limited internal validity. In contrast, in the columns "±5 Points" and "±2 Points" in Panel B, none of the sorting correction terms of the correction function estimator are statistically significant, which suggests the absence of either adverse selection or cream skimming. In this sense, the RD robust estimator provides the estimated ATE for P-900, which is roughly 0.3 standard deviations for a population of schools close to the selection threshold. The correction terms of the correction function estimator in Panel B, case (3), confirm the validity of this ATE interpretation.

Fourth, the correction function estimator proposed by Wooldridge (2002), in Panel B, case (4), forces the sorting to be globally linear, either positive (cream skimming) or negative (adverse selection). The contrast in the P-900 effect estimates appears for both mathematics and language in cases (3) and (4) in the "Full Sample" column. The proposed correction function estimator detects a U-shaped sorting pattern, based on the positive sign of the quadratic sorting correction term. Furthermore, the cubic sorting correction term is insignificant for both mathematics and language, which confirms the U-shaped sorting pattern. In contrast, Wooldridge's (2002) estimator detects the negative sorting only, which leads to a possibly over-estimated program effect when the positive sorting is ignored. For both the mathematics and language gain scores, the magnitude of Wooldridge's (2002) P-900 estimates is approximately 55%-60% greater than those given by the proposed correction function estimator with cubic correction terms, which can take into account the nonlinear sorting.

The correction function estimator also has greater external validity than the RD robust estimator, because the former is aimed at the ATE for the entire population, which is the effect for a treatment that is supposed to be mandatory for every subject in the population. The latter instead applies to a population close to the selection

cutoff. However, this gain in external validity comes at the cost of internal validity, because the parametric assumptions imposed by the correction function estimator make it susceptible to specification errors. In contrast, the RD robust estimator trades external validity for greater internal validity, because it can avoid certain specification errors and therefore gives compelling estimates for a chosen population near the threshold. In Table 2, the estimated program effect given by the RD robust estimator shows that the 1988-1992 gain score of P-900 schools is approximately 0.3 standard deviations higher than the non-P-900 schools. Therefore, the average P-900 school achieves approximately 62% of the non-P900 school distribution, which suggests a moderate improvement for a population of schools close to the selection threshold. On the basis of this estimate, P-900 effectiveness can be used to construct some cost-benefit measures, such as the per-student expenditure necessary to raise average test score by 0.1 standard deviation. To conduct a cost-benefit analysis for the entire population of schools, not just those close to the selection threshold, we would need to use the correction function estimator. In the absence of specification errors, this estimator gives the program's impact on a randomly selected school from the entire population, which is the approximately 0.7 standard deviations gain shown in Table 2.

The RD robust estimator also gives compelling ATE estimates for a population near the selection threshold, whereas the correction function estimator with correct parameterization offers ATE estimates for the entire population. The plausibility of the ATE identification using the correction function estimator depends on the point in the distribution at which the discontinuity occurs. Choosing between the RD robust estimator and the correction function estimator represents the balance between a study's internal and external validity, which should be guided by the research question or the population of interest, well before any estimations take place. The RD robust estimator further allows for the efficient estimation of an average effect in a range

of observations local to the discontinuity. Within that range, the correction function estimator suggests specification tests for the validity of an ATE interpretation, which will be invalidated in the presence of sorting. In the P-900 example, comparing the gains of schools just above and just below the RD's selection cutoff effectively eliminates the direct impact of mean-reversion noises. The RD robust estimator thus appears able to improve the efficiency-bias trade-off that arises in the presence of treatment effect heterogeneity. On the one hand, these benefits depend on the validity of the RD design's borderline experiments; on the other hand, the RD design's instrumental nature provides a specification check for the plausibility of this quasi-experiment and the validity of an ATE interpretation for the chosen population. The strategies illustrated herein, which integrate the RD design's dual nature for compelling inference, should be applicable whenever tests or other "prescores" with observable selection cutoffs serve to allocate a program and the cutoffs themselves induce variations that are exclusive to the selection process.

5 Conclusion

Previous work on the RD design emphasizes the identification and estimation of an effect at the selection threshold, which pinpoints the measurement of the size of the discontinuity. This paper discusses treatment effect evaluations with the virtue of the RD design's dual nature—that is, a borderline experiment provided near the threshold and a strong and valid exclusion restriction provided in the selection equation for the choice of treatment. Focusing on the fuzzy RD design, this paper proposes two new estimators to deal with multiple selection biases. The first, RD robust estimator, is applicable to a population close to the threshold, where selection on observables can be justified by RD's borderline experiment. This estimator avoids specification errors in the conditional expectation of the untreated potential outcome. The efficiency gain

associated with this estimator is due largely to RD's exclusion restriction, as provided in the selection equation. It also allows for the interaction between the observables and the treatment, which requires parametric assumptions about the treatment effect heterogeneity. To deal with selection on unobservables, this paper attempts to integrate RD literature with broader literature on selection biases when one has a valid exclusion restriction and thus proposes a second estimator—the correction function estimator—that takes into account the heterogeneous sorting through moment restrictions between unobservables in the selection and potential outcome equations. This proposed estimator shares the same parametric nature with Wooldridge's (2002) correction function estimator, but it extends the existing approach to the case in which a nonlinear sorting pattern (with both cream skimming and adverse selection) is allowed and can be detected by adding the cubic polynomial correction terms to Wooldridge's (2002) linear specification. Both proposed estimators are easy to implement using standard software. In addition, this research examines their large and small sample performances through a series of Monte Carlo experiments. We reexamine an existing empirical study to show the improvement that both estimators bring to the efficiency-bias trade-off in evaluating an education program assigned on the basis of test score cutoffs. Choosing between the RD robust estimator and the correction function estimator involves a careful balance between a study's internal validity and its external validity, and this balance should be guided by the research question or the population of interest. For example, evaluating the program impact for a population close to the initial cutoff point would be necessary if the cutoff might be changed later, and the RD robust estimator can be useful in this assessment. If the program is likely to be made mandatory for the entire population, identifying ATE in the presence of selection bias due to both observables and unobservables would be necessary, and the correction function estimator therefore should be considered.

A Proof of Theorem 1

In Section 2.1, the observed outcome is written as

$$\begin{aligned} y &= g_0(z^*) + \alpha d + (\lambda(z^*) - \mathbb{E}(\lambda(z^*)))d + \widetilde{e},\ \widetilde{e} \equiv e + d(\eta - \mathbb{E}(\eta)),\ e \equiv u_0 + d(u_1 - u_0) \\ &\Rightarrow\ y - \mathbb{E}(y|z, z^*) = \alpha(d - \mathbb{E}(d|z, z^*)) + (\lambda(z^*) - \mathbb{E}(\lambda(z^*))(d - \mathbb{E}(d|z, z^*)) + \widetilde{e}. \quad (21) \end{aligned}$$

Note that with selection on observables, Assumption 2, and $\eta \amalg (d, z^*)$, we have[24]

$$\begin{aligned} \mathbb{E}(\widetilde{e}|z, z^*) &= \mathbb{E}(u_0 + d(u_1 - u_0)|z, z^*) + \mathbb{E}(d|z, z^*)\mathbb{E}(\eta - \mathbb{E}(\eta)) \\ &= \mathbb{E}(d|z, z^*)\mathbb{E}(u_1|z, z^*) - \mathbb{E}(d|z, z^*)\mathbb{E}(u_0|z, z^*) \text{ (because } u_0 \perp d|z^* \text{ and } u_1 \perp d|z^*) \\ &= 0. \quad (22) \end{aligned}$$

We next verify two moment equations:

$$\begin{aligned} 0 &= \mathbb{E}[(d - \mathbb{E}(d|z, z^*))\widetilde{e}] \text{ and} \quad (23) \\ 0 &= \mathbb{E}[(d - \mathbb{E}(d|z, z^*))(\lambda(z^*) - \mathbb{E}(\lambda(z^*))\widetilde{e}]. \quad (24) \end{aligned}$$

With selection on observables and Assumption 2, we have

$$\begin{aligned} \mathbb{E}[(d - \mathbb{E}(d|z, z^*))\widetilde{e}|z^*] &= \mathbb{E}[(d - \mathbb{E}(d|z, z^*))\,(u_0 + d(u_1 - u_0) + d(\eta - \mathbb{E}(\eta)))\,|z^*] \\ &= 0 + \mathbb{E}[(\eta - \mathbb{E}(\eta))\mathbb{E}(\mathbb{V}(d|z, z^*))] = 0, \quad (25) \end{aligned}$$

which implies

$$\begin{aligned} &\mathbb{E}[(d - \mathbb{E}(d|z, z^*))\widetilde{e}] = \mathbb{E}\{\mathbb{E}[(d - \mathbb{E}(d|z, z^*))\widetilde{e}|z^*]\} = 0, \text{ and} \\ &\mathbb{E}[(d - \mathbb{E}(d|z, z^*))(\lambda(z^*) - \mathbb{E}(\lambda(z^*))\widetilde{e}] = 0. \end{aligned} \quad (26)$$

[24] We use "$\perp$" for orthogonality.

Under Assumption 3, we have:

$$
\begin{aligned}
0 &= \mathbb{E}[(d - \mathbb{E}(d|z, z^*))\widetilde{e}], && (27)\\
0 &= \mathbb{E}[(d - \mathbb{E}(d|z, z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))'\gamma\widetilde{e}], \text{ and} && (28)\\
\widetilde{e} &= y - \mathbb{E}(y|z, z^*) - \alpha(d - \mathbb{E}(d|z, z^*)) - (d - \mathbb{E}(d|z, z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))'\gamma. && (29)
\end{aligned}
$$

To simplify notations, we define the following:

$$
\begin{aligned}
x_1 &\equiv d - \mathbb{E}(d|z, z^*),\\
\mathbf{x}_2 &\equiv (d - \mathbb{E}(d|z, z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w})), \text{ and}\\
\widetilde{y} &\equiv y - \mathbb{E}(y|z, z^*),
\end{aligned}
$$

and we have $\widetilde{y} = \alpha x_1 + \mathbf{x}_2'\gamma + \widetilde{e}$, with the following orthogonality conditions:

$$
\mathbb{E}(x_1\widetilde{e}) = 0 \text{ and } \mathbb{E}(\mathbf{x}_2\widetilde{e}) = 0. \tag{30}
$$

Therefore,

$$
\begin{bmatrix} \mathbb{E}(x_1^2) & \mathbb{E}(x_1\mathbf{x}_2') \\ \mathbb{E}(x_1\mathbf{x}_2) & \mathbb{E}(\mathbf{x}_2\mathbf{x}_2') \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} = \begin{bmatrix} \mathbb{E}(x_1\widetilde{y}) \\ \mathbb{E}(\mathbf{x}_2\widetilde{y}) \end{bmatrix}, \tag{31}
$$

where

$$\begin{aligned}
\mathbb{E}(x_1^2) &= \mathbb{E}[\,(d-\mathbb{E}(d|z,z^*))^2] = \mathbb{E}(\mathbb{V}(d|z,z^*)),\\
\mathbb{E}(x_1\mathbf{x}_2) &= \mathbb{E}[\,(d-\mathbb{E}(d|z,z^*))^2\,(\mathbf{w}-\mathbb{E}(\mathbf{w}))] = \mathbb{E}[\mathbb{V}(d|z,z^*)\,(\mathbf{w}-\mathbb{E}(\mathbf{w}))],\\
\mathbb{E}(\mathbf{x}_2\mathbf{x}_2') &= \mathbb{E}[\mathbb{V}(d|z,z^*)\,(\mathbf{w}-\mathbb{E}(\mathbf{w}))\,(\mathbf{w}-\mathbb{E}(\mathbf{w}))'],\\
\mathbb{E}(x_1\widetilde{y}) &= \mathbb{E}\left[(d-\mathbb{E}(d|z,z^*))\,(y-\mathbb{E}(y|z,z^*))\right] = \mathbb{E}\left[(d-\mathbb{E}(d|z,z^*))\,y\right]\\
&= \mathbb{E}(x_1y), \text{ and}\\
\mathbb{E}(\mathbf{x}_2\widetilde{y}) &= \mathbb{E}\left[(d-\mathbb{E}(d|z,z^*))(\mathbf{w}-\mathbb{E}(\mathbf{w}))\,(y-\mathbb{E}(y|z,z^*))\right] = \mathbb{E}\left[(d-\mathbb{E}(d|z,z^*))(\mathbf{w}-\mathbb{E}(\mathbf{w}))y\right]\\
&= \mathbb{E}(x_2y).
\end{aligned}$$

Then we define $\theta \equiv (\alpha,\gamma')'$ and $\mathbf{x} \equiv (x_1,\mathbf{x}_2')'$, and we attain the following "least squares" estimator:

$$\theta = \mathbb{E}^{-1}(\mathbf{x}\mathbf{x}')\,\mathbb{E}(\mathbf{x}y). \tag{32}$$

$$\mathbf{x} = (d-\mathbb{E}(d|z,z^*),(d-\mathbb{E}(d|z,z^*))(\mathbf{w}-\mathbb{E}(\mathbf{w}))')'. \tag{33}$$

We next solve for α and γ separately, using results from Amemiya (1985, p. 460):

$$\begin{bmatrix} \mathbb{E}(x_1^2) & \mathbb{E}(x_1\mathbf{x}_2') \\ \mathbb{E}(x_1\mathbf{x}_2) & \mathbb{E}(\mathbf{x}_2\mathbf{x}_2') \end{bmatrix}^{-1} = \begin{bmatrix} E^{-1} & -E^{-1}BD^{-1} \\ -D^{-1}CE^{-1} & F^{-1} \end{bmatrix}, \tag{34}$$

where

$$\begin{aligned}
A &\equiv \mathbb{E}(x_1^2),\ B \equiv \mathbb{E}(x_1\mathbf{x}_2'),\ C \equiv \mathbb{E}(x_1\mathbf{x}_2),\ D \equiv \mathbb{E}(\mathbf{x}_2\mathbf{x}_2'),\\
E &= A - BD^{-1}C = \mathbb{E}(x_1^2) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2), \text{ and}\\
F &= D - CA^{-1}B = \mathbb{E}(\mathbf{x}_2\mathbf{x}_2') - \mathbb{E}(x_1\mathbf{x}_2)\mathbb{E}^{-1}(x_1^2)\mathbb{E}(x_1\mathbf{x}_2').
\end{aligned}$$

Thus, we have

$$\begin{aligned}\alpha &= E^{-1}\left[\mathbb{E}(x_1 y) - BD^{-1}\mathbb{E}(\mathbf{x}_2 y)\right] \\ &= \frac{\mathbb{E}(x_1 y) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(\mathbf{x}_2 y)}{\mathbb{E}(x_1^2) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)} \equiv \text{ATE},\end{aligned} \tag{35}$$

and

$$\begin{aligned}\gamma &= -D^{-1}CE^{-1}\mathbb{E}(x_1 y) + F^{-1}\mathbb{E}(\mathbf{x}_2 y) \\ &= \left[\mathbb{E}(\mathbf{x}_2\mathbf{x}_2') - \mathbb{E}(x_1\mathbf{x}_2)\mathbb{E}^{-1}(x_1^2)\mathbb{E}(x_1\mathbf{x}_2')\right]^{-1}\mathbb{E}(\mathbf{x}_2 y) \\ &\quad -\frac{\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)\mathbb{E}(x_1 y)}{\mathbb{E}(x_1^2) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)}.\end{aligned} \tag{36}$$

B Proof of Theorem 2

This proof is largely based on Wooldridge's (2002) Chapter 6, Appendix 6A. The regressors included in the model defined at the population are:

$$\mathbf{x} \equiv \left[(d - p(z, z^*; \lambda)), (d - p(z, z^*; \lambda))(\mathbf{w} - \boldsymbol{\mu})'\right]' \equiv \mathbf{f}(d, z, z^*, \mathbf{w}; \lambda, \boldsymbol{\mu}). \tag{37}$$

Some of the regressors included in the actual model are generated from a random sample, $i = 1, 2, \cdots, N$:

$$\widehat{\mathbf{x}} \equiv [(d - p(z, z^*; \widehat{\lambda})), (d - p(z, z^*; \widehat{\lambda}))(\mathbf{w} - \widehat{\boldsymbol{\mu}})']' \equiv \mathbf{f}(d, z, z^*, \mathbf{w}; \widehat{\lambda}, \widehat{\boldsymbol{\mu}}) \tag{38}$$

The actual model used for estimation, based on a random sample, is

$y_i = \widehat{\mathbf{x}}_i'\theta + h(z_i^*) + \widetilde{e}_i = \widehat{\mathbf{x}}_i'\theta + (\mathbf{x}_i - \widehat{\mathbf{x}}_i)'\theta + h(z_i^*) + \widetilde{e}_i$, $\mathbb{E}(\widetilde{e}_i|z_i^*) = \mathbb{E}(\widetilde{e}_i|d_i, z_i^*) = 0$, and $h(z_i^*) = g_0(z_i^*) + \alpha p(z_i, z_i^*; \lambda) + p(z_i, z_i^*; \lambda)(\mathbf{w}_i - \boldsymbol{\mu})'\gamma$, where $p(z, z^*; \lambda) \equiv \mathbb{E}(d|z, z^*)$.

The RD robust estimator is:

$$\widehat{\theta}_{\text{RD-robust}} = \left(\sum_{i=1}^{N} \widehat{\mathbf{x}}_i \widehat{\mathbf{x}}_i'\right)^{-1} \left(\sum_{i=1}^{N} \widehat{\mathbf{x}}_i y_i\right). \tag{39}$$

We next show the consistency and asymptotic normality of $\widehat{\theta}_{\text{RD-robust}}$.

- Consistency

Because $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{w}}$ and $\boldsymbol{\mu} \equiv \mathbb{E}(\mathbf{w})$, the consistency $\widehat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$ holds because of the law of large numbers. If $\widehat{\lambda} \xrightarrow{p} \lambda$ also holds, then, by Slutsky's theorem, we have $p(z, z^*; \widehat{\lambda}) \xrightarrow{p} p(z, z^*; \lambda)$ and $\widehat{\mathbf{x}} \xrightarrow{p} \mathbf{x}$. Therefore,

$$\frac{1}{N} \sum_{i=1}^{N} \widehat{\mathbf{x}}_i \widehat{\mathbf{x}}_i' \xrightarrow{p} \mathbb{E}(\mathbf{x}\mathbf{x}') \equiv A_0. \tag{40}$$

Given that $\mathbb{E}(\widetilde{e}|z^*) = \mathbb{E}(\widetilde{e}|d, z^*) = 0$, we have

$$\begin{aligned} \widehat{\theta}_{\text{RD-robust}} &= \theta + \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{x}}_i \widehat{\mathbf{x}}_i'\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{x}}_i (\mathbf{x}_i - \widehat{\mathbf{x}}_i)'\theta + \frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{x}}_i h(z_i^*) + \frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{x}}_i \widetilde{e}_i\right) \\ &\xrightarrow{p} \theta + \mathbb{E}^{-1}(\mathbf{x}\mathbf{x}') \left[\mathbb{E}(\mathbf{x}h(z^*)) + \mathbb{E}(\mathbf{x}\widetilde{e})\right] = \theta. \end{aligned} \tag{41}$$

Consistency is established straightforwardly.

- Asymptotic Normality

For the RD robust estimator, we have:

$$\sqrt{N}\left(\widehat{\theta}_{\text{RD-robust}} - \theta\right) = \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{x}}_i \widehat{\mathbf{x}}_i'\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \widehat{\mathbf{x}}_i [(\mathbf{x}_i - \widehat{\mathbf{x}}_i)'\theta + y_i - \mathbf{x}_i'\theta]\right). \tag{42}$$

Consider a first-order Taylor expansion for $\widehat{\mathbf{x}} \equiv \mathbf{f}(d, z, z^*, \mathbf{w}; \widehat{\lambda}, \widehat{\boldsymbol{\mu}})$ at $(\lambda', \boldsymbol{\mu}')'$:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \widehat{\mathbf{x}}_i (\mathbf{x}_i - \widehat{\mathbf{x}}_i)'\theta = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} (\theta \otimes \widehat{\mathbf{x}}_i')' (\mathbf{x}_i - \widehat{\mathbf{x}}_i), \tag{43}$$

where

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\theta\otimes\widehat{\mathbf{x}}_i'\right)'\left(\mathbf{x}_i-\widehat{\mathbf{x}}_i\right)\\
=\;&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\theta\otimes\widehat{\mathbf{x}}_i'\right)'\left(-\frac{\partial\mathbf{x}_i}{\partial\widehat{\lambda}'}(\widehat{\lambda}-\lambda)-\frac{\partial\mathbf{x}_i}{\partial\widehat{\boldsymbol{\mu}}'}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})+o_p(1)\right)\\
=\;&-B_0\sqrt{N}(\widehat{\lambda}-\lambda)+o_p(1) \qquad (44)
\end{aligned}$$

with the following definition:

$$B_0\equiv\mathbb{E}\left(\left(\theta\otimes\mathbf{x}'\right)'\frac{\partial\mathbf{x}}{\partial\lambda'}\right). \qquad (45)$$

Now, we have

$$\begin{aligned}
\sqrt{N}\left(\widehat{\theta}_{\text{RD-robust}}-\theta\right) &= \left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i[(\mathbf{x}_i-\widehat{\mathbf{x}}_i)'\theta+h(z_i^*)+\widetilde{e}_i]\right)\\
&= A_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\mathbf{x}_i\left(y_i-\mathbf{x}_i'\theta\right)-B_0\mathbf{r}_i(\lambda)\right)+o_p(1), \text{ and}\\
\sqrt{N}(\widehat{\lambda}-\lambda) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{r}_i(\lambda)+o_p(1), \qquad (46)
\end{aligned}$$

where $\mathbf{r}_i(\lambda)$ is the influence function with $\mathbb{E}(\mathbf{r}_i(\lambda))=\mathbf{0}$. With selection on observables, we also have $\mathbb{E}(\mathbf{x}_i\left(y_i-\mathbf{x}_i'\theta\right))=\mathbb{E}(\mathbf{x}_i(h(z_i^*)+\widetilde{e}_i))=\mathbf{0}$. Applying the central limit theorem to

$$\sqrt{N}\left(\widehat{\theta}_{\text{RD-robust}}-\theta\right)=A_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\mathbf{x}_i\left(y_i-\mathbf{x}_i'\theta\right)-B_0\mathbf{r}_i(\lambda)\right)+o_p(1), \qquad (47)$$

we obtain

$$\sqrt{N}\left(\widehat{\theta}_{\text{RD-robust}}-\theta\right)\xrightarrow{d}N\left(\mathbf{0},A_0^{-1}\Omega A_0^{-1}\right),\text{ where }\Omega\equiv\mathbb{V}\left(\mathbf{x}\left(y-\mathbf{x}'\theta\right)-B_0\mathbf{r}(\lambda)\right), \tag{48}$$

together with the influence function for $\sqrt{N}(\widehat{\lambda}-\lambda)$:

$$\sqrt{N}(\widehat{\lambda}-\lambda)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{r}_i(\lambda)+o_p(1),\ \mathbb{E}(\mathbf{r}_i(\lambda))=\mathbf{0}. \tag{49}$$

Asymptotic normality is thus established.

C Proof of Theorem 3

Apply the law of iterated expectation:

$$\begin{aligned}\mathbb{E}(d(u_1-u_0)|z,z^*) &= \mathbb{E}_v[d\mathbb{E}((u_1-u_0)|z,z^*,v)|z,z^*]\\ &= \mathbb{E}_v[d(\xi_1 v+\xi_2 v^2+\xi_3 v^3)|z,z^*]\\ &= \xi_1\mathbb{E}_v(dv|z,z^*)+\xi_2\mathbb{E}_v(dv^2|z,z^*)+\xi_3\mathbb{E}_v(dv^3|z,z^*).\end{aligned} \tag{50}$$

Compute the following:

$$\begin{aligned}\xi_1\mathbb{E}_v(dv|z,z^*) &= \xi_1\int_{-\infty}^{+\infty}1\{\pi_0+\pi_1 z+\pi_2 z^*+s>0\}s\phi(s)ds\\ &= \xi_1\int_{-\pi_0-\pi_1 z-\pi_2 z^*}^{+\infty}s\phi(s)ds=\xi_1\int_{-\pi_0-\pi_1 z-\pi_2 z^*}^{+\infty}-\phi'(s)ds\\ &= \xi_1\phi(\pi_0+\pi_1 z+\pi_2 z^*)\end{aligned} \tag{51}$$

and

$$\begin{aligned}\xi_2\mathbb{E}_v(dv^2|z,z^*) &= \xi_2\int_{-\infty}^{+\infty}1\{\pi_0+\pi_1 z+\pi_2 z^*+s>0\}s^2\phi(s)ds \qquad (52)\\ &= \xi_2\int_{-\pi_0-\pi_1 z-\pi_2 z^*}^{+\infty}s^2\phi(s)ds=\xi_2\int_{-\pi_0-\pi_1 z-\pi_2 z^*}^{+\infty}(\phi''(s)+\phi(s))ds\\ &= \xi_2\left[\Phi(\pi_0+\pi_1 z+\pi_2 z^*)-(\pi_0+\pi_1 z+\pi_2 z^*)\phi(\pi_0+\pi_1 z+\pi_2 z^*)\right]\end{aligned}$$

and

$$\begin{aligned}\xi_3\mathbb{E}_v(dv^3|z,z^*) &= \xi_3\int_{-\infty}^{+\infty}1\{\pi_0+\pi_1 z+\pi_2 z^*+s>0\}s^3\phi(s)ds \qquad (53)\\ &= \xi_3\int_{-\pi_0-\pi_1 z-\pi_2 z^*}^{+\infty}s^3\phi(s)ds=\xi_3\int_{-\pi_0-\pi_1 z-\pi_2 z^*}^{+\infty}(3s\phi(s)-\phi'''(s))ds\\ &= \xi_3\left[2\phi(\pi_0+\pi_1 z+\pi_2 z^*)+(\pi_0+\pi_1 z+\pi_2 z^*)^2\phi(\pi_0+\pi_1 z+\pi_2 z^*)\right].\end{aligned}$$

D Proof of Theorem 4

This proof is largely based on Wooldridge's (2002) Chapter 6, Appendix 6A. The regressors included in the model defined at the population are:

$$\begin{aligned}\mathbf{x} &\equiv (1,\mathbf{w}',d,d(\mathbf{w}-\boldsymbol{\mu})',\phi(\widetilde{\mathbf{z}}'\pi),\Phi(\widetilde{\mathbf{z}}'\pi)-(\widetilde{\mathbf{z}}'\pi)\phi(\widetilde{\mathbf{z}}'\pi),2\phi(\widetilde{\mathbf{z}}'\pi)+(\widetilde{\mathbf{z}}'\pi)^2\phi(\widetilde{\mathbf{z}}'\pi))'\\ &\equiv \mathbf{f}(d,\widetilde{\mathbf{z}},\mathbf{w};\pi,\boldsymbol{\mu}). \qquad (54)\end{aligned}$$

Some of the regressors included in the actual model are generated from a random sample, $i=1,2,\cdots,N$:

$$\begin{aligned}\widehat{\mathbf{x}}_i &\equiv (1,\mathbf{w}_i',d_i,d_i(\mathbf{w}_i-\widehat{\boldsymbol{\mu}})',\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}),\Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi})-(\widetilde{\mathbf{z}}_i'\widehat{\pi})\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}),2\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi})+(\widetilde{\mathbf{z}}_i'\widehat{\pi})^2\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}))'\\ &\equiv \mathbf{f}(d_i,\widetilde{\mathbf{z}}_i,\mathbf{w}_i;\widehat{\pi},\widehat{\boldsymbol{\mu}}). \qquad (55)\end{aligned}$$

The instruments (both included and excluded) used in the population model are

$$\begin{aligned}\mathbf{z} &\equiv (1, \mathbf{w}', \Phi(\widetilde{\mathbf{z}}'\pi), \Phi(\widetilde{\mathbf{z}}'\pi)(\mathbf{w}-\boldsymbol{\mu})', \phi(\widetilde{\mathbf{z}}'\pi), \Phi(\widetilde{\mathbf{z}}'\pi) - (\widetilde{\mathbf{z}}'\pi)\phi(\widetilde{\mathbf{z}}'\pi), 2\phi(\widetilde{\mathbf{z}}'\pi) + (\widetilde{\mathbf{z}}'\pi)^2\phi(\widetilde{\mathbf{z}}'\pi))' \\ &\equiv \mathbf{g}(\widetilde{\mathbf{z}}, \mathbf{w}; \pi, \boldsymbol{\mu}). \end{aligned} \tag{56}$$

Similarly, some of the instruments included in the actual model are generated from a random sample, $i = 1, 2, \cdots, N$:

$$\begin{aligned}\widehat{\mathbf{z}}_i &\equiv (1, \mathbf{w}_i', \Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}), \Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi})(\mathbf{w}_i-\widehat{\boldsymbol{\mu}})', \phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}), \Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}) - (\widetilde{\mathbf{z}}_i'\widehat{\pi})\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}), 2\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}) + (\widetilde{\mathbf{z}}_i'\widehat{\pi})^2\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}))' \\ &\equiv \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\boldsymbol{\mu}}). \end{aligned} \tag{57}$$

The actual model used for estimation, based on a random sample, is

$$\begin{aligned} &y_i = \widehat{\mathbf{x}}_i'\theta + \widetilde{\widetilde{e}}_i = \widehat{\mathbf{x}}_i'\theta + (\mathbf{x}_i - \widehat{\mathbf{x}}_i)'\theta + \widetilde{e}_i,\ \mathbb{E}(\widetilde{e}_i|\widetilde{\mathbf{z}}_i) = 0. \\ &d_i = 1\{\widetilde{\mathbf{z}}_i'\pi + v_i > 0\},\ \text{where } v_i \sim \text{i.i.d. } N(0,1). \end{aligned}$$

The correction function estimator is:

$$\widehat{\theta}_{\text{crrf}} = \left(\sum_{i=1}^{N} \widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\left(\sum_{i=1}^{N} \widehat{\mathbf{z}}_i y_i\right). \tag{58}$$

We next show the consistency and asymptotic normality of $\widehat{\theta}_{\text{crrf}}$.

- Consistency

Because $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{w}}$ and $\boldsymbol{\mu} \equiv \mathbb{E}(\mathbf{w})$, the consistency $\widehat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$ holds because of the law of large numbers. If $\widehat{\pi} \xrightarrow{p} \pi$ also holds, then, by Slutsky's theorem, we have

$\Phi(\widetilde{\mathbf{z}}'\widehat{\pi}) \xrightarrow{p} \Phi(\widetilde{\mathbf{z}}'\pi)$ and $\phi(\widetilde{\mathbf{z}}'\widehat{\pi}) \xrightarrow{p} \phi(\widetilde{\mathbf{z}}'\pi)$. Therefore, we have

$$\begin{aligned} &\widehat{\mathbf{z}}_i \equiv \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\boldsymbol{\mu}}) \xrightarrow{p} \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu}) \equiv \mathbf{z}_i. \\ &\widehat{\mathbf{x}}_i \equiv \mathbf{f}(d_i, \widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\boldsymbol{\mu}}) \xrightarrow{p} \mathbf{f}(d_i, \widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu}) \equiv \mathbf{x}_i. \end{aligned} \tag{59}$$

$$\frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{z}}_i \widehat{\mathbf{x}}_i' \xrightarrow{p} \mathbb{E}(\mathbf{z}\mathbf{x}') \equiv A_0. \tag{60}$$

Given that $\mathbb{E}(\widetilde{e}|\mathbf{z}) = 0$,

$$\widehat{\theta}_{\text{crrf}} = \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{z}}_i \widehat{\mathbf{x}}_i'\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{z}}_i y_i\right) \xrightarrow{p} \mathbb{E}^{-1}(\mathbf{z}\mathbf{x}')\mathbb{E}(\mathbf{z}\mathbf{x}')\theta + \mathbb{E}(\mathbf{z}\widetilde{e}) = \theta. \tag{61}$$

Consistency is established straightforwardly.

- Asymptotic Normality

Since $v \sim N(0,1)$, $\widehat{\pi}$ is obtained from a probit model, and $\widehat{\pi} \xrightarrow{p} \pi$. For the correction on the asymptotic variance of $\widehat{\theta}_{\text{crrf}}$, recall the influence function representation of a probit model:

$$\begin{aligned} \sqrt{N}(\widehat{\pi} - \pi) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{r}_i(\pi) + o_p(1). \\ \mathbf{r}_i(\pi) &\equiv \mathbb{E}^{-1}\left(\frac{\phi^2(\widetilde{\mathbf{z}}_i'\pi)\widetilde{\mathbf{z}}_i\widetilde{\mathbf{z}}_i'}{\Phi(\widetilde{\mathbf{z}}_i'\pi)\left(1-\Phi(\widetilde{\mathbf{z}}_i'\pi)\right)}\right) \frac{\phi(\widetilde{\mathbf{z}}_i'\pi)\widetilde{\mathbf{z}}_i(d_i - \Phi(\widetilde{\mathbf{z}}_i'\pi))}{\Phi(\widetilde{\mathbf{z}}_i'\pi)\left(1-\Phi(\widetilde{\mathbf{z}}_i'\pi)\right)}, \text{ and } \mathbb{E}\left(\mathbf{r}_i(\pi)\right) = \mathbf{0}. \end{aligned} \tag{62}$$

Similarly, for $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{w}}$, $\widehat{\boldsymbol{\mu}}$ has the following asymptotic properties:

$$\widehat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu} \text{ and } \sqrt{N}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \Sigma_{\mathbf{w}}), \text{ where } \Sigma_{\mathbf{w}} \equiv \mathbb{V}(\mathbf{w}). \tag{63}$$

For the correction function estimator, we have:

$$\sqrt{N}(\widehat{\theta}_{\text{crrf}} - \theta) = \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{z}}_i \widehat{\mathbf{x}}_i'\right)^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \widehat{\mathbf{z}}_i[(\mathbf{x}_i - \widehat{\mathbf{x}}_i)'\theta + \widetilde{e}_i]. \tag{64}$$

Consider a first-order Taylor expansion for $\widehat{\mathbf{z}}_i \equiv \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\boldsymbol{\mu}})$ at $(\pi', \boldsymbol{\mu}')'$:

$$\begin{aligned} & \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \widehat{\mathbf{z}}_i \widetilde{e}_i \\ = \; & \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{z}_i \widetilde{e}_i + \\ & \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu})}{\partial \widehat{\pi}'} \widetilde{e}_i \sqrt{N} \left(\widehat{\pi} - \pi \right) + \frac{\partial \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu})}{\partial \widehat{\boldsymbol{\mu}}'} \widetilde{e}_i \sqrt{N} \left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} \right) \right) + o_p(1). \end{aligned} \tag{65}$$

Because $\mathbb{E}(\widetilde{e}_i | \widetilde{\mathbf{z}}_i) = 0$, we obtain the following results:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu})}{\partial \widehat{\pi}'} \widetilde{e}_i \xrightarrow{p} \mathbb{E} \left(\frac{\partial \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu})}{\partial \widehat{\pi}'} \widetilde{e}_i \right) &= \mathbf{0} \Rightarrow \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu})}{\partial \widehat{\pi}'} \widetilde{e}_i = o_p(1). \\ \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu})}{\partial \widehat{\boldsymbol{\mu}}'} \widetilde{e}_i \xrightarrow{p} \mathbb{E} \left(\frac{\partial \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu})}{\partial \widehat{\boldsymbol{\mu}}'} \widetilde{e}_i \right) &= \mathbf{0} \Rightarrow \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu})}{\partial \widehat{\boldsymbol{\mu}}'} \widetilde{e}_i = o_p(1). \end{aligned}$$

Because $\sqrt{N} \left(\widehat{\pi} - \pi \right) = O_p(1)$, $\sqrt{N} \left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} \right) = O_p(1)$, and $o_p(1) O_p(1) = o_p(1)$,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \widehat{\mathbf{z}}_i \widetilde{e}_i = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{z}_i \widetilde{e}_i + o_p(1). \tag{66}$$

Similarly, we consider a first-order Taylor expansion for $\widehat{\mathbf{x}}_i \equiv \mathbf{f}(d_i, \widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\boldsymbol{\mu}})$ at $(\pi', \boldsymbol{\mu}')'$:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \widehat{\mathbf{z}}_i (\mathbf{x}_i - \widehat{\mathbf{x}}_i)' \theta = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\theta \otimes \widehat{\mathbf{z}}_i')' (\mathbf{x}_i - \widehat{\mathbf{x}}_i), \tag{67}$$

where

$$\begin{aligned} & \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\theta \otimes \widehat{\mathbf{z}}_i')' (\mathbf{x}_i - \widehat{\mathbf{x}}_i) \\ = \; & \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\theta \otimes \widehat{\mathbf{z}}_i')' \left(- \frac{\partial \mathbf{f}(d_i, \widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu})}{\partial \widehat{\pi}'} \left(\widehat{\pi} - \pi \right) - \frac{\partial \mathbf{f}(d_i, \widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \boldsymbol{\mu})}{\partial \widehat{\boldsymbol{\mu}}'} \left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} \right) + o_p(1) \right) \\ = \; & - B_0 \sqrt{N} \left(\widehat{\pi} - \pi \right) - B_1 \sqrt{N} \left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} \right) + o_p(1), \end{aligned} \tag{68}$$

with the following definitions:

$$\begin{aligned} B_0 &\equiv \mathbb{E}\left((\theta \otimes \mathbf{z}')' \frac{\partial \mathbf{f}(d, \widetilde{\mathbf{z}}, \mathbf{w}; \pi, \boldsymbol{\mu})}{\partial \pi'}\right) = \mathbb{E}\left[\left(2\xi_3 - (\widetilde{\mathbf{z}}'\pi)^2 + \xi_2 \widetilde{\mathbf{z}}'\pi - \xi_1\right)(\widetilde{\mathbf{z}}'\pi)\phi(\widetilde{\mathbf{z}}'\pi)\mathbf{z}\widetilde{\mathbf{z}}'\right], \text{ and} \\ B_1 &\equiv \mathbb{E}\left((\theta \otimes \mathbf{z}')' \frac{\partial \mathbf{f}(d, \widetilde{\mathbf{z}}, \mathbf{w}; \pi, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}'}\right) = -\mathbb{E}\left(d\mathbf{z}\right)\gamma'. \end{aligned} \tag{69}$$

Combining the expansion results for both $\widehat{\mathbf{z}}_i \equiv \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\boldsymbol{\mu}})$ and $\widehat{\mathbf{x}}_i \equiv \mathbf{f}(d_i, \widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\boldsymbol{\mu}})$, we attain

$$\sqrt{N}(\widehat{\theta}_{\text{crrf}} - \theta) = A_0^{-1}\left(-B_0\sqrt{N}\left(\widehat{\pi} - \pi\right) - B_1\sqrt{N}\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{z}_i\widetilde{e}_i\right) + o_p(1). \tag{70}$$

We next derive the influence function representation for $\widehat{\theta}_{\text{crrf}}$, substituting the results from the probit model:

$$\sqrt{N}(\widehat{\theta}_{\text{crrf}} - \theta) = A_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\mathbf{z}_i\widetilde{e}_i - B_0\mathbf{r}_i(\pi) - B_1(\mathbf{w}_i - \boldsymbol{\mu})\right] + o_p(1). \tag{71}$$

According to the condition $\mathbb{E}(\widetilde{e}|\mathbf{z}) = 0$, $\mathbb{E}\left(\mathbf{r}_i(\pi)\right) = \mathbf{0}$, we note

$$\mathbb{E}(\mathbf{z}_i\widetilde{e}_i - B_0\mathbf{r}_i(\pi) - B_1(\mathbf{w}_i - \boldsymbol{\mu})) = \mathbb{E}(\mathbf{z}_i\widetilde{e}_i) - \mathbb{E}(B_0\mathbf{r}_i(\pi)) - \mathbb{E}(B_1(\mathbf{w}_i - \boldsymbol{\mu})) = \mathbf{0}. \tag{72}$$

Applying the central limit theorem to

$$\sqrt{N}(\widehat{\theta}_{\text{crrf}} - \theta) = A_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[\mathbf{z}_i\widetilde{e}_i - B_0\mathbf{r}_i(\pi) - B_1(\mathbf{w}_i - \boldsymbol{\mu})] + o_p(1), \tag{73}$$

we obtain

$$\sqrt{N}(\widehat{\theta}_{\text{crrf}} - \theta) \xrightarrow{d} N\left(\mathbf{0}, A_0^{-1}\Omega A_0'^{-1}\right), \text{ where } \Omega \equiv \mathbb{V}(\mathbf{z}\widetilde{e} - B_0\mathbf{r}(\pi) - B_1(\mathbf{w} - \boldsymbol{\mu})). \tag{74}$$

Asymptotic normality is thus established.

References

Amemiya, T., 1985, Advanced econometrics. Harvard University Press, Cambridge, MA.

Angrist, J.D., G.W. Imbens and D.B. Rubin, 1996, Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444-455.

Angrist, J.D. and V. Lavy, 1999, Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Quarterly Journal of Economics 114, 533-575.

Berk, R.A. and J. de Leeuw, 1999, An evaluation of California's inmate classification system using a generalized regression discontinuity design. Journal of the American Statistical Association 94, 1045-1052.

Black, D.A., J. Galdo and J.A. Smith, 2007, Evaluating the worker profiling and reemployment services system using a regression discontinuity approach. American Economic Review 97, 104-107.

Black, S.E., 1999, Do better schools matter? Parental valuation of elementary education. Quarterly Journal of Economics 114, 577-599.

Chay, K.Y. and M. Greenstone, 2003, Air quality, infant mortality, and the clean air act of 1970. NBER Working Paper 10053.

Chay, K.Y., P.J. McEwan and M. Urquiola, 2005, The central role of noise in evaluating interventions that use test scores to rank schools. American Economic Review 95, 1237-1258.

Cobb-Clark, D.-A. and T. Crossley, 2003, Econometrics for evaluations: An Introduction to recent developments. Economic Record 79, 491-511.

Cook, T.D., 2008, "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. Journal of Econometrics 142, 636-654.

DiNardo, J. and D.S. Lee, 2004, Economic impacts of new unionization on private sector employers: 1984-2001. Quarterly Journal of Economics 119, 1383-1441.

Goldberger, A.S., 1972a, Selection bias in evaluating treatment effects: some formal illustrations. Madison, WI, unpublished manuscript.

Goldberger, A.S., 1972b, Selection bias in evaluating treatment effects: the case of interaction. Madison, WI, unpublished manuscript.

Hahn, J., P. Todd and W. Van der Klaauw, 2001, Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69, 201-209.

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

Hirano, K., G.W. Imbens and G. Ridder, 2003, Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161-1189.

Hogan, J.W. and T. Lancaster, 2004, Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. Statistical Methods in Medical Research 13, 17-48.

Holland, P.W., 1986, Statistics and causal inference. Journal of the American Statistical Association 81, 945-960.

Imbens, G.W. and J.D. Angrist, 1994, Identification and estimation of local average treatment effects. Econometrica 62, 467-475.

Imbens, G.W. and T. Lemieux, 2008, Regression discontinuity designs: A guide to practice. Journal of Econometrics 142, 615-635.

Lee, D.S., 2008, Randomized experiments from non-random selection in U.S. house elections. Journal of Econometrics 142, 675-697.

Lee, D.S. and D. Card, 2008, Regression discontinuity inference with specification error. Journal of Econometrics 142, 655-674.

Lee, M.-J., 2005, Micro-econometrics for policy, program, and treatment effects. Oxford University Press, Oxford.

Lemieux, T. and K. Milligan, 2008, Incentive effects of social assistance: A regression discontinuity approach. Journal of Econometrics 142, 807-828.

Ludwig, J. and D.L. Miller, 2007, Does head start improve children's life chances? Evidence from a regression discontinuity design. Quarterly Journal of Economics 122, 159-208.

Malani, A., 2006, Identifying placebo effects with data from clinical trials. The Journal of Political Economy 114, 236-256.

McCrary, J., 2008, Manipulation of the running variable in the regression discontinuity design: A density test. Journal of Econometrics 142, 698-714.

Robinson, P.M., 1988, Root-N-consistent semiparametric regression. Econometrica 56, 931-954.

Rosenbaum, P.R. and D.B. Rubin, 1983a, Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society, Series B (Methodological) 45, 212-218.

Rosenbaum, P.R. and D.B. Rubin, 1983b, The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41-55.

Rubin, D.B., 1974, Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology 66, 688-701.

Thistlethwaite, D. and D. Campbell, 1960, Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology 51, 309-317.

Trochim, W.M.K, 1984, Research design for program evaluation: The regression-discontinuity approach. Sage Publications, Beverly Hills, CA.

Van der Klaauw, W., 2002, Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. International Economic Review 43, 1249-1287.

Winship, C. and S.L. Morgan, 1999, The estimation of causal effects from observational data. Annual Review of Sociology 25, 659-706.

Wooldridge, J.M., 1997, On two stage least squares estimation of the average treatment effect in a random coefficient model. Economics Letters 56, 129-133.

Wooldridge, J.M., 2002, Econometric analysis of cross section and panel data. MIT Press, Cambridge, MA.

Wooldridge, J.M., 2003, Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. Economics Letters 79, 185-191.

Wooldridge, J.M., 2007, Inverse probability weighted estimation for general missing data problems. Journal of Econometrics 141, 1281-1301.

Table 1: Estimating ATE in the Presence of Multiple Selection Biases

100 observations			Mean Bias					Median Bias		
1,000 replications	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Selection biases: part (b) to part (e) in equation (*)	Robust	Robinson	Corr Func	Ctrl Func	OLS	Robust	Robinson	Corr Func	Ctrl Func	OLS
Model I (h = 0, ATE = 2):										
part (b) and (c): eligibility indicator not used in d-equation	-0.0289	-0.0355	-0.4948	-0.1695	-0.5875	-0.0923	-0.0346	-0.4342	-0.0565	-0.5742
part (b) and (c): eligibility indicator used in d-equation	0.0600	0.0177	-0.4948	-0.1695	-0.5875	0.0316	0.0003	-0.4342	-0.0565	-0.5742
part (b), (c), and (d): OVB and no sorting	1.9880	1.8957	0.9826	-0.3382	0.9911	1.9636	1.8905	0.7736	-0.0810	0.9975
part (b), (c), (d), and (e): OVB and nonlinear sorting bias	4.9471	4.7455	-2.1060	-0.2735	3.2545	4.8245	4.6439	2.7350	0.5569	3.2298
part (b), (c), (d), and (e): OVB and linear sorting bias	2.9462	2.8456	0.8983	-0.5721	1.7829	2.9318	2.8206	1.3526	-0.0897	1.7892
Model II (h = 1, ATE = 1):										
part (b) and (c): eligibility indicator not used in d-equation	0.0386	0.0462	1.0394	-0.0491	0.0120	0.0310	0.0301	-0.0736	-0.0239	0.0107
part (b) and (c): eligibility indicator used in d-equation	0.0458	0.0351	1.0394	-0.0491	0.0120	0.0428	0.0245	-0.0736	-0.0239	0.0107
part (b), (c), and (d): OVB and no sorting	1.9047	1.8784	2.0628	-0.0743	0.8115	1.9127	1.8954	-0.0138	-0.0326	0.8117
part (b), (c), (d), and (e): OVB and nonlinear sorting bias	4.5638	4.5124	-0.4807	0.3376	2.2141	4.4952	4.4672	-0.0001	0.3868	2.2186
part (b), (c), (d), and (e): OVB and linear sorting bias	2.8240	2.7889	1.4917	-0.0955	1.2019	2.8308	2.7867	0.0928	-0.0502	1.1882
100 observations			RMSE					Median Absolute Error		
1,000 replications	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Selection biases: part (b) to part (e) in equation (*)	Robust	Robinson	Corr Func	Ctrl Func	OLS	Robust	Robinson	Corr Func	Ctrl Func	OLS
Model I (h = 0, ATE = 2):										
part (b) and (c): eligibility indicator not used in d-equation	4.7466	0.6957	80.6383	3.3987	0.7497	0.6117	0.3426	3.2847	0.9785	0.5786
part (b) and (c): eligibility indicator used in d-equation	4.9457	1.2013	80.6383	3.3987	0.7497	0.5983	0.3498	3.2847	0.9785	0.5786
part (b), (c), and (d): OVB and no sorting	5.4217	2.2156	143.4809	3.3286	1.1281	2.0164	1.8968	4.9119	1.1340	0.9990
part (b), (c), (d), and (e): OVB and nonlinear sorting bias	7.4264	5.0734	145.8808	6.4366	3.3715	4.8596	4.6470	8.0034	2.3219	3.2298
part (b), (c), (d), and (e): OVB and linear sorting bias	5.9474	3.1365	157.4204	4.8014	1.8877	2.9806	2.8218	6.0211	1.3537	1.7892
Model II (h = 1, ATE = 1):										
part (b) and (c): eligibility indicator not used in d-equation	0.3862	0.3745	37.0338	1.0361	0.2600	0.2560	0.2525	0.9038	0.2504	0.1726
part (b) and (c): eligibility indicator used in d-equation	0.4094	0.3970	37.0338	1.0361	0.2600	0.2725	0.2681	0.9038	0.2504	0.1726
part (b), (c), and (d): OVB and no sorting	1.9543	1.9279	64.3438	1.4231	0.8721	1.9127	1.8954	1.1769	0.2949	0.8117
part (b), (c), (d), and (e): OVB and nonlinear sorting bias	4.6374	4.5885	21.4470	1.5451	2.2740	4.4952	4.4672	1.9441	0.5508	2.2186
part (b), (c), (d), and (e): OVB and linear sorting bias	2.8681	2.8344	41.1054	1.4485	1.2645	2.8308	2.7867	1.4154	0.3757	1.1882

1,000 observations			Mean Bias					Median Bias		
1,000 replications	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Selection biases: part (b) to part (e) in equation (*)	Robust	Robinson	Corr Func	Ctrl Func	OLS	Robust	Robinson	Corr Func	Ctrl Func	OLS
Model I (h = 0, ATE = 2):										
part (b) and (c): eligibility indicator not used in d-equation	-0.1268	-0.0562	2.3446	0.0077	-0.5608	-0.1282	-0.0603	-0.1677	0.0212	-0.5558
part (b) and (c): eligibility indicator used in d-equation	0.0078	0.0041	2.3446	0.0077	-0.5608	0.0022	0.0024	-0.1677	0.0212	-0.5558
part (b), (c), and (d): OVB and no sorting	1.9199	1.9142	2.5725	-0.0024	1.0280	1.9193	1.9134	0.4549	0.0326	1.0199
part (b), (c), (d), and (e): OVB and nonlinear sorting bias	4.7720	4.7614	4.5130	0.6272	3.3182	4.7813	4.7564	0.7100	0.7258	3.3195
part (b), (c), (d), and (e): OVB and linear sorting bias	2.8768	2.8700	2.3658	-0.0142	1.8234	2.8782	2.8709	0.4865	0.0419	1.8162
Model II (h = 1, ATE = 1):										
part (b) and (c): eligibility indicator not used in d-equation	0.0056	0.0272	0.0192	-0.0016	0.0013	0.0020	0.0254	0.0177	-0.0054	-0.0001
part (b) and (c): eligibility indicator used in d-equation	0.0046	0.0033	0.0192	-0.0016	0.0013	0.0040	0.0037	0.0177	-0.0054	-0.0001
part (b), (c), and (d): OVB and no sorting	1.8614	1.8590	0.0668	0.0002	0.8119	1.8570	1.8521	0.0377	0.0037	0.8113
part (b), (c), (d), and (e): OVB and nonlinear sorting bias	4.5318	4.5270	0.0875	0.4177	2.2190	4.5366	4.5321	0.0594	0.4230	2.2213
part (b), (c), (d), and (e): OVB and linear sorting bias	2.7913	2.7880	0.0953	0.0002	1.2170	2.7887	2.7851	0.0630	0.0047	1.2171
1,000 observations			RMSE					Median Absolute Error		
1,000 replications	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
multiple selection biases	Robust	Robinson	Corr Func	Ctrl Func	OLS	Robust	Robinson	Corr Func	Ctrl Func	OLS
Model I (h = 0, ATE = 2):										
part (b) and (c): eligibility indicator not used in d-equation	0.2383	0.1347	66.7925	0.3876	0.5770	0.1570	0.0957	2.1130	0.2546	0.5558
part (b) and (c): eligibility indicator used in d-equation	0.2093	0.1241	66.7925	0.3876	0.5770	0.1295	0.0875	2.1130	0.2546	0.5558
part (b), (c), and (d): OVB and no sorting	1.9318	1.9189	56.2045	0.4598	1.0397	1.9193	1.9134	2.9991	0.2974	1.0199
part (b), (c), (d), and (e): OVB and nonlinear sorting bias	4.7801	4.7672	104.0578	1.1280	3.3278	4.7813	4.7564	5.0853	0.8800	3.3195
part (b), (c), (d), and (e): OVB and linear sorting bias	2.8845	2.8739	69.7028	0.5888	1.8322	2.8782	2.8709	3.7616	0.3940	1.8162
Model II (h = 1, ATE = 1):										
part (b) and (c): eligibility indicator not used in d-equation	0.1110	0.1122	0.3521	0.1052	0.0775	0.0754	0.0759	0.2210	0.0664	0.0488
part (b) and (c): eligibility indicator used in d-equation	0.1174	0.1151	0.3521	0.1052	0.0775	0.0842	0.0822	0.2210	0.0664	0.0488
part (b), (c), and (d): OVB and no sorting	1.8657	1.8632	0.4750	0.1367	0.8173	1.8570	1.8521	0.3085	0.0956	0.8113
part (b), (c), (d), and (e): OVB and nonlinear sorting bias	4.5385	4.5338	0.7774	0.4732	2.2245	4.5366	4.5321	0.4884	0.4230	2.2213
part (b), (c), (d), and (e): OVB and linear sorting bias	2.7952	2.7919	0.5994	0.1699	1.2227	2.7887	2.7851	0.3733	0.1182	1.2171

Table 2: P-900 Effects on 1988-1992 Gain Scores within Bands of the Selection Threshold

	Full Sample		±5 Points		±2 Points	
	Math	Language	Math	Language	Math	Language
Panel A:						
(1) 2SLS	2.51**	2.35**	1.82	1.58	1.90	1.44
(existing study)	(1.07)	(0.93)	(1.31)	(1.20)	(2.20)	(1.98)
Standard deviation gain	0.32	0.32	0.23	0.21	0.24	0.19
(2) RD robust	2.38***	2.32***	1.74**	2.05***	2.34**	2.51***
(this paper)	(0.68)	(0.61)	(0.83)	(0.71)	(1.17)	(0.93)
Standard deviation gain	0.31	0.31	0.22	0.28	0.30	0.34
Panel B:						
(3) Correction function	5.42*	5.56**	4.80	4.94	−0.51	−1.73
(this paper)	[2.95]	[2.30]	(3.81)	(3.55)	(7.09)	(6.94)
Standard deviation gain	0.69	0.75	0.62	0.67	−0.06	−0.23
Correction term 1	−16.24	−7.32	−3.24	2.86	26.44	22.86
(linear correction)	[10.48]	[8.99]	(12.22)	(11.52)	(23.33)	(21.93)
Correction term 2	14.01**	9.52*	−0.93	−1.99	−5.89	−5.80
(quadratic correction)	[7.04]	[5.34]	(5.91)	(5.06)	(11.40)	(10.32)
Correction term 3	−1.90	−5.77	−0.01	−3.62	−11.85	−8.59
(cubic correction)	[6.74]	[5.30]	(5.81)	(5.15)	(9.99)	(8.55)
$\chi^2(2)$	4.34	3.77	0.01	0.38	1.14	0.83
$\Pr > \chi^2$	0.114	0.152	0.987	0.687	0.321	0.435
(4) Correction function	8.74***	8.62***	4.66	5.99**	5.97	2.97
(Wooldridge, 2002)	[2.54]	[2.14]	(3.05)	(2.89)	(5.76)	(5.45)
Standard deviation gain	1.12	1.16	0.60	0.81	0.77	0.40
Correction term	−10.12***	−10.67***	−4.05	−6.08	−4.78	−1.44
(linear correction)	[3.58]	[3.02]	(4.38)	(4.16)	(7.89)	(7.19)
Sample size	2,591		938		392	

* Significance at 10% level; ** Significance at 5% level; *** Significance at 1% level.

Notes: To be consistent with Chay, McEwan, and Urquiola (2005), the sample includes urban schools with 15 or more students in the fourth grade in 1988. The dependent variables are the 1988-1992 gain scores in math and language. Regressors, in addition to the P-900 dummy, include cubic polynomials for the 1988 average test score, SES in 1990, and the changes in SES between 1990 and 1992. The columns correspond to subsamples of schools with 1988 test scores relative to the cutoff score in the chosen range. The 2SLS in Panel A is proposed by Chay, McEwan, and Urquiola (2005). The RD robust estimator in Panel A is the one proposed by this paper, which uses the first-stage residual, which is the deviation between the actual treatment status and the estimated treatment probability, as the instrument for the actual treatment status. The correction function estimator with the cubic polynomial sorting correction terms in Panel B is proposed by this paper. The correction function estimator with only the linear sorting correction term in Panel B is proposed by Wooldridge (2002). A χ^2-test and associated p-value for the joint significance of the quadratic and the cubic correction terms are provided. Standard errors robust to heteroskedasticity are in parentheses; bootstrapped standard errors based on 2,000 replications are in square brackets.

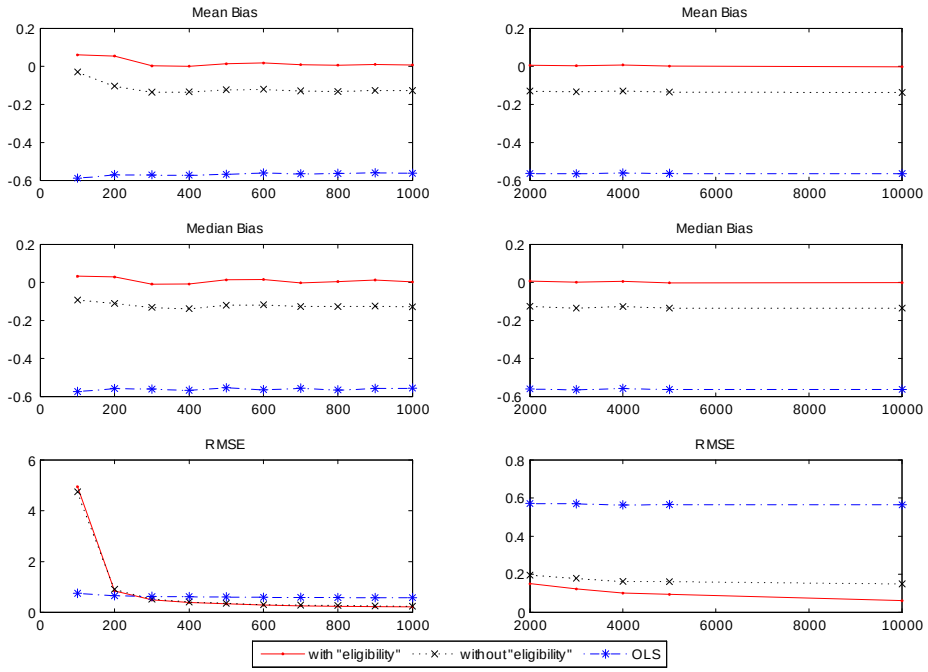

Figure 1: Selection bias due to observables (Model I)

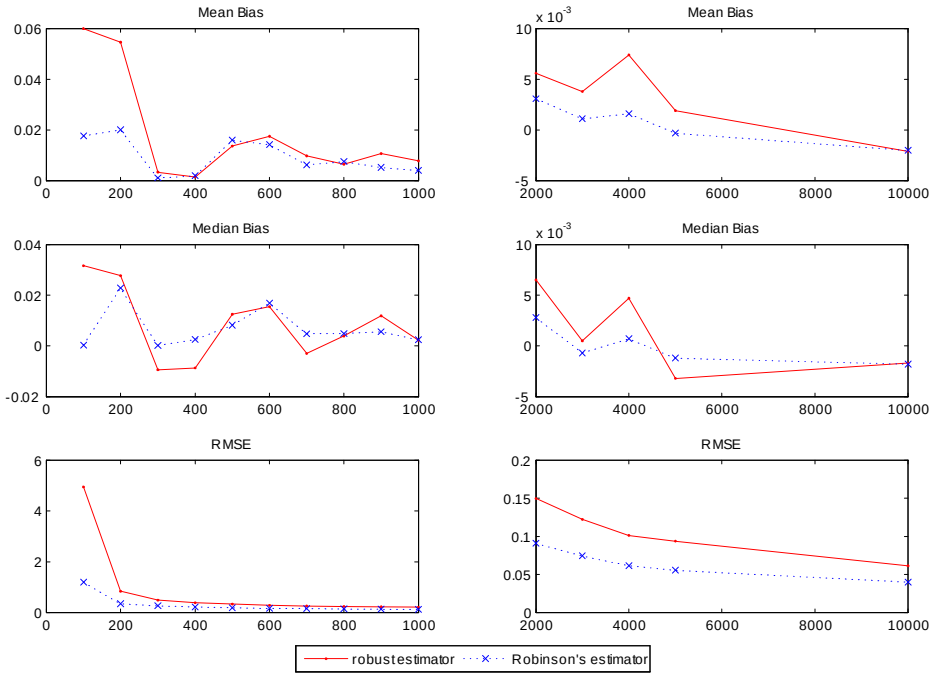

Figure 2: Selection bias due to observables with eligibility used in d-equation (Model I)

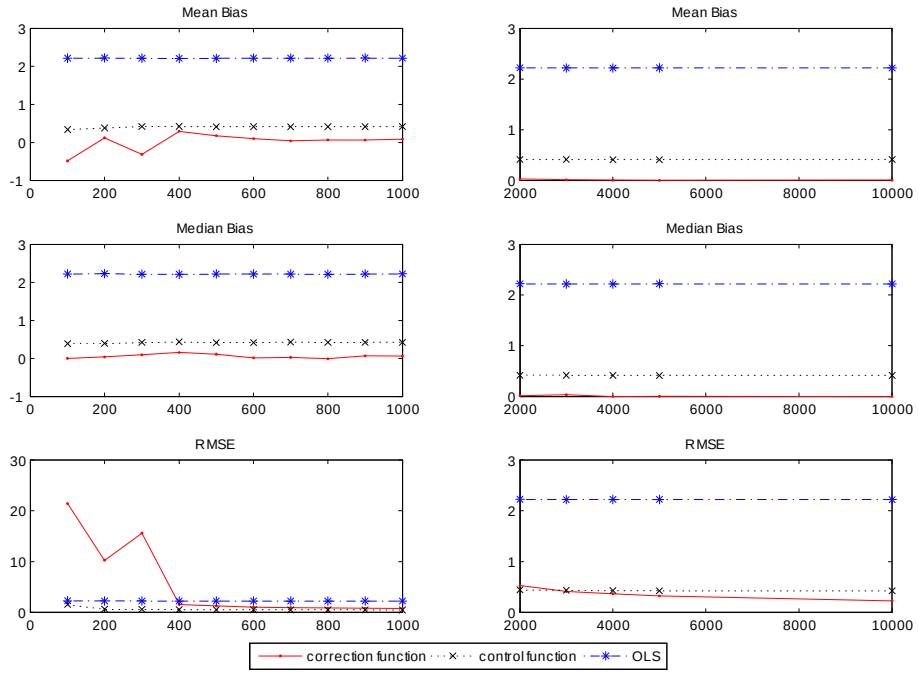

Figure 3: Selection bias due to OVB and nonlinear sorting (Model II)

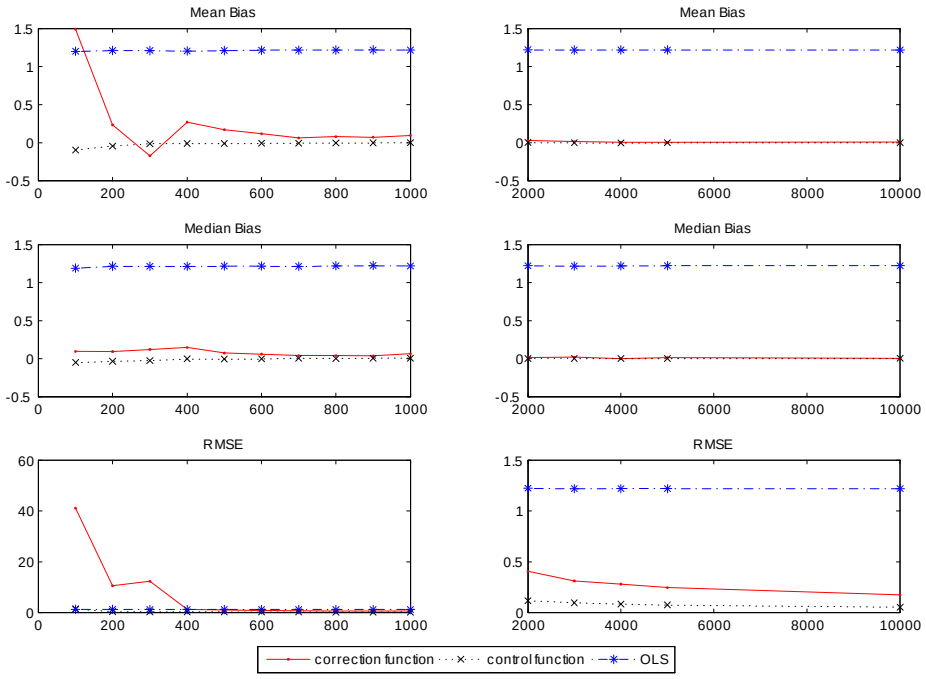

Figure 4: Selection bias due to OVB and linear sorting (Model II)