Regression Discontinuity Design: Identification and Estimation of Treatment Effects with Multiple Selection Biases[1]

Muzhe Yang[2]
Department of Agricultural and Resource Economics
University of California, Berkeley

Job Market Paper (semi-final version)

October 2007

[1]I thank my advisors, Ken Chay and Jeffrey Perloff, for their encouragement as well as numerous insightful comments and suggestions. I am grateful to Jeffrey Greenbaum and Maria Carolina Caetano for their valuable inputs. The paper also benefited from discussions with, among many others, James Powell, Arnold Zellner and the participants at the ARE summer 2007 workshop and the participants at the econometrics seminar at Brown University. I thank Ken Chay for generously providing the data for the empirical applications of this paper. Financial support provided by UC-Berkeley Chancellor's Dissertation-Year Fellowship is also gratefully acknowledged. The usual disclaimers apply.

[2]Correspondence: Department of Agricultural and Resource Economics, 207 Giannini Hall #3310, University of California, Berkeley, CA 94720-3310. E-mail: yang@are.berkeley.edu. Website: http://are.berkeley.edu/~yang.

Abstract

Previous work on the regression discontinuity (RD) design has emphasized identification and estimation of an effect *at* the selection threshold (discontinuity). Focusing on the so-called "fuzzy" RD design, this paper examines identification and estimation of the *average treatment effect* (ATE) under various forms of selection bias—selection on the observables, selection on the unobservables, and selection based on heterogeneity in the effects of the treatment. Easy to implement estimators that are root-N consistent and asymptotically normal are derived. They allow for general functional forms for the selection biases and imply specification tests for the plausibility of the statistical assumptions. This paper also investigates the trade-off between efficiency and bias in estimating the average treatment effect (and average effects local to the discontinuity) when the effects covary with the observables and the unobservables. The theoretical results leverage the dual nature of the RD design—both the "borderline experiment" provided near the threshold and the strong and valid exclusion restriction provided in the selection equation for the choice of treatment. This point is demonstrated through Monte-Carlo experiments and empirical applications.

Keywords: fuzzy regression discontinuity design, selection bias, heterogeneous treatment effects, average treatment effect, efficiency-bias trade-off.

1 Introduction

The fundamental problem of causal inference on a treatment effect is the unobservability of the *same* individual in both the treated and untreated states.[1] Since it is impossible to *observe* both the factual and counterfactual states at the same time for a given individual, identification of treatment effects must rely on comparing the outcomes of different individuals with different treatment status. This leads to questions of differences in outcomes arising 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 viewed as unethical.[3]

One procedure that society and governments follow to allocate resources—and that is not viewed as unethical—is the assignment of resources based on merit or need. Often, this assignment is based on strict eligibility cutoffs for the program in which the odds of qualifying for the intervention change substantially at these cutoffs. While the efficiency of such policy design can and should be debated, it provides a unique opportunity for the researcher to evaluate the effect of the intervention while leveraging some of the features of random assignment. For example, researchers have noted that as long as there is some "noise" or arbitrariness in the eligibility criteria—that is, the criteria are not perfectly related to individual outcomes—then near the cutoffs, the assignment of resources is "close to random". This allows for a transparent way to rule out competing hypotheses while testing the assumption of randomness.

The regression discontinuity (RD) design has been developed to utilize these "discontinuous" changes in the probability of treatment at the eligibility cutoff(s).[4] The "sharp" RD design occurs when the probability of treatment goes from zero to one at the cutoff. This study focuses on the so-called "fuzzy" RD design (Trochim 1984), in which the change in the probability of treatment is less than one but still substantial, as this design is closer to the design of most policy interventions.

Much of the theoretical research on the RD design (Hahn, Todd and Van der Klaauw 2001; Porter 2003) has emphasized measurement of effects *at* the eligibility thresholds.[5] In the limit—that is, as one approaches the discontinuity—potential biases disappear while the probability of treatment changes significantly. Its similarity to a "borderline experiment" at

[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, 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] Besides issues of feasibility, random assignment is still subject to threats to both internal and external validity, such as substitution bias (noncompliance), randomization bias (different participants), and Hawthorne effects (measurement errors) (Winship and Morgan 1999; Cobb-Clark and Crossley 2003).

[4] For a history and overview of the RD design, see Thistlethwaite and Campbell (1960), Goldberger (1972a, 1972b) and Cook (2007).

[5] Porter (2003) derives the optimal rate of convergence for estimating treatment effects at the selection threshold. Sun (2005) generalizes the local polynomial estimator that Porter (2003) proposes to allow the order of the polynomials to be jointly determined by the data through an adaptive procedure. Lee and Card (2007) further extends the applicability of an RD design to the case where the selection variable has a discrete support.

the eligibility threshold has also been recognized and exploited by many empirical studies.[6] However, two concerns have arisen about the measurement of the effects at the threshold. First, empirically, the definition of "the limit" can be ad hoc and, if made narrow enough, will preclude inference due to a paucity of data. Indeed, almost all applied studies implementing the RD design use data away from the discontinuity or assume a functional form for the selection bias due to observables to derive estimates and to form confidence intervals.[7] In some of these applications, such as Angrist and Lavy (1999) and Chay, McEwan and Urquiola (2005), as data away from the discontinuity are trimmed, the confidence intervals grow large enough to disallow the rejection of many hypotheses.

Second, in the presence of heterogeneity in the effects of the treatment, near the discontinuity, only the average effect for a particular population can be identified under certain conditions (such as monotonicity),[8] instead of the average effect for a randomly selected member of the population—also known as the *average treatment effect* (ATE). While the former is still useful for policy analysis as it measures the impact of the eligibility criteria that *were* used by the program, the latter is often viewed as more useful for forecasting the relative benefits of policies under consideration.

This study attempts to address these concerns by integrating the results from the literature on selection biases with the RD design literature. The theoretical results derived in this paper leverage the dual nature of the RD design—that it provides both a "borderline experiment" near the discontinuity and a strong and valid exclusion restriction in the selection equation for the choice of treatment. The second point allows one to theoretically examine identification and estimation of the *average treatment effect* under various forms of selection bias—selection on the observables, selection on the unobservables, and selection based on heterogeneity in the effects of the treatment. In particular, since the probability of selection changes significantly due to the eligibility cutoff, which can be excluded from the potential outcome equation, the discontinuity effectively provides a strong and valid instrumental variable for treatment choice.

Given that 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), this paper formalizes this potential outcome problem in terms of the following regression model:

[6] For example, Berk and de Leeuw (1999), Black (1999), Buddelmeyer and Skoufias (2003), Lee (2007), Lemieux and Milligan (2007), Ludwig and Miller (2006), and Van der Klaauw (2002). Some studies further offer empirical tests of the validity of an RD design (Black, Galdo and Smith 2005; Cook and Wong 2007) using various parametric and nonparametric estimators.

[7] For example, Angrist and Lavy (1999), Chay and Greenstone (2003), Chay, McEwan and Urquiola (2005), DiNardo and Lee (2004), Ludwig and Miller (2006) and McCrary and Royer (2003).

[8] This is the local average treatment effect (LATE) (Imbens and Angrist 1994; Angrist, Imbens and Rubin 1996).

$y_0 = g_0(z^*) + u_0, \mathbb{E}(u_0|z^*) = 0$
$y_1 = g_1(z^*) + \eta + u_1, \mathbb{E}(u_1|z^*) = 0, g_j(z^*)$ continuous in z^*, $j \in \{0, 1\}$
$y = dy_1 + (1-d)y_0$ where $\begin{cases} y_1 \text{ is potential outcome in the presence of a treatment} \\ y_0 \text{ is potential outcome in the absence of a treatment} \end{cases}$
$u = du_1 + (1-d)u_0$: unobservables that affect observed outcome y
j: potential state indicator $= \begin{cases} 1 & \text{in the presence of a treatment intervention} \\ 0 & \text{in the absence of a treatment intervention} \end{cases}$
d: observed treatment status $= \begin{cases} 1 & \text{receiving treatment} \\ 0 & \text{not receiving treatment} \end{cases}$
η: the pure randomness induced by the treatment and η is independent of (d, z^*)
u_j: unobserved heterogeneities that affect potential outcome y_j

In this model, the average treatment effect (ATE) is defined as:

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

The observed average outcome difference between the treatment and the control groups contains the following five differences:

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

Thus, in order to identify ATE, defined as $\mathbb{E}\left[\eta + (g_1(z^*) - g_0(z^*))\right]$, the researcher needs to identify (a) and takes into account several sources of selection bias, (b) through (e), which correspond to $\mathbb{E}\left[(g_1(z^*) - g_0(z^*))\right]$.

For the specific case of the RD design, we can write the selection equation as follows:

$d = 1\{\pi_0 + \pi_1 z + \pi_2 z^* + v > 0\}$
$z = 1\{z^* \leqslant 0\}, \pi_1 \neq 0, v \sim F_v(\cdot)$
z^*: observable selection variable used by a selection rule for treatment assignment
$z = 1\{z^* \leqslant 0\}$: eligibility indicator specified by a selection rule with zero as the cutoff point
v: unobservables that affect selection process

The above equation for the selection process implies that the eligibility discontinuity provides a valid exclusion restriction from the outcome equation as long as the cutoff point used by the selection is unexpected or does not affect potential outcomes. Further, this exclusion restriction will be a powerful predictor of selection (i.e., a strong instrument). This is the first paper to theoretically examine how using this as an instrument can allow one to investigate the potential for selection on the unobservables, i.e. (d) in (*), and self-selection due to heterogeneous treatment effects, i.e. (e) in (*).

Current theoretical RD papers focus on identification and estimation of treatment effects

at the threshold. They assume selection only on the observables near the threshold, in which case biases due to (d) and (e) in (*) vanish close to the threshold. Emphasis is therefore placed on how to control for $g_0(\cdot)$ (Hahn, Todd and Van der Klaauw 2001; Porter 2003; Ai 2007) and $g_1(\cdot)$ (Ai 2007) in order to minimize the bias in estimating the effect at the cutoff point. Although they are important contributions, the estimators derived in these papers except for Ai (2007) do not achieve a root-N rate of convergence due to the need for smoothing. Imbens and Lemieux (2007) suggests an easy to implement two-stage least squares (2SLS) estimator which is numerically equivalent to a nonparametric regression with a uniform kernel. However, this 2SLS estimator is aimed for the effect at the cutoff point, and it does not work if selection on the unobservables occurs. In contrast, my paper reformulates RD based on the switching regression model, and focuses on identification and estimation of treatment effects for a predefined population away from the threshold, in which case multiple selection biases, i.e. (d) through (e) in (*), are taken into account. My contribution to the previous theoretical RD papers can be summarized by the following table.

Table 1: Biases Arising in RD design

Selection Bias due to	Hahn, Todd and Van der Klaauw (2001)	Porter (2003)	Imbens and Lemieux (2007)	Ai (2007)	this paper
(b) in (*)	✓	✓	✓	✓	✓
(c) in (*)				✓	✓
(d) in (*)					✓
(e) in (*)					✓

As shown in the above table, this paper aims to account for all four biases listed, in addition to solving two technical problems. The first problem is how to control for (b) and (c) with the least restrictive assumptions in the absence of (d) and (e). The second problem is how to correct for the impacts of (d) and (e) when they are present.

In response to the first problem, I propose a new estimator—RD robust estimator—for the ATE of a predefined population which is not restrictively at the threshold. It is based on the moment conditions derived from the conditional mean independence between (u_1, u_0) and v when selection is on the observables. This estimator is robust in the sense that it does not require estimating the conditional expectation of the outcome. It is instead based on the orthogonality conditions that are functions of the conditional probability of selection—also known as the propensity score.[9] Thus it avoids smoothing and achieves the root-N rate of convergence. This proposed estimator is shown to be consistent, asymptotically normal and easy to implement using standard software, compare with nonparametric alternatives proposed by Hahn, Todd and Van der Klaauw (2001) and Porter (2003) and series estimators proposed by Ai (2007). Furthermore, under selection-on-observables, the exclusion restriction in the RD design brings efficiency gains to the proposed estimator, which suggests over-identification tests for the added moment conditions.[10]

[9]Note that estimators under 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 under a fuzzy RD because the "overlapping or common support" identification assumption is difficult to meet, and it is completely violated under a sharp RD design.

[10]Battistin and Rettore (2007) also investigates the potential for an RD design to offer specification tests for treatment effects for program participants away from the threshold when individuals self-select into

To address the second problem, I propose another estimator—correction function estimator—for the ATE of a predefined population which is not restrictively at the threshold and when selection-on-unobservables is concerned. It uses the eligibility discontinuity as the instrument for (d) in (*) and it requires correction terms to be added back to the outcome equation. These correction terms are constructed from the exclusion restriction and exogenous variables to account for (e) in (*). This correction function estimator allows one to estimate ATE even in the presence of heterogeneous sorting. In this way, I attempt to integrate the literature on RD designs with the larger literature on selection biases when one has a valid exclusion restriction. Further, these approaches allow one to test the assumption of selection on the observables for estimating ATE. This proposed correction function estimator is based on Wooldridge (2002), but it extends the existing results to allow for nonlinear selection due to differential sorting. This improvement is reflected by adding a quadratic selection term to Wooldridge (2002)'s specification, and it is useful in many cases where comparative advantages matter such as unions and sorting based on potential gains.

In summary, my paper makes the following contributions: first, I propose two estimators which extend RD's applicability to the cases of selection not only on the observables, but also on the unobservables and on the returns to the treatment; second, I investigate the efficiency-bias trade-off in estimating various average effects when the effects covary with the observables and the unobservables; third, the proposed estimators suggest specification tests for more restrictive models which can be used as falsification tests in empirical applications; fourth, I rewrite the selection-on-observables problem using moment conditions and derive estimators which are root-N consistent, asymptotically normal and easy to implement using standard software; fifth, the proposed estimators can also provide a transparent link between the economics of the problem and the estimation.

The rest of the paper is organized as follows. Section 2 presents identification results of average treatment effects with the RD design. Section 3 discusses the proposed estimators' large sample properties and lays out estimation procedures. Section 4 evaluates finite sample performances of the proposed estimators using a series of Monte Carlo experiments. To illustrate the implementation of the proposed estimators for either removing or correcting for selection biases, Section 5 offers two empirical applications. Both applications investigate the trade-off between efficiency and bias in estimating the average treatment effect (and average effects local to the discontinuity). The first one, using data from Chay, McEwan and Urquiola (2005), highlights how the efficiency gain associated with the RD robust estimator better evaluate an education intervention that uses test scores to allocate resources. The other one, using data from Chay and Greenstone (2003), demonstrates how a specification test implied by RD's instrumentality can better analyze the impacts from air quality on infant mortality. This example also presents an extension of RD's instrumental nature to the case of a continuous treatment, which in the data is the different level of air pollution. Finally, Section 6 concludes. All proofs, supplemental discussions, and additional tables and figures are in the appendices.

---

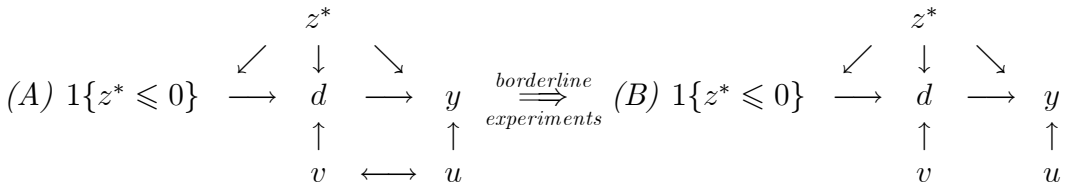
participation conditional on some eligibility criteria. However, their results directly apply to a sharp RD design.

2 Identification of Treatment Effects

To exploit RD's dual nature, on one hand, I use its "borderline experiment" to set up the RD robust estimator under selection-on-observables where the impacts of (d) and (e) in (*) can be plausibly removed close to the threshold; on the other hand, I use its "instrumentality" implied by the selection rule to deal with (d) in (*) and to construct correction terms for the correction function estimator that accounts for (e) in (*), which also suggests a specification test to falsify selection-on-observables when it is suspected.

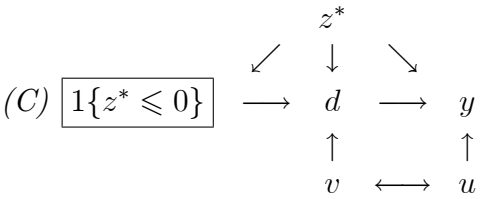
The following graphs help to visualize the dual nature of an RD design, where an arrow denotes "causing" and an $\boxed{x}$ refers to an instrument for d.

Case 1 ("borderline experiment") *Using an RD design to remove impacts of (d) and (e) close to the selection threshold:*



We have selection-on-unobservables in (A) because the outcome (y) and the treatment status (d) are mutually dependent through the dependence between u and v; as a result d is not exogenous to y. However, near the threshold as in (B), individuals' positioning just above or just below the cutoff point is likely to be randomly determined. In this situation, u and v become independent because it is probably the luck that plays a key role in determining each individual's treatment status at the margin. This case characterizes selection-on-observables, where d is exogenous to y. However, in order for this "borderline experiment" to work, individuals cannot manipulate z^* perfectly to qualify themselves for the treatment. A sufficient condition for this situation is that the selection threshold or the cutoff point is *unexpected* prior to its implementation. This can, to a large extent, avoid behavioral changes near the threshold, which make z^* endogenous and therefore invalidate this "borderline experiment".[11]

Case 2 ("instrumentality") *Using an RD design to instrument for (d) and correct for (e):*



In this situation, $1\{z^* \leqslant 0\}$ serves as an instrument for d because it affects d directly, and it affects y only through d. A non-smoothness arises only in the selection process. To integrate the RD's dual nature, the strategy is to turn (A) into (B) and use (C) as a falsification

[11] Such a situation may be detectable by checking whether the density of the selection variable is discontinuous at the cutoff point (McCrary 2007).

check to detect any dependence between u and v, which will invalidate (B).[12] Since either removing or correcting for selection biases hinges upon moment restrictions imposed on the selection process and the pre- and post-treatment outcomes, unlike "smoothing" estimators root-N consistent, asymptotically normal and moment-based estimators are available. They allow for general functional forms for the selection biases and imply specification tests for the plausibility of statistical assumptions. One contribution of the paper is to investigate the trade-off between efficiency and bias in estimating the average treatment effect (and average effects local to the discontinuity) in the presence of multiple selection biases. This trade-off intensifies upon adding more observations away from the selection threshold.

An RD design reveals to a researcher a more informative and transparent selection process. The selection rule, an indicator for eligibility, is a known and deterministic function of selection variables.[13] The criterion can be generically specified as $z = 1\{z^* \leqslant 0\}$ where the cutoff point (threshold) is normalized to zero. Such a criterion changes a potential treatment status exogenously if the selection threshold is unexpected and the selection variable (z^*) cannot be manipulated perfectly. Building on the potential outcomes framework, we have two potential treatment states (d_j) corresponding to being intervened ($j = 1$, eligible) and not being intervened ($j = 0$, ineligible). Given the discreteness of treatment status, a binary response model can be used to describe the selection process. If the selection criterion has no "discrimination" in nature, then such an intervention in the selection process should have a homogeneous impact on potential treatment status. A model for this idea is specified as follows:

$$\begin{cases} d_j = 1\{d_j^* > 0\} \\ d_j^* = \pi_0 + \pi_1 j + \pi_2 z^* + v \\ \mathbb{E}(d_j^* | z^*) = \pi_0 + \pi_1 j + \pi_2 z^*, \text{ if } \mathbb{E}(v|z^*) = 0 \end{cases}$$

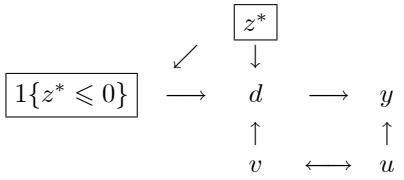
Because the eligibility is a deterministic function of the selection variable, $z = 1\{z^* \leqslant 0\}$, there is no variation in z after conditioning on z^*. Accordingly z is independent of d_j^* conditional on z^*. And, conditional on z^*, z is also independent ("$\amalg$") of d_j. Therefore, we have:

$$(d_0^*, d_1^*) \amalg z | z^* \Rightarrow \mathbb{E}(d_j^* | z, z^*) = \mathbb{E}(d_j^* | z^*) \text{ and } (d_0, d_1) \amalg z | z^*$$

Lee (2005) provides the following two useful results.

---

[12] The supplemental Web Appendix available at http://are.berkeley.edu/~yang/research.html discusses the following case, where over-identification can take place because two instruments, z and z^*, are available for one endogenous treatment status (d):



In this situation, the "positioning" z^* (selection variable) has no direct impact on y, and it affects d directly together with $1\{z^* \leqslant 0\}$. So both z^* and $1\{z^* \leqslant 0\}$ are valid instruments for d, and the program effect $d \to y$ can be over-identified.

[13] They are simplified to be a scalar in our analysis.

Lemma 1 *If two potential regression functions, $\mathbb{E}(d_j^*|z^*)$ where $j \in \{0,1\}$ only differ by a constant, π_1, then we can replace a potential state indicator (j) with an indicator for observed treatment status (z) if and only if the selection, z, is based only on the observables, z^*, i.e., $\mathbb{E}(d_j^*|z,z^*) = \mathbb{E}(d_j^*|z^*)$.*

Proof. See Lee (2005, page 34). ■

Lemma 2 *A symmetric version of mean-independence of y_j and d refers to $Cov\,(y_j,d) = 0 \Leftrightarrow \mathbb{E}(y_j d) = \mathbb{E}(y_j)\mathbb{E}\,(d)$. An asymmetric version of mean-independence of y_j from d refers to $\mathbb{E}(y_j|d) = \mathbb{E}(y_j)$. When d is binary and $0 < \Pr(d=1) < 1$, the symmetric version of mean-independence is equivalent to the asymmetric version. Analogously, when d is binary, conditional on x where $0 < \Pr(d=1|x) < 1$, we have*

$$\begin{aligned}\mathbb{E}(y_j d|x) &= \mathbb{E}(y_j|x)\mathbb{E}(d|x)\\ &\Leftrightarrow \mathbb{E}(y_j|d,x) = \mathbb{E}(y_j|x)\\ &\Leftrightarrow \mathbb{E}(y_j|d=1,x) = \mathbb{E}(y_j|d=0,x)\end{aligned}$$

Proof. See Lee (2005, page 36).[14] ■

From Lemma 1, $\mathbb{E}(d^*|z,z^*) = \pi_0 + \pi_1 z + \pi_2 z^*$ where $d^* = z d_1^* + (1-z)d_0^*$. The observed treatment status can be written as $d = 1\{d^* > 0\} = 1\{\pi_0 + \pi_1 z + \pi_2 z^* + v > 0\}$. A fuzzy RD design enforces $\pi_1 \neq 0$ (and $\pi_1 > 0$ without loss of generality), making the eligible *more likely* to be selected into treatment. For a sharp RD design, there is no randomness in the selection process so that $d = z = 1\{z^* \leqslant 0\}$ where the eligibility criterion is used without exception as the actual treatment assignment rule.

We next model two potential outcomes, together with the selection process, in a switching regression context. We assume that y_j is separably additive in a selection variable z^* and unobserved heterogeneity u_j.[15]

Assumption 1 (Model) *A switching regression model with a fuzzy RD design is specified as follows:*

$$\begin{aligned}&y_j = g_j(z^*) + u_j, \mathbb{E}(u_j|z^*) = 0, g_j(z^*) \text{ continuous in } z^*,\ j \in \{0,1\}\\ &d = 1\{\pi_0 + \pi_1 z + \pi_2 z^* + v > 0\}, z = 1\{z^* \leqslant 0\}, \pi_1 > 0, v \sim F_v(\cdot)\end{aligned}$$

The average treatment effect (ATE) is:

$$ATE \equiv \mathbb{E}(y_1 - y_0) = \mathbb{E}(g_1(z^*) - g_0(z^*))$$

Identifying ATE hinges upon u_j and v. As previously mentioned, when u_j and v are independent, we have selection-on-observables;[16] when u_j and v are not independent, we

[14] In general, unless y_j is binary with $0 < \Pr(y_j = 1) < 1$, the mean-independence $\mathbb{E}(y_j d) = \mathbb{E}(y_j)\mathbb{E}(d)$ does not necessarily imply the other asymmetric mean-independence $\mathbb{E}(d|y_j) = \mathbb{E}(d)$, which can be called the "mean-independence of d from y_j".

[15] The following discussions on identification and estimation of treatment effects consider no additional covariates ($\mathbf{x}$'s) to avoid unnecessary complications. Imbens and Lemieux (2007) discusses adjustment for additional covariates.

[16] In program evaluation literature, this is also called ignorability or unconfoundedness of the treatment.

have selection-on-unobservables. Their formal definitions are given by Lee (2005).

Definition 1 (Selection-on-Observables) *For y_j with a density or probability function $f(\cdot)$, $f(y_j|d) \neq f(y_j)$ but $f(y_j|d, z^*) = f(y_j|z^*)$ for the observables z^*.*

A weaker version of this definition is: $\mathbb{E}(y_j|d) \neq \mathbb{E}(y_j)$ and $\mathbb{E}(y_j|d, z^*) = \mathbb{E}(y_j|z^*)$, which is sufficient for identifying various average effects of a well-defined treatment. Under selection-on-observables, the treatment status does not affect the potential outcomes directly conditional on the observables, z^*. Therefore, once the observables are controlled for, the treatment status becomes exogenous to potential outcomes and functions just like a "curve shifter". While it is possible that d is determined by some unobservables, i.e. v, selection-on-observables requires $(u_0, u_1) \amalg v|z^*$.

Definition 2 (Selection-on-Unobservables) *For y_j with a density or probability function $f(\cdot)$, $f(y_j|d, z^*) \neq f(y_j|z^*)$ but $f(y_j|d, z^*, \epsilon) = f(y_j|z^*, \epsilon)$ for the observables z^* and some unobservables ϵ.*

A weaker version of this definition is $\mathbb{E}(y_j|d, z^*) \neq \mathbb{E}(y_j|z^*)$ but $\mathbb{E}(y_j|d, z^*, \epsilon) = \mathbb{E}(y_j|z^*, \epsilon)$ for observables z^* and some unobservables ϵ, which is sufficient for identifying various average effects of a well-defined treatment. Under selection-on-unobservables, u_0 and u_1 are not independent of v conditional only on the observables, z^*. Conditional on z^*, we still encounter omitted variables bias because either u_0 or u_1 is not independent of v. We also need to deal with selectivity or sorting bias because the potential gain, $u_1 - u_0$, is not independent of v. This occurs under "cream-skimming", i.e., when a program or treatment is assigned to individuals according to their potential benefits in order to maximize the *prima facie* program effectiveness.

As previously mentioned, $z = 1\{z^* \leqslant 0\}$ is a valid instrument for the treatment status, which satisfies the following two requirements:

(1) redundancy (exclusion restriction): $\mathbb{E}(y_j|z, z^*) = \mathbb{E}(y_j|z^*)$
(2) relevancy (inclusion restriction): $\pi_1 \neq 0 \Rightarrow \mathbb{E}(d|z = 1, z^*) \neq \mathbb{E}(d|z = 0, z^*)$

For the ATE, we consider the cases of both homogeneous and heterogeneous effects. The homogeneous ATE implies a constant distance between $g_1(z^*)$ and $g_0(z^*)$ for all z^* under Assumption 1. On the other hand the heterogeneous ATE has an observable (explicit) part and an unobservable (implicit) part, which are idiosyncratic and vary with both observables and unobservables. For example, a weight training program may be more effective for young people than old people although both can benefit from it. Those who have greater motivation may benefit even more, which is one unobservable component of treatment effect heterogeneity. We next discuss how to identify ATE in the presence of both the observable and unobservable part of treatment effect heterogeneity. This discussion includes the homogeneous ATE as a special case.

Assumption 2 (Homogeneity) *ATE$\equiv$ $\mathbb{E}(y_1 - y_0) = y_1 - y_0 \equiv \alpha, g_1(z^*) = g_0(z^*)$ and $u_1 = u_0$.*

Here, ATE is only a constant distance between $\mathbb{E}(y_1|z^*)$ and $\mathbb{E}(y_0|z^*)$. The treatment effect is unconditional.

Assumption 3 (Heterogeneity) *ATE*$\equiv \mathbb{E}(y_1 - y_0) = \mathbb{E}(\eta + \lambda(z^*)) \equiv \alpha$, *where* $\mathbb{E}(y_1 - y_0|z^*) = \eta + \lambda(z^*)$; $\lambda(z^*)$ *is the explicit part of treatment effect heterogeneity and* η *is the unobserved and innocuous heterogeneity with* $\eta \amalg (d, z^*)$.

Under Assumption 1 and Assumption 3, observed outcomes y can be written as

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

It can be rewritten in terms of ATE as

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

Under Assumption 2, the above model can be simplified to

$$y = g_0(z^*) + \alpha d + u_0$$

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 e and v. If e and v are independent, we are in the situation of selection-on-observables, where the only hindrance is $g_0(z^*)$ and $\lambda(z^*)$. We next discuss different identification strategies under selection-on-observables and selection-on-unobservables when there is no independence between e and v.

2.1 Identification under Selection-on-Observables

Identification of ATE in the presence of $\lambda(z^*)$ requires the following additional assumption on its structure.

Assumption 4 (Heterogeneity Parameterization) $\lambda(z^*)$ *can be parameterized as* $\lambda(z^*) = \lambda(\mathbf{w}; \gamma) = \mathbf{w}'\gamma$, *where* $\mathbf{w}$ *is a function of* z^* *and corresponding higher order terms.*

Under Assumption 1 and Assumption 3, the central idea of identifying α is to utilize the conditional moment restrictions granted by selection-on-observables, namely $v \amalg (u_0, u_1)$, to generate orthogonality conditions that will purge overt biases from $g_0(z^*)$ and $\lambda(z^*)$.

Under selection-on-observables, we have:

$$\mathbb{E}(\widetilde{e}|d, z, z^*) = 0 = \mathbb{E}(\widetilde{e}|z, z^*) \Rightarrow \mathbb{E}[(d - \mathbb{E}(d|z, z^*))\widetilde{e}] = 0$$

and, under Assumption 1, we have:

$$\mathbb{E}[(d - \mathbb{E}(d|z, z^*))|z^*] = 0 \Rightarrow \mathbb{E}[g_0(z^*)(d - \mathbb{E}(d|z, z^*))] = 0$$

The first-stage residual, $v = d - \mathbb{E}(d|z, z^*)$, plays a crucial role in orthogonalizing both $g_0(z^*)$ and $\widetilde{e}$. We have the following:

$$\begin{aligned} y - \alpha d &= g_0(z^*) + d(\lambda(z^*) - \mathbb{E}(\lambda(z^*))) + \widetilde{e} \\ \mathbb{E}[(y - \alpha d)(d - \mathbb{E}(d|z^*))] &= \mathbb{E}[d(\lambda(z^*) - \mathbb{E}(\lambda(z^*)))(d - \mathbb{E}(d|z^*))] \\ \alpha &= \frac{\mathbb{E}[y(d - \mathbb{E}(d|z^*))] - \mathbb{E}[(\lambda(z^*) - \mathbb{E}(\lambda(z^*)))Var(d|z^*)]}{\mathbb{E}[d(d - \mathbb{E}(d|z^*))]} \end{aligned}$$

Identification of α is complicated by $\lambda(z^*)$. However, if there are only "innocuous" treatment effect heterogeneities, i.e., $\lambda(z^*) = 0$, then $\alpha = \mathbb{E}(\eta)$ can be identified:

$$\text{ATE} \equiv \alpha = \mathbb{E}(\eta) = \frac{\mathbb{E}[y(d - \mathbb{E}(d|z^*))]}{\mathbb{E}[d(d - \mathbb{E}(d|z^*))]}$$

The following theorem shows that ATE still can be identified in the presence of the explicit treatment effect heterogeneity, i.e., $\lambda(z^*) \neq 0$, with additional Assumption 4.

Theorem 1 *Under Assumption 1, Assumption 3 and Assumption 4 and Definition 1,* $\theta \equiv (\alpha, \gamma')'$ *can be identified by the following:*

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

where[17]

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

Furthermore, ATE (α) *is identified 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 identified and 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

$$\begin{aligned} &\mathbb{E}(x_1^2) = \mathbb{E}(Var(d|z^*)) \\ &\mathbb{E}(x_1\mathbf{x}_2) = \mathbb{E}[Var(d|z^*)(\mathbf{w} - \mathbb{E}(\mathbf{w}))] \\ &\mathbb{E}(\mathbf{x}_2\mathbf{x}_2') = \mathbb{E}[Var(d|z^*)(\mathbf{w} - \mathbb{E}(\mathbf{w}))(\mathbf{w} - \mathbb{E}(\mathbf{w}))'] \\ &\mathbb{E}(x_1 y) = \mathbb{E}[(d - \mathbb{E}(d|z^*))y] \\ &\mathbb{E}(\mathbf{x}_2 y) = \mathbb{E}[(d - \mathbb{E}(d|z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))y] \end{aligned}$$

Proof. See Appendix B.1. ■

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

Corollary 1 *For a fuzzy RD design,*

(1) when $z^* \to 0$, the identification result in Theorem 1 accommodates Hahn, Todd and Van der Klaauw (2001)'s result on identifying ATE at the cutoff point:

$$ATE(0) \equiv \lim_{z^* \to 0} \mathbb{E}(\eta + \lambda(z^*)|z^*) = \mathbb{E}(\eta|z^* = 0) = \frac{\lim_{z^* \downarrow 0} \mathbb{E}(y|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(y|z^*)}{\lim_{z^* \downarrow 0} \mathbb{E}(d|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(d|z^*)}$$

(2) when $z^* \to 0$, if we replace Assumption 3 and Assumption 4 with Assumption 2, then the identification result in Theorem 1 accommodates Hahn, Todd and Van der Klaauw (2001)'s result on identifying ATE at the cutoff point:

$$ATE = \frac{\mathbb{E}[y(d - \mathbb{E}(d|z^*))]}{\mathbb{E}[d(d - \mathbb{E}(d|z^*))]} = \frac{\lim_{z^* \downarrow 0} \mathbb{E}(y|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(y|z^*)}{\lim_{z^* \downarrow 0} \mathbb{E}(d|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(d|z^*)}$$

Proof. See Appendix B.2. ■

It is evident that the identification of treatment effects with an RD design is similar to "matching estimators". They both hinge upon the "propensity score", $\mathbb{E}(d|z^*)$, because they are both built upon selection-on-observables.

2.2 Identification under Selection-on-Unobservables

It is worth emphasizing that identification of treatment effects with an RD design requires that the selection variable should not be perfectly manipulated by individuals. Otherwise, individuals may self-select into the treatment, which induces correlations between the observable z^* and the unobservable u_0, u_1 and v. Such correlations make the eligibility indicator z endogenous and thus invalidate an RD design's instrumentality. In practice, several cases effectively prevent manipulation. First, the cutoff point is unexpected. If the cutoff point for eligibility comes entirely as a surprise, and no one has ever anticipated such a selection rule, then manipulation of the selection variable is completely avoided. Second, the eligibility criterion is determined by predetermined variables that are simply not manipulable. Third, individuals only have partial control over selection variables. In some cases, manipulation is possible, but impossible to be perfect. Even for an anticipated threshold, within a small neighborhood, there will still be some randomness that prevents perfect manipulation (Lee 2007).

In the absence of manipulating the selection variable, regarding identification of ATE under selection-on-unobservables, a first reaction may be to apply Theorem 1, but instrument d by $z = 1\{z^* \leqslant 0\}$. However, this strategy does not work simply because once conditional on z^*, there is no variation in z. In this situation, z and the first-stage residual, $d - \mathbb{E}(d|z^*)$, must be independent, which invalidates the rank condition for identification with an instrumental

variable (IV). To clarify this point, consider a case with a homogeneous effect:

$$
\begin{aligned}
&y = g_0(z^*) + \alpha d + u_0 \\
&y - \mathbb{E}(y|z^*) = \alpha(d - \mathbb{E}(d|z^*)) + u_0 \\
&\text{instrumenting } d \text{ by } z \text{ and using the moment condition } \mathbb{E}(zu_0) = 0 \\
&\mathbb{E}[z(y - \mathbb{E}(y|z^*))] = \alpha\mathbb{E}[z(d - \mathbb{E}(d|z^*))] \\
&\mathbb{E}[z(y - \mathbb{E}(y|z^*))] = \mathbb{E}(zy - z\mathbb{E}(y|z^*)) = \mathbb{E}(zy) - \mathbb{E}(\mathbb{E}(zy|z^*)) = 0 \\
&\mathbb{E}[z(d - \mathbb{E}(d|z^*))] = \mathbb{E}(zd - z\mathbb{E}(d|z^*)) = \mathbb{E}(zd) - \mathbb{E}(\mathbb{E}(zd|z^*)) = 0 \\
&\Rightarrow \alpha \text{ unidentifiable}
\end{aligned}
$$

This example stresses that identification strategies under selection-on-unobservables inevitably require making untestable assumptions on the functional form of $g_0(z^*)$. In contrast, such assumptions can be avoided under selection-on-observables using a partially linear model.

Under Assumption 1 and Definition 2, we have the following results:

(1) d is endogenous: $Cov(d, u_0 + d(u_1 - u_0)) \neq 0$
(2) omitted variables bias: $Cov(u_0, v|z^*) \neq 0, Cov(u_1, v|z^*) \neq 0$
(3) selectivity (or sorting) bias: $Cov(u_1 - u_0, v|z^*) \neq 0$

There are two main streams of analyses on heterogeneous treatment effects under selection-on-unobservables: the control functions approach and the instrumental variables approach. The former, aimed at ATE, is in line with Heckman (1979), which is based on figuring out $\mathbb{E}(y|d, z^*)$ directly through assumptions on the joint distribution of (u_0, u_1, v). The latter, aimed instead for LATE, replaces distributional assumptions with monotonicity restrictions on the relationship between instruments and potential treatment status. The identification strategy discussed in this paper under selection-on-unobservables focuses on the case $Cov(u_1 - u_0, v|z^*) \neq 0$ because it is often the case that sorting behavior is rational: i.e. and individuals do $(d = 1)$ the best $(Cov(u_1 - u_0, v|z^*) > 0)$ for themselves. The proposed correction function approach comes as a middle ground between the control functions approach and the IV approach. It allows for identifying ATE but only requires that the conditional moment restriction $\mathbb{E}(u_1 - u_0|v)$ takes a polynomial form instead of assuming a joint distribution for (u_1, u_0, v). The proposed correction function approach is derived from a correlated random coefficient (CRC) model used by Wooldridge (2002, 2007), which extend Garen (1984), Heckman and Vytlacil (1998) and Wooldridge (1997, 2003) to the case of a switching regression model with a binary treatment. This paper further relaxes Wooldridge (2002, 2007), which assumes 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, namely quadratic, in v. This extension to nonlinear relationships between the treatment selection and treatment gains accommodates cases with important economic implications. For example, an adverse selection occurs when $Cov(u_1 - u_0, v) < 0$ and "cream-skimming" exists when $Cov(u_1 - u_0, v) > 0$.

Theorem 2 *Under Assumption 1, Assumption 3, Definition 2, and the following two additional assumptions:*

(A1) $\mathbb{E}(u_1 - u_0|v) = \mathbb{E}(u_1 - u_0|v, z^*) = \xi_1 v + \xi_2 v^2$
(A2) $F_v(\cdot) \sim N(0, 1)$

the observed outcome can be rewritten as:

$$\begin{aligned} y &= g_0(z^*) + \alpha d + d(\lambda(z^*) - \mathbb{E}(\lambda(z^*)) + \widetilde{e} \\ &\quad + \xi_1 \phi(\pi_0 + \pi_1 z + \pi_2 z^*) \\ &\quad + \xi_2 \left[\Phi(\pi_0 + \pi_1 z + \pi_2 z^*) - (\pi_0 + \pi_1 z + \pi_2 z^*)\phi(\pi_0 + \pi_1 z + \pi_2 z^*)\right] \\ \widetilde{e} &\equiv u_0 + d(u_1 - u_0) - \mathbb{E}(d(u_1 - u_0)|z, z^*) + d(\eta - \mathbb{E}(\eta)) \end{aligned}$$

With two correction functions $\phi(\cdot)$ and $\Phi(\cdot) - (\cdot)\phi(\cdot)$ added back in, d instrumented by z, ATE$\equiv \mathbb{E}(\eta + \lambda(z^*)) \equiv \alpha$ is identified by using $\mathbb{E}(d|z, z^*)$ and $\mathbb{E}(d|z, z^*)\left[\lambda(z^*) - \mathbb{E}(\lambda(z^*))\right]$ as the instruments for d and $d\left[\lambda(z^*) - \mathbb{E}(\lambda(z^*))\right]$ respectively.

Proof. See Appendix B.3. ■

Because $\widetilde{e}$ is heteroskedastic due to $d(\eta - \mathbb{E}(\eta))$, we may just use z as the instrument for d in estimation. Also notice that a sufficient condition for (A1) is a bivariate normal distribution:

$$\begin{pmatrix} u_1 - u_0 \\ v \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & \xi \\ \xi & 1 \end{pmatrix}\right), \text{ where } \xi \equiv Cov(u_1 - u_0, v)$$

In Appendix B.4, we give identification results using two control functions to deal with both omitted variables bias and selectivity bias simultaneously in the same spirit as Heckman (1979).

Note that if $u_1 = u_0$, selection-on-unobservables only causes omitted variables bias. There is no sorting based on treatment gains. In this situation, an IV estimator is able to identify ATE.

Theorem 3 *Under Assumption 1, Assumption 3 with $u_1 = u_0$ and Definition 2, z is an instrument for d, and ATE$\equiv \mathbb{E}(\eta + \lambda(z^*)) \equiv \alpha$ is identified by using $\mathbb{E}(d|z, z^*)$ and $\mathbb{E}(d|z, z^*)\left[\lambda(z^*) - \mathbb{E}(\lambda(z^*))\right]$ as the instruments for d and $d\left[\lambda(z^*) - \mathbb{E}(\lambda(z^*))\right]$ respectively.*

Proof. We can rewrite the y-equation:

$$\begin{aligned} &y = g_0(z^*) + \alpha d + d\left[\lambda(z^*) - \mathbb{E}(\lambda(z^*))\right] + \widetilde{e} \\ &u_1 = u_0 \Rightarrow e = u_0 = u_1 \\ &\Rightarrow \mathbb{E}\left[\widetilde{e}|\mathbb{E}(d|z, z^*)\right] = \mathbb{E}\left[e + d(\eta - \mathbb{E}(\eta))|\mathbb{E}(d|z, z^*)\right] \\ &= \mathbb{E}\{\mathbb{E}\left[e + d(\eta - \mathbb{E}(\eta))|z, z^*\right]|\mathbb{E}(d|z, z^*)\} \text{ (generalized law of iterated expectation)} \\ &= \mathbb{E}[\mathbb{E}(u_0|z^*) + \mathbb{E}(d|z, z^*)\mathbb{E}(\eta - \mathbb{E}(\eta))|\mathbb{E}(d|z, z^*)] = 0 \end{aligned}$$

The redundancy restriction of $\mathbb{E}(d|z, z^*)$ is verified. The relevancy restriction of $\mathbb{E}(d|z, z^*)$ holds trivially. ■

3 Estimation of Treatment Effects

Based on identification results discussed in Section 2, estimators can be constructed straightforwardly by the analogy principle (Bera and Bilias 2002).

3.1 Estimation under Selection-on-Observables

The following two-stage estimator is constructed according to the identification results given by Theorem 1.

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

where

$y_i = g_0(z_i^*) + \alpha d_i + d(\mathbf{w}_i - \mu)'\gamma + \widetilde{e}_i$
$\widetilde{e} \equiv u_0 + d(u_1 - u_0) + d(\eta - \mathbb{E}(\eta))$, $\mathbb{E}(\widetilde{e}|z^*) = \mathbb{E}(\widetilde{e}|d, z^*) = 0$
$\widehat{\mathbf{x}}_i \equiv \left[\left(d_i - p(z_i^*; \widehat{\lambda})\right), \left(d_i - p(z_i^*; \widehat{\lambda})\right)(\mathbf{w}_i - \widehat{\mu})'\right]'$, $\mathbf{w}_i$ is a vector including polynomials of $\mathbf{z}_i^*$
$\widehat{\mu} \equiv \widehat{\mathbb{E}}(\mathbf{w}_i), p(z_i^*; \widehat{\lambda}) \equiv \widehat{\mathbb{E}}(d_i|z_i^*)$
$\theta \equiv (\alpha, \gamma')', \alpha \equiv \mathbb{E}(\eta + \lambda(z^*)) \equiv \text{ATE}, \lambda(z^*) \equiv \mathbf{w}'\gamma$

This estimator takes into account both ATE and the explicit part of treatment effect heterogeneity. Since there is no first-stage for y, which differentiates it from Robinson (1988)'s two-stage estimator, we will not run the risk of misspecifying $\mathbb{E}(y|z^*)$. This estimator is therefore robust to various functional forms of overt biases caused by $g_0(z^*)$. To implement this estimator, we only need a consistent estimator for $\mathbb{E}(d|z^*)$ in the first-stage. The intuition behind this estimator is to use a first-stage to clean up overt biases caused by the observables, and then use the "cleansed" residual obtained in the first-stage to generate orthogonality conditions for a moment-based estimation for θ. Note that these orthogonality conditions are implied by selection-on-observables. We next present the large sample properties for $\widehat{\theta}_{\text{RD\_robust}}$ with its asymptotic variance adjusted for generated regressors which are the first-stage residuals.

Theorem 4 (Consistency and Asymptotic Normality) *Under Assumption 1, Assumption 3, Assumption 4, Definition 1 and the parametric assumption $\mathbb{E}(d|z^*) = p(z^*; \lambda)$, we have*

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

where

$A_0 \equiv \mathbb{E}(\mathbf{x}\mathbf{x}')$
$\Omega \equiv Var\left(\mathbf{x}\left(y - \mathbf{x}'\theta\right) - B_0\mathbf{r}(\lambda)\right)$
$B_0 \equiv \mathbb{E}\left((\theta \otimes \mathbf{x}')' \frac{\partial \mathbf{x}}{\partial \lambda'}\right) = \mathbb{E}\left((\theta \otimes \widehat{\mathbf{x}}')' \frac{\partial \mathbf{f}(d, z^*, \mathbf{w}; \lambda, \mu)}{\partial \lambda'}\right)$
$\mathbf{x} \equiv \left[\left(d - p(z^*; \lambda)\right), \left(d - p(z^*; \lambda)\right)(\mathbf{w} - \mu)'\right]' \equiv \mathbf{f}(d, z^*, \mathbf{w}; \lambda, \mu)$

together with the influence function for $\widehat{\lambda}$:

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

Proof. See Appendix C.1. ■

The supplemental Web Appendix[18] discusses a simplification of $\widehat{\theta}_{\text{RD\_robust}}$ in the absence of the interactions between the treatment and the observables, i.e. $\lambda(z^*) = 0$.

3.2 Estimation under Selection-on-Unobservables

Based on Theorem 2, the following estimator is constructed to deal with both omitted variables bias and selectivity bias. We also provide its asymptotic properties in the presence of generated instruments (used for eliminating omitted variables bias) and generated regressors (used for eliminating selectivity bias).

Under selection-on-unobservables with heterogeneous treatment effects, a parameterized model with two correction functions added back in is given by the following:

$$\begin{aligned}
y &= \beta_0 + \mathbf{w}'\beta_1 + \alpha d + d(\mathbf{w}-\mu)'\gamma + \widetilde{e} \\
&\quad + \xi_1\phi(\pi_0 + \pi_1 z + \pi_2 z^*) \\
&\quad + \xi_2\left[\Phi(\pi_0 + \pi_1 z + \pi_2 z^*) - (\pi_0 + \pi_1 z + \pi_2 z^*)\phi(\pi_0 + \pi_1 z + \pi_2 z^*)\right] \\
d &= 1\{\pi_0 + \pi_1 z + \pi_2 z^* + v > 0\} \\
\widetilde{e} &\equiv u_0 + d(u_1 - u_0) - \mathbb{E}(d(u_1 - u_0)|z, z^*) + d(\eta - \mathbb{E}(\eta)) \\
\mu &\equiv \mathbb{E}(\mathbf{w})
\end{aligned}$$

where $\mathbb{E}(\widetilde{e}|z, z^*) = 0, v \sim N(0,1), \phi(\cdot)$ is normal pdf, $\Phi(\cdot)$ is normal cdf. In addition, we use the following definitions and parameterization:

$$\begin{aligned}
&\alpha \equiv \mathbb{E}(\eta + \lambda(z^*)) \equiv \text{ATE} \\
&g_0(z^*) \equiv \beta_0 + \mathbf{w}'\beta_1 \\
&\lambda(z^*) \equiv \mathbf{w}'\gamma, \text{ where } \mathbf{w} \text{ is a vector including polynomials of } z^*
\end{aligned}$$

We also give the following definitions to simply notation: $\theta \equiv (\beta_0, \beta_1', \alpha, \gamma', \xi_1, \xi_2)', \pi \equiv (\pi_0, \pi_1, \pi_2)', \widetilde{\mathbf{z}} \equiv (1, z, z^*)$. The regressors included in the model defined at the population are:

$$\begin{aligned}
\mathbf{x} &\equiv (1, \mathbf{w}', d, d(\mathbf{w}-\mu)', \phi(\widetilde{\mathbf{z}}'\pi), \Phi(\widetilde{\mathbf{z}}'\pi) - (\widetilde{\mathbf{z}}'\pi)\phi(\widetilde{\mathbf{z}}'\pi))' \\
&\equiv \mathbf{f}(d, \widetilde{\mathbf{z}}, \mathbf{w}; \pi, \mu)
\end{aligned}$$

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

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

The instruments, both included and excluded, used in the population model are:

$$\begin{aligned}
\mathbf{z} &\equiv (1, \mathbf{w}', \Phi(\widetilde{\mathbf{z}}'\pi), \Phi(\widetilde{\mathbf{z}}'\pi)(\mathbf{w}-\mu)', \phi(\widetilde{\mathbf{z}}'\pi), \Phi(\widetilde{\mathbf{z}}'\pi) - (\widetilde{\mathbf{z}}'\pi)\phi(\widetilde{\mathbf{z}}'\pi))' \\
&\equiv \mathbf{g}(\widetilde{\mathbf{z}}, \mathbf{w}; \pi, \mu)
\end{aligned}$$

[18] It is available at http://are.berkeley.edu/~yang/research.html.

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

$$\begin{aligned}\widehat{\mathbf{z}}_i &\equiv (1, \mathbf{w}_i', \Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}), \Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi})(\mathbf{w}_i - \widehat{\mu})', \phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}), \Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}) - (\widetilde{\mathbf{z}}_i'\widehat{\pi})\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}))' \\ &\equiv \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\mu})\end{aligned}$$

The problem of generated regressors and generated instruments is caused by $\widehat{\pi}$ and $\widehat{\mu}$. The actual model used for estimation, based on a random sample, is:

$$\begin{aligned}&y_i = \widehat{\mathbf{x}}_i'\theta + \text{error}_i \\ &d_i = 1\{\widetilde{\mathbf{z}}_i'\pi + v_i > 0\} \\ &v \sim N(0,1), \phi(\cdot) \text{ normal pdf, } \Phi(\cdot) \text{ normal cdf}\end{aligned}$$

To analyze asymptotic properties, it is useful to rewrite the model in the following way:

$$y_i = \widehat{\mathbf{x}}_i'\theta + \text{error}_i = \widehat{\mathbf{x}}'\theta + (\mathbf{x}_i - \widehat{\mathbf{x}}_i)'\theta + \widetilde{e}_i, \mathbb{E}(\widetilde{e}_i|\widetilde{\mathbf{z}}_i) = 0$$

Given the distributional assumption that $v \sim N(0,1)$, $\mathbf{z}_i$ defined in the population model are the optimal instruments if conditional homoskedasticity for $Var(\widetilde{e}_i|\widetilde{\mathbf{z}}_i)$ holds. Since we have equal number of endogenous variables and instruments, the model is just-identified. An IV estimator for θ with generated regressors and instruments including two correction functions is given by the following:

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

Theorem 5 (Consistency and Asymptotic Normality) *Under Assumption 1, Assumption 3, Definition 2, (A1), (A2) and the following two additional assumptions:*

(A3) $g_0(z^*) \equiv \beta_0 + \mathbf{w}'\beta_1$
(A4) $\lambda(z^*) \equiv \mathbf{w}'\gamma$
where $\mathbf{w}$ *is a vector including polynomials of* z^* *and* $\theta \equiv (\beta_0, \beta_1', \alpha, \gamma', \xi_1, \xi_2)'$

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

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

where

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

Proof. See Appendix C.2. ■

The variance in the limit distribution of $\sqrt{N}(\widehat{\theta}_{\text{crrf}} - \theta)$ can be estimated by the following:

$$
\begin{aligned}
\widehat{A}_0 &= \frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i' \\
\widehat{\Omega} &= \frac{1}{N}\sum_{i=1}^{N}(\widehat{\mathbf{z}}_i\widehat{\widetilde{e}}_i - \widehat{B}_0\mathbf{r}_i(\widehat{\pi}) - \widehat{B}_1(\mathbf{w}_i - \overline{\mathbf{w}}))(\widehat{\mathbf{z}}_i\widehat{\widetilde{e}}_i - \widehat{B}_0\mathbf{r}_i(\widehat{\pi}) - \widehat{B}_1(\mathbf{w}_i - \overline{\mathbf{w}}))' \\
\widehat{B}_0 &= \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\xi}_2\widetilde{\mathbf{z}}_i'\widehat{\pi} - \widehat{\xi}_1\right)(\widetilde{\mathbf{z}}_i'\widehat{\pi})\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi})\widehat{\mathbf{z}}_i\widetilde{\mathbf{z}}_i' \\
\widehat{B}_1 &= -\frac{1}{N}\sum_{i=1}^{N}d_i\widehat{\mathbf{z}}_i\widehat{\gamma}' \\
\mathbf{r}(\widehat{\pi}) &= \left(\frac{1}{N}\sum_{i=1}^{N}\frac{\phi_i^2(\widetilde{\mathbf{z}}_i'\widehat{\pi})\widetilde{\mathbf{z}}_i\widetilde{\mathbf{z}}_i'}{\Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi})\,(1-\Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}))}\right)^{-1}\frac{\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi})\widetilde{\mathbf{z}}_i(d_i-\Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}))}{\Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi})\,(1-\Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}))} \\
\widehat{\widetilde{e}}_i &= y_i - \widehat{\mathbf{x}}_i'\widehat{\theta}_{\text{crrf}}
\end{aligned}
$$

In contrast to this semi-structural correction functions approach, a more structural approach is to obtain $\mathbb{E}(y|d,z,z^*)$ directly, which involves adding two control functions—one for omitted variables bias and the other one for selectivity bias. Detailed discussions on this control functions estimator are provided in the supplemental Web Appendix.[19]

4 Monte Carlo Experiments

To investigate the trade-off between efficiency and bias in estimating the average treatment effect (and average effects local to the discontinuity) when the effects covary with observables and unobservables, I conduct a series of Monte Carlo experiments with various sample sizes of 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000 and 10000. All simulations are based on 1000 trials.[20]

[19] It is available at http://are.berkeley.edu/~yang/research.html.

[20] A complete set of simulation results are available upon request.

4.1 Design and Estimators

Based on Assumption 1, the data generating process (DGP) used by Monte Carlo experiments is specified as follows ($i = 1, 2, \cdots, N$):

(a) Selection process:

$$
\begin{aligned}
& z_i^* \sim \text{i.i.d. } U[-1,1] \Rightarrow \mathbb{E}(z^*) = 0, Var(z^*) = \frac{1}{3} \\
& z_i = 1\{z_i^* \leqslant 0\} \\
& d_i = 1\{\pi_0 + \pi_1 z_i + \pi_2 z_i^* + v_i > 0\} \\
& v_i \sim \text{i.i.d. } N(0,1), \Phi(\cdot) \text{ normal cdf} \\
& \quad \Rightarrow \begin{array}{l} \mathbb{E}(d_i | z_i^* \leqslant 0) = \Phi(\pi_0 + \pi_1 + \pi_2 z_i^* | z_i^* \leqslant 0) \\ \mathbb{E}(d_i | z_i^* > 0) = \Phi(\pi_0 + \pi_2 z_i^* | z_i^* > 0) \end{array}
\end{aligned}
$$

(b) Potential and observed outcomes:

$$
\begin{aligned}
g_0(z^*) \equiv \mathbb{E}(y_0 | z^*) &= \begin{cases} \beta_0 + \beta_1 \cos(h) + \beta_2 h^2 & (-h \leqslant z^* \leqslant h) \\ \beta_0 + \beta_1 \cos(z^*) + \beta_2 z^{*2} & (\text{else}) \end{cases} \\
g_1(z^*) \equiv \mathbb{E}(y_1 | z^*) &= \begin{cases} \beta_0 + \beta_1 \cos(h) + \beta_2 h^2 + \mathbb{E}(\eta) & (-h \leqslant z^* \leqslant h) \\ \beta_0 + \beta_1 \cos(z^*) + \gamma_1 z^* + (\beta_2 + \gamma_2) z^{*2} + \mathbb{E}(\eta) & (\text{else}) \end{cases} \\
\eta &\sim N(1,1), \eta \amalg z^* \\
u_0 &= \xi_0 v + \varepsilon, \varepsilon \sim N(0,1), \varepsilon \amalg v \\
u_1 &= (\xi_0 + \xi_1) v + \xi_2 v^2 + \varepsilon \Rightarrow \mathbb{E}(u_1 - u_0 | v) = \xi_1 v + \xi_2 v^2 \\
y_0 &= g_0(z^*) + u_0 \\
y_1 &= g_1(z^*) + u_1 + \eta - \mathbb{E}(\eta) \\
y &= d y_1 + (1-d) y_0
\end{aligned}
$$

(c) Fixed parameters:

$$
\begin{aligned}
& \text{“jump”} = 0.5 \equiv \text{“discontinuity” in the selection process} \\
& \pi_0 = -\Phi^{-1}\left(\frac{\text{jump}+1}{2}\right), \pi_1 = 2\Phi^{-1}\left(\frac{\text{jump}+1}{2}\right), \pi_2 = -1 \\
& \beta_0 = 1, \beta_1 = 1, \beta_2 = 1 \\
& \gamma_1 = 1, \gamma_2 = 3
\end{aligned}
$$

(d) Varying parameters:

$h = 0$ (Model I: potential outcomes affected by the selection variable)
$h = 1$ (Model II: potential outcomes are mean-independent from the selection variable)
$\xi_0 = 0, \xi_1 = \xi_2 = 0$ (selection-on-observables)
$\xi_0 = 1 \neq 0, \xi_1 = \xi_2 = 0$ (omitted variables bias, OVB, only)
$\xi_0 = 1 \neq 0, \xi_1 = 1 \neq 0, \xi_2 = 1 \neq 0$ (OVB and nonlinear selectivity bias)
$\xi_0 = 1 \neq 0, \xi_1 = 1 \neq 0, \xi_2 = 0$ (OVB and linear selectivity bias)

(e) ATE:

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

Model I: $\text{ATE} = \mathbb{E}(y_1 - y_0) = \mathbb{E}(\gamma_1 z^* + \gamma_2 z^{*2} + 1) = 2$
$\text{ATE}(0) = \mathbb{E}(y_1 - y_0|z^* = 0) = \mathbb{E}(y_1 - y_0) - \gamma_1 \mathbb{E}(z^*) - \gamma_2 \mathbb{E}(z^{*2}) = 1$

Model II: $\text{ATE}(-h \leqslant z^* \leqslant h) = \mathbb{E}(y_1 - y_0| - h \leqslant z^* \leqslant h) = 1$
$\text{ATE}(0) = \mathbb{E}(y_1 - y_0|z^* = 0) = 1$

In summary, we consider the following ten cases, with the following five cases for both Model I and Model II:

- selection-on-observables without using the "eligibility indicator" (z) in estimating $\mathbb{E}(d|z^*)$
- selection-on-observables using the "eligibility indicator" (z) in estimating $\mathbb{E}(d|z^*)$
- selection-on-unobservables with omitted variables bias only
- selection-on-unobservables with both omitted variables bias and selectivity bias
- selection-on-unobservables with omitted variables bias and selectivity bias and the joint distribution of unobservables consistent with joint normality

For each trial, based on these four criteria—mean bias, median bias, root mean squared error (RMSE) and median absolute error—we evaluate the performance of the following six estimators, where $\pi \equiv (\pi_0, \pi_1, \pi_2)', \widetilde{\mathbf{z}} \equiv (1, z, z^*)$, under both Model I and Model II.

(1) proposed robust RD estimator ("robust")

Model I: regress y on $(d - \Phi(\pi'\widetilde{\mathbf{z}})), (d - \Phi(\pi'\widetilde{\mathbf{z}})), (d - \Phi(\pi'\widetilde{\mathbf{z}}))\mathbf{w}))$ by least squares, where $\mathbf{w} = (z^*, z^{*2})$.
Model II: regress y on d by least squares using $d - \Phi(\pi'\widetilde{\mathbf{z}})$ as the instrument for d.

(2) Robinson's (1988) two-stage estimator ("robinson")

Model I and II: in the first-stage, regress y on $\mathbf{w}$ and obtain the residual $\widetilde{y}$; in the second-stage, regress $\widetilde{y}$ on $(d - \Phi(\pi'\widetilde{\mathbf{z}})), (d - \Phi(\pi'\widetilde{\mathbf{z}})), (d - \Phi(\pi'\widetilde{\mathbf{z}}))\mathbf{w}))$ by least squares, where $\mathbf{w} = (z^*, z^{*2})$.

(3) revised correction-function estimator ("corr func")

Model I: regress y on $(1, d, \mathbf{w}_1, d(\mathbf{w}_2 - \overline{\mathbf{w}}_2), \phi(\pi'\widetilde{\mathbf{z}}), \Phi(\pi'\widetilde{\mathbf{z}}) - (\pi'\widetilde{\mathbf{z}})\phi(\pi'\widetilde{\mathbf{z}}))$ by lV using $(1, \Phi(\pi'\widetilde{\mathbf{z}}), \mathbf{w}_1, \Phi(\pi'\widetilde{\mathbf{z}})(\mathbf{w}_2 - \overline{\mathbf{w}}_2), \phi(\pi'\widetilde{\mathbf{z}}), \Phi(\pi'\widetilde{\mathbf{z}}) - (\pi'\widetilde{\mathbf{z}})\phi(\pi'\widetilde{\mathbf{z}}))$ as the instrument, where $\mathbf{w}_1 = (\cos(z^*), z^*), \mathbf{w}_2 = (z^*, z^{*2})$.

Model II: regress y on $(1, d, \phi(\pi'\widetilde{\mathbf{z}}), \Phi(\pi'\widetilde{\mathbf{z}}) - (\pi'\widetilde{\mathbf{z}})\phi(\pi'\widetilde{\mathbf{z}}))$ by lV using $(1, \Phi(\pi'\widetilde{\mathbf{z}}), \phi(\pi'\widetilde{\mathbf{z}}), \Phi(\pi'\widetilde{\mathbf{z}}) - (\pi'\widetilde{\mathbf{z}})\phi(\pi'\widetilde{\mathbf{z}}))$ as the instrument.

(4) control-function estimator ("ctrl func")

Model I: regress y on $(1, d, \mathbf{w}_1, d(\mathbf{w}_2 - \overline{\mathbf{w}}_2), d(\phi(\pi'\widetilde{\mathbf{z}})/\Phi(\pi'\widetilde{\mathbf{z}})), (1-d)(\phi(-\pi'\widetilde{\mathbf{z}})/\Phi(-\pi'\widetilde{\mathbf{z}})))$ by least squares, where $\mathbf{w}_1 = (\cos(z^*), z^*), \mathbf{w}_2 = (z^*, z^{*2})$.

Model II: regress y on $(1, d, d(\phi(\pi'\widetilde{\mathbf{z}})/\Phi(\pi'\widetilde{\mathbf{z}})), (1-d)(\phi(-\pi'\widetilde{\mathbf{z}})/\Phi(-\pi'\widetilde{\mathbf{z}})))$ by least squares.

(5) OLS

Model I: regress y on $(1, d, z^*, d(z^* - \overline{z^*}))$.

Model II: regress y on $(1, d)$.

(6) 2SLS (Imbens and Lemieux 2007)

We regress y on $(1, d, z^*, zz^*)$ using z as the instrument for d to get the estimated ATE at the cutoff point.

4.2 Results and Discussion

The series of Monte Carlo experiments reveal the efficiency-bias trade-off in estimating ATE under both selection-on-observables and selection-on-unobservables. It is also evident that such a trade-off intensifies as the sample size grows. Our following discussions are based on both a small sample (with 100 observations) and a relatively large sample size (with 1000 observations). Detailed results are presented in Appendix D.

4.2.1 Selection-on-observables

Appendix Table D.1 reveals that with a fuzzy RD design, an OLS estimator, maybe efficient under selection-on-observables, suffers from severe attenuation bias. The downward bias is about 30% in terms of mean bias and 29% in terms of median bias with a small sample of 100 observations. Such a downward bias persists and stays about 28% in terms of both mean bias and median bias when the sample size reaches 1000 observations. Furthermore, this attenuation bias remains about 30% in terms of both mean bias and median bias even with a sample of 10000 observations. This can be seen in Appendix Figure 1 (for the case of selection-on-observables in Model I) in Appendix E. In contrast, our proposed robust RD estimator effectively reduces the attenuation bias to 3% in terms of mean bias and 4% in terms of median bias with only 100 observations. The usefulness in removing such an attenuation bias of this robust RD estimator becomes prominent only with a relatively large sample of 1000 observations: it gives a slightly upward bias of 0.4% in terms of mean bias

and 0.1% in terms of median bias. The improvement on bias reduction is due to the fact that we use the first stage residual, $d - \mathbb{E}(d|z, z^*)$, as an instrument to orthogonalize $g_0(z^*)$ and the error terms in the second-stage. Those orthogonality conditions are free of both specification errors and measurement errors. In addition, the "eligibility" instrument (z) helps to further reduce the attenuation bias. This can be verified in Appendix Figure 1 in Appendix E. Furthermore, compared with the Robinson's two-stage estimator under correct specification, the efficiency loss in using the robust RD estimator is fairly negligible even with a moderate sample size (with 200 observations). This efficiency issue is reflected in Appendix Figure 2 (for the case of selection-on-observables in Model I) in Appendix E. It is worth mentioning that the bias of OLS suggests a situation of measurement errors that occurs under a sharp RD with unknown threshold. For more discussions on this, please see the discussion on "sharp RD with unknown threshold and measurement errors" in the supplemental Web Appendix[21]. It is also noticeable that there is no much difference, in terms of the bias-efficiency trade-off, between the robust RD estimator and Robinson's two-stage estimator. Appendix Figure 3 (for the case of selection-on-observables in Model II) in Appendix E illustrates this point.

When estimating ATE at the cutoff point in the presence of interactions between the treatment and observables (i.e. in the presence of explicit treatment effect heterogeneity), 2SLS works poorly in terms of mean and median bias, compared with the proposed robust RD estimator. See Table D.2 in Appendix D for details. This is probably due to misspecification of the bandwidth that contains the cutoff point. The robust RD estimator is less efficient than Robinson's two-stage estimator only when the latter has no specification errors in the first-stage. Therefore, the proposed RD estimator seems still preferable when ATE at the cutoff point is the parameter of interest. Appendix Figure 4 (for the case of selection-on-observables and identification at the cutoff point) in Appendix E verifies this argument.

4.2.2 Selection-on-unobservables

In the presence of explicit treatment heterogeneity, i.e. interactions between the treatment and observables, the correction function estimator doesn't work well in terms of both mean and median bias with moderate sample sizes unless the sample size gets very large. For more discussions on finite sample bias reduction with over-identification, please refer to The supplemental Web Appendix[22]. Besides, it is shown in Table D.1 in Appendix D that both the proposed robust RD estimator and Robinson's two-stage estimator work worse, in terms of bias, than OLS under selection-on-unobservables. Furthermore, Table D.1 in Appendix D shows that the control function estimator encounters greater bias once the underlying joint distribution of (u_0, u_1, v) is inconsistent with joint-normality.

When potential outcomes are mean-independent from the selection variable, the correction function estimator is uniformly more robust, in terms of bias, than the control function estimator. Appendix Figure 5 (for the case of selection-on-unobservables with omitted variables bias and nonlinear selectivity bias in Model II) in Appendix E illustrates this. Appendix Figure 6 (for the case of selection-on-unobservables with omitted variables bias and linear

[21]It is available at http://are.berkeley.edu/~yang/research.html.

[22]It is available at http://are.berkeley.edu/~yang/research.html.

selectivity bias) in Appendix E further shows that the efficiency loss of a correction function estimator due to its gain in robustness can be well compensated only by a sample of moderate size (with 400 observations).

Also note that under selection-on-unobservables with omitted variables bias and positive sorting bias, the OLS estimator suffers from a large upward bias of 111% in terms of both mean and median bias, and this magnitude stays constant with both a small sample of 100 observations and a relatively large sample of 1000 observations. In sharp contrast, with a small sample of 100 observations, the correction function estimator reduces the upward bias to -24% in terms of mean bias and further down to -0.005% in terms of median bias. This suggests that a correction function estimator is sensitive to outliers with a small sample. When the sample gets large, with 1000 observations, the correction function estimator manages to cut the upward bias down to 4% in terms of mean bias and 3% in terms of median bias. This suggests that the correction function estimator becomes more reliable when the sample size increases. The above argument is supported by Table D.1 in Appendix D.

5 Empirical Applications

To investigate the trade-off between efficiency and bias that arises in the presence of heterogeneous treatment effects, we offer two empirical applications. The first example uses data from Chay, McEwan and Urquiola (2005) to show the improvement of using the RD robust estimator proposed in Section 3.1 upon bias reduction and efficiency gain relative to 2SLS that Chay, McEwan and Urquiola (2005) uses. Under selection-on-observables, the RD robust estimator reduces bias because it is shown to be free of specification errors that occur in estimating the conditional expectations of the outcome using 2SLS. It brings efficiency gains because of the over-identifying restriction imposed by the eligibility criterion in the selection process and the estimated propensity score (Hirano, Imbens and Ridder 2003) in the first stage. It is further shown that such an improvement under selection-on-observables can be checked by the presence of sorting or selectivity bias. The correction function estimator proposed in Section 3.2 suggests a specification test based on the significance of the two correction terms.[23] If sorting or selectivity bias is detected in a chosen range, then the assumption of selection-on-observables should be rejected. Such a specification check is implied by the RD design's instrumental nature, and the construction of those correction terms is detailed in Section 3.2. In addition to being able to falsify selection-on-observables, we may also be interested to know any existence of "cream-skimming" or adverse selection, which are indicated by the two correction terms, that will either overstate or understate the program effects.

The efficiency-bias trade-off also arises in the case of continuous treatment. The second example uses data from Chay and Greenstone (2003) to demonstrate the improvement upon this trade-off. It also shows how to extend the RD design's applicability from a binary treatment case to a continuous treatment case such as different levels of air pollution. We

[23] Because these two correction terms are constructed from the data, we have the problem of generated regressors. Wooldridge (2007) shows that inferences based on the usual t-statistic, under the null hypothesis, is 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.

reexamine the impacts from air quality on infant mortality, utilizing the over-identifying restrictions implied by the RD design to obtain the efficiency gain in a range local to the discontinuity. We also use these over-identifying restrictions as a check for the validity of the instruments for the chosen range. The over-identifying restrictions can be seen as the counterparts for the correction terms with a binary treatment, both of which can be used to detect the sorting bias that invalidates an estimate's ATE interpretation. It is worth emphasizing that in the second example, such specification checks are not possible in conventional cross-sectional and fixed effects analyses in the absence of the RD design.

5.1 Chile's 900-School Program Evaluation

To improve school performance, in 1990, Chile's government initiated the "900 School Program" ("P-900", henceforth), a country-wide intervention to target low-performing and publicly-funded schools (Chay, McEwan and Urquiola 2005).[24] Eligibility of 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 value in its region chosen by the Ministry of Education.[25] As Chay, McEwan and Urquiola (2005) emphasizes, the schools' 1988 test scores were collected under a different political regime before the fall of Pinochet, at which time there was no evidence showing that such an intervention was ever contemplated. Therefore, it is plausible that schools had no incentive to manipulate their test performance in 1988 to qualify the P-900.

5.1.1 Program assignment

The actual assignment of P-900 program involved two stages. During the first stage in 1988, the Ministry of Education administered country-wide achievement tests to the population of the fourth graders. Officials of the Ministry then calculated each school's mean test scores in language and mathematics and then 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.[26] Schools whose overall average fell below the within-region's cutoff value were eligible for participating into this P-900 program.

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 another parallel program (MECE-Rural) was designed to accommodate them. Secondly, schools were removed from the pre-selected list if they had managerial

[24] There are four interventions associated with this program: (1) infrastructure improvement such as building repairs; (2) new instructional materials including textbooks for students from grade 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; (4) after-school tutoring workshops for the 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. The first two 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.

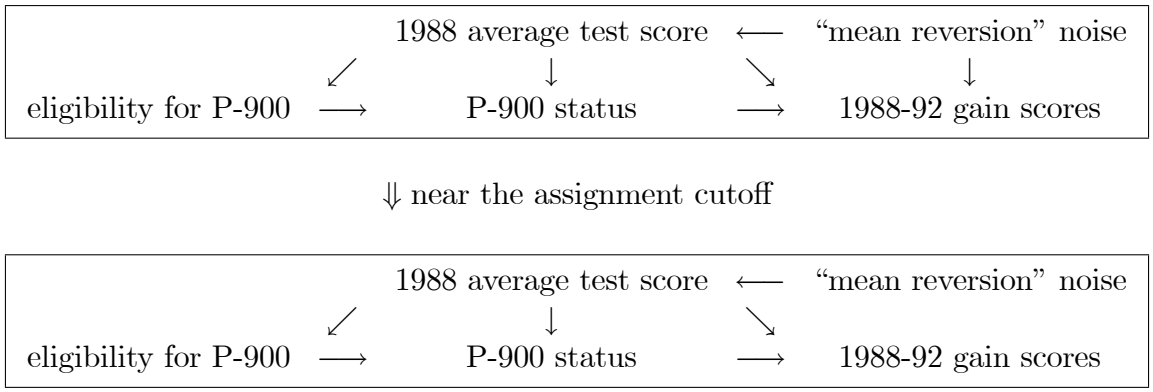
[25] There are 13 administrative regions in Chile.

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

problems, such as misreported enrollment, that were uncovered. In addition, the regional teams also introduced unobserved criteria to select certain schools, some of which were ineligible by the first-stage criteria.[27] Therefore, "selection-on-unobservables" existed in P-900's program assignment because this final set of characteristics is unobserved to an econometrician. Meanwhile, from a school's perspective, there was no incentive for them to forgo participation because the national government covered all of the costs. Accordingly, the imperfect program's assignment according to schools' eligibility is largely due to the unobservables created by the program's administers.[28]

5.1.2 Impact of the P-900 program

To gauge P-900's impact, we focus on whether the P-900 program had significant effects on test score (mathematics and language) gains of the fourth-graders over the period 1988–1992 to be consistent with Chay, McEwan and Urquiola (2005). The identification strategy using an RD design is explained by the following mechanism:



The above mechanism highlights the case where transitory testing noise, possibly due to luck, is mean reverting. Average test scores in 1988 may provide a noisy measure of school performances and a misleading ranking of schools (based on this noise measure) since the transitory noise contaminated the 1988 test scores. For example, a school's appearance at the bottom of a ranking, and therefore being selected into P-900, may be the result of transitory bad luck in the testing year. Unless the bad luck is persistent, test scores in this school would rebound in the next period even in the absence of the P-900 intervention. Thus, ignoring the mean-reversion noise will overstate the effect of P-900 that uses test-based rankings to select schools.[29]

Using RD's quasi-experimental nature, as Chay, McEwan and Urquiola (2005) suggests, we can effectively remove the direct impact of "mean-reversion" noises close to the cutoff

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

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

[29] Chay, McEwan and Urquiola (2005) finds 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 (1984–1988)."

point. Since schools near the assignment cutoff are likely to be randomized into the treatment, as a result, on average, mean-reversion noises experienced by these schools are also likely to be common to these schools. Thus, the direct impact of a common mean-reversion noise can be absorbed by the constant term included in the outcome equation. We now only have to deal with the indirect impact of the mean-reversion noise, and in this case (shown in the above mechanism) the 1988 test score becomes the proxy for that transitory noise. Therefore, the mean-reversion noise turns into a classical measurement error in the actual 1988 test score. The selection threshold that is set up by an RD design provides a powerful tool to remove the direct impacts of unobservable confounders such as mean-reversion noises.[30]

However, as we focus on a neighborhood around the cutoff, although we can be free of omitted variables bias by removing the direct impact of mean-reversion noise close to the cutoff point, we still need to watch out the sorting bias due to the interaction between the treatment and the unobservables. Ignoring such an interaction in the form of either "cream-skimming" or adverse selection, we will either overstate or understate the program's effectiveness. To detect such sorting biases, we use a t-test, as suggested by Theorem 5 and Wooldridge (2002, 2007).

The following table gives the estimates of P-900 effects. To reduce overt bias, we follow Chay, McEwan and Urquiola (2005), controlling for school-level socioeconomic status (SES)[31] 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 two correction terms follows Theorem 5.

[30] Figure 5 in Chay, McEwan and Urquiola (2005) provides evidence of mean-reversion noises and the program's impact.

[31] The SES index measures student socioeconomic status (SES), as reported by JUNAEB (*Junta Nacional de Auxilio Escolar y Becas*). It is scaled 0-100, with higher values indicating higher SES.

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
	(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***
	(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	5.85**	6.89***	4.80	6.44**	5.96	2.96
(with quadratic term)	[2.64]	[2.31]	(3.11)	(2.93)	(5.80)	(5.48)
standard deviation gain	0.75	0.93	0.62	0.87	0.76	0.40
correction term 1	−19.12***	−16.06***	−3.26	−3.60	6.03	8.05
	[6.17]	[4.87]	(6.86)	(6.32)	(19.60)	(16.19)
correction term 2	13.54**	8.11	−0.93	−2.92	−10.04	−8.81
	[6.49]	[5.35]	(5.70)	(5.11)	(12.93)	(10.13)
(4) correction function	8.74***	8.62***	4.66	5.99**	5.97	2.97
(with only linear term)	[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
	[3.58]	[3.02]	(4.38)	(4.16)	(7.89)	(7.19)
Sample size	2,591		938		392	

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, besides 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 point) in the specified range. The 2SLS in Panel A is proposed by Chay, McEwan and Urquiola (2005). The RD robust estimator in Panel A is proposed by this paper, which uses the first stage residual, the deviation between the treatment status and the estimated propensity score, as the instrument for the treatment. The correction function estimator with the quadratic correction terms in Panel B is proposed by this paper. The correction function estimator with only the linear term in Panel B is proposed by Wooldridge (2002). Standard errors robust to heteroskedasticity are in parentheses; Bootstrapped standard errors based on 2,000 replications are in [brackets]; "*" indicates significance at 10% level; "**" indicates significance at 5% level; "***" indicates significance at 1% level.

The 2SLS in Panel A is proposed by Chay, McEwan and Urquiola (2005), which uses the eligibility indicator as the instrument for the P-900 treatment status. The RD robust estimator in Panel A is proposed by this paper in Theorem 4, where a probit model for the P-900 treatment status is adopted. The regressors included in the probit model, besides the excluded instrument—the eligibility indicator, 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, the deviation between the treatment status and the estimated propensity score, as the instrument for the treatment. The correction function estimator with the quadratic correction terms in Panel B is proposed by this paper in Theorem 5. The correction function estimator with only the linear term in Panel B is proposed by Wooldridge (2002). For all cases (1)–(4) in Panel A and B, for the outcome equation we include, besides the P-900 treatment dummy, cubic polynomials for the 1988 average test score, SES in 1990, the changes in SES between 1990 and 1992 to be consistent with Chay, McEwan and Urquiola (2005, Table 5).

The results in Table 2 highlight the following:

First, in Panel A case (1) the efficiency-bias trade-off shows up when 2SLS is applied. When the full sample is used, the P-900 effect, about 0.32 standard deviations in both mathematics and language, is shown to be statistically significant. However, in the presence of heterogeneous treatment effects, this estimate loses the ATE interpretation because the sorting bias has been detected by the correction terms in the correction function estimator, which is shown in the "Full Sample" column in Table 2. As we focus on the schools close to the selection threshold, the selectivity bias can be effectively removed because the direct impact of the mean-reversion noise is homogenized between schools just above and just below the threshold. Therefore, estimates of P-900 effects will regain a valid ATE interpretation, which, however, is at the cost of efficiency. As columns "±5 Points" and "±2 Points" show, the 2SLS estimates become statistically insignificant.

Second, in Panel A case (2) the proposed RD robust estimator has made improvement upon this 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 when the full sample is used. Such effects (LATE) are only valid for an underlying (and unidentifiable) population of "compliers" because the sorting bias has been detected by the correction terms of the correction function estimators. When schools near the threshold are focused on, where ATE can be regained, in contrast to 2SLS, the RD robust estimator is able to detect P-900 effects and quantifies them between 0.22 and 0.34 standard deviations at 1% to 5% significance level. Compared with conventional estimators focusing on the effect only at the threshold, this RD estimator obtains the ATE for a predefined population local to the threshold, which has greater external validity. Compared with 2SLS, this RD estimator also has greater internal validity because it is free of specification errors. This point has been made in Section 3.1 and has been confirmed in Section 4.2.1.

Third, the correction terms in Panel B case (3) for the sorting bias are useful for checking the validity of the ATE interpretation for either the 2SLS or the RD robust estimator. When the full sample is used and the sorting bias is detected, the estimated P-900 effect in Panel A, statistically significant and roughly of 0.3 standard deviations, only measures the impact of P-900's eligibility criteria that *were* used by the program. To forecast P-900's effectiveness on a randomly selected school of a well defined population, we need to obtain ATE, not LATE, for a population predefined by its distance to the selection threshold. In Table 2, the RD robust estimator gives P-900's estimated ATE, which is roughly 0.3 standard deviations for a population of schools close to the selection threshold, and the validity of this ATE interpretation is checked by two correction terms in Panel B case (3): neither of them is statistically significant, which suggests the absence of adverse selection and "cream-skimming".

Fourth, the correction function estimator proposed by Wooldridge (2002) in Panel B case (4) forces the sorting bias to be either positive or negative. When the full sample is used, Wooldridge (2002)'s estimator can only detect a countervailed sorting bias when both adverse selection (negative sorting) and "cream-skimming" (positive sorting) are present. This point is shown in the gain score estimates in mathematics in case (3) and case (4) when the full sample is used. Our proposed correction function estimator detects a "U-shaped" sorting pattern because of the positive sign of the second correction term. In contrast, Wooldridge (2002)'s estimator can only detect the negative sorting, which also leads to a possibly over-estimated program effect when the positive sorting is neglected: the point

estimate of Wooldridge (2002)'s estimator is about 50% greater than our proposed correction function estimator with two correction terms, which takes into account the sorting in both directions.

It is worth mentioning that the correction function estimator has a lot more external validity than the RD robust estimator because the former is aimed for the treatment effects applied to the entire population while the latter is only applicable to a population close to the threshold. However, this externality is at the cost of internal validity because the parametric nature makes the correction function estimator susceptible to specification errors. In contrast, the RD robust estimator trades external validity for greater internal validity because it is free of specification errors. Therefore, the RD robust estimator gives compelling estimates for the specified population near the threshold. The estimated program effect given by the RD robust estimator shows that the 1988–1992 gain score of P-900 schools is 0.3 standard deviations higher than the non-P-900 schools. This indicates that the average P-900 school is at about the 62% of the non-P900 school distribution, which suggests a moderate improvement for a population of school close to the selection threshold. On the basis of this estimate, the P-900 effectiveness can be used to construct some cost-benefit measure, such as 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 close to the selection threshold, we need to use the correction function estimator. In the absence of specification errors, it gives the program's impact on a randomly selected school from the whole population.

The RD robust estimator gives compelling ATE estimates for a predefined population near the selection threshold while the correction function estimator gives ATE estimates under correct specifications for the entire population. The plausibility of the ATE identification using this correction function estimator will also depend upon the point in the distribution where the discontinuity occurs. Choosing between the RD robust estimator and the correction estimator is the balancing between internal and external validity, which should be guided by the research question or the population of policy interest before estimations take place. In summary, our proposed RD robust estimator allows for efficient estimation of an average effect in a range of observations local to the discontinuity, and within the range the correction function estimator suggests specification checks for the validity of an ATE interpretation, which will be violated in the presence of sorting biases. In the P-900 example, comparing the gains of schools just above and just below the assignment cutoff, which is set by the RD design, effectively eliminates the direct impact of mean-reversion noises. The RD robust estimator has been shown able to improve the efficiency-bias trade-off that arises in the presence of heterogeneous treatment effects. On one hand, these benefits depends upon 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 range. The strategies illustrated herein which integrate the RD design's dual nature for compelling inference should be applicable whenever tests or other "prescores"—in concert with assignment cutoffs—are used to allocate a program.

5.2 Air Quality, Infant Mortality and the Clean Air Act of 1970

Chay and Greenstone (2003) examines the effects of total suspended particles (TSPs) air pollution on infant mortality rates using the air quality improvement induced by the 1970 Clean Air Act Amendments (CAAA) in the first year that they were in force as the source for identifying the causal relationships between TSPs and infant mortality rates (IMRs). The 1970 CAAA imposed strict regulations on industrial polluters in "nonattainment" counties. Specifically, for TSPs pollution the Environmental Protection Agency (EPA) was required to designate a county as nonattainment if its TSPs concentrations exceed either of these two thresholds: (1) the annual geometric mean concentration exceeded $75\mu g/m^3$, or (2) the second highest daily concentration exceeded $260\mu g/m^3$.[32] "Ideally", a random assignment of pollution exposure across mothers or infants can ensure that pollution is independent of other confounding factors and the causal impacts of TSPs on IMRs can therefore be identified straightforwardly. In the absence of such a random assignment, Chay and Greenstone (2003) provides evidence of the impacts of TSPs on IMRs from an RD design analysis. The credibility of the findings based on the RD approach depends on the following:

First, the 1970 CAAA regulation was federally mandated, and this regulatory pressure is plausibly orthogonal to changes in IMRs except through its impacts on air pollution.[33]

Second, significant reductions in TSPs occurred immediately after 1970 CAAA. The greater reductions in nonattainment counties near the federal ceiling relative to the attainment counties narrowly below the ceiling suggest that the CAAA regulation had a separate impact on TSPs reduction.[34]

Third, near the regulatory threshold, discrete changes in TSPs and IMRs are likely due to the regulation and not competing factors. In the neighborhood of the regulatory ceiling, i.e. $75\mu g/m^3$, transitory shocks to TSPs levels in 1970 are orthogonal to unobserved shocks to infant mortality rates between 1971 and 1972. This implies that the impacts of omitted variables are homogenized between nonattainment and attainment counties within this neighborhood.[35]

The first two points verify the RD design's instrumental nature, and the third point stresses its quasi-experimental nature near the threshold. Chay and Greenstone (2003) combines these two and has made improvement upon the conventional cross-sectional and fixed effect estimates. Their paper uses nonattainment status as an instrumental variable for 1971–1972 changes in TSPs to estimate their impact on infant mortality rates changes in the first year that the 1970 CAAA was in force. Robustness checks are provided by focusing on

[32] This standard prevailed from 1971 until 1987, when the EPA shifted its focus to the regulation of finer particles.

[33] Consistent with this, Chay and Greenstone (2003) finds little association between nonattainment status and other observable variables, including parents' characteristics, prenatal care utilization and transfer payments from social programs. For additional evidence, see Figure 2 in Chay and Greenstone (2003).

[34] Chay and Greenstone (2003) find that "the entire decline in TSPs during the early 1970s occurred in nonattainment counties and that two-thirds of the 1971–1974 decline in these counties occurred between 1971 and 1972, the first year that the 1970 CAAA was in force. Consequently, this study uses nonattainment status as an instrumental variable for 1971–1972 changes in TSPs to estimate their impact on changes in infant mortality."

[35] Table 2 in Chay and Greenstone (2003) shows that the observable characteristics of nonattainment and attainment counties near the federal TSPs ceiling are very similar.

counties close to the regulatory threshold. They estimate that a 1% decline in TSPs results in a 0.5% decline in the infant mortality rate at the county level.

However, in the presence of heterogeneous effects of a continuous treatment, an ATE interpretation for the IV estimate in Chay and Greenstone (2003), which bears policy implications, is bounded by the existence of the sorting bias, which probably results from the county-level residential choice. The IV estimate only has a local ATE interpretation for "compliers" if the sorting bias exists. In Chay and Greenstone (2003), "for simplicity, it is assumed that the 'true' effect of exposure to particulates pollution is homogeneous across infants and over time." Focusing on the counties near the threshold, the IV estimate in Chay and Greenstone (2003) will regain the ATE interpretation because the unobservables and their interactions with the treatment are likely to be precluded by the "borderline experiment", but this gain is at the cost of efficiency. The IV estimates lose statistical significance for the chosen range near the threshold in Chay and Greenstone (2003). Following and on the basis of Chay and Greenstone (2003), I extend the RD design's applicability to the continuous treatment case and exploit over-identifying restrictions that are implied by its dual nature. The proposed GMM estimates demonstrate improvement upon the efficiency-bias trade-off. Besides, these over-identifying restrictions are shown to be useful in detecting non-random sorting, which occurs when the treatment effect heterogeneity is correlated with the selection variable, i.e. TSPs in 1970. The over-identifying restrictions are essentially the counterpart for the correction terms with a binary treatment, both of which can be used to detect the sorting bias that invalidates an estimate's ATE interpretation.

5.2.1 Assignment rule for 1972 county-level TSPs nonattainment status

Chay and Greenstone (2003)'s determination of the 1972 TSPs attainment/nonattainment designations is based on the dates associated with the passage and enforcement of the 1970 CAAA. On December 31, 1970, President Richard Nixon signed the 1970 CAAA, which was followed by the EPA's final publication of the National Ambient Air Quality Standards (NAAQS) four months later on April 30, 1971, that specified the national TSPs standards. On August 14, 1971, the EPA published "Requirements for Preparation, Adoption, and Submittal of Implementation Plans (SIPs)" in the Code of Federal Regulations (CFR). This set forth how states were to write their SIPs to achieve compliance with the NAAQS by 1975. Finally, the SIPs were due to the EPA in January, 1972.[36]

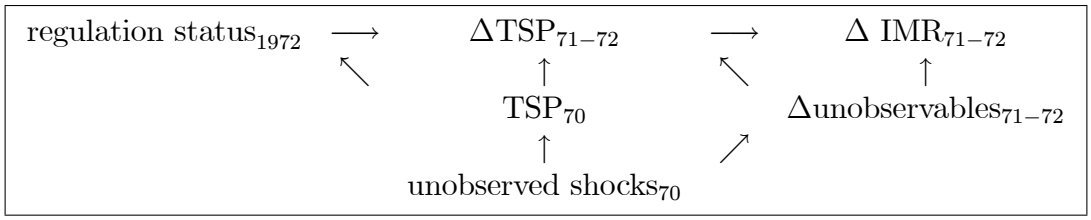
The timeline conveys important information that counties and plants were not likely to know what the EPA air quality standards would be when making decisions on emissions in 1970. This fact precludes non-random sorting, such as "avoidance", of counties near the regulatory threshold, which invalidates the RD analysis.[37]

[36] Appendix Table 1 in Chay and Greenstone (2003) summarizes these dates.

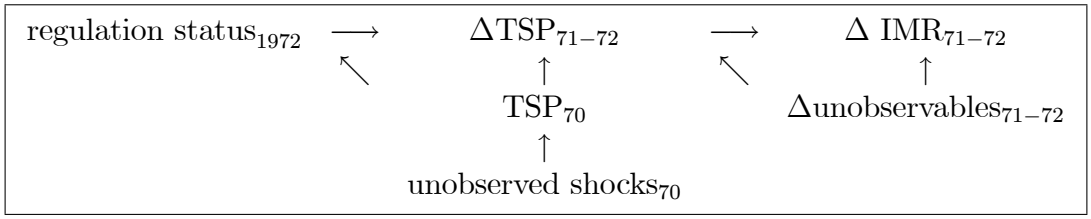
[37] More details about the assignment rule for 1972 county-level TSPs attainment/nonattainment status are provided in the Data Appendix in Chay and Greenstone (2003).

5.2.2 Identification strategy based on an RD design

Chay and Greenstone (2003)'s identification strategy based on an RD design using 1971–1972 first-differenced data can be summarized as follows ($\Delta x_{71-72} = x_{1972} - x_{1971}$).



$\Downarrow$ near regulation cutoff ("borderline experiment")



The above mechanism reveals that the inclusion of county fixed effects or the use of the first-differenced data can still be biased if Δunobservables$_{71-72}$ affect both ΔTSP$_{71-72}$ and ΔIMR$_{71-72}$.[38] Furthermore, the regulation status in 1972, which is determined by TSP$_{70}$, will not be a valid instrument for ΔTSP$_{71-72}$ is there are unobserved shocks in 1970 that affect both TSP$_{70}$ and Δunobservables$_{71-72}$ and eventually ΔTSP$_{71-72}$ and Δ IMR$_{71-72}$. The exclusion restriction on the instrumentality of the regulation status in 1972 is therefore violated. However, near the regulation cutoff, the validity of the instrumentality of 1972's regulation status can be guaranteed to a large extent. Since the regulation status in 1972 is a discontinuous function of 1970 pollution levels,[39] near the cutoff point the "borderline experiment" will preclude the impacts of the unobserved shocks in 1970 on Δunobservables$_{71-72}$. Chay and Greenstone (2003) uses the regulation status in 1972 as the instrument for ΔTSP$_{71-72}$ and conducts an IV estimation for the effect of TSPs on IMRs based on the ratio of two reduced-form effects: the impact of 1972 nonattainment status on improvements in air quality, and its association with declines in infant mortality. Their estimator is illustrated as follows:

$$\frac{\Delta\text{IMR}_{71-72} \leftarrow \text{regulation status}_{1972}}{\Delta\text{TSP}_{71-72} \leftarrow \text{regulation status}_{1972}} = \Delta\text{IMR}_{71-72} \leftarrow \Delta\text{TSP}_{71-72}$$

[38] Table 4 in Chay and Greenstone (2003) shows that the fixed effects association between TSPs and infant mortality is small and sensitive to specification, which is consistent with potential biases due to omitted variables.

[39] Footnote 13 in Chay and Greenstone (2003): "If T_{c70}^{avg} and T_{c70}^{max} are the annual geometric mean and the 2nd highest daily TSPs concentrations, respectively, then the actual regulatory instrument used is $1\{T_{c70}^{\text{avg}} > 75\mu g/m^3$ or $T_{c70}^{\text{max}} > 260\mu g/m^3\}$. Only six counties were nonattainment in 1972 for exceeding the 2nd highest daily concentration threshold, but not the annual geometric mean ceiling."

As an extension to Chay and Greenstone (2003), we herein point out that in the above mechanism TSP_{70} and its interaction with the regulation status in 1972 are additional valid instruments for ΔTSP_{71-72} because both of their inclusion and exclusion restrictions are satisfied. Thus the impact of TSPs on IMRs can be over-identified and a GMM estimator can be employed. In Table 3 we show the efficiency gain of the GMM estimates over the IV estimates in Chay and Greenstone (2003). The test, based on the Hansen J-statistic, provides checks for the validity of the GMM estimate's ATE interpretation for a predefined range through the checks for the validity of instruments for the chosen range. This specification check is also valid if there exist heterogeneous treatment effects as long as the heterogeneity is uncorrelated with the selection rule in the chosen range.

Table 3: GMM and IV Estimates for the Counties with Their 1970 Geometric Mean TSPs near the EPA's Regulation Threshold

	1970 geometric mean TSPs between					
	45 and 105 $\mu g/m^3$		60 and 90 $\mu g/m^3$		65 and 85 $\mu g/m^3$	
	(1)	(2)	(3)	(4)	(5)	(6)
A: IV estimates						
Mean TSPs	5.53	7.41	12.98	13.56*	18.05	19.68
	(8.32)	(5.83)	(8.06)	(7.08)	(15.43)	(15.38)
B: GMM estimates						
Mean TSPs	14.09*	10.11*	13.46*	10.14*	10.96*	12.82*
	(8.12)	(5.78)	(7.88)	(5.99)	(8.43)	(7.54)
p-value for J-stat. (dof=2)	0.393	0.097	0.954	0.376	0.637	0.688
Change in natality vars.	N	Y	N	Y	N	Y
1971 level in natality vars.	N	Y	N	Y	N	Y
Sample Size	316	315	171	170	116	115

Notes: The dependent variables are the 1971–1972 first-differences in the number of infant deaths due to internal causes within one-year per 100,000 live births. The columns correspond to subsamples of counties with annual geometric mean readings of TSPs in 1970 in the specified range, where nonattainment counties account for about half of the sample. Results are from (A) the IV estimation with 1971–1972 changes in mean TSPs instrumented by nonattainment status in 1972; (B) the two-step efficient GMM estimation, with 1971–1972 changes in mean TSPs instrumented by nonattainment status in 1972, 1970 geometric mean TSPs and their interaction in columns (1) through (6). Both the IV and the GMM estimator allows for heteroskedasticity and state-level clustering in calculating the standard errors. The standard errors in parentheses are robust to heteroskedasticity and clustering on the state-level. Also presented are the p-values and degrees of freedom (dof) from the Hansen J-statistic for testing the over-identifying restrictions based on the GMM criterion function evaluated at the optimal weighting matrix. Analyses are weighted by the sum of total births in 1971 and 1972; "*" indicates significance at 10% level; "**" indicates significance at 5% level; "***" indicates significance at 1% level.

For columns (1)–(6), we use the same specification to be consistent with Chay and Greenstone (2003). In Panel A of Table 3, the IV estimator is proposed by Chay and Greenstone (2003). Panel B of Table 3 uses the GMM estimator proposed by this paper. The GMM estimates show efficiency gain relative to the IV estimates in all cases, and the GMM point estimates become statistically significant. Furthermore, the GMM estimates stay robust for various specified ranges. It is confirmed that a 1 $\mu g/m^3$ TSPs reduction results in 10-14 fewer infants deaths per 100,000 births for the counties intervened by the EPA's TSP regulation. The estimates are about 11%–17% higher than the findings in Chay and Greenstone (2003). Aided by the RD design, the efficiency gains of using GMM and the resulted statistical significance of these estimates bear important implications. If we multiply the estimates by the 1.52 million births that occurred in nonattainment counties in 1972, then a 1 $\mu g/m^3$ enforced TSPs reduction can result in 150–220 fewer infants deaths for a randomly selected nonattainment county. This ATE interpretation will be checked by the p-value of the Hansen

J-statistic under over-identification. For example, column (2) suggests the invalidation of the ATE interpretation, which is possibly due to non-random sorting biases.

As we extend the RD design's applicability from a binary treatment (P-900 intervention) to a continuous treatment (different levels of TSPs), in addition to exploiting the over-identifying restrictions for efficiency gains, we can also use these over-identifying restrictions as a check for the validity of the instruments for a chosen range. As previously mentioned, the over-identifying restrictions are essentially the counterparts for the correction terms with a binary treatment, both of which can be used to detect the sorting bias that invalidates an estimate's ATE interpretation. It is worth emphasizing that in the CAAA example, such specification checks are not possible in conventional cross-sectional and fixed effects analyses in the absence of the RD design.

6 Conclusion

This paper discusses how to conduct program evaluations in virtue of the dual nature of a regression discontinuity design—both the "borderline experiment" and the instrumentality are implied by the selection rule. Focusing on the fuzzy RD design, this paper aims to identify and estimate the average treatment effect under both selection-on-observables and selection-on-unobservables. Root-N consistent and asymptotically normal estimators are derived. Both of their large and small sample properties are investigated. The proposed estimators allow for general functional forms for the selection biases. Specification tests for the plausibility of statistical assumptions are also suggested by these estimators. Empirical examples show that proposed estimators help to balance a study's internal validity with its external validity and they are easy to implement using standard software.

Previous work on the regression discontinuity (RD) design has emphasized identification and estimation of an effect at the selection threshold, which pinpoints the measurement of the size of the discontinuity. In contrast, one contribution of this paper is to investigate the trade-off between efficiency and bias in estimating the average treatment effect (and average effects local to the discontinuity) when the effects covary with the observables and the unobservables. The plausibility of the average treatment effect identification will depend on the point in the distribution where the discontinuity occurs. At a minimum, our approach allows for efficient estimation of an average effect in a range of data local to the discontinuity as well as specification tests of the assumptions necessary for the validity of the chosen range. The choice of the range local to the discontinuity should be guided by the research question or the population of policy interest before estimations take place.

References

Ai, C. (2007). An Alternative Estimation of Regression Discontinuity Models, Department of Economics, University of Florida, Gainesville, FL.

Amemiya, T. (1985). *Advanced econometrics*. Cambridge, Massachusetts, Harvard University Press.

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(434): 444-455.

Angrist, J. D. and V. Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114(2): 533-575.

Battistin, E. and E. Rettore (2007). "Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs." *Journal of Econometrics* In Press, Corrected Proof.

Bera, A. K. and Y. Bilias (2002). "The MM, ME, ML, EL, EF and GMM approaches to estimation: a synthesis." *Journal of Econometrics* 107(1-2): 51-86.

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(448): 1045-1052.

Black, D., J. Galdo and J. Smith (2005). Evaluating the Regression Discontinuity Design Using Experimental Data.

Black, S. E. (1999). "Do Better Schools Matter? Parental Valuation of Elementary Education." *The Quarterly Journal of Economics* 114(2): 577-599.

Buddelmeyer, H. and E. Skoufias (2003). An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA.

Chay, K. Y. and M. Greenstone (2003). Air Quality, Infant Mortality, and the Clean Air Act of 1970, NBER.

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(4): 1237-58.

Cobb-Clark, D.-A. and T. Crossley (2003). "Econometrics for Evaluations: An Introduction to Recent Developments." *Economic Record* 79(247): 491-511.

Cook, T. D. (2007). "'Waiting for Life to Arrive': A history of the regression-discontinuity design in Psychology, Statistics and Economics." *Journal of Econometrics* In Press, Corrected Proof.

Cook, T. D. and V. C. Wong (2007). Empirical Tests of the Validity of the Regression Discontinuity Design.

DiNardo, J. and D. S. Lee (2004). "Economic Impacts of New Unionization on Private Sector Employers: 1984–2001." *Quarterly Journal of Economics* 119(4): 1383-1441.

Garen, J. (1984). "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable." *Econometrica* 52(5): 1199-1218.

Goldberger, A. S. (1972a). Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations, University of Wisconsin, Madison, WI.

Goldberger, A. S. (1972b). Selection Bias in Evaluating Treatment Effects: The Case of Interaction, University of Wisconsin, Madison, WI.

Hahn, J. and J. A. Hausman (2003). IV Estimation with Valid and Invalid Instruments, SSRN.

Hahn, J., P. Todd, and Van der Klaauw (2001). "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1): 201-209.

Heckman, J. (1979). "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153-161.

Heckman, J. and E. Vytlacil (1998). "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling." *The Journal of Human Resources* 33(4): 974-987.

Hirano, K., G. W. Imbens and G. Ridder. (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71(4): 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(1): 17-48.

Holland, P. W. (1986). "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945-960.

Imbens, G. W. and J. D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467-475.

Imbens, G. W. and T. Lemieux (2007). "Regression discontinuity designs: A guide to practice." *Journal of Econometrics* In Press, Corrected Proof.

Lee, D. S. (2007). "Randomized experiments from non-random selection in U.S. House elections." *Journal of Econometrics* In Press, Corrected Proof.

Lee, D. S. and D. Card (2007). "Regression discontinuity inference with specification error." *Journal of Econometrics* In Press, Corrected Proof.

Lee, M.-j. (2005). *Micro-econometrics for policy, program, and treatment effects*. Oxford ; New York, Oxford University Press.

Lemieux, T. and K. Milligan (2007). "Incentive effects of social assistance: A regression discontinuity approach." *Journal of Econometrics* In Press, Corrected Proof.

Ludwig, J. and D. L. Miller (2006). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design.

McCrary, J. (2007). "Manipulation of the running variable in the regression discontinuity design: A density test." *Journal of Econometrics* In Press, Corrected Proof.

McCrary, J. and H. Royer (2003). Does Maternal Education Affect Infant Health? A Regression Discontinuity Approach Based on School Age Entry Laws.

Porter, J. (2003). Estimation in the Regression Discontinuity Model.

Robinson, P. M. (1988). "Root-N-Consistent Semiparametric Regression." *Econometrica* 56(4): 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(2): 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(1): 41-55.

Rubin, D. B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5): 688-701.

Sun, Y. (2005). Adaptive Estimation of the Regression Discontinuity Model.

Thistlethwaite, D. and D. Campbell (1960). "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology* 51(6): 309-317.

Trochim, W. M. K. (1984). *Research Design for Program Evaluation: The Regression-discontinuity Approach*. Sage Publications.

Van der Klaauw, W. (2002). "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." *International Economic Review* 43(4): 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. (2007). Instrumental Variables Estimation of the Average Treatment Effect in the Correlated Random Coefficient Model.

Wooldridge, J. M. (2007). "Inverse probability weighted estimation for general missing data problems." *Journal of Econometrics* In Press, Corrected Proof.

Wooldridge, J. M. (1997). "On two stage least squares estimation of the average treatment effect in a random coefficient model." *Economics Letters* 56(2): 129-133.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, Mass., MIT Press.

Wooldridge, J. M. (2003). "Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model." *Economics Letters* 79(2): 185-191.

A Properties of the Second Moments

The following results are frequently invoked in our analysis. Consider four random variables, x, y, z and d, where d is binary.[40]

1. $Cov(x, y) = Cov_x(x, \mathbb{E}(y|x))$

Proof. RHS is

$$\begin{aligned} Cov_x(x, \mathbb{E}(y|x)) &= \mathbb{E}[(x - \mathbb{E}(x))(\mathbb{E}(y|x) - \mathbb{E}(y))] \\ &= \mathbb{E}[\mathbb{E}(xy|x) - \mathbb{E}(x)\mathbb{E}(y|x)] = Cov(x, y) \end{aligned}$$

which equals LHS. ■

2. $Cov(d, y) = Var(d)\left[\mathbb{E}(y|d = 1) - \mathbb{E}(y|d = 0)\right]$ for binary d

Proof. LHS is

$$\begin{aligned} Cov(d, y) &= \Pr(d = 1)\mathbb{E}(y|d = 1) - \Pr(d = 1)\left[\Pr(d = 0)\mathbb{E}(y|d = 0) + \Pr(d = 1)\mathbb{E}(y|d = 1)\right] \\ &= \Pr(d = 1)\left[\mathbb{E}(y|d = 1) - \Pr(d = 0)\mathbb{E}(y|d = 0) - \Pr(d = 1)\mathbb{E}(y|d = 1)\right] \\ &= Var(d)\left[\mathbb{E}(y|d = 1) - \mathbb{E}(y|d = 0)\right] \end{aligned}$$

which equals RHS. ■

3. $Cov(x, y) = Cov_z(\mathbb{E}(x|z), \mathbb{E}(y|z)) + \mathbb{E}_z(Cov(x, y|z))$

Proof. RHS is

$$\begin{aligned} & Cov_z(\mathbb{E}(x|z), \mathbb{E}(y|z)) + \mathbb{E}_z(Cov(x, y|z)) \\ &= \mathbb{E}_z(\mathbb{E}(x|z)\mathbb{E}(y|z)) - \mathbb{E}_z(\mathbb{E}(x|z))\mathbb{E}_z(\mathbb{E}(y|z)) + \mathbb{E}_z\left(\mathbb{E}(xy|z) - \mathbb{E}(x|z)\mathbb{E}(y|z)\right) \\ &= \mathbb{E}_z(\mathbb{E}(x|z)\mathbb{E}(y|z)) - \mathbb{E}(x)\mathbb{E}(y) + \mathbb{E}(xy) - \mathbb{E}_z\left(\mathbb{E}(x|z)\mathbb{E}(y|z)\right) \\ &= \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y) \end{aligned}$$

which equals LHS. ■

4. $\mathbb{E}\left[(y - \mathbb{E}(y|x))y\right] = \mathbb{E}\left[Var(y|x)\right] = \mathbb{E}\left[(y - \mathbb{E}(y|x))^2\right]$

Proof. LHS is

$$\begin{aligned} \mathbb{E}\left[(y - \mathbb{E}(y|x))y\right] &= \mathbb{E}\{\mathbb{E}\left[(y - \mathbb{E}(y|x))y\right]|x\} = \mathbb{E}[\mathbb{E}(y^2|x) - \mathbb{E}^2(y|x)] = \mathbb{E}\left[Var(y|x)\right] \\ &= \mathbb{E}\{\mathbb{E}\left[(y - \mathbb{E}(y|x))^2|x\right]\} = \mathbb{E}\left[(y - \mathbb{E}(y|x))^2\right] \end{aligned}$$

which equals RHS. ■

5. $\mathbb{E}\left[(x - \mathbb{E}(x|z))(y - \mathbb{E}(y|z))\right] = \mathbb{E}\left[(x - \mathbb{E}(x|z))y\right] = \mathbb{E}\left[x(y - \mathbb{E}(y|z))\right]$

Proof. LHS is

$$\begin{aligned} \mathbb{E}\left[(x - \mathbb{E}(x|z))(y - \mathbb{E}(y|z))\right] &= \mathbb{E}[(x - \mathbb{E}(x|z))y] - \mathbb{E}\{\mathbb{E}\left[(x - \mathbb{E}(x|z))\mathbb{E}(y|z)\right]|z\} \\ &= \mathbb{E}[(x - \mathbb{E}(x|z))y] - \mathbb{E}\{[\mathbb{E}(x|z) - \mathbb{E}(x|z)]\mathbb{E}(y|z)\} \\ &= \mathbb{E}[(x - \mathbb{E}(x|z))y] \\ \mathbb{E}\left[(x - \mathbb{E}(x|z))(y - \mathbb{E}(y|z))\right] &= \mathbb{E}[x(y - \mathbb{E}(y|z))] - \mathbb{E}\{\mathbb{E}\left[\mathbb{E}(x|z)(y - \mathbb{E}(y|z))\right]|z\} \\ &= \mathbb{E}[x(y - \mathbb{E}(y|z))] - \mathbb{E}\{\mathbb{E}(x|z)[\mathbb{E}(y|z) - \mathbb{E}(y|z)]\} \\ &= \mathbb{E}\left[x(y - \mathbb{E}(y|z))\right] \end{aligned}$$

which equals RHS. ■

[40] LHS and RHS refer to "left-hand side" and "right-hand side", respectively.

B Identification of Treatment Effects

B.1 Proof of Theorem 1

Under Assumption 1 and Assumption 3, observed outcomes can written as

$$\begin{aligned} y &= y_0 + (y_1 - y_0)d \\ &= g_0(z^*) + (\eta + \lambda(z^*))d + e, \text{ where } e \equiv u_0 + d(u_1 - u_0) \end{aligned}$$

ATE is defined as $\mathbb{E}(\eta + \lambda(z^*))$ here. So the observed outcomes can be rewritten as

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

Under Definition 1 and Assumption 3, we have

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

Therefore, we have

$$\mathbb{E}(y|z^*) = g_0(z^*) + \mathbb{E}(\eta + \lambda(z^*))\mathbb{E}(d|z^*) + (\lambda(z^*) - \mathbb{E}(\lambda(z^*))\mathbb{E}(d|z^*)$$

We treat $g_0(z^*)$ as a nuisance parameter, which is to be differenced out.

$$y - \mathbb{E}(y|z^*) = \mathbb{E}(\eta + \lambda(z^*))(d - \mathbb{E}(d|z^*)) + (\lambda(z^*) - \mathbb{E}(\lambda(z^*))(d - \mathbb{E}(d|z^*)) + \widetilde{e}$$

We next verify two moment equations:

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

Under Definition 1 and Assumption 3, we have

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

Under Assumption 4, we have the following moment equations:

$$\begin{aligned} 0 &= \mathbb{E}[(d - \mathbb{E}(d|z^*))\widetilde{e}] \\ 0 &= \mathbb{E}[(d - \mathbb{E}(d|z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))'\widetilde{e}] \\ \widetilde{e} &= y - \mathbb{E}(y|z^*) - \mathbb{E}(\eta + \lambda(z^*))(d - \mathbb{E}(d|z^*)) - (d - \mathbb{E}(d|z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))'\gamma \end{aligned}$$

This implies:

$$\begin{aligned} 0 &= \mathbb{E}\{(d - \mathbb{E}(d|z^*))[y - \mathbb{E}(y|z^*) - \mathbb{E}(\eta + \lambda(z^*))(d - \mathbb{E}(d|z^*)) - (d - \mathbb{E}(d|z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))'\gamma]\} \\ 0 &= \mathbb{E}\{(d - \mathbb{E}(d|z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))'[y - \mathbb{E}(y|z^*) - \mathbb{E}(\eta + \lambda(z^*))(d - \mathbb{E}(d|z^*)) - (d - \mathbb{E}(d|z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))'\gamma]\} \end{aligned}$$

To simplify notations, we define the following:

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

and, we have:

$$\widetilde{y} = \mathbb{E}(\eta + \lambda(z^*))x_1 + \mathbf{x}_2'\gamma + \widetilde{e}, \text{ and } \mathbb{E}(\widetilde{e}|x_1, \mathbf{x}_2) = 0$$

This gives:

$$\begin{bmatrix} \mathbb{E}(x_1^2) & \mathbb{E}(x_1\mathbf{x}_2') \\ \mathbb{E}(x_1\mathbf{x}_2) & \mathbb{E}(\mathbf{x}_2\mathbf{x}_2') \end{bmatrix} \cdot \begin{bmatrix} \mathbb{E}(\eta + \lambda(z^*)) \\ \gamma \end{bmatrix} = \begin{bmatrix} \mathbb{E}(x_1\widetilde{y}) \\ \mathbb{E}(\mathbf{x}_2\widetilde{y}) \end{bmatrix}$$

where

$$\begin{aligned}
\mathbb{E}(x_1^2) &= \mathbb{E}[\,(d - \mathbb{E}(d|z^*))^2] = \mathbb{E}(Var(d|z^*)) \\
\mathbb{E}(x_1\mathbf{x}_2) &= \mathbb{E}[\,(d - \mathbb{E}(d|z^*))^2 (\mathbf{w} - \mathbb{E}(\mathbf{w}))] = \mathbb{E}[Var(d|z^*)(\mathbf{w} - \mathbb{E}(\mathbf{w}))] \\
\mathbb{E}(\mathbf{x}_2\mathbf{x}_2') &= \mathbb{E}[\,(d - \mathbb{E}(d|z^*))^2 (\mathbf{w} - \mathbb{E}(\mathbf{w}))(\mathbf{w} - \mathbb{E}(\mathbf{w}))'] = \mathbb{E}[Var(d|z^*)(\mathbf{w} - \mathbb{E}(\mathbf{w}))(\mathbf{w} - \mathbb{E}(\mathbf{w}))'] \\
\mathbb{E}(x_1\widetilde{y}) &= \mathbb{E}\left[(d - \mathbb{E}(d|z^*))(y - \mathbb{E}(y|z^*))\right] = \mathbb{E}\left[(d - \mathbb{E}(d|z^*))\, y\right] \\
&= \mathbb{E}(x_1 y) \\
\mathbb{E}(\mathbf{x}_2\widetilde{y}) &= \mathbb{E}\left[(d - \mathbb{E}(d|z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))(y - \mathbb{E}(y|z^*))\right] = \mathbb{E}\left[(d - \mathbb{E}(d|z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))y\right] \\
&= \mathbb{E}(x_2 y)
\end{aligned}$$

Solve for $\mathbb{E}(\eta + \lambda(z^*))$ and γ:

$$\begin{aligned}
\begin{bmatrix} \mathbb{E}(\eta + \lambda(z^*)) \\ \gamma \end{bmatrix} &= \begin{bmatrix} \mathbb{E}(x_1^2) & \mathbb{E}(x_1\mathbf{x}_2') \\ \mathbb{E}(x_1\mathbf{x}_2) & \mathbb{E}(\mathbf{x}_2\mathbf{x}_2') \end{bmatrix}^{-1} \cdot \begin{bmatrix} \mathbb{E}(x_1 y) \\ \mathbb{E}(\mathbf{x}_2 y) \end{bmatrix} \\
&= \mathbb{E}^{-1}\left[\begin{pmatrix} x_1 \\ \mathbf{x}_2 \end{pmatrix}\begin{pmatrix} x_1 & \mathbf{x}_2' \end{pmatrix}\right]\mathbb{E}\left[\begin{pmatrix} x_1 \\ \mathbf{x}_2 \end{pmatrix} y\right]
\end{aligned}$$

Define $\theta \equiv (\mathbb{E}(\eta + \lambda(z^*)), \gamma')'$ and $\mathbf{x} \equiv (x_1, \mathbf{x}_2')'$, and we have the following "least squares" estimator:

$$\begin{aligned}
\theta &= \mathbb{E}^{-1}(\mathbf{x}\mathbf{x}')\,\mathbb{E}(\mathbf{x}y) \\
\mathbf{x} &= (d - \mathbb{E}(d|z^*), (d - \mathbb{E}(d|z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))')'
\end{aligned}$$

We next solve for $\mathbb{E}(\eta + \lambda(z^*))$ and γ individually. Using results from Amemiya (1985, page 460), we have

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

where

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

Thus,

$$\begin{aligned}
&\mathbb{E}(\eta + \lambda(z^*)) \\
&= E^{-1}\left[\mathbb{E}(x_1 y) - BD^{-1}\mathbb{E}(\mathbf{x}_2 y)\right] \\
&= \left[\mathbb{E}(x_1^2) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)\right]^{-1} \cdot \left[\mathbb{E}(x_1 y) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(\mathbf{x}_2 y)\right] \\
&= \frac{\mathbb{E}(x_1 y) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(\mathbf{x}_2 y)}{\mathbb{E}(x_1^2) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)}
\end{aligned}$$

Similarly,

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

Therefore,

$$\begin{aligned}
\text{ATE} \equiv \mathbb{E}(\eta + \lambda(z^*)) &= \frac{\mathbb{E}(x_1 y) - \mathbb{E}(x_1 \mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(\mathbf{x}_2 y)}{\mathbb{E}(x_1^2) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)} \\
\gamma &= \left[\mathbb{E}(\mathbf{x}_2\mathbf{x}_2') - \mathbb{E}(x_1\mathbf{x}_2)\mathbb{E}^{-1}(x_1^2)\mathbb{E}(x_1\mathbf{x}_2')\right]^{-1}\mathbb{E}(\mathbf{x}_2 y) \\
&\quad - \frac{\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)\mathbb{E}(x_1 y)}{\mathbb{E}(x_1^2) - \mathbb{E}(x_1\mathbf{x}_2')\mathbb{E}^{-1}(\mathbf{x}_2\mathbf{x}_2')\mathbb{E}(x_1\mathbf{x}_2)} \\
\mathbb{E}(x_1^2) &= \mathbb{E}(Var(d|z^*)) \\
\mathbb{E}(x_1\mathbf{x}_2) &= \mathbb{E}[Var(d|z^*)\,(\mathbf{w} - \mathbb{E}(\mathbf{w}))] \\
\mathbb{E}(\mathbf{x}_2\mathbf{x}_2') &= \mathbb{E}[Var(d|z^*)\,(\mathbf{w} - \mathbb{E}(\mathbf{w}))\,(\mathbf{w} - \mathbb{E}(\mathbf{w}))'] \\
\mathbb{E}(x_1 y) &= \mathbb{E}\left[(d - \mathbb{E}(d|z^*))\,y\right] \\
\mathbb{E}(\mathbf{x}_2 y) &= \mathbb{E}\left[(d - \mathbb{E}(d|z^*))(\mathbf{w} - \mathbb{E}(\mathbf{w}))y\right]
\end{aligned}$$

B.2 Proof of Corollary 1

Proof of Part (1). According to Appendix B.1,

$$\begin{aligned}
\lim_{z^*\to 0}\begin{bmatrix} \mathbb{E}(x_1 y) \\ \mathbb{E}(\mathbf{x}_2 y)\end{bmatrix} &= \lim_{z^*\to 0}\begin{bmatrix} \mathbb{E}(x_1^2) & \mathbb{E}(x_1\mathbf{x}_2') \\ \mathbb{E}(x_1\mathbf{x}_2) & \mathbb{E}(\mathbf{x}_2\mathbf{x}_2')\end{bmatrix}\cdot \lim_{z^*\to 0}\begin{bmatrix} \mathbb{E}(\eta + \lambda(z^*)) \\ \gamma\end{bmatrix} \\
\begin{bmatrix} \lim_{z^*\to 0}\mathbb{E}(x_1 y) \\ \mathbf{0}\end{bmatrix} &= \begin{bmatrix} \lim_{z^*\to 0}\mathbb{E}(x_1^2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\cdot\begin{bmatrix} \lim_{z^*\to 0}\mathbb{E}(\eta + \lambda(z^*)) \\ \gamma\end{bmatrix} \\
\lim_{z^*\to 0}\mathbb{E}(\eta + \lambda(z^*)) &= \frac{\lim_{z^*\to 0}\mathbb{E}\left[(d - \mathbb{E}(d|z^*))\,y\right]}{\lim_{z^*\to 0}\mathbb{E}(Var(d|z^*))}
\end{aligned}$$

Under Assumption 3,

$$\begin{aligned}
\lim_{z^*\to 0}\mathbb{E}(\eta + \lambda(z^*)) = \lim_{z^*\to 0}\lambda(z^*) &= \frac{\lim_{z^*\to 0}\mathbb{E}\left[(d - \mathbb{E}(d|z^*))\,y|z^*\right]}{\lim_{z^*\to 0}\mathbb{E}[(d - \mathbb{E}(d|z^*))d|z^*]} \\
&= \frac{\lim_{z^*\to 0}\mathbb{E}(yd|z^*) - \mathbb{E}(y|z^*)\mathbb{E}(d|z^*)}{\lim_{z^*\to 0}\mathbb{E}(d^2|z^*) - \mathbb{E}^2(d|z^*)} = \lim_{z^*\to 0}\frac{Cov(d, y|z^*)}{Var(d|z^*)} \\
&= \lim_{z^*\to 0}\mathbb{E}(y|d=1, z^*) - \mathbb{E}(y|d=0, z^*) \quad \text{(see Appendix A)} \\
&= \lim_{z^*\to 0}\mathbb{E}(y_1|d=1, z^*) - \mathbb{E}(y_0|d=0, z^*) \\
&= \lim_{z^*\to 0}\mathbb{E}(y_1|z^*) - \mathbb{E}(y_0|z^*) \text{ (selection-on-observables)} \\
&= \text{ATE}(0) \\
&= \lim_{z^*\to 0}\mathbb{E}(y_1 - y_0|z^*) = \lim_{z^*\downarrow 0}\mathbb{E}(y_1 - y_0|z^*) = \lim_{z^*\uparrow 0}\mathbb{E}(y_1 - y_0|z^*)
\end{aligned}$$

Recall that $y = y_0 + d(y_1 - y_0)$, and we have:

$$\begin{aligned}
\mathbb{E}(y|z^*) &= \mathbb{E}(y_0|z^*) + \mathbb{E}(d|z^*)\mathbb{E}(y_1 - y_0|z^*) \text{ (selection-on-observables)} \\
\lim_{z^*\downarrow 0}\mathbb{E}(y|z^*) &= \lim_{z^*\downarrow 0}\mathbb{E}(y_0|z^*) + \lim_{z^*\downarrow 0}\mathbb{E}(d|z^*)\lim_{z^*\downarrow 0}\mathbb{E}(y_1 - y_0|z^*) \\
\lim_{z^*\uparrow 0}\mathbb{E}(y|z^*) &= \lim_{z^*\uparrow 0}\mathbb{E}(y_0|z^*) + \lim_{z^*\uparrow 0}\mathbb{E}(d|z^*)\lim_{z^*\uparrow 0}\mathbb{E}(y_1 - y_0|z^*)
\end{aligned}$$

Given the continuity of $\mathbb{E}(y_0|z^*)$, discontinuity of $\mathbb{E}(d|z^*)$ at $z^* = 0$ in Assumption 1 and $\lim_{z^* \to 0} \mathbb{E}(y_1 - y_0|z^*) = \lim_{z^* \downarrow 0} \mathbb{E}(y_1 - y_0|z^*) = \lim_{z^* \uparrow 0} \mathbb{E}(y_1 - y_0|z^*)$ we have

$$\begin{aligned}\text{ATE}(0) = \lim_{z^* \to 0} \mathbb{E}(y_1 - y_0|z^*) &= \frac{\lim_{z^* \downarrow 0} \mathbb{E}(y|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(y|z^*)}{\lim_{z^* \downarrow 0} \mathbb{E}(d|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(d|z^*)} \\ &= \lim_{z^* \to 0} \mathbb{E}(\eta + \lambda(z^*)) \text{ (given by Theorem 1)}\end{aligned}$$

Therefore, with a fuzzy RD design, the identification result in Theorem 1 accommodates the conventional nonparametric result at the cutoff point at the limit $z^* \to 0$. ■

Proof of Part (2). First, recall that under a fuzzy RD design, we have:

$$\begin{aligned}\text{ATE} &= \mathbb{E}(\eta + \lambda(z^*)) \\ &= \frac{\mathbb{E}[y(d - \mathbb{E}(d|z^*))] - \mathbb{E}[(\lambda(z^*) - \mathbb{E}(\lambda(z^*)))Var(d|z^*)]}{\mathbb{E}[d(d - \mathbb{E}(d|z^*))]}\end{aligned}$$

If we replace Assumption 3 and Assumption 4 with Assumption 2, then $\lambda(z^*) - \mathbb{E}(\lambda(z^*)) = 0$. We therefore have:

$$\text{ATE} = \frac{\mathbb{E}[y(d - \mathbb{E}(d|z^*))]}{\mathbb{E}[d(d - \mathbb{E}(d|z^*))]}$$

Using results from Appendix A and under selection-on-observables, we have

$$\begin{aligned}\frac{\mathbb{E}[y(d - \mathbb{E}(d|z^*))]}{\mathbb{E}[d(d - \mathbb{E}(d|z^*))]} &= \frac{\mathbb{E}\left(Cov(d, y|z^*)\right)}{\mathbb{E}\left(Var(d|z^*)\right)} \\ &= \frac{\mathbb{E}\left[Var(d|z^*)\left(\mathbb{E}(y|d = 1, z^*) - \mathbb{E}(y|d = 0, z^*)\right)\right]}{\mathbb{E}\left(Var(d|z^*)\right)} \\ &= \frac{\mathbb{E}\left[Var(d|z^*)\left(\mathbb{E}(y_1|d = 1, z^*) - \mathbb{E}(y_0|d = 0, z^*)\right)\right]}{\mathbb{E}\left(Var(d|z^*)\right)} \\ &= \frac{\mathbb{E}\left[Var(d|z^*)\mathbb{E}(y_1 - y_0|z^*)\right]}{\mathbb{E}\left(Var(d|z^*)\right)} \text{ (selection-on-observables)} \\ &= \frac{\mathbb{E}\left[Var(d|z^*)\mathbb{E}(y_1 - y_0)\right]}{\mathbb{E}\left(Var(d|z^*)\right)} \text{ (Assumption 2)} \\ &= \mathbb{E}(y_1 - y_0)\end{aligned}$$

Recall that $y = y_0 + d(y_1 - y_0)$, and we have:

$$\begin{aligned}\mathbb{E}(y|z^*) &= \mathbb{E}(y_0|z^*) + \mathbb{E}(d|z^*)\mathbb{E}(y_1 - y_0|z^*) \text{ (selection-on-observables)} \\ &= \mathbb{E}(y_0|z^*) + \mathbb{E}(d|z^*)\mathbb{E}(y_1 - y_0) \text{ (Assumption 2)} \\ \lim_{z^* \downarrow 0} \mathbb{E}(y|z^*) &= \lim_{z^* \downarrow 0} \mathbb{E}(y_0|z^*) + \lim_{z^* \downarrow 0} \mathbb{E}(d|z^*)\mathbb{E}(y_1 - y_0) \\ \lim_{z^* \uparrow 0} \mathbb{E}(y|z^*) &= \lim_{z^* \uparrow 0} \mathbb{E}(y_0|z^*) + \lim_{z^* \uparrow 0} \mathbb{E}(d|z^*)\mathbb{E}(y_1 - y_0)\end{aligned}$$

Given the continuity of $\mathbb{E}(y_0|z^*)$ and discontinuity of $\mathbb{E}(d|z^*)$ at $z^* = 0$ in Assumption 1, we have

$$\mathbb{E}(y_1 - y_0) = \frac{\lim_{z^* \downarrow 0} \mathbb{E}(y|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(y|z^*)}{\lim_{z^* \downarrow 0} \mathbb{E}(d|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(d|z^*)}$$

Under a sharp RD design:

$$\lim_{z^* \downarrow 0} \mathbb{E}(d|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(d|z^*) = 1 \Rightarrow \mathbb{E}(y_1 - y_0) = \lim_{z^* \downarrow 0} \mathbb{E}(y|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(y|z^*)$$

Therefore, we have the result for a fuzzy RD design:

$$\text{ATE} = \frac{\mathbb{E}[y(d - \mathbb{E}(d|z^*))]}{\mathbb{E}[d(d - \mathbb{E}(d|z^*))]} = \frac{\lim_{z^* \downarrow 0} \mathbb{E}(y|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(y|z^*)}{\lim_{z^* \downarrow 0} \mathbb{E}(d|z^*) - \lim_{z^* \uparrow 0} \mathbb{E}(d|z^*)}$$

Therefore, with a fuzzy RD design, the identification result in Theorem 1 accommodates the conventional nonparametric result with a constant treatment effect at the cutoff point at the limit $z^* \to 0$. ■

B.3 Proof of Theorem 2

Consider the observed outcome (y-equation):

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

where

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

Compute

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

and

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

Thus

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

Now we have a new y-equation with two correction functions $\phi(\cdot)$ and $\Phi(\cdot) - (\cdot)\phi(\cdot)$ for selectivity bias:

$$\begin{aligned} y &= g_0(z^*) + \mathbb{E}(\eta + \lambda(z^*))d + d(\lambda(z^*) - \mathbb{E}(\lambda(z^*)) + \widetilde{e} \\ &\quad + \xi_1 \phi(\pi_0 + \pi_1 z + \pi_2 z^*) \\ &\quad + \xi_2 \left[\Phi(\pi_0 + \pi_1 z + \pi_2 z^*) - (\pi_0 + \pi_1 z + \pi_2 z^*)\phi(\pi_0 + \pi_1 z + \pi_2 z^*)\right] \\ \widetilde{e} &\equiv u_0 + d(u_1 - u_0) - \mathbb{E}(d(u_1 - u_0)|z, z^*) + d(\eta - \mathbb{E}(\eta)) \\ &\quad \Rightarrow \mathbb{E}(\widetilde{e}|z, z^*) = \mathbb{E}(u_0 + d(u_1 - u_0) - \mathbb{E}(d(u_1 - u_0)|z, z^*) + \mathbb{E}(d(\eta - \mathbb{E}(\eta))|z, z^*) = 0 \end{aligned}$$

However,

$$\begin{aligned} \mathbb{E}(\widetilde{e}|d, z, z^*) &= \mathbb{E}(u_0 + d(u_1 - u_0) - \mathbb{E}(d(u_1 - u_0)|z, z^*) + d(\eta - \mathbb{E}(\eta))|d, z, z^*) \\ &= \mathbb{E}(u_0|d = 0, z, z^*) + \mathbb{E}(u_1|d = 1, z, z^*) \\ &\quad - \xi_1 \phi(\pi_0 + \pi_1 z + \pi_2 z^*) \\ &\quad - \xi_2 \left[\Phi(\pi_0 + \pi_1 z + \pi_2 z^*) - (\pi_0 + \pi_1 z + \pi_2 z^*)\phi(\pi_0 + \pi_1 z + \pi_2 z^*)\right] \\ &\neq 0 \end{aligned}$$

Therefore, with two correction functions $\phi(\cdot)$ and $\Phi(\cdot)-(\cdot)\phi(\cdot)$ added back in, we still need to use z as an instrument for d, and ATE$\equiv\ \mathbb{E}(\eta+\lambda(z^*))$ is just-identified by an IV estimator using $\mathbb{E}(d|z,z^*)$ and $\mathbb{E}(d|z,z^*)\left[\lambda(z^*)-\mathbb{E}(\lambda(z^*))\right]$ as the instruments for d and $d\left[\lambda(z^*)-\mathbb{E}(\lambda(z^*))\right]$ respectively.

B.4 Control Functions Approach

Theorem 6 *Under Assumption 1, Assumption 3, Definition 2, and the following additional assumption:*

$$(A3)\ \begin{pmatrix} u_0 \\ u_1 \\ v \end{pmatrix} \amalg z^* \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{0v} \\ \sigma_{01} & \sigma_1^2 & \sigma_{1v} \\ \sigma_{0v} & \sigma_{1v} & 1 \end{pmatrix}\right)$$

the observed outcome can be rewritten as:

$$\begin{aligned} y &= g_0(z^*)+\mathbb{E}(\eta+\lambda(z^*))d+d(\lambda(z^*)-\mathbb{E}(\lambda(z^*)))+\widetilde{\widetilde{e}} \\ &\quad+\sigma_{1v}d\frac{\phi(\pi_0+\pi_1 z+\pi_2 z^*)}{\Phi(\pi_0+\pi_1 z+\pi_2 z^*)}-\sigma_{0v}(1-d)\frac{\phi(-\pi_0-\pi_1 z-\pi_2 z^*)}{\Phi(-\pi_0-\pi_1 z-\pi_2 z^*)} \\ \widetilde{\widetilde{e}} &\equiv u_0+d(u_1-u_0)-\mathbb{E}(u_0+d(u_1-u_0)|d,z,z^*)+d(\eta-\mathbb{E}(\eta)),\mathbb{E}(\widetilde{\widetilde{e}}|d,z,z^*)=0 \end{aligned}$$

With two control functions added back in,

$$\sigma_{1v}\frac{\phi(\pi_0+\pi_1 z+\pi_2 z^*)}{\Phi(\pi_0+\pi_1 z+\pi_2 z^*)}\ \textit{and}\ -\sigma_{0v}\frac{\phi(-\pi_0-\pi_1 z-\pi_2 z^*)}{\Phi(-\pi_0-\pi_1 z-\pi_2 z^*)}$$

ATE$\equiv\ \mathbb{E}(\eta+\lambda(z^*))$ can be identified under selection-on-observables due to $\mathbb{E}(\widetilde{\widetilde{e}}|d,z,z^*)=0$.

Proof. Consider the observed outcome (y-equation):

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

Compute $\mathbb{E}(u_0+d(u_1-u_0)|d,z,z^*)$ under (A3):

$$\begin{aligned} &\mathbb{E}(u_1|d=1,z,z^*)=\mathbb{E}(u_1|v>-\pi_0-\pi_1 z-\pi_2 z^*) \\ &=\sigma_{1v}\mathbb{E}(v|v>-\pi_0-\pi_1 z-\pi_2 z^*) \\ &=\sigma_{1v}\tfrac{\phi(-\pi_0-\pi_1 z-\pi_2 z^*)}{1-\Phi(-\pi_0-\pi_1 z-\pi_2 z^*)}=\sigma_{1v}\tfrac{\phi(\pi_0+\pi_1 z+\pi_2 z^*)}{\Phi(\pi_0+\pi_1 z+\pi_2 z^*)} \\ &\mathbb{E}(u_0|d=0,z,z^*)=\mathbb{E}(u_0|v\leqslant-\pi_0-\pi_1 z-\pi_2 z^*) \\ &=\sigma_{0v}\mathbb{E}(v|v\leqslant-\pi_0-\pi_1 z-\pi_2 z^*) \\ &=-\sigma_{0v}\tfrac{\phi(-\pi_0-\pi_1 z-\pi_2 z^*)}{\Phi(-\pi_0-\pi_1 z-\pi_2 z^*)}=-\sigma_{0v}\tfrac{\phi(-\pi_0-\pi_1 z-\pi_2 z^*)}{\Phi(-\pi_0-\pi_1 z-\pi_2 z^*)} \\ &\mathbb{E}(u_0+d(u_1-u_0)|d,z,z^*)=\sigma_{1v}d\tfrac{\phi(\pi_0+\pi_1 z+\pi_2 z^*)}{\Phi(\pi_0+\pi_1 z+\pi_2 z^*)}-\sigma_{0v}(1-d)\tfrac{\phi(-\pi_0-\pi_1 z-\pi_2 z^*)}{\Phi(-\pi_0-\pi_1 z-\pi_2 z^*)} \end{aligned}$$

Now we have a new y-equation with two control functions for both omitted variable bias and selectivity bias:

$$\begin{aligned} y &= g_0(z^*)+\mathbb{E}(\eta+\lambda(z^*))d+d(\lambda(z^*)-\mathbb{E}(\lambda(z^*)))+\widetilde{\widetilde{e}} \\ &\quad+\sigma_{1v}d\frac{\phi(\pi_0+\pi_1 z+\pi_2 z^*)}{\Phi(\pi_0+\pi_1 z+\pi_2 z^*)}-\sigma_{0v}(1-d)\frac{\phi(-\pi_0-\pi_1 z-\pi_2 z^*)}{\Phi(-\pi_0-\pi_1 z-\pi_2 z^*)} \\ \widetilde{\widetilde{e}} &= u_0+d(u_1-u_0)-\mathbb{E}(u_0+d(u_1-u_0)|d,z,z^*)+d(\eta-\mathbb{E}(\eta)),\mathbb{E}(\widetilde{\widetilde{e}}|d,z,z^*)=0 \end{aligned}$$

With two control functions added back in, $\mathbb{E}(\eta+\lambda(z^*))$ can be identified under the case of selection-on-observables ($\mathbb{E}(\widetilde{\widetilde{e}}|d,z,z^*)=0$). ■

C Estimation of Treatment Effects

C.1 Asymptotics for the RD robust Estimator

Based on Theorem 1, a two-stage estimator dealing with both ATE and the explicit part of treatment effect heterogeneity under a partially linear model is given by

$$\widehat{\theta}_{\mathrm{RD\_robust}}=\left(\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\left(\sum_{i=1}^{N}\widehat{\mathbf{x}}_i y_i\right)$$

where

$$
\begin{aligned}
y_i &= g_0(z_i^*) + \alpha d_i + d(\mathbf{w}_i - \mu)'\gamma + \widetilde{e}_i \\
\widetilde{e} &\equiv u_0 + d(u_1 - u_0) + d(\eta - \mathbb{E}(\eta)),\ \mathbb{E}(\widetilde{e}|z^*) = \mathbb{E}(\widetilde{e}|d, z^*) = 0 \\
\widehat{\mathbf{x}}_i &= \left[\left(d_i - p(z_i^*; \widehat{\lambda})\right), \left(d_i - p(z_i^*; \widehat{\lambda})\right)(\mathbf{w}_i - \widehat{\mu})'\right]', \text{ where } \mathbf{w}_i \text{ is a vector including polynomials of } \mathbf{z}_i^* \\
\widehat{\mu} &\equiv \widehat{\mathbb{E}}(\mathbf{w}_i), p(z_i^*; \widehat{\lambda}) \equiv \widehat{\mathbb{E}}(d_i|z_i^*) \\
\theta &\equiv (\alpha, \gamma')' \\
\alpha &\equiv \text{ATE} \equiv \mathbb{E}(\eta + \lambda(z^*)), \lambda(z^*) \equiv \mathbf{w}'\gamma
\end{aligned}
$$

In the first stage, we assume that a consistent estimator for $\mathbb{E}(d|z^*)$ can be obtained parametrically, that is, $\mathbb{E}(d|z^*)$ is known up to a finite dimension. Next, define the following

$$
\begin{aligned}
\mathbf{x} &\equiv \left[(d - p(z^*; \lambda)), (d - p(z^*; \lambda))(\mathbf{w} - \mu)'\right]' \equiv \mathbf{f}(d, z^*, \mathbf{w}; \lambda, \mu) \\
\widehat{\mathbf{x}} &\equiv \left[\left(d - p(z^*; \widehat{\lambda})\right), \left(d - p(z^*; \widehat{\lambda})\right)(\mathbf{w} - \widehat{\mu})'\right]' \equiv \mathbf{f}(d, z^*, \mathbf{w}; \widehat{\lambda}, \widehat{\mu}) \\
h(z^*) &\equiv \mathbb{E}(y|z^*) = g_0(z^*) + \alpha p(z^*; \lambda) + p(z^*; \lambda)(\mathbf{w} - \mu)'\gamma
\end{aligned}
$$

The model defined at the population is:

$$
y = \mathbf{x}'\theta + h(z^*) + \widetilde{e}, \mathbb{E}(\widetilde{e}|z^*) = \mathbb{E}(\widetilde{e}|d, z^*) = 0
$$

The model in the second stage is:

$$
\begin{aligned}
y &= \widehat{\mathbf{x}}'\theta + h(z^*) + \text{error} \\
&= \widehat{\mathbf{x}}'\theta + (\mathbf{x} - \widehat{\mathbf{x}})'\theta + h(z^*) + \widetilde{e}
\end{aligned}
$$

and the partially linear estimator is:

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

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

Proof (Consistency). Given the consistency $\widehat{\mu} \xrightarrow{p} \mu$ and $\widehat{\lambda} \xrightarrow{p} \lambda$, by Slutsky theorem, $p(z^*; \widehat{\lambda}) \xrightarrow{p} p(z^*; \lambda)$ and $\widehat{\mathbf{x}} \xrightarrow{p} \mathbf{x}$. Therefore,

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

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

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

Consistency is established straightforwardly. ■

Proof (Asymptotic Normality). Recall that

$$\begin{aligned}
\widehat{\theta}_{\mathrm{RD_robust}} &= \left(\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\left(\sum_{i=1}^{N}\widehat{\mathbf{x}}_i y_i\right) \\
&= \left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i[\widehat{\mathbf{x}}_i'\theta+(\mathbf{x}_i-\widehat{\mathbf{x}}_i)'\theta+h(z_i^*)+\widetilde{e}_i]\right) \\
&= \theta+\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i[(\mathbf{x}_i-\widehat{\mathbf{x}}_i)'\theta+y_i-\mathbf{x}_i'\theta]\right) \\
&\Rightarrow \sqrt{N}\left(\widehat{\theta}_{\mathrm{RD_robust}}-\theta\right) = \left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i[(\mathbf{x}_i-\widehat{\mathbf{x}}_i)'\theta+y_i-\mathbf{x}_i'\theta]\right)
\end{aligned}$$

Given the consistency $\widehat{\mu}\xrightarrow{p}\mu$ and $\widehat{\lambda}\xrightarrow{p}\lambda$, by Slutsky theorem, $p(z^*;\widehat{\lambda})\xrightarrow{p}p(z^*;\lambda)$ and $\widehat{\mathbf{x}}\xrightarrow{p}\mathbf{x}$. Therefore,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i'\xrightarrow{p}\mathbb{E}(\mathbf{x}\mathbf{x}')\equiv A_0 \\
&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\left(y_i-\mathbf{x}_i'\theta\right)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x}_i\left(y_i-\mathbf{x}_i'\theta\right)+o_p(1)\quad\text{(under selection-on-observables)}
\end{aligned}$$

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

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

where

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

with the following definition

$$B_0\equiv\mathbb{E}\left(\left(\theta\otimes\mathbf{x}_i'\right)'\frac{\partial\mathbf{x}}{\partial\lambda'}\right)=\mathbb{E}\left(\left(\theta\otimes\widehat{\mathbf{x}}_i'\right)'\frac{\partial\mathbf{f}(d,z^*,\mathbf{w};\lambda,\mu)}{\partial\lambda'}\right)$$

Now we have:

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

where we use the influence function representation of $\widehat{\gamma}$:

$$\sqrt{N}(\widehat{\lambda}-\lambda)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{r}_i(\lambda)+o_p(1)$$
$$\mathbb{E}(\mathbf{r}_i(\lambda))=\mathbf{0} \text{ and } \mathbb{E}(\mathbf{x}_i\left(y_i-\mathbf{x}_i'\theta\right))=\mathbf{0}$$

Apply the central limit theorem, we obtain:

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

where

$$\begin{aligned}
A_0 &\equiv \mathbb{E}(\mathbf{x}\mathbf{x}')\\
\Omega &\equiv Var\left(\mathbf{x}\left(y-\mathbf{x}'\theta\right)-B_0\mathbf{r}(\lambda)\right)\\
B_0 &\equiv \mathbb{E}\left(\left(\theta\otimes\mathbf{x}'\right)'\frac{\partial\mathbf{x}}{\partial\lambda'}\right)=\mathbb{E}\left(\left(\theta\otimes\widehat{\mathbf{x}}'\right)'\frac{\partial\mathbf{f}(d,z^*,\mathbf{w};\lambda,\mu)}{\partial\lambda'}\right)\\
\mathbf{x} &\equiv \left[\left(d-p(z^*;\lambda)\right),\left(d-p(z^*;\lambda)\right)\left(\mathbf{w}-\mu\right)'\right]'\equiv\mathbf{f}(d,z^*,\mathbf{w};\lambda,\mu)
\end{aligned}$$

together with the influence function for $\widehat{\lambda}$:

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

Asymptotic normality is established. ■

C.2 Asymptotics for the Correction Function Estimator

Under selection-on-unobservables with heterogeneous treatment effects, a parameterized model with two correction functions added back in is:

$$\begin{aligned}
y &= \beta_0+\mathbf{w}'\beta_1+\alpha d+d(\mathbf{w}-\mu)'\gamma+\widetilde{e}\\
&\quad+\xi_1\phi(\pi_0+\pi_1 z+\pi_2 z^*)\\
&\quad+\xi_2\left[\Phi(\pi_0+\pi_1 z+\pi_2 z^*)-(\pi_0+\pi_1 z+\pi_2 z^*)\phi(\pi_0+\pi_1 z+\pi_2 z^*)\right]\\
d &= 1\{\pi_0+\pi_1 z+\pi_2 z^*+v>0\}\\
\widetilde{e} &\equiv u_0+d(u_1-u_0)-\mathbb{E}(d(u_1-u_0)|z,z^*)+d(\eta-\mathbb{E}(\eta))\\
\mu &\equiv \mathbb{E}(\mathbf{w})
\end{aligned}$$

where $\mathbb{E}(\widetilde{e}|z,z^*)=0, v\sim N(0,1), \phi(\cdot)$ is normal pdf, $\Phi(\cdot)$ is normal cdf. We use the following definitions and parameterization:

$$\begin{aligned}
\text{ATE} &\equiv \alpha\equiv\mathbb{E}(\eta+\mathbf{w}'\gamma)\\
g_0(z^*) &\equiv \beta_0+\mathbf{w}'\beta_1\\
\lambda(z^*) &\equiv \mathbf{w}'\gamma, \text{ where } \mathbf{w} \text{ is a vector including polynomials of } z^*
\end{aligned}$$

We also give the following definitions to simplify notation: $\theta\equiv(\beta_0,\beta_1',\alpha,\gamma',\xi_1,\xi_2)', \pi\equiv(\pi_0,\pi_1,\pi_2)', \widetilde{\mathbf{z}}\equiv(1,z,z^*)'$. The regressors included in the model defined at the population are:

$$\begin{aligned}
\mathbf{x} &\equiv (1,\mathbf{w}',d,d(\mathbf{w}-\mu)',\phi(\widetilde{\mathbf{z}}'\pi),\Phi(\widetilde{\mathbf{z}}'\pi)-(\widetilde{\mathbf{z}}'\pi)\phi(\widetilde{\mathbf{z}}'\pi))'\\
&\equiv \mathbf{f}(d,\widetilde{\mathbf{z}},\mathbf{w};\pi,\mu)
\end{aligned}$$

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

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

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

$$\begin{aligned}
\mathbf{z} &\equiv (1,\mathbf{w}',\Phi(\widetilde{\mathbf{z}}'\pi),\Phi(\widetilde{\mathbf{z}}'\pi)(\mathbf{w}-\mu)',\phi(\widetilde{\mathbf{z}}'\pi),\Phi(\widetilde{\mathbf{z}}'\pi)-(\widetilde{\mathbf{z}}'\pi)\phi(\widetilde{\mathbf{z}}'\pi))'\\
&\equiv \mathbf{g}(\widetilde{\mathbf{z}},\mathbf{w};\pi,\mu)
\end{aligned}$$

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

$$\begin{aligned}\widehat{\mathbf{z}}_i &\equiv (1, \mathbf{w}_i', \Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}), \Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi})(\mathbf{w}_i - \widehat{\mu})', \phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}), \Phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}) - (\widetilde{\mathbf{z}}_i'\widehat{\pi})\phi(\widetilde{\mathbf{z}}_i'\widehat{\pi}))' \\ &\equiv \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\mu})\end{aligned}$$

The problem of generated regressors and generated instruments is caused by $\widehat{\pi}$ and $\widehat{\mu}$. The actual model used for estimation, based on a random sample, is:

$$\begin{aligned}&y_i = \widehat{\mathbf{x}}_i'\theta + \text{error}_i \\ &d_i = 1\{\widetilde{\mathbf{z}}_i'\pi + v_i > 0\} \\ &v \sim N(0,1), \phi(\cdot) \text{ normal pdf, } \Phi(\cdot) \text{ normal cdf}\end{aligned}$$

To analyze asymptotic properties, it is useful to rewrite the model in the following way:

$$y_i = \widehat{\mathbf{x}}_i'\theta + \text{error}_i = \widehat{\mathbf{x}}'\theta + (\mathbf{x}_i - \widehat{\mathbf{x}}_i)'\theta + \widetilde{e}_i, \mathbb{E}(\widetilde{e}_i|\widetilde{\mathbf{z}}_i) = 0$$

Given the distributional assumption that $v \sim N(0,1)$, $\mathbf{z}_i$ defined in the population model are the optimal instruments if conditional homoskedasticity for $Var(\widetilde{e}_i|\widetilde{\mathbf{z}}_i)$ holds. Since we have equal number of endogenous variables and instruments, the model is just-identified. An instrumental variable (IV) estimator for θ with generated regressors and instruments including two correction functions is:

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

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

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

$$\left.\begin{aligned}&\widehat{\mathbf{z}}_i \equiv \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\mu}) \xrightarrow{p} \mathbf{g}(\widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \mu) \equiv \mathbf{z}_i \\ &\widehat{\mathbf{x}}_i \equiv \mathbf{f}(d_i, \widetilde{\mathbf{z}}_i, \mathbf{w}_i; \widehat{\pi}, \widehat{\mu}) \xrightarrow{p} \mathbf{f}(d_i, \widetilde{\mathbf{z}}_i, \mathbf{w}_i; \pi, \mu) \equiv \mathbf{x}_i\end{aligned}\right\} \Rightarrow \left(\frac{1}{N}\sum_{i=1}^N \widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i'\right)^{-1} \xrightarrow{p} \mathbb{E}^{-1}(\mathbf{z}\mathbf{x}')$$

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

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

Consistency is established straightforwardly. ■

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

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

$$\mathbf{r}_i(\pi) \equiv \mathbb{E}^{-1}\left(\frac{\phi^2(\widetilde{\mathbf{z}}_i'\pi)\widetilde{\mathbf{z}}_i\widetilde{\mathbf{z}}_i'}{\Phi(\widetilde{\mathbf{z}}_i'\pi)\left(1 - \Phi(\widetilde{\mathbf{z}}_i'\pi)\right)}\right)\frac{\phi(\widetilde{\mathbf{z}}_i'\pi)\widetilde{\mathbf{z}}_i(d_i - \Phi(\widetilde{\mathbf{z}}_i'\pi))}{\Phi(\widetilde{\mathbf{z}}_i'\pi)\left(1 - \Phi(\widetilde{\mathbf{z}}_i'\pi)\right)} \text{ and we have } \mathbb{E}\left(\mathbf{r}_i(\pi)\right) = \mathbf{0}$$

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

$$\widehat{\mu} \xrightarrow{p} \mu \text{ and } \sqrt{N}(\widehat{\mu} - \mu) \xrightarrow{d} N(\mathbf{0}, \Sigma_{\mathbf{w}}) \text{ where } \Sigma_{\mathbf{w}} \equiv Var(\mathbf{w})$$

Consider the IV estimator under just-identification with generated regressors and instruments:

$$\begin{aligned}\widehat{\theta}_{\text{crrf}} &= \left(\sum_{i=1}^{N}\widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i y_i = \left(\sum_{i=1}^{N}\widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i[\widehat{\mathbf{x}}_i'\theta + (\mathbf{x}_i-\widehat{\mathbf{x}}_i)'\theta + \widetilde{e}_i] \\ &= \theta + \left(\sum_{i=1}^{N}\widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i[(\mathbf{x}_i-\widehat{\mathbf{x}}_i)'\theta + \widetilde{e}_i] \\ &\Rightarrow \sqrt{N}(\widehat{\theta}_{\text{crrf}}-\theta) = \left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i[(\mathbf{x}_i-\widehat{\mathbf{x}}_i)'\theta + \widetilde{e}_i]\end{aligned}$$

Given that $\widehat{\pi}\xrightarrow{p}\pi$ and $\widehat{\mu}\xrightarrow{p}\mu$, we have $\Phi(\widetilde{\mathbf{z}}'\widehat{\pi})\xrightarrow{p}\Phi(\widetilde{\mathbf{z}}'\pi)$ and $\phi(\widetilde{\mathbf{z}}'\widehat{\pi})\xrightarrow{p}\phi(\widetilde{\mathbf{z}}'\pi)$ by Slutsky theorem. And the consistency also holds for:

$$\begin{aligned}&\widehat{\mathbf{z}}_i \equiv \mathbf{g}(\widetilde{\mathbf{z}}_i,\mathbf{w}_i;\widehat{\pi},\widehat{\mu})\xrightarrow{p}\mathbf{g}(\widetilde{\mathbf{z}}_i,\mathbf{w}_i;\pi,\mu)\equiv\mathbf{z}_i \\ &\widehat{\mathbf{x}}_i \equiv \mathbf{f}(d_i,\widetilde{\mathbf{z}}_i,\mathbf{w}_i;\widehat{\pi},\widehat{\mu})\xrightarrow{p}\mathbf{f}(d_i,\widetilde{\mathbf{z}}_i,\mathbf{w}_i;\pi,\mu)\equiv\mathbf{x}_i \\ &\quad\Rightarrow \frac{1}{N}\textstyle\sum_{i=1}^{N}\widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i'\xrightarrow{p}\mathbb{E}(\mathbf{z}\mathbf{x}')\equiv A_0\end{aligned}$$

To apply the central limit theorem, consider a first-order Taylor expansion for $\widehat{\mathbf{z}}_i\equiv\mathbf{g}(\widetilde{\mathbf{z}}_i,\mathbf{w}_i;\widehat{\pi},\widehat{\mu})$ at $(\pi',\mu')'$:

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

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

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

Because $\sqrt{N}(\widehat{\pi}-\pi)=O_p(1)$, $\sqrt{N}(\widehat{\mu}-\mu)=O_p(1)$, and $o_p(1)O_p(1)=o_p(1)$, we have

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

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

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

where

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

with the following definitions:

$$
\begin{aligned}
B_0&\equiv\mathbb{E}\left((\theta\otimes\mathbf{z}_i')'\frac{\partial\mathbf{f}(d_i,\widetilde{\mathbf{z}}_i,\mathbf{w}_i;\pi,\mu)}{\partial\pi'}\right)=\mathbb{E}\left[(\xi_2\widetilde{\mathbf{z}}_i'\pi-\xi_1)(\widetilde{\mathbf{z}}_i'\pi)\phi(\widetilde{\mathbf{z}}_i'\pi)\mathbf{z}_i\widetilde{\mathbf{z}}_i'\right]\\
B_1&\equiv\mathbb{E}\left((\theta\otimes\mathbf{z}_i')'\frac{\partial\mathbf{f}(d_i,\widetilde{\mathbf{z}}_i,\mathbf{w}_i;\pi,\mu)}{\partial\mu'}\right)=-\mathbb{E}\left(d_i\mathbf{z}_i\right)\gamma'
\end{aligned}
$$

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

$$
\begin{aligned}
\sqrt{N}(\widehat{\theta}_{\text{crrf}}-\theta)&=\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i[(\mathbf{x}_i-\widehat{\mathbf{x}}_i)'\theta+\widetilde{e}_i]\\
&=\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i\widehat{\mathbf{x}}_i'\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i(\mathbf{x}_i-\widehat{\mathbf{x}}_i)'\theta+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\widehat{\mathbf{z}}_i\widetilde{e}_i\right)\\
&=A_0^{-1}\left(-B_0\sqrt{N}(\widehat{\pi}-\pi)-B_1\sqrt{N}(\widehat{\mu}-\mu)+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{z}_i\widetilde{e}_i\right)+o_p(1)
\end{aligned}
$$

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

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

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

$$
\mathbb{E}(\mathbf{z}_i\widetilde{e}_i-B_0\mathbf{r}_i(\pi)-B_1(\mathbf{w}_i-\mu))=\mathbb{E}(\mathbf{z}_i\widetilde{e}_i)-\mathbb{E}(B_0\mathbf{r}_i(\pi))-\mathbb{E}[B_1(\mathbf{w}_i-\mu)]=\mathbf{0}
$$

Apply the central limit theorem to:

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

and we obtain:

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

where

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

Asymptotic normality is established. ■

D Additional Tables

D.1 Average Treatment Effect

Mean Bias (1000 replications, N = 100)					
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**
model I:					
selection-on-observables: no IV	-0.0289	-0.0355	-0.4948	-0.1695	-0.5875
selection-on-observables: with IV	0.0600	0.0177	-0.4948	-0.1695	-0.5875
omitted variables bias (OVB) only	1.9880	1.8957	0.9826	-0.3382	0.9911
OVB + selectivity bias	4.9471	4.7455	-2.1060	-0.2735	3.2545
OVB + selectivity bias + joint normality	2.9462	2.8456	0.8983	-0.5721	1.7829
model II:					
selection-on-observables: no IV	0.0386	0.0462	1.0394	-0.0491	0.0120
selection-on-observables: with IV	0.0458	0.0351	1.0394	-0.0491	0.0120
omitted variables bias (OVB) only	1.9047	1.8784	2.0628	-0.0743	0.8115
OVB + selectivity bias	4.5638	4.5124	-0.4807	0.3376	2.2141
OVB + selectivity bias + joint normality	2.8240	2.7889	1.4917	-0.0955	1.2019

RMSE (1000 replications, N = 100)					
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**
model I:					
selection-on-observables: no IV	4.7466	0.6957	80.6383	3.3987	0.7497
selection-on-observables: with IV	4.9457	1.2013	80.6383	3.3987	0.7497
omitted variables bias (OVB) only	5.4217	2.2156	143.4809	3.3286	1.1281
OVB + selectivity bias	7.4264	5.0734	145.8808	6.4366	3.3715
OVB + selectivity bias + joint normality	5.9474	3.1365	157.4204	4.8014	1.8877
model II:					
selection-on-observables: no IV	0.3862	0.3745	37.0338	1.0361	0.2600
selection-on-observables: with IV	0.4094	0.3970	37.0338	1.0361	0.2600
omitted variables bias (OVB) only	1.9543	1.9279	64.3438	1.4231	0.8721
OVB + selectivity bias	4.6374	4.5885	21.4470	1.5451	2.2740
OVB + selectivity bias + joint normality	2.8681	2.8344	41.1054	1.4485	1.2645

Median Bias (1000 replications, N = 100)					
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**
model I:					
selection-on-observables: no IV	-0.0923	-0.0346	-0.4342	-0.0565	-0.5742
selection-on-observables: with IV	0.0316	0.0003	-0.4342	-0.0565	-0.5742
omitted variables bias (OVB) only	1.9636	1.8905	0.7736	-0.0810	0.9975
OVB + selectivity bias	4.8245	4.6439	2.7350	0.5569	3.2298
OVB + selectivity bias + joint normality	2.9318	2.8206	1.3526	-0.0897	1.7892
model II:					
selection-on-observables: no IV	0.0310	0.0301	-0.0736	-0.0239	0.0107
selection-on-observables: with IV	0.0428	0.0245	-0.0736	-0.0239	0.0107
omitted variables bias (OVB) only	1.9127	1.8954	-0.0138	-0.0326	0.8117
OVB + selectivity bias	4.4952	4.4672	-0.0001	0.3868	2.2186
OVB + selectivity bias + joint normality	2.8308	2.7867	0.0928	-0.0502	1.1882

Median Absolute Error (1000 replications, N = 100)					
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**
model I:					
selection-on-observables: no IV	0.6117	0.3426	3.2847	0.9785	0.5786
selection-on-observables: with IV	0.5983	0.3498	3.2847	0.9785	0.5786
omitted variables bias (OVB) only	2.0164	1.8968	4.9119	1.1340	0.9990
OVB + selectivity bias	4.8596	4.6470	8.0034	2.3219	3.2298
OVB + selectivity bias + joint normality	2.9806	2.8218	6.0211	1.3537	1.7892
model II:					
selection-on-observables: no IV	0.2560	0.2525	0.9038	0.2504	0.1726
selection-on-observables: with IV	0.2725	0.2681	0.9038	0.2504	0.1726
omitted variables bias (OVB) only	1.9127	1.8954	1.1769	0.2949	0.8117
OVB + selectivity bias	4.4952	4.4672	1.9441	0.5508	2.2186
OVB + selectivity bias + joint normality	2.8308	2.7867	1.4154	0.3757	1.1882

Mean Bias (1000 replications, N = 1000)					
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**
model I:					
selection-on-observables: no IV	-0.1268	-0.0562	2.3446	0.0077	-0.5608
selection-on-observables: with IV	0.0078	0.0041	2.3446	0.0077	-0.5608
omitted variables bias (OVB) only	1.9199	1.9142	2.5725	-0.0024	1.0280
OVB + selectivity bias	4.7720	4.7614	4.5130	0.6272	3.3182
OVB + selectivity bias + joint normality	2.8768	2.8700	2.3658	-0.0142	1.8234
model II:					
selection-on-observables: no IV	0.0056	0.0272	0.0192	-0.0016	0.0013
selection-on-observables: with IV	0.0046	0.0033	0.0192	-0.0016	0.0013
omitted variables bias (OVB) only	1.8614	1.8590	0.0668	0.0002	0.8119
OVB + selectivity bias	4.5318	4.5270	0.0875	0.4177	2.2190
OVB + selectivity bias + joint normality	2.7913	2.7880	0.0953	0.0002	1.2170

RMSE (1000 replications, N = 1000)					
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**
model I:					
selection-on-observables: no IV	0.2383	0.1347	66.7925	0.3876	0.5770
selection-on-observables: with IV	0.2093	0.1241	66.7925	0.3876	0.5770
omitted variables bias (OVB) only	1.9318	1.9189	56.2045	0.4598	1.0397
OVB + selectivity bias	4.7801	4.7672	104.0578	1.1280	3.3278
OVB + selectivity bias + joint normality	2.8845	2.8739	69.7028	0.5888	1.8322
model II:					
selection-on-observables: no IV	0.1110	0.1122	0.3521	0.1052	0.0775
selection-on-observables: with IV	0.1174	0.1151	0.3521	0.1052	0.0775
omitted variables bias (OVB) only	1.8657	1.8632	0.4750	0.1367	0.8173
OVB + selectivity bias	4.5385	4.5338	0.7774	0.4732	2.2245
OVB + selectivity bias + joint normality	2.7952	2.7919	0.5994	0.1699	1.2227

Median Bias (1000 replications, N = 1000)					
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**
model I:					
selection-on-observables: no IV	-0.1282	-0.0603	-0.1677	0.0212	-0.5558
selection-on-observables: with IV	0.0022	0.0024	-0.1677	0.0212	-0.5558
omitted variables bias (OVB) only	1.9193	1.9134	0.4549	0.0326	1.0199
OVB + selectivity bias	4.7813	4.7564	0.7100	0.7258	3.3195
OVB + selectivity bias + joint normality	2.8782	2.8709	0.4865	0.0419	1.8162
model II:					
selection-on-observables: no IV	0.0020	0.0254	0.0177	-0.0054	-0.0001
selection-on-observables: with IV	0.0040	0.0037	0.0177	-0.0054	-0.0001
omitted variables bias (OVB) only	1.8570	1.8521	0.0377	0.0037	0.8113
OVB + selectivity bias	4.5366	4.5321	0.0594	0.4230	2.2213
OVB + selectivity bias + joint normality	2.7887	2.7851	0.0630	0.0047	1.2171

Median Absolute Error (1000 replications, N = 1000)					
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**
model I:					
selection-on-observables: no IV	0.1570	0.0957	2.1130	0.2546	0.5558
selection-on-observables: with IV	0.1295	0.0875	2.1130	0.2546	0.5558
omitted variables bias (OVB) only	1.9193	1.9134	2.9991	0.2974	1.0199
OVB + selectivity bias	4.7813	4.7564	5.0853	0.8800	3.3195
OVB + selectivity bias + joint normality	2.8782	2.8709	3.7616	0.3940	1.8162
model II:					
selection-on-observables: no IV	0.0754	0.0759	0.2210	0.0664	0.0488
selection-on-observables: with IV	0.0842	0.0822	0.2210	0.0664	0.0488
omitted variables bias (OVB) only	1.8570	1.8521	0.3085	0.0956	0.8113
OVB + selectivity bias	4.5366	4.5321	0.4884	0.4230	2.2213
OVB + selectivity bias + joint normality	2.7887	2.7851	0.3733	0.1182	1.2171

D.2 Average Treatment Effect at the Cutoff Point

Mean Bias (1000 replications, N = 100)						
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**	**2SLS**
model I:						
selection-on-observables: no IV	0.1408	-0.0914	0.9599	-0.1154	0.4125	-1.0020
selection-on-observables: with IV	-0.0897	0.0184	0.9599	-0.1154	0.4125	-1.0020
omitted variables bias (OVB) only	1.6631	1.7695	2.7790	-0.1835	1.9911	-1.2100
OVB + selectivity bias	3.9383	4.0706	-0.9400	-0.2617	4.2545	-1.4393
OVB + selectivity bias + joint normality	2.5425	2.6481	2.8939	-0.3200	2.7829	-1.2917

RMSE (1000 replications, N = 100)						
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**	**2SLS**
model I:						
selection-on-observables: no IV	1.1711	0.4765	87.5649	2.9278	0.6221	2.1466
selection-on-observables: with IV	2.1830	0.7404	87.5649	2.9278	0.6221	2.1466
omitted variables bias (OVB) only	2.7367	1.9444	164.4096	2.4751	2.0627	5.0884
OVB + selectivity bias	4.7186	4.2660	114.5888	4.7444	4.3447	9.0264
OVB + selectivity bias + joint normality	3.4220	2.8091	173.3461	3.7545	2.8512	6.7821

Median Bias (1000 replications, N = 100)						
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**	**2SLS**
model I:						
selection-on-observables: no IV	0.1904	-0.0982	-0.1799	-0.0260	0.4258	-0.8646
selection-on-observables: with IV	0.0039	0.0098	-0.1799	-0.0260	0.4258	-0.8646
omitted variables bias (OVB) only	1.7695	1.7523	0.7246	-0.0159	1.9975	-0.8255
OVB + selectivity bias	4.0605	3.9780	2.3025	0.4216	4.2298	-0.6515
OVB + selectivity bias + joint normality	2.6747	2.6118	1.2668	0.0151	2.7892	-0.7470

Median Absolute Error (1000 replications, N = 100)						
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**	**2SLS**
model I:						
selection-on-observables: no IV	0.7786	0.3174	2.5319	0.7523	0.4732	0.9483
selection-on-observables: with IV	0.7733	0.3419	2.5319	0.7523	0.4732	0.9483
omitted variables bias (OVB) only	1.8150	1.7523	3.6196	0.8803	1.9975	1.0174
OVB + selectivity bias	4.0700	3.9780	6.2017	1.7324	4.2298	1.4128
OVB + selectivity bias + joint normality	2.6823	2.6118	4.4077	1.1284	2.7892	1.1235

Mean Bias (1000 replications, N = 1000)						
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**	**2SLS**
model I:						
selection-on-observables: no IV	0.1327	-0.0625	1.7126	0.0088	0.4392	-0.8949
selection-on-observables: with IV	-0.0003	-0.0036	1.7126	0.0088	0.4392	-0.8949
omitted variables bias (OVB) only	1.7429	1.7392	1.8772	0.0063	2.0280	-0.8910
OVB + selectivity bias	4.0409	4.0366	3.3763	0.3381	4.3182	-0.8351
OVB + selectivity bias + joint normality	2.6171	2.6128	1.7395	0.0019	2.8234	-0.8894

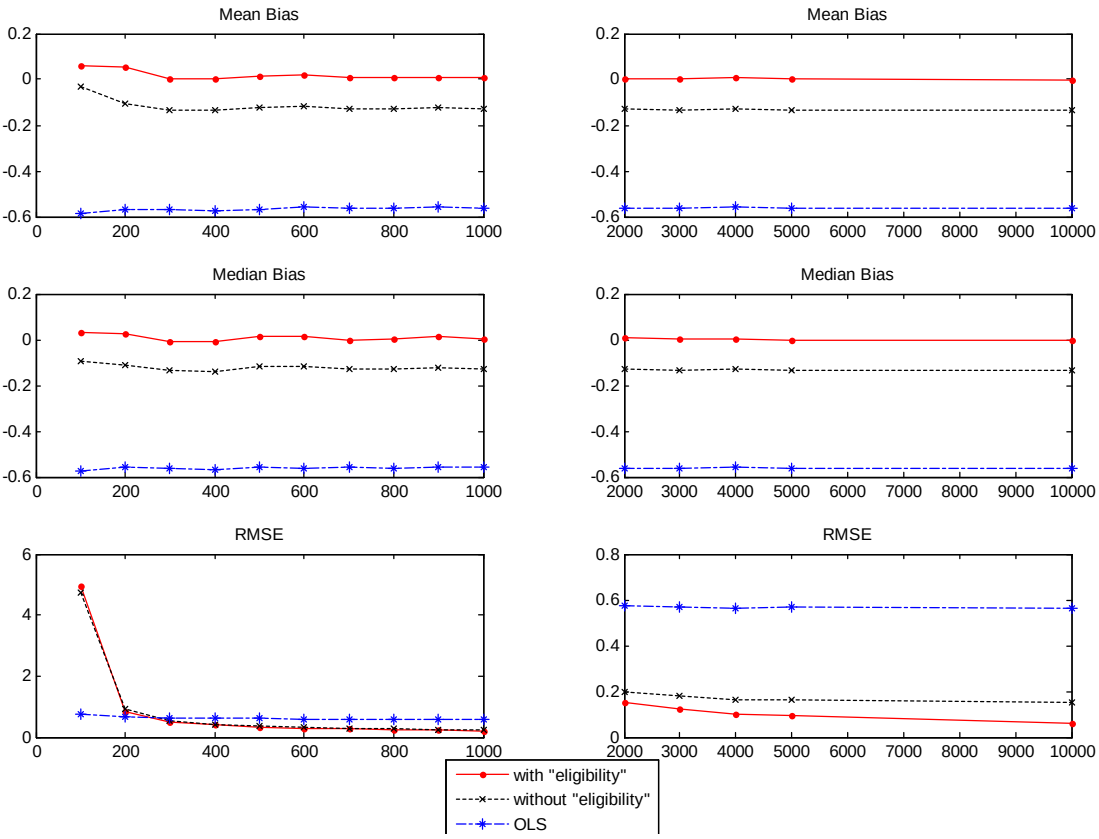
RMSE (1000 replications, N = 1000)						
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**	**2SLS**
model I:						
selection-on-observables: no IV	0.3245	0.1568	53.4452	0.3043	0.4596	0.9498
selection-on-observables: with IV	0.3170	0.1576	53.4452	0.3043	0.4596	0.9498
omitted variables bias (OVB) only	1.7739	1.7479	50.8568	0.3771	2.0340	0.9976
OVB + selectivity bias	4.0684	4.0473	76.4218	0.8119	4.3256	1.1277
OVB + selectivity bias + joint normality	2.6425	2.6202	54.6139	0.4891	2.8291	1.0474

Median Bias (1000 replications, N = 1000)						
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**	**2SLS**
model I:						
selection-on-observables: no IV	0.1353	-0.0631	-0.0578	0.0057	0.4442	-0.8761
selection-on-observables: with IV	0.0179	-0.0070	-0.0578	0.0057	0.4442	-0.8761
omitted variables bias (OVB) only	1.7574	1.7321	0.2352	0.0303	2.0199	-0.8552
OVB + selectivity bias	4.0515	4.0260	0.3011	0.4279	4.3195	-0.7922
OVB + selectivity bias + joint normality	2.6304	2.6128	0.2730	0.0418	2.8162	-0.8498

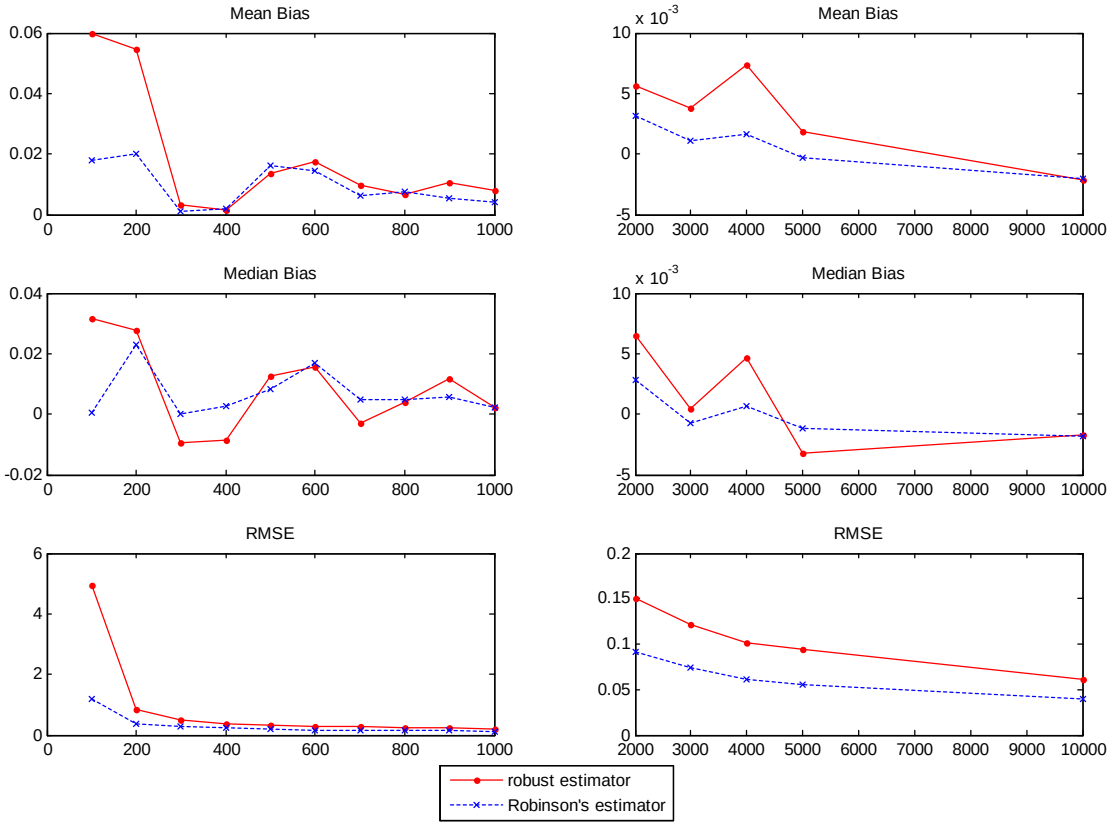
Median Absolute Error (1000 replications, N = 1000)						
Cases	**Robust**	**Robinson**	**Corr Func**	**Ctrl Func**	**OLS**	**2SLS**
model I:						
selection-on-observables: no IV	0.2258	0.1066	1.4563	0.1968	0.4442	0.8761
selection-on-observables: with IV	0.2107	0.1127	1.4563	0.1968	0.4442	0.8761
omitted variables bias (OVB) only	1.7574	1.7321	2.0610	0.2509	2.0199	0.8552
OVB + selectivity bias	4.0515	4.0260	3.5682	0.5913	4.3195	0.8004
OVB + selectivity bias + joint normality	2.6304	2.6128	2.6607	0.3241	2.8162	0.8498

E Additional Figures

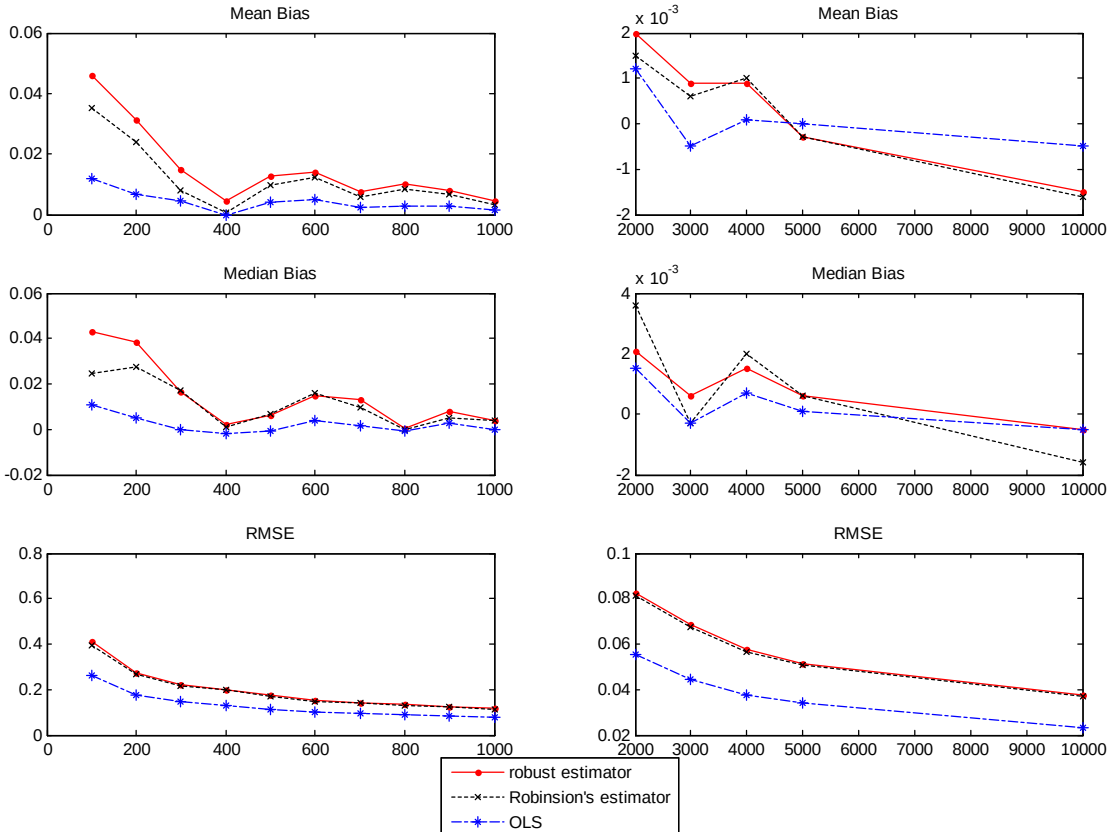
Appendix Figure 1



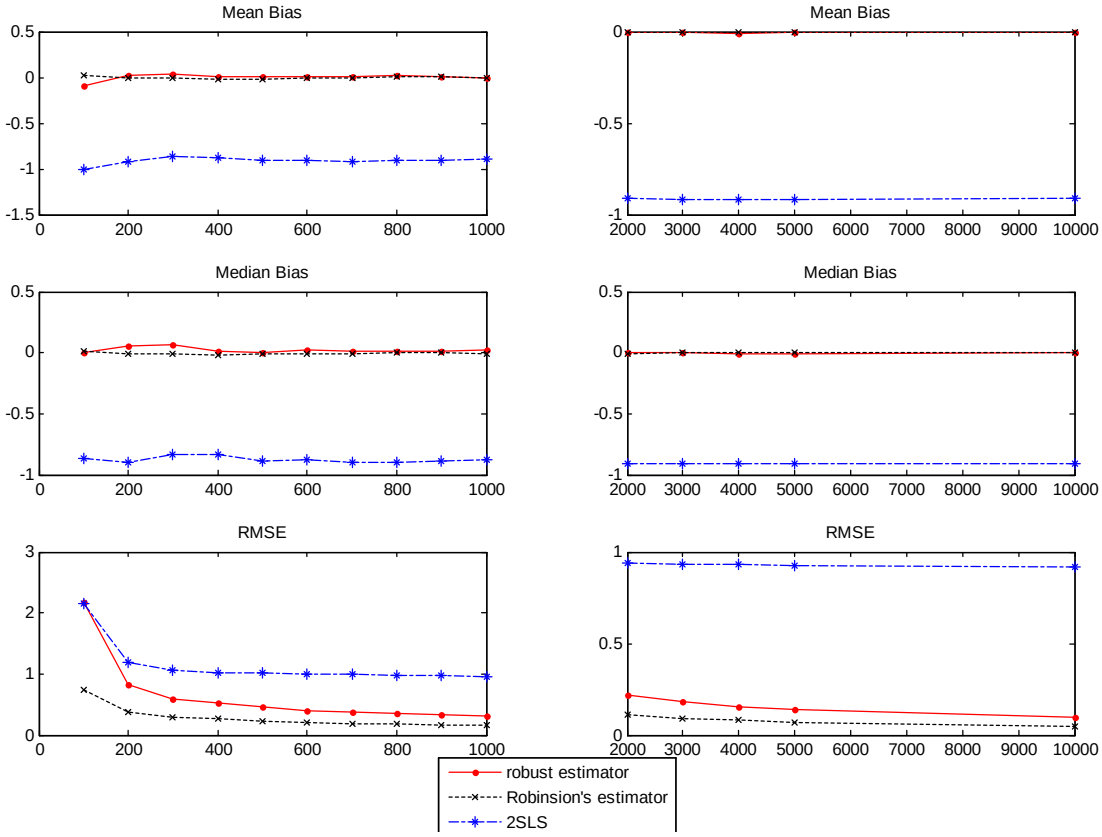
Appendix Figure 2



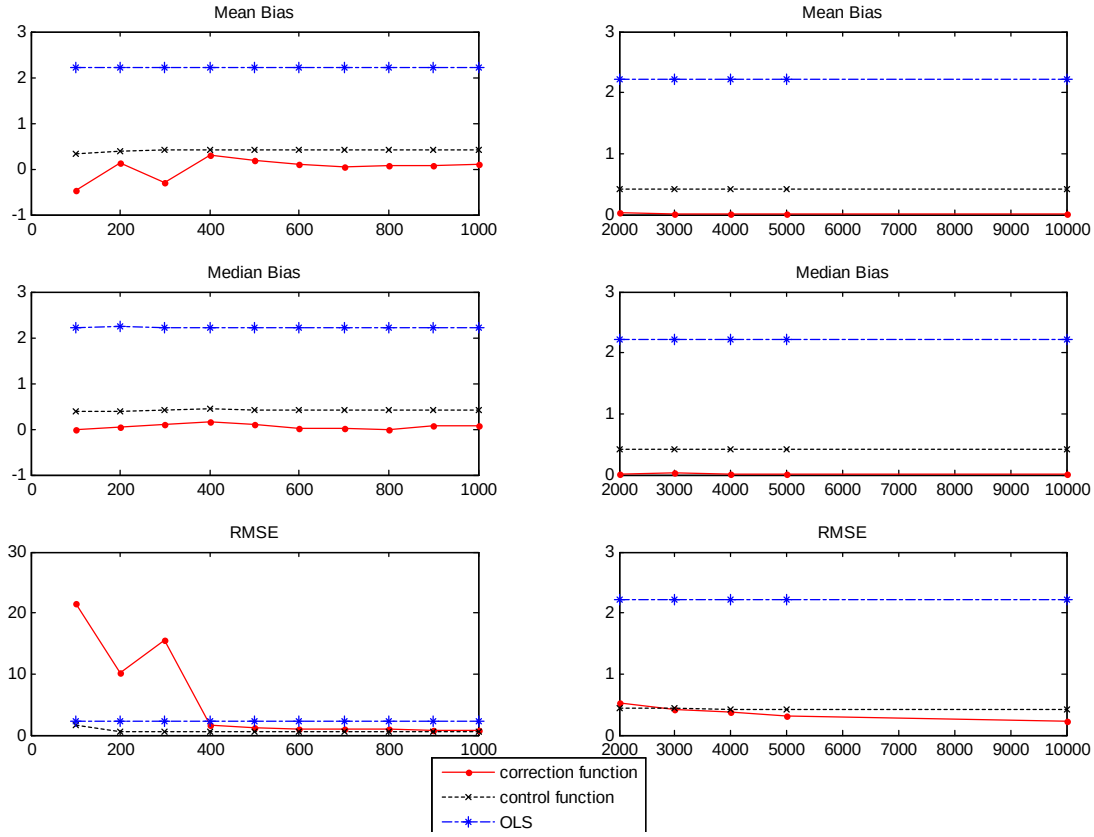
Appendix Figure 3



Appendix Figure 4



Appendix Figure 5



Appendix Figure 6