Testing for Treatment Effect Heterogeneity in Regression Discontinuity Design

Yu-Chin Hsu[†]

Institute of Economics

Academia Sinica

Shu Shen[‡]

Department of Economics

University of California, Davis

[†] E-mail: ychsu@econ.sinica.edu.tw. 128 Academia Road, Section 2, Nankang, Taipei, 115 Taiwan.

[‡] E-mail: shushen@ucdavis.edu. One Shields Avenue, Davis, CA 95616.

Acknowledgements: Yu-Chin Hsu gratefully acknowledges the research support from Ministry of Science and Technology of Taiwan (MOST103-2628-H-001-001-MY4) and Career Development Award of Academia Sinica, Taiwan. Shu Shen gratefully acknowledges the research support from the Institute of Social Science of the University of California, Davis and the Hellman Fellows Award. All errors are our own.

Abstract

Treatment effect heterogeneity is frequently studied in regression discontinuity (RD) applications. This paper is the first to propose tests for treatment effect heterogeneity under the RD setup. The proposed tests study whether a policy treatment is 1) beneficial for at least some subpopulations defined by covariate values, 2) has any impact on at least some subpopulations, and 3) has a heterogeneous impact across subpopulations. Compared with other methods currently adopted in applied RD studies, such as the subsample regression method and the interaction term method, our tests have the advantage of being fully nonparametric, robust to weak inference and powerful. Monte Carlo simulations show that our tests perform very well in small samples. We apply the tests to study the impact of attending a better high school and discover interesting patterns of treatment effect heterogeneity that were neglected by classic mean RD analyses.

JEL classification: C21, C31

Keywords: Sharp regression discontinuity, fuzzy regression discontinuity, treatment effect heterogeneity.

1 Introduction

Regression discontinuity (RD) has gained increasing popularity in the field of applied economics in the past two decades for providing credible and straightforward identification of the causal effect of policies.[1] The identification strategy uses the fact that the probability of an individual receiving a policy treatment changes discontinuously with an underlying variable, often referred to as the running variable. Researchers compare the response outcome above and below the point of the underlying variable where the discontinuity occurs and identify the average treatment effect of individuals at the margin of policy treatment. If researchers are only interested in the average effect, controlling for additional covariates other than the running variable is not necessary under the RD setup. However, when researchers are further interested in treatment effect heterogeneity, or the variation of the policy impact among individuals, it is important to extract information from additional controls and to consider the conditional average policy effects for individuals with different observed characteristics.

This paper considers the inference of conditional average policy effects under the RD setup. Specifically, we propose (uniform) tests for treatment effect heterogeneity that test whether a policy treatment is 1) beneficial to at least some subpopulations defined by covariate values, 2) has any impact on at least some subpopulations, and has heterogeneous impact across all subpopulations. Both sharp and fuzzy RD designs are considered.

The tests we propose are useful because applied researchers are often interested in treatment effect heterogeneity under the RD setup. A survey of recent publications in top general interest journals in economics finds that 15 out of 17 papers that adopted the RD framework analyze treatment effect heterogeneity.[2] The common practice is to

[1] Pioneering work and early applications in RD include van der Klaauw (2002), Angrist and Lavy (1999), and Black (1999), among others. Imbens and Lemieux (2008) and Lee and Lemieux (2010) provide excellent reviews of the topic.

[2] The survey includes all 2015 issues of Quarterly Journal of Economics, Journal of Political Economy, Review of Econommic Studies, American Economic Review, American Economic Journal: Applied Economics and American Economic Journal: Economic Policy as well as 2016 issues of these Journals published before April. A total of 17 papers (0 in QJE, 1 in JPE, 0 in RESTUD, 5 in AER, 5 in AEJ: AE, and 6 in AEJ: EP) use the RD method, among which 15 address the issue of treatment effect heterogene-

accompany the primary RD regression with some subsample regressions or to build linear regression models with interaction terms between the indicator of whether a running variable exceeds the threshold and additional controls of interest.

The interaction term method, as we will show, is parametric and severely over-rejects under model misspecification, even if researchers use data only close to the cut-off of the running variable for estimation. This is in sharp contrast with the classic RD regression method which is nonparametric and robust to misspecification as long as the estimation window, or bandwidth, is properly chosen.

The subsample regression method, on the other hand, is nonparametric. To implement the method correctly it is essential to adjust all inference results for multiple testing (see, for example, Romano and Shaikh, 2010; Anderson, 2008). However, all papers in our survey using the subsample regression method ignore the issue of multiple testing. Moreover, even if multiple testing is correctly accounted for, the subsample regression method is suboptimal. First, it can produce over-rejected tests and under-covered confidence intervals under the fuzzy RD design if the sample size and the proportion of compliers are small for some subsamples (Feir, Lemieux, and Marmer, 2015). The main reason is that the subsample regression method relies on subsample local average treatment effect estimators that can have non-classical inference with non-normal distributions under weak first stage (classic articles on weak inference includes Stock and Yogo, 2005; Staiger and Stock, 1997; Moreira, 2003, among others). Second, the subsample regression method often requires categorizing continuous covariates into discrete groups. If the groups are coarsely defined, important information on treatment effect heterogeneity can be lost. If the groups are finely discretized, the subsample regression method can lose power for having subsamples of small sizes.

We characterize the hypotheses of interest using nonparametric conditional moment equalities/inequalities conditional on both the running variable of the RD model and other additional covariates of interest, and then use the idea of the instrument function method developed in Andrews and Shi (2013, 2015) to transform the hypotheses to (an

ity. 2 of the 15 papers carry out the heterogeneity analysis using linear regressions with interaction terms. All of the other 13 papers use subsample RD regressions. None of the 13 papers using the subsample regression method correct for multiple testing.

infinite number of) instrumented conditional moment equalities/inequalities conditional only on the running variable. This transformation of hypotheses is without loss of information, and each of the transformed moments can be estimated by nonparametric local linear estimator at the boundary. The tests we propose have statistics of order $(nh)^{-1/2}$, which means that although we are looking at conditional average policy effects conditional on multiple control variables, the statistic has the same rate of convergence as the classic mean RD estimators that do not control covariates other than the running variable. Moreover, the proposed tests do not rely on plug-in estimators of conditional average treatment effects, meaning that the proposed tests are robust to the weak inference problem discussed above. As we demonstrate in the Monte Carlo simulation section, the proposed tests have very good small sample performance compared to both the interaction term method and the subsample regression method currently adopted in the applied literature.

The tests proposed in this paper are related to Andrews and Shi (2013, 2015), and other conditional moment equality/inequality tests that apply the instrument function method (e.g. Hsu, 2015; Bugni, Canay, and Shi, 2015). Since estimation of the nonparametric RD model involves boundary estimators in the local polynomial class, and such estimators have not been previously used in conjunction with the instrument function method, our paper contributes to the literature in developing a new testing method for conditional moment equality/inequalities that require nonparametric estimation on the boundary. In addition, we propose a new multiplier bootstrap method for simulating critical values in proposed testing approaches.

We apply the proposed tests to study the impact of attending a better high school in Romania following Pop-Eleches and Urquiola (2013). Mean RD analysis in Pop-Eleches and Urquiola (2013) found that going to a more selective high school significantly improves the average Baccalaureate exam grade among marginal students but does not seem to affect the probability of a student taking the Baccalaureate exam. Pop-Eleches and Urquiola (2013) carry out an analysis of the heterogeneity of the estimated effect using the subsample regression method and again find little evidence supporting the effect of going to a better school on the exam-taking rate. In contrast, our proposed tests detect a clear signal that attending a more selective school has a significant effect on the

exam-taking rate for at least some subpopulations. Our tests also find strong evidence supporting treatment effect heterogeneity. A closer look at the testing results suggest that the insignificant mean effect found in Pop-Eleches and Urquiola (2013) results from the cancellation of opposite-signed effects among different subpopulations.

The paper is organized as follows. Section 2 sets up the model and identifies the conditional treatment effects of interest under sharp and fuzzy RD designs. Section 3 proposes three uniform tests for treatment effect heterogeneity under the sharp RD design. Section 4 extends the tests to the fuzzy RD design. Section 5 examines the small sample performance of the proposed tests and compares the performance with other naive tests currently adopted in the applied literature. Section 6 applies the proposed tests to study the heterogeneous effect of going to a better school using the Romanian dataset published by Pop-Eleches and Urquiola (2013). Proofs and technical assumptions are provided in the Appendix.

2 Model Framework

Let Y_i denote the outcome of interest and T_i the dummy variable indicating treatment of individual i if $T_i = 1$. Use $Y_i(0)$ and $Y_i(1)$ to denote potential outcomes when $T_i = 0$ and $T_i = 1$, respectively. Whether individual i receives treatment depends at least partially on the running variable Z_i. A policy intervention encourages an individual i to receive treatment if the running variable Z_i is larger than or equal to c. Let $T_i(1)$, $T_i(0)$ be the potential treatment decisions of individual i depending on whether he/she is encouraged (i.e. $Z_i \geq c$) or not (i.e. $Z_i < c$). Let X_i denote a set of covariates with compact support $\mathcal{X} \subset R^{d_x}$. Without loss of generality, assume that $\mathcal{X} = \times_{j=1}^{d_x}[0,1]$ and use $\mathcal{X}_c \subset \mathcal{X}$ to denote the support of X_i conditional on $Z_i = c$. For notational simplicity, we assume that X_i includes only continuous variables. In the next section, we will discuss how to implement our test when X_i contains discrete variables.

Assumption 2.1 *For a running variable Z_i continuously distributed in a neighborhood of the threshold value c, assume that*

(i) *$E[Y_i(t)|X_i = x, Z_i = z]$ is continuous with respect to z in the neighborhood of c for both $t = 0, 1$ and $x \in \mathcal{X}_c$.*

(ii) The distribution function of $X_i|Z_i = z$ is continuous with respect to z in a neighborhood of c.

Assumption 2.1(i) requires that the conditional means of the potential outcomes and the conditional distribution of the additional controls are continuous with respect to the running variable in a neighborhood of c. At first appearance, Assumption 2.1(i) is stronger than the standard continuity assumption of $E[Y_i(t)|Z_i = z]$ required in the literature (c.f. Imbens and Lemieux, 2008). However, as we will illustrate, both parts of the assumption are simply direct consequences of having "no precise control over the running variable", a rule for identification introduced by Lee and Lemieux (2010) and well-accepted in the RD literature. To be specific, suppose without loss of generality that the potential outcomes $Y_i(t) = g_t(X_i, Z_i, V_i)$ are a function of predetermined observed and unobserved characteristics X_i and V_i, as well as the running variable Z_i, for both $t = 0, 1$. Following Lee and Lemieux (2010), an individual is said to have imprecise control over the running variable if the conditional density $Z_i = z|X_i, V_i$ is continuous in z around c. As shown by Lee and Lemieux (2010), this no perfect control requirement implies that the density of $X_i, V_i|Z_i = z$ is continuous in z around c, which further implies continuity of the density $X_i|Z_i = z$ around $z = c$, the condition imposed in Assumption 2.1(ii). Further, the conditional mean

$$E[Y_i(t)|X_i = x, Z_i = z] = \int g_t(x, z, V) \frac{f(x, V|z)}{f(x|z)} dV$$

is continuous in z around c because both conditional densities in the formula are continuous in z around c.

When the treatment decision T_i is a deterministic function of the running variable Z_i such that $T_i = 1(Z_i \geq c)$, the model follows a *sharp RD design*. Under Assumption 2.1, the conditional average policy effect (ATE) conditional on $Z_i = c$ is defined and identified as

$$ATE = E[Y_i(1)|Z_i = c] - E[Y_i(0)|Z_i = c] = \lim_{z \searrow c} E[Y_i|Z_i = z] - \lim_{z \nearrow c} E[Y_i|Z_i = z],$$

while the conditional average policy effect conditional on $Z_i = c$ and $X_i = x$, $CATE(x)$, is defined and identified as

$$\begin{aligned} CATE(x) &= E[Y_i(1)|X_i = x, Z_i = c] - E[Y_i(0)|X_i = x, Z_i = c] \\ &= \lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z]. \end{aligned}$$

More generally, when the treatment status T_i is a probabilistic function of Z_i, the RD model follows a *fuzzy* design. For identification in this general case we require the following additional assumption.

Assumption 2.2 *Assume that*

(i) $E[T_i(1)|X_i = x, Z_i = z]$ *and* $E[T_i(0)|X_i = x, Z_i = z]$ *are continuous with respect to* z *in a neighborhood of* c *for all* $x \in \mathcal{X}_c$*;*

(ii) $T_i(1) \geq T_i(0)$*;*

(iii) $E[T_i(1)|X_i = x, Z_i = c] - E[T_i(0)|X_i = x, Z_i = c] > 0$ *for all* $x \in \mathcal{X}_c$*.*

Assumption 2.2(i) requires the continuity of compliance. It is stronger than the standard continuity of compliance assumption in the literature (c.f. Imbens and Lemieux, 2008) but is again implied by the well-accepted "no perfect control assumption". Assumption 2.2(ii) assumes away the presence of defiers and is a common identifying restriction in models with fuzzy RD designs. Assumption 2.2(iii) requires the non-trivial presence of compliers. It is stronger than the standard assumption that only requires the existence of compliers unconditional on X_i. This stronger condition is required for the identification of a conditional local average treatment effect that conditions on the value of X_i.

Under fuzzy RD design, the local average treatment effect (LATE) and the conditional local average treatment effect (CLATE) for compliers are defined and identified

respectively as

$$
\begin{aligned}
LATE &= E[Y_i(1) - Y_i(0) | Z_i = c, T_i(1) - T_i(0) = 1] \\
&= \frac{E[(Y_i(1) - Y_i(0))\,(T_i(1) - T_i(0))\,|Z_i = z]}{E[T_i(1) - T_i(0)|Z_i = c]} \\
&= \frac{\lim_{z \searrow c} E[Y_i|Z_i = z] - \lim_{z \nearrow c} E[Y_i|Z_i = z]}{\lim_{z \searrow c} E[T_i|Z_i = z] - \lim_{z \nearrow c} E[T_i|Z_i = z]}, \text{ and} \\
CLATE(x) &= E[Y_i(1) - Y_i(0) | X_i = x, Z_i = c, T_i(1) - T_i(0) = 1] \\
&= \frac{E[(Y_i(1) - Y_i(0))\,(T_i(1) - T_i(0))\,|X_i = x, Z_i = z]}{E[T_i(1) - T_i(0)|X_i = x, Z_i = c]} \\
&= \frac{\lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z]}{\lim_{z \searrow c} E[T_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[T_i|X_i = x, Z_i = z]}. \qquad (2.1)
\end{aligned}
$$

The identification of $LATE$ is standard. The identification of $CLATE(x)$ follows from Assumptions 2.1 and 2.2 and is given in the Appendix. The numerators of the $LATE$ and $CLATE(x)$ are the average reduced-form effect and the conditional average reduced-form effect of the treatment. When the proportion of always-takers is zero (i.e. $T_i(0) = 0$), they also represent the intent-to-treat effects. All identified treatment effects, including ATE, $LATE$, $CATE$, and $CLATE$, can be estimated by standard local linear estimation methods.

When the support of X_i at $Z_i = c$, or $\mathcal{X}_c$ is large, it is possible that the proportion of compliers, $E[T_i(1) - T_i(0)|X_i = x, Z_i = c]$, is small for some subset of x values. Since the estimation of $CLATE(x)$ relies on the estimation of proportion of compliers in the denominator, it can have poor finite sample performance analogous to the concerns raised in the weak IV literature. Therefore, testing procedures relying on plug-in estimators of $CLATE(x)$ such as the subsample regression method are suboptimal. We avoid using plug-in estimators of $CLATE(x)$, as well as $LATE$, in all proposed tests in Section 4. As we demonstrate in the Monte Carlo simulation section, our testing procedure is much more robust than the subsample regression method, even if the latter correctly corrects for multiple testing.

3 Testing Under the Sharp RD Design

Researchers are often interested in knowing whether a policy treatment is beneficial to at least some subpopulations defined by covariate values, whether a policy treatment

has any impact on at least some subpopulations, and whether its effect is heterogeneous across all subpopulations. In this section, we develop uniform tests for these purposes under the sharp RD design. We extend the tests to the fuzzy RD design in the next section.

3.1 Testing if the Treatment is Beneficial for At Least Some Subpopulations

Hypotheses Formation

To test if a policy treatment is beneficial to at least some subpopulations defined by covariate values or equivalently to test if the conditional average treatment effects is strictly positive for some covariate values, the null and alternative hypotheses can be formulated as

$$\begin{aligned} H_{0,ate}^{neg} &: CATE(x) = E[Y_i(1) - Y_i(0)|X_i = x, Z_i = c] \leq 0, \ \forall \ x \in \mathcal{X}_c, \\ H_{1,ate}^{neg} &: CATE(x) = E[Y_i(1) - Y_i(0)|X_i = x, Z_i = c] > 0, \ \text{for some} \ \ x \in \mathcal{X}_c. \end{aligned} \tag{3.1}$$

If $H_{0,ate}^{neg}$ is rejected, then one can conclude that the policy is beneficial to at least some subpopulations defined by covariate values, with some pre-specified confidence level. Note that $H_{0,ate}^{neg}$ and $H_{1,ate}^{neg}$ are defined in a form of conditional moment inequality and we apply the instrument function approach in Andrews and Shi (2013) and Andrews and Shi (2015) to transform them to an infinite number of instrumented conditional moment inequalities without loss of information. We first introduce the set of instrument functions we will use. Let $\mathcal{G}$ be the set of the indicator functions of countable hyper cubes C_ℓ such that

$$\begin{aligned} \mathcal{G} &= \{g_\ell(\cdot) = 1(\cdot \in C_\ell) : \ell \equiv (x, r) \in \mathcal{L}\}, \text{ where} \\ C_\ell &= \left(\times_{j=1}^{d_x}[x_j, x_j + r]\right) \cap \mathcal{X} \text{ and} \\ \mathcal{L} &= \left\{(x, (2q)^{-1}) : \ (2q) \cdot x \in \{0, 1, 2, \cdots, (2q-1)\}^{d_x}, \text{ and } q = 1, 2, \cdots\right\}. \end{aligned} \tag{3.2}$$

For each $\ell \in \mathcal{L}$, we define the instrumented conditional moment condition $\nu(\ell) = E[g_\ell(X_i)CATE(X_i)|Z_i = z]$ as in Andrews and Shi (2015). $\nu(\ell)$ is also the average treatment effect for individuals with $X_i \in C_\ell$. When $\ell = (\mathbf{0}, 1)$, $C_\ell = \times_{j=1}^{d_x}[0, 1]$, then $\nu(\ell)$ reduces to $\nu(\mathbf{0}, 1) = E[CATE(X_i)|Z_i = z]$, the ATE under the sharp RD design

and the reduced-form effect in the fuzzy RD design. As in Andrews and Shi (2013, 2015), the hypotheses $H_{0,ate}^{neg}$ and $H_{1,ate}^{neg}$ in (3.1) are equivalent to

$$\begin{aligned} H_{0,ate}^{neg}:&\ \nu(\ell) = E[g_\ell(X_i)CATE(X_i)|Z_i = z] \le 0,\ \forall\ \ell \in \mathcal{L}, \\ H_{1,ate}^{neg}:&\ \nu(\ell) = E[g_\ell(X_i)CATE(X_i)|Z_i = z] > 0,\ \text{for some } \ell \in \mathcal{L}. \end{aligned} \tag{3.3}$$

As a result, the hypotheses $H_{0,ate}^{neg}$ and $H_{1,ate}^{neg}$ can be characterized by infinitely many instrumented conditional moment inequalities without loss of information. Furthermore, in Appendix, we show that $\nu(\ell)$ is identified by

$$\begin{aligned} \nu(\ell) =& E[g_\ell(X_i)CATE(X_i)|Z_i = z] \\ =& \lim_{z \searrow c} E[g_\ell(X_i)Y_i|Z_i = z] - \lim_{z \nearrow c} E[g_\ell(X_i)Y_i|Z_i = z]. \end{aligned} \tag{3.4}$$

Test Statistic and Asymptotic Results

Based on the identification result in (3.4), for each $\ell \in \mathcal{L}$, $\nu(\ell)$ can be estimated by a difference between two local linear estimators. To be specific, let $m_+(\ell) = \lim_{z \searrow c} E[g_\ell(X_i)Y_i|Z_i = z]$ and $m_-(\ell) = \lim_{z \nearrow c} E[g_\ell(X_i)Y_i|Z_i = z]$. The estimators $\hat{m}_+(\ell)$ and $\hat{m}_-(\ell)$ for $m_+(\ell)$ and $m_-(\ell)$ are the constant terms $\hat{a}_+(\ell)$ and $\hat{a}_-(\ell)$ in regressions of the form

$$\begin{aligned} &\min_{\hat{a}_+(\ell), \hat{b}_+(\ell)} \sum_{i=1}^n 1(Z_i \ge c) \cdot K\left(\frac{Z_i - c}{h}\right) \Big[g_\ell(X_i)Y_i - \hat{a}_+(\ell) - \hat{b}_+(\ell)(Z_i - c)\Big]^2, \\ &\min_{\hat{a}_-(\ell), \hat{b}_-(\ell)} \sum_{i=1}^n 1(Z_i < c) \cdot K\left(\frac{Z_i - c}{h}\right) \Big[g_\ell(X_i)Y_i - \hat{a}_-(\ell) - \hat{b}_-(\ell)(Z_i - c)\Big]^2. \end{aligned}$$

where $K(\cdot)$ is a symmetric kernel function and h is the bandwidth. In the Monte Carlo simulation and the empirical application of this paper, we follow the RD literature and use the triangular kernel for boundary local linear estimators.

An estimator for $\nu(\ell)$ is given by $\hat{\nu}(\ell) = \hat{m}_+(\ell) - \hat{m}_-(\ell)$. Following Fan and Gijbels (1992), for $j = 0, 1, 2, \ldots$, define

$$S_{n,j}^+ = \sum_{i=1}^n 1(Z_i \ge c) \cdot K\left(\frac{Z_i - c}{h}\right)(Z_i - c)^j, \quad S_{n,j}^- = \sum_{i=1}^n 1(Z_i < c) \cdot K\left(\frac{Z_i - c}{h}\right)(Z_i - c)^j,$$

For all $\ell \in \mathcal{L}$, the local linear estimators can also be written as

$$\begin{aligned} \hat{m}_+(\ell) &= \frac{\sum_{i=1}^n 1(Z_i \ge c) \cdot K(\frac{Z_i - c}{h})[S_{n,2}^+ - S_{n,1}^+(Z_i - c)]g_\ell(X_i)Y_i}{\sum_{i=1}^n 1(Z_i \ge c) \cdot K(\frac{Z_i - c}{h})[S_{n,2}^+ - S_{n,1}^+(Z_i - c)]} \equiv \sum_{i=1}^n w_{ni}^+ \cdot g_\ell(X_i)Y_i, \\ \hat{m}_-(\ell) &= \frac{\sum_{i=1}^n 1(Z_i < c) \cdot K(\frac{Z_i - c}{h})[S_{n,2}^- - S_{n,1}^-(Z_i - c)]g_\ell(X_i)Y_i}{\sum_{i=1}^n 1(Z_i < c) \cdot K(\frac{Z_i - c}{h})[S_{n,2}^- - S_{n,1}^-(Z_i - c)]} \equiv \sum_{i=1}^n w_{ni}^- \cdot g_\ell(X_i)Y_i, \end{aligned}$$

where

$$w_{ni}^{+} = \frac{1(Z_i \geq c) \cdot K(\frac{Z_i - c}{h})[S_{n,2}^{+} - S_{n,1}^{+}(Z_i - c)]}{\sum_{i=1}^{n} 1(Z_i \geq c) \cdot K(\frac{Z_i - c}{h})[S_{n,2}^{+} - S_{n,1}^{+}(Z_i - c)]}$$
$$w_{ni}^{-} = \frac{1(Z_i < c) \cdot K(\frac{Z_i - c}{h})[S_{n,2}^{+} - S_{n,1}^{+}(Z_i - c)]}{\sum_{i=1}^{n} 1(Z_i < c) \cdot K(\frac{Z_i - c}{h})[S_{n,2}^{+} - S_{n,1}^{+}(Z_i - c)]}.$$

For $j = 0, 1, 2, \ldots,$ let $\vartheta_j = \int_0^\infty u^j K(u) du$. Let $\sigma_+^2(\ell_1, \ell_2) = Cov(g_{\ell_1}(X)Y(1), g_{\ell_2}(X)Y(1))|Z = c]$ be the conditional covariance of $g_{\ell_1}(X)Y(1)$ and $g_{\ell_2}(X)Y(1)$. Define $\sigma_-^2(\ell_1, \ell_2)$ similarly.

We summarize the asymptotics of $\sqrt{nh}(\hat{\nu}(\ell) - \nu(\ell))$ in the following lemma.

Lemma 3.1 *Under Assumption 2.1 and Assumptions A.1 and A.2 described in the appendix, we have*

$$\left| \sqrt{nh} \left(\hat{\nu}(\ell) - \nu(\ell) \right) - \sum_{i=1}^{n} \phi_{\nu,ni}(\ell) \right| = o_p(1),$$
$$\phi_{\nu,ni}(\ell) = \sqrt{nh} \left(w_{ni}^{+} \cdot (g_\ell(X_i)Y_i - m_+(\ell)) - w_{ni}^{-} \cdot (g_\ell(X_i)Y_i - m_-(\ell)) \right) \tag{3.5}$$

where the $o_p(1)$ result holds uniformly over $\ell \in \mathcal{L}$. Also,

$$\sqrt{nh}(\hat{\nu}(\ell) - \nu(\ell)) \Rightarrow \Phi_{h_{2,\nu}}(\ell),$$

where $\Phi_{h_2}(\ell)$ denote a mean zero Gaussian process with covariance kernel

$$h_{2,\nu}(\ell_1, \ell_2) = \frac{\int_0^\infty (\vartheta_2 - u\vartheta_1)^2 K^2(u) du}{(\vartheta_2\vartheta_0 - \vartheta_1^2)^2} \frac{\sigma_+^2(\ell_1, \ell_2) + \sigma_-^2(\ell_1, \ell_2)}{f_z(c)}$$

for $\ell_1, \ell_2 \in \mathcal{L}$.

The proof is given in the Appendix. Lemma 3.1 shows that $\sqrt{nh} \left(\hat{\nu}(\ell) - \nu(\ell) \right)$ weakly converges to a mean zero Gaussian process. $\phi_{\nu,ni}(\ell)$ in (3.5) denotes the influence function for each observation that contributes to the the limiting distribution of $\widehat{\Phi}_{\nu,n}(\ell) = \sqrt{nh} \left(\hat{\nu}(\ell) - \nu(\ell) \right)$.

The Kolmogorov-Smirnov (KS) type test statistic is then defined as

$$\widehat{S}_{neg} = \sqrt{nh} \sup_{\ell \in \mathcal{L}} \hat{\nu}(\ell).$$

Decision Rule and Simulated Critical Value

Given the influence function representation in (3.5), we can use the multiplier bootstrap method in Hsu (2016) to approximate the whole empirical process. To be specific, let U_1, U_2,... be i.i.d. pseudo random variables with $E[U] = 0$, $E[U^2] = 1$ and $E[U^4] < \infty$ that are independent of the sample path. In the Monte Carlo simulation and empirical application of the paper, the pseudo random variables are simulated following the standard normal distribution. Let the simulated process $\widehat{\Phi}^u_{\nu,n}(\ell)$ be

$$\widehat{\Phi}^u_{\nu,n}(\ell) = \sum_{i=1}^{n} U_i \cdot \hat{\phi}_{\nu,ni}(\ell),$$
$$\hat{\phi}_{\nu,ni}(\ell) = \sqrt{nh}\left(w^+_{ni} \cdot (g_\ell(X_i)Y_i - \hat{m}_+(\ell)) - w^-_{ni} \cdot (g_\ell(X_i)Y_i - \hat{m}_-(\ell))\right).$$

$\hat{\phi}_{\nu,ni}(\ell)$ is called the estimated influence function because the unknown functions $m_+(\ell)$ and $m_-(\ell)$ need to be estimated. The following lemma shows that the multiplier bootstrapped process $\widehat{\Phi}^u_{\nu,n}(\ell)$ can approximate the empirical process $\widehat{\Phi}_{\nu,n}(\ell)$ well and the proof is given in Appendix.

Lemma 3.2 *Under Assumption 2.1 and Assumptions A.1, A.2 and A.3 described in the appendix, we have* $\widehat{\Phi}^u_n(\ell) \overset{p}{\Rightarrow} \Phi_{h_{2,\nu}}(\ell)$.[3]

Let P^u denote the multiplier probability measure. For significance level $\alpha < 1/2$, define the simulated critical value $\hat{c}^{neg}_{n,ate}(\alpha)$ as

$$\hat{c}^{neg}_{n,ate}(\alpha) = \sup\left\{q \middle| P^u\Big(\sup_{\ell\in\mathcal{L}} \widehat{\Phi}^u_{\nu,n}(\ell) \le q\Big) \le 1-\alpha\right\},$$

i.e., $\hat{c}^{neg}_{n,ate}(\alpha)$ is the $(1-\alpha)$-th quantile of the simulated null distribution, $\sup_{\ell\in\mathcal{L}} \widehat{\Phi}^u_{\nu,n}(\ell)$.

Finally, let the *decision rule* be: "Reject $H^{neg}_{0,ate}$ if $\widehat{S}^{neg}_{ate} > \hat{c}^{neg}_{n,ate}(\alpha)$."

[3]The conditional weak convergence is in the sense of Section 2.9 of van der Vaart and Wellner (1996) and Chapter 2 of Kosorok (2008). To be more specific, $\Psi^u_n \overset{p}{\Rightarrow} \Psi$ in the metric space $(\mathbb{D}, d)$ if and only if $\sup_{f\in BL_1} |E_u f(\Psi^u_n) - Ef(\Psi)| \overset{p}{\to} 0$ and $E_u f(\Psi^u_n)^* - E_u f(\Psi^u_n)_* \overset{p}{\to} 0$, where the subscript u in E_u indicates conditional expectation over the weights U_i's given the remaining data, BL_1 is the space of functions $f : \mathbb{D} \to R$ with Lipschitz norm bounded by 1, and $f(\Psi^u_n)^*$ and $f(\Psi^u_n)_*$ denote measurable majorants and minorants with respect to the joint data including the U_i's. The notation $\Psi^u_n \overset{a.s.}{\Rightarrow} \Psi$ is defined similarly, with all the $\overset{p}{\to}$ requirements used in the definition for $\Psi^u_n \overset{p}{\Rightarrow} \Psi$ replaced by $\overset{a.s.}{\to}$. Note that by Lemma 1.9.2 (ii) of van der Vaart and Wellner (1996), it is true that $\Psi^u_N \overset{p}{\Rightarrow} \Psi$ if and only if every subsequence k_N of N has a further subsequence ℓ_N of k_N such that $\Psi^u_{\ell_N} \overset{a.s.}{\Rightarrow} \Psi$.

Size and Power Properties

We summarize the size and power properties of our test in the following theorem. The regularity conditions and the proof are given in Appendix B.

Theorem 3.1 *Under Assumption 2.1 and Assumptions A.1, A.2 and A.3 described in the appendix, if we reject* $H_{0,ate}^{neg}$ *when* $\widehat{S}_{ate}^{neg} > \hat{c}_{n,neg}(\alpha)$*, then*

(1) under $H_{0,ate}^{neg}$, $\lim_{n\to\infty} P(\widehat{S}_{ate}^{neg} > \hat{c}_{n,ate}^{neg}(\alpha)) \leq \alpha$, *and*

(2) under $H_{1,ate}^{neg}$, $\lim_{n\to\infty} P(\widehat{S}_{ate}^{neg} > \hat{c}_{n,ate}^{neg}(\alpha)) = 1$.

Theorem 3.1 shows that our test for $H_{0,ate}^{neg}$ can control size well asymptotically and is consistent. The asymptotic size is less than or equal to α as a result of adopting the least favorable configuration case (LFC) in constructing the critical value. One can use the moment selection or recentering method to avoid using LFC to improve the power of the test as in Andrews and Shi (2013, 2015) and Donald and Hsu (2016). The implementation of such a test and the result are standard in the literature, so we omit the details. In addition, in this paper we focus on KS type tests, but all results can be extended to Cramér-von Mises type tests fairly easily given the asymptotic results of $\hat{\nu}(\ell)$ and the simulated process $\widehat{\Phi}_n^u(\ell)$.

Note that Lemmas 3.1 and 3.2 can be extended to other classes of functions that satisfy the Pollard's entropy condition defined in (4.2) of Andrews (1994). That is, let $\{f_t;\ t \in \mathcal{T}\}$ be a collection of functions with envelope function $\mathcal{F}$ such that the Pollard's entropy condition holds. Then under suitable conditions, Lemmas 3.1 and 3.2 would still hold with $g_\ell(X_i)Y_i$'s being replaced with f_t's. Andrews (1994) gives examples of classes of functions that satisfy the Pollard's entropy condition and discuss how one can generate classes of functions that satisfy Pollard's entropy condition from sets of functions that are known to satisfy Pollard's entropy condition. See Andrews (1994) for details.

Also notice that the proposed test $H_{0,ate}^{neg}$ can be trivially extended to study the hypotheses

$$H_{0,ate}^{po} : CATE(x) \geq 0,\ \forall\ x \in \mathcal{X}_c,$$
$$H_{1,ate}^{po} : CATE(x) < 0,\ \text{for some}\ \ x \in \mathcal{X}_c.$$

We just need to replace Y_i in the test for $H_{0,ate}^{neg}$ with $-Y_i$.

Adding Discrete Covariates to the Control Set

Although in this section the X_i variable is restricted to be continuous, the tests we propose can be easily adapted to the case in which X_i includes discrete covariates. Without loss of generality, we consider the case in which in addition to X_i, there is one binary variable, X_{di}, taking values in $\{0, 1\}$ in the conditioning set. Let $\mathcal{G}$ be defined as before and let $\mathcal{G}_1 \equiv \{1(X_d = 1) \cdot g_\ell(\cdot) : \ \ell \in \mathcal{L}\}$. We define $\mathcal{G}_0$ similarly. Let $\widetilde{\mathcal{G}} = \mathcal{G}_1 \cup \mathcal{G}_0$. It is straightforward to show that

$$H_{0,ate}^{neg} : CATE(x, x_d) \leq 0, \ \forall \ x \in \mathcal{X}_c \text{ and } x_d = 0, 1$$
$$H_{1,ate}^{neg} : CATE(x, x_d) > 0, \text{ for some } \ x \in \mathcal{X}_c \text{ and } \ x_d = 0, 1.$$

are equivalent to

$$H_{0,ate}^{neg} : \ \nu(\tilde{g}) = E[\tilde{g}(X_i, X_{di})CATE(X_i, X_{di})] \leq 0, \ \forall \ \tilde{g} \in \widetilde{\mathcal{G}},$$
$$H_{1,ate}^{neg} : \ \nu(\tilde{g}) = E[\tilde{g}(X_i, X_{di})CATE(X_i, X_{di})] > 0, \text{ for some } \tilde{g} \in \widetilde{\mathcal{G}}.$$

Therefore, we can carry out the uniform sign test in the same way as is discussed in this section replacing $\mathcal{G}$ with $\widetilde{\mathcal{G}}$, and all results of the test will remain valid.

3.2 Testing if the Treatment Has Any Impact

To test if a policy treatment has any impact on at least some subpopulations defined by covariate values, the null and alternative hypotheses can be formulated as

$$H_{0,ate}^{zero} : CATE(x) = 0, \ \forall \ x \in \mathcal{X}_c,$$
$$H_{1,ate}^{zero} : CATE(x) \neq 0, \text{ for some } \ x \in \mathcal{X}_c. \tag{3.6}$$

Similar to the previous subsection, we can transform the hypotheses in (3.6) to

$$H_{0,ate}^{zero} : \ \nu(\ell) = 0, \ \forall \ \ell \in \mathcal{L},$$
$$H_{1,ate}^{zero} : \ \nu(\ell) \neq 0, \text{ for some } \ell \in \mathcal{L}. \tag{3.7}$$

The KS type test statistic is then defined as

$$\widehat{S}^{zero}_{ate} = \sqrt{nh} \sup_{\ell \in \mathcal{L}} \big|\hat{\nu}(\ell)\big|.$$

and for significance level $\alpha < 1/2$, define the simulated critical value $\hat{c}^{zero}_{n,ate}(\alpha)$ as

$$\hat{c}^{zero}_{n,ate}(\alpha) = \sup\left\{ q \Big| P^u\Big(\sup_{\ell \in \mathcal{L}} \big|\widehat{\Phi}^u_{\nu,n}(\ell)\big| \le q \Big) \le 1 - \alpha \right\},$$

i.e., $\hat{c}^{zero}_{n,ate}(\alpha)$ is the $(1-\alpha)$-th quantile of the simulated null distribution, $\sup_{\ell \in \mathcal{L}} \big|\widehat{\Phi}^u_{\nu,n}(\ell)\big|$.

Let the decision rule be: "Reject $H^{zero}_{0,ate}$ if $\widehat{S}^{zero}_{ate} > \hat{c}^{zero}_{n,ate}(\alpha)$." We summarize the size and power property of our test in the following theorem and the proof is given in Appendix.

Theorem 3.2 *Under Assumption 2.1 and Assumptions A.1, A.2 and A.3, if we reject* $H^{zero}_{0,ate}$ *when* $\widehat{S}^{zero}_{ate} > \hat{c}^{zero}_{n,ate}(\alpha)$*, then*

(1) under $H^{zero}_{0,ate}$, $\lim_{n\to\infty} P(\widehat{S}^{zero}_{ate} > \hat{c}^{zero}_{n,ate}(\alpha)) = \alpha$, *and*

(2) under $H^{zero}_{1,ate}$, $\lim_{n\to\infty} P(\widehat{S}^{zero}_{ate} > \hat{c}^{zero}_{n,ate}(\alpha)) = 1$.

3.3 Testing if the Treatment Effect is Heterogenous

To test if the treatment effect is heterogenous over covariate values, we define the hypotheses as

$$\begin{aligned} &H^{hetero}_{0,ate} : CATE(x) = \gamma,\ \forall\ x \in \mathcal{X}_c \text{ and some } \gamma \in R, \\ &H^{hetero}_{1,ate} :\ H^{hetero}_{0,ate} \text{ does not hold.} \end{aligned} \tag{3.8}$$

If $CATE(x) = \gamma$ for all $x \in \mathcal{X}_c$ for some $\gamma \in R$, then it would hold with $\gamma = ATE = \nu((\mathbf{0}, 1))$ so that

$$\nu(\ell) = E[g_\ell(X_i) CATE(X_i) | Z_i = z] = E[g_\ell(X_i)\nu((\mathbf{0},1))|Z_i = z] = p(\ell) \cdot \nu((\mathbf{0},1))$$

where $p(\ell) = E[g_\ell(X_i)|Z_i = c]$ is the conditional probability of $X_i \in C_\ell$. Therefore, the hypotheses in (3.8) are equivalent to

$$\begin{aligned} &H^{hetero}_{0,ate} :\ \nu_{hetero,ate}(\ell) = \nu(\ell) - \nu((\mathbf{0},1)) \cdot p(\ell) = 0,\ \forall\ \ell \in \mathcal{L}, \\ &H^{hetero}_{1,ate} :\ \nu_{hetero,ate}(\ell) = \nu(\ell) - \nu((\mathbf{0},1)) \cdot p(\ell) \neq 0,\ \text{for some } \ell \in \mathcal{L}. \end{aligned} \tag{3.9}$$

Let the estimator for $p(\ell)$ be $\hat{p}(\ell)$ such that

$$\hat{p}(\ell) = \frac{\sum_{i=1}^n K(\frac{Z_i - c}{h})[S_{n,2} - S_{n,1}(Z_i - c)]g_\ell(X_i)}{\sum_{i=1}^n K(\frac{Z_i - c}{h})[S_{n,2} - S_{n,1}(Z_i - c)]} \equiv \sum_{i=1}^n w_{ni} g_\ell(X_i),$$

where

$$w_{ni} = \frac{K(\frac{Z_i - c}{h})[S_{n,2} - S_{n,1}(Z_i - c)]}{\sum_{i=1}^n K(\frac{Z_i - c}{h})[S_{n,2} - S_{n,1}(Z_i - c)]},$$
$$S_{n,j} = \sum_i K\left(\frac{Z_i - c}{h}\right)(Z_i - c)^j, \text{ for all } j = 0, 1, 2, ..$$

Hence, the estimator for $\nu_{hetero,ate}(\ell)$ would be $\hat{\nu}_{hetero,ate}(\ell) = \hat{\nu}(\ell) - \hat{\nu}((\mathbf{0}, 1)) \cdot \hat{p}(\ell)$ and the test statistic would be

$$\widehat{S}_{ate}^{hetero} = \sqrt{nh} \sup_{\ell \in \mathcal{L}} \left|\hat{\nu}_{hetero,ate}(\ell)\right|.$$

To obtain the influence function representation of $\hat{\Phi}_{n,ate}^{hetero}(\ell) = \sqrt{nh}(\hat{\nu}_{hetero,ate}(\ell) - \nu_{hetero,ate}(\ell))$, similar to Lemma 3.1, we have

$$\left|\sqrt{nh}(\hat{p}(\ell) - p(\ell)) - \sum_{i=1}^n \phi_{p,ni}(\ell)\right| = o_p(1),$$
$$\phi_{p,ni}(\ell) = \sqrt{nh}\Big(w_{ni}(g_\ell(X_i) - p(\ell))\Big)$$

and in the Appendix, we show that

$$\left|\sqrt{nh}(\hat{\nu}_{hetero,ate}(\ell) - \nu_{hetero,ate}(\ell)) - \sum_{i=1}^n \phi_{\nu,ni}^{hetero}(\ell)\right| = o_p(1),$$
$$\phi_{ate,ni}^{hetero}(\ell) = \phi_{\nu,ni}(\ell) - p(\ell)\phi_{\nu,ni}((\mathbf{0}, 1)) - \nu((\mathbf{0}, 1)) \cdot \phi_{p,ni}(\ell). \tag{3.10}$$

Let the simulated process $\widehat{\Phi}_{n,ate}^{hetero,u}(\ell)$ be

$$\widehat{\Phi}_{n,ate}^{hetero,u}(\ell) = \sum_{i=1}^n U_i \cdot \hat{\phi}_{ate,ni}^{hetero}(\ell),$$
$$\hat{\phi}_{ate,ni}^{hetero}(\ell) = \hat{\phi}_{\nu,ni}(\ell) - \hat{p}(\ell)\hat{\phi}_{\nu,ni}((\mathbf{0}, 1)) - \hat{\nu}((\mathbf{0}, 1)) \cdot \hat{\phi}_{p,ni}(\ell),$$
$$\hat{\phi}_{p,ni}(\ell) = \sqrt{nh}\Big(w_{ni}(g_\ell(X_i) - \hat{p}(\ell))\Big).$$

For significance level $\alpha < 1/2$, define the simulated critical value $\hat{c}_{n,ate}^{hetero}(\alpha)$ as

$$\hat{c}_{n,ate}^{hetero}(\alpha) = \sup\left\{q \middle| P^u\Big(\sup_{\ell \in \mathcal{L}} |\widehat{\Phi}_{n,ate}^{hetero,u}(\ell)| \leq q\Big) \leq 1 - \alpha\right\}.$$

Let the decision rule be: "Reject $H_{0,ate}^{hetero}$ if $\widehat{S}_{ate}^{hetero} > \hat{c}_{n,ate}^{hetero}(\alpha)$."

Theorem 3.3 *Under Assumption 2.1 and Assumptions A.1, A.2 and A.3, if we reject* $H_{0,ate}^{hetero}$ *when* $\widehat{S}_{ate}^{hetero} > \hat{c}_{n,ate}^{hetero}(\alpha)$*, then*

(1) under $H_{0,ate}^{hetero}$, $\lim_{n\to\infty} P(\widehat{S}_{ate}^{hetero} > \hat{c}_{n,ate}^{hetero}(\alpha)) = \alpha$*, and*

(2) under $H_{1,ate}^{hetero}$, $\lim_{n\to\infty} P(\widehat{S}_{ate}^{hetero} > \hat{c}_{n,ate}^{hetero}(\alpha)) = 1$.

Notice that this test can also be directly applied to test for first stage heterogeneity in a fuzzy RD model as the selection equation in any fuzzy RD model follows a sharp RD design. This further implies that our proposed test could be used to check the validity of two-sample RD regressions (see He, 2016, for an application), where the outcome of interest is not included in the same dataset as the treatment assignment variable. The validity of the two-sample RD regression is similar to two sample IV regression approaches developed in Angrist and Krueger (1992) and Angrist and Krueger (1995), who assume that the first stage parameters are the same across datasets. In the RD setting, this means that the proportion of compliers needs to be the same across datasets. Since the first stage selection to treatment decision is not observed in both datasets, this condition is infeasible to test directly. One sufficient testable assumption is that the fuzzy RD model has a homogeneous first stage. Our test could be applied to test this sufficient assumption and hence be used as a simple check for the validity of any two-sample RD regression.

4 Testing in Fuzzy RD Design

In this section, we extend the tests to the fuzzy RD design. All tests proposed in this section do not rely on plug-in estimators of LATE or CLATE and are robust to weak inference.

We are interested in testing the following three null hypotheses:

$$H_{0,late}^{neg} : CLATE(x) \leq 0, \ \forall \ x \in \mathcal{X}_c, \tag{4.1}$$

for testing whether the treatment is beneficial for some subpopulations;

$$H_{0,late}^{zero} : CLATE(x) = 0, \ \forall \ x \in \mathcal{X}_c, \tag{4.2}$$

for testing whether the treatment has any impact on at least some subpopulations; and

$$H_{0,late}^{hetero} : CLATE(x) = \tau, \ \forall \ x \in \mathcal{X}_c \ and \ some \ \tau \in R, \tag{4.3}$$

for testing whether the treatment effect is heterogeneous. Recall that

$$CLATE(x) = \frac{\lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z]}{E[T_i(1) - T_i(0)|X_i = x, Z_i = c]}$$

and Assumption 2.2(iii) requires that $E[T_i(1) - T_i(0)|X_i = x, Z_i = c] > 0$ for all $x \in \mathcal{X}_c$. Therefore, $CLATE(x) \leq 0$ if and only if $\lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z] \leq 0$, and $CLATE(x) = 0$ if and only if $\lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z] = 0$. The first two hypotheses $H_{0,late}^{neg}$ and $H_{0,late}^{zero}$ hold if and only if $\lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z]$, is uniformly negative or uniformly zero, respectively, implying that these two hypothesis can be tested by applying the procedures developed for testing $H_{0,ate}^{neg}$ and $H_{0,ate}^{zero}$ in Section 3.

For the third hypotheses, the null hypothesis $CLATE(x) = \tau$ holds for all $x \in \mathcal{X}_c$ for some $\tau \in R$ if and only if $CLATE(x) = LATE$ for all $x \in \mathcal{X}_c$. Let

$$\mu(\ell) = E[g_\ell(X_i)(T_i(1) - T_i(0))|Z_i = c].$$

Then it is clear that $LATE = \nu((\mathbf{0}, 1))/\mu((\mathbf{0}, 1))$ and $\nu(\ell)/\mu(\ell)$ is the local average treatment effect for those people with $X_i \in C_\ell$. In the Appendix, we show that the null hypothesis in (4.3) is equivalent to

$$H_{0,late}^{hetero} : \ \nu_{hetero,late}(\ell) = \nu(\ell) \cdot \mu((\mathbf{0}, 1)) - \nu((\mathbf{0}, 1)) \cdot \mu(\ell) = 0, \ \forall \ \ell \in \mathcal{L}. \tag{4.4}$$

It is straightforward to see that $\mu(\ell)$ is identified by

$$\mu(\ell) = \lim_{z \searrow c} E[g_\ell(X_i)T_i|Z_i = z] - \lim_{z \nearrow c} E[g_\ell(X_i)T_i|Z_i = z].$$

Let $\hat{\mu}(\ell)$ be the estimator for $\mu(\ell)$ that is defined in the same way as $\hat{\nu}(\ell)$ except that we replace the Y_i's with T_i's. Let $\nu_{hetero,late}(\ell)$ be estimated by $\hat{\nu}(\ell) \cdot \hat{\mu}((\mathbf{0}, 1)) - \hat{\nu}((\mathbf{0}, 1)) \cdot \hat{\mu}(\ell)$. The test statistic for $H_{0,late}^{hetero}$ is

$$\widehat{S}_{late}^{hetero} = \sqrt{nh} \sup_{\ell \in \mathcal{L}} \left| \hat{\nu}_{hetero,late}(\ell) \right|.$$

Let $\hat{\phi}_{\mu,ni}(\ell)$ be the estimated influence function for $\sqrt{nh}(\hat{\mu}(\ell) - \mu(\ell))$ that is defined in the same way as $\hat{\phi}_{\mu,ni}(\ell)$ except that we replace Y_i's with T_i's. Let the simulated process $\widehat{\Phi}^{hetero,u}_{n,late}(\ell)$ be

$$\widehat{\Phi}^{hetero,u}_{n,late}(\ell) = \sum_{i=1}^{n} U_i \cdot \hat{\phi}^{hetero}_{late,ni}(\ell),$$

$$\hat{\phi}^{hetero}_{ate,ni}(\ell) = \hat{\mu}((\mathbf{0},1)) \cdot \hat{\phi}_{\nu,ni}(\ell) + \hat{\nu}(\ell) \cdot \hat{\phi}_{\mu,ni}((\mathbf{0},1)) - \hat{\nu}((\mathbf{0},1)) \cdot \hat{\phi}_{\mu,ni}(\ell) - \hat{\mu}(\ell) \cdot \hat{\phi}_{\nu,ni}((\mathbf{0},1)).$$

For significance level $\alpha < 1/2$, define the simulated critical value $\hat{c}^{hetero}_{n,late}(\alpha)$ as

$$\hat{c}^{hetero}_{n,late}(\alpha) = \sup\left\{ q \Big| P^u\Big(\sup_{\ell \in \mathcal{L}} \big|\widehat{\Phi}^{hetero,u}_{n,late}(\ell)\big| \leq q \Big) \leq 1 - \alpha \right\}.$$

Finally, the decision rule would be: "Reject $H^{hetero}_{0,late}$ if $\widehat{S}^{hetero}_{late} > \hat{c}^{hetero}_{n,late}(\alpha)$." We omit the details of the size and power properties for brevity.

5 Simulations

In this section we carry out Monte Carlo simulations. First, we use four data generating processes (DGPs) to investigate the small sample size and power performance of the proposed tests. Then we use another three DGPs to demonstrate the size distortion of the two naive methods for heterogeneity analysis that are popular in applied RD literature: the interaction term method and the subsample regression method. The first four DGPs are described below.

DGP 1: Sharp RD, Homogeneous Zero Effect

$$Z \sim 2Beta(2,2) - 1;\ X \sim U[0,1];\ T = 1(Z \geq 0);\ u \sim N(0,1);$$

$$Y = -0.708 + 0.607X + 0.481Z + 0.441XZ + 0.038Z^2 - 0.085X^2 + 0.1u;$$

DGP 2: Sharp RD, Heterogeneous Treatment Effect

$$Z \sim 2Beta(2,2) - 1;\ X \sim U[0,1];\ T = 1(Z \geq 0);\ u \sim N(0,1);$$

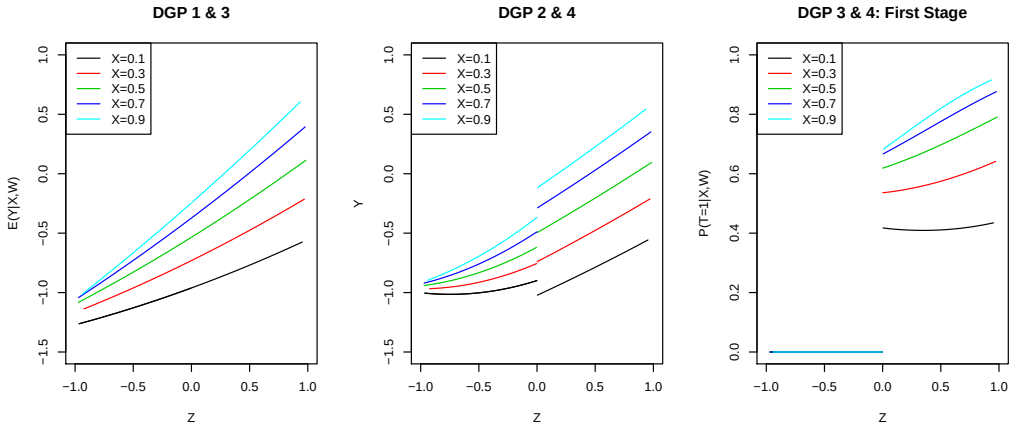
$$Y = \begin{cases} -0.753 + 0.905X + 0.506Z + 0.224XZ + 0.022Z^2 - 0.225X^2 + 0.1u & \text{if } Z \geq 0 \\ -0.577 + 0.011X + 0.634Z + 0.131XZ + 0.233Z^2 + 0.255X^2 + 0.1u & \text{if } Z < 0 \end{cases}$$

DGP 3: Fuzzy RD, Homogeneous Zero Effect

Let DGP 3 be the same as DGP 1 except that

$$T = \begin{cases} 1(-0.140 + 1.307X + 0.957Z - 0.074XZ - 0.223Z^2 - 0.611X^2 + u > 0) & \text{if } Z \geq 0 \\ 0 & \text{if } Z < 0 \end{cases}$$

Figure 1: The Data Generating Processes



Note: The DGPs are estimated from the data of the empirical section. To get the model in DGP 1 (left graph), we first rescale the running variable (i.e. transition score) in that dataset to $[-1, 1]$ to match the support of the generated X variable and then regress the outcome (i.e. score in Baccalaureate exam) on the running variable, the additional control of interest (i.e. the admission score cut-off), as well as their interaction term and second order polynomial terms. To get the model in DGP 2 (middle), we fit the same regression model separately for the subsamples to the left and the right of the cutoff value (i.e. 0). DGPs 3 and 4 share the same date generating processes for (X, Y, Z) with DGPs 1 and 2, respectively, while modeling an additional layer of first stage treatment decision. To get the model for the first stage (right graph), we run two separate probit regressions of the treatment status (i.e. a dummy for attending a more selective high school) on the running variable, the additional control of interest, their interaction term and second order polynomial terms with the subsamples to the left and the right of zero, the threshold.

DGP 4: Fuzzy RD, Heterogeneous Treatment Effect

Let DGP 4 be the same as DGP 2 except that

$$T = \begin{cases} 1(-0.140 + 1.307X + 0.957Z - 0.074XZ - 0.223Z^2 - 0.611X^2 + u > 0) & \text{if } Z \geq 0 \\ 0 & \text{if } Z < 0 \end{cases}$$

We use these four DGPs to study the performance of the uniform sign test with the null hypothesis $H_0 : CATE(x) \leq 0, \ \forall x \in [0, 1]$ and the heterogeneity test with the null hypothesis $H_0 : CATE(x) = ATE, \ \forall x \in [0, 1]$. We also report the results for the standard mean test $H_0 : ATE = 0$ as a benchmark comparison. We omit the results from the uniform significance test (i.e. $H_0 : CATE(x) = 0, \ \forall x \in [0, 1]$) as they are similar to those for the uniform sign test.

DGPs 1 and 2 follow the sharp RD design. The functional forms are estimated from

the empirical example (details described in the footnote of Figure 1). Figure 1 visually represents the data generating processes. Under DGPs 1 and 3, all three tests of interest are under the null, and the uniform sign test is under the least favorable condition. Under DGPs 2 and 4, all three tests are under the alternative with the treatment effect varying with the additional control X.

Four different sample sizes ($n = 1000$, 2000, 4000 and 8000) are used and 1000 samples are simulated for each DGP and sample size combination. With each simulated dataset, tests are carried out with three different bandwidths while the bootstrap critical values are calculated from 1000 bootstrap simulations each time. The three bandwidths are selected according to the formula $h_{IK} \times n^{1/5-1/c}$, where h_{IK} is the optimal bandwidth following Imbens and Kalyanaraman (2012) (IK), and c is the undersmoothing constant. In Table 1, we report results with $c = 4.5, 4.75$ and 5. When $c < 5$, the bandwidth undersmooths and satisfies the condition in Assumption A.2. When $c = 5$, the bandwidth reduces to the IK bandwidth which is used for comparison purposes. The cubes defined in Equation (3.2) have side-lengths $1/(2q)$ for $q = 1, ..., Q$. Simulations reported in Table 1 uses $Q = 3$, which includes a total of 12 overlapping intervals (since the dimension of X is one). Among the 12 intervals, 2 have length 1/2, 4 have length 1/4, and 6 have length 1/6). When $n = 1000$, the average bandwidth ranges from around 0.35 to 0.40, which means that the average effective sample size (i.e. those with $X_i \in C_\ell$) for each local linear regression (on one-side of the RD cut-off) is around 30 when the smallest cubes are used. Robustness checks with $Q = 5$ are reported in Table 2, where the average effective sample size for each local linear regression is around 19 when the smallest cubes are used and $n = 1000$. We see from the tables that our tests control size very well and have good power performances. When the bandwidth is not undersmoothed and the IK bandwidth ($c = 5$) is used, the tests have some very slight tendencies of overrejection. In the empirical application, we use a benchmark bandwidth with $c = 4.5$ but also include other bandwidth selection rules for robustness checks. The result is again very robust to the bandwidth choice. Last but not least, the size and power performance of the proposed tests are not sensitive to the variation in Q.

Next, we compare the proposed tests with the interaction term method and the sub-sample regression method that are commonly used in the RD literature. We demonstrate

Table 1: Small Sample Performance of Proposed Tests, $Q = 3$

	$H_0 : ATE = 0$			$H_0 : CATE(x) \leq 0$			$H_0 : CATE(x) = ATE$		
	c=4.5	c=4.75	c=5	c=4.5	c=4.75	c=5	c=4.5	c=4.75	c=5
DGP 1: Sharp RD, Homogeneous Zero Effect									
n=1000	0.060	0.063	0.061	0.061	0.059	0.062	0.063	0.067	0.058
n=2000	0.059	0.063	0.059	0.064	0.064	0.065	0.053	0.053	0.052
n=4000	0.066	0.067	0.065	0.062	0.057	0.055	0.052	0.055	0.060
n=8000	0.055	0.058	0.059	0.056	0.062	0.065	0.050	0.056	0.057
DGP 2: Sharp RD, Heterogeneous Treatment Effect									
n=1000	0.395	0.417	0.427	0.151	0.175	0.205	0.117	0.124	0.127
n=2000	0.633	0.662	0.688	0.428	0.500	0.554	0.190	0.205	0.221
n=4000	0.864	0.891	0.908	0.835	0.880	0.916	0.322	0.355	0.389
n=8000	0.972	0.979	0.985	0.995	0.998	1.000	0.589	0.632	0.677
DGP 3: Fuzzy RD, Homogeneous Zero Effect									
n=1000	0.060	0.063	0.061	0.061	0.059	0.062	0.057	0.057	0.058
n=2000	0.059	0.063	0.059	0.064	0.064	0.065	0.055	0.055	0.049
n=4000	0.066	0.067	0.065	0.062	0.057	0.055	0.049	0.054	0.058
n=8000	0.055	0.058	0.059	0.056	0.062	0.065	0.050	0.057	0.062
DGP 4: Fuzzy RD, Heterogeneous Treatment Effect									
n=1000	0.395	0.417	0.427	0.151	0.175	0.205	0.109	0.112	0.117
n=2000	0.633	0.662	0.688	0.428	0.500	0.554	0.156	0.163	0.182
n=4000	0.864	0.891	0.908	0.835	0.880	0.916	0.237	0.273	0.293
n=8000	0.972	0.979	0.985	0.995	0.998	1.000	0.448	0.489	0.523

Table 2: Small Sample Performance of Proposed Tests, $Q = 5$

	$H_0 : ATE = 0$			$H_0 : CATE(x) \leq 0$			$H_0 : CATE(x) = ATE$		
	c=4.5	c=4.75	c=5	c=4.5	c=4.75	c=5	c=4.5	c=4.75	c=5
DGP 1: Sharp RD, Homogeneous Zero Effect									
n=1000	0.060	0.063	0.061	0.061	0.060	0.059	0.068	0.065	0.064
n=2000	0.059	0.063	0.059	0.066	0.062	0.066	0.052	0.054	0.055
n=4000	0.066	0.067	0.065	0.057	0.055	0.055	0.048	0.053	0.056
n=8000	0.055	0.058	0.059	0.058	0.062	0.061	0.051	0.047	0.054
DGP 2: Sharp RD, Heterogeneous Treatment Effect									
n=1000	0.395	0.417	0.427	0.142	0.165	0.184	0.116	0.115	0.128
n=2000	0.633	0.662	0.688	0.401	0.471	0.518	0.186	0.194	0.213
n=4000	0.864	0.891	0.908	0.816	0.865	0.902	0.296	0.338	0.370
n=8000	0.972	0.979	0.985	0.994	0.998	1.000	0.570	0.617	0.661
DGP 3: Fuzzy RD, Homogenous Zero Effect									
n=1000	0.060	0.063	0.061	0.061	0.060	0.059	0.063	0.058	0.058
n=2000	0.059	0.063	0.059	0.066	0.062	0.066	0.056	0.057	0.053
n=4000	0.066	0.067	0.065	0.057	0.055	0.055	0.044	0.053	0.053
n=8000	0.055	0.058	0.059	0.058	0.062	0.061	0.053	0.049	0.054
DGP 4: Fuzzy RD, Heterogeneous Treatment Effect									
n=1000	0.395	0.417	0.427	0.142	0.165	0.184	0.110	0.113	0.117
n=2000	0.633	0.662	0.688	0.401	0.471	0.518	0.151	0.162	0.175
n=4000	0.864	0.891	0.908	0.816	0.865	0.902	0.227	0.263	0.277
n=8000	0.972	0.979	0.985	0.994	0.998	1.000	0.432	0.473	0.505

that when the model is misspecified, the interaction term method severely over rejects. Further, when the sample size is small, the subsample regression method shows sizable over-rejection due to weak inference.

We consider the following three DGPs with a control parameter η varying in the interval $[-1, 1]$. With this set of DGPs we compare three tests for treatment effect heterogeneity: 1) the proposed heterogeneity test (Hetero) that is also studied in DGP 1-3, 2) a naive heterogeneity test based on an RD regression with interaction term (Hetero-INT), and 3) a naive heterogeneity test based on subsample RD regressions (Hetero-SUB). The Hetero-INT test is carried out by testing the slope coefficient on the interaction term $X1(Z > 0)$ in the linear regression of Y on X, Z, $1(Z > 0)$, $X1(Z > 0)$, and $Z1(Z > 0)$, using data inside the estimation window determined by the bandwidth. The Hetero-SUB test is carried out by testing whether the local linear regression of Y on Z for any of the five subsamples with $X = [0, 0.2], X = (0.2, 0.4], X = (0.4, 0.6], X = (0.6, 0.8], X = (0.8, 1]$ is different from the true average treatment effect. The Hetero-SUB adjusts for multiple testing using the Bonferroni method and plugs in the unknown true ATE for computational simplicity.

DGP 5: Sharp RD, Homogeneous Zero Effect

$$Z \sim 2Beta(2,2) - 1;\ X \sim U[0,1];\ u \sim N(0,1);$$
$$Y = -0.708 + 0.607X + 0.481Z + \eta(0.441XZ + 0.038Z^2 - 0.085X^2) + 0.1u;$$

DGP 6: Fuzzy RD, Homogeneous Zero Effect

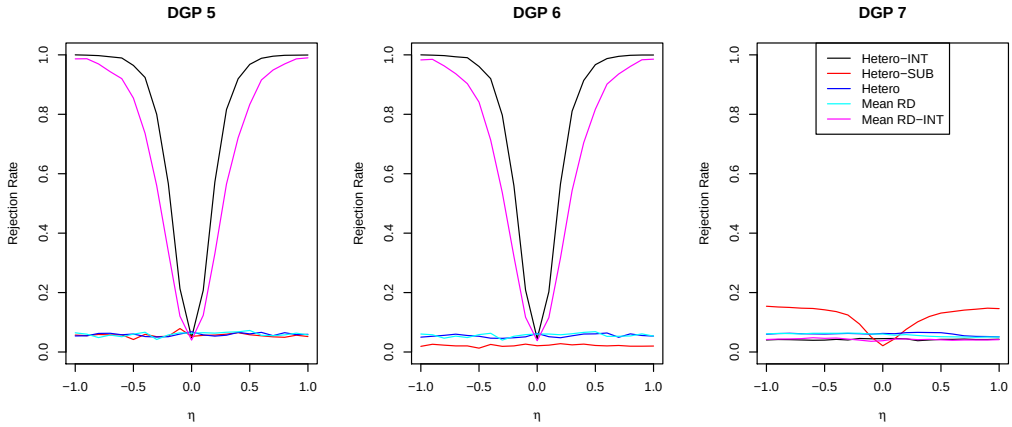
Let DGP 6 be the same as DGP 5 except that

$$T = \begin{cases} 1(0.357 + 0.921Z - 0.240Z^2 + u > 0) & \text{if } Z \geq 0 \\ 0 & \text{if } Z > 0 \end{cases}$$

DGP 7: Fuzzy RD, Homogeneous Effect

$$Z \sim 2Beta(2,2) - 1;\ X \sim U[0,1];\ u \sim N(0,1);$$
$$Y = \begin{cases} (-0.708 + \eta) + 0.607X + 0.481Z + 0.1u & \text{if } Z \geq 0 \\ -0.708 + 0.607X0.481Z + u & \text{if } Z < 0 \end{cases}$$
$$T = \begin{cases} 1(0.357 + 0.921Z - 0.240Z^2 + u > 0) & \text{if } Z \geq 0 \\ 0 & \text{if } Z > 0 \end{cases}$$

Figure 2: Performance of Naive and Proposed Testing Methods, $n = 1000$



In DGPs 5 and 6, the treatment effect is homogenous and zero, and the control parameter η determines the degree of model misspecification. When $\eta = 1$, DGPs 5 and 6 reduce to DGPs 1 and 3. When $\eta = 0$, the linear regression model with the additional interaction term is correctly specified. As η deviates from 0, the model becomes increasingly misspecified. The left and middle graphs of Figure 2 summarize the size control of all three tests. The Hetero-INT test is correctly specified and controls size at 5% only when $\eta = 0$. On the other hand, the proposed test Hetero and the subsample test Hetero-SUB control size well irrespective of the value of η. Besides the three tests for treatment effect heterogeneity, we also report in Figure 2 two tests for the standard ATE/LATE estimates. Test Mean RD is the standard t-test following the classic local linear estimation method. Test Mean RD-INT is the t-test for the slope coefficient of $1(Z > 0)$ in the linear regression with the interaction term. We see that Mean RD-INT over-rejects as well when the model is misspecified. This raises a question about the empirical strategy of adding interaction terms into RD regression models for analyses of heterogeneity.

In DGP 7 the treatment effect is again homogeneous, but the control parameter η now determines the size of the treatment effect. The mean tests Mean RD and Mean RD-INT test the null hypothesis that the ATE is equal to the true value. The heterogeneity tests Hetero, Hetero-INT and Hetero-SUB again test the null hypothesis of

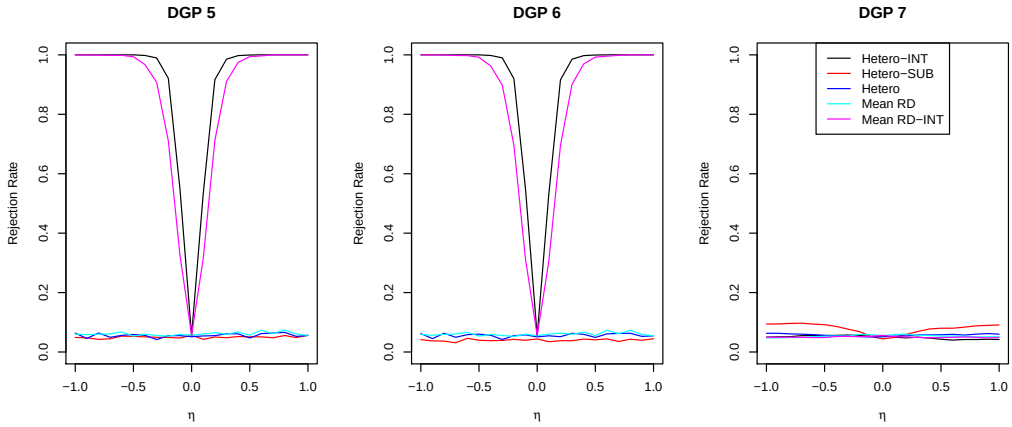
homogeneous treatment effect. For DGP 7, Test Hetero-INT and Mean RD-INT have good size properties because the linear regression model with interaction term is in fact correctly specified. What is interesting is that the Hetero-SUB test starts to over-reject when the true effect deviates from zero. This problem is due to weak inference. As is discussed in Feir, Lemieux, and Marmer (2015), when the proportion of compliers is small, the standard t-test for LATE over-rejects unless the null hypothesis is imposed in the standard error calculation. This explains why the infeasible Hetero-SUB test, after correcting for multiple testing, controls size properly when $\eta = 0$ but has sizable over-rejection when η deviates away from zero. For these three DGPs, the bandwidth is selected following the formula $h_{IK} \times n^{1/5}(n/5)^{-1/4.5}$, where h_{IK} is the optimal IK bandwidth for the whole sample and $n/5$ is used, since the subsample regression method involves five subsamples with equal sample size. If the bandwidth is selected following the undersmoothed IK formula for the full sample, or if the proportion of compliers in the first stage is lower (notice it is around 55% in DGP 7 but can be substantially lower in empirical applications), the over-rejection problem for the Hetero-SUB test is even worse.

Figure 3 repeats the simulation experiment reported in Figure 2 with $n = 4000$. We see that the over-rejection problem is mitigated for the Hetero-SUB test. The main reason is that when the proportion of compliers is fixed, the weak inference problem is a small sample problem. However, the over-rejection problem of the Hetero-INT test does not improve with sample size because the root of over-rejection for that test is model misspecification.

6 The Heterogeneous Effect of Going to a Better High School

In Romania, a typical elementary school student takes a nationwide test in the last year of elementary school (8th grade) and applies to a list of high schools and tracks. The admission decision is entirely dependent on the student's transition score, an average of the student's performance on the nationwide test and grade point average, as well as a student's preference for schools. A student with a transition score above a school's

Figure 3: Performance of Naive and Proposed Testing Methods, $n = 4000$



cutoff is admitted to the most selective school for which he or she qualifies. Pop-Eleches and Urquiola (2013) use an administrative dataset from Romania to study the impact of attending a more selective high school. They find that attending a better school significantly improves a student's performance on the Baccalaureate exam, although the effect is not statistically significant if the more selective high school has a low admission score cut-off. A marginal student attending a more selective high school is also more likely to face negative peer interactions and perceive himself as weak. Shen and Zhang (2015) conduct a distributional analysis using the same dataset and find that the insignificant result among selective schools with a low admission score is due to a heterogeneous distributional effect – a marginal student attending a selective school with lower admission score cut-offs is more likely to have both relatively low scores and relatively high scores on the Baccalaureate exam. In this section, we revisit Pop-Eleches and Urquiola (2013) and investigate the treatment effect heterogeneity of attending a better high school based on the admission score cut-off.

Following Pop-Eleches and Urquiola (2013), we use the RD approach to identify and estimate the effect of attending a higher-ranked school. In this study, we restrict our attention to two-school towns[4] because we notice that score cutoffs within a town are

[4]Pop-Eleches and Urquiola (2013) also report results of all towns including towns with more than two high schools. In such towns, there is more than one selective high school. For example, if a town has three high schools, then there is one school that is not selective and two selective schools with different admission cutoffs.

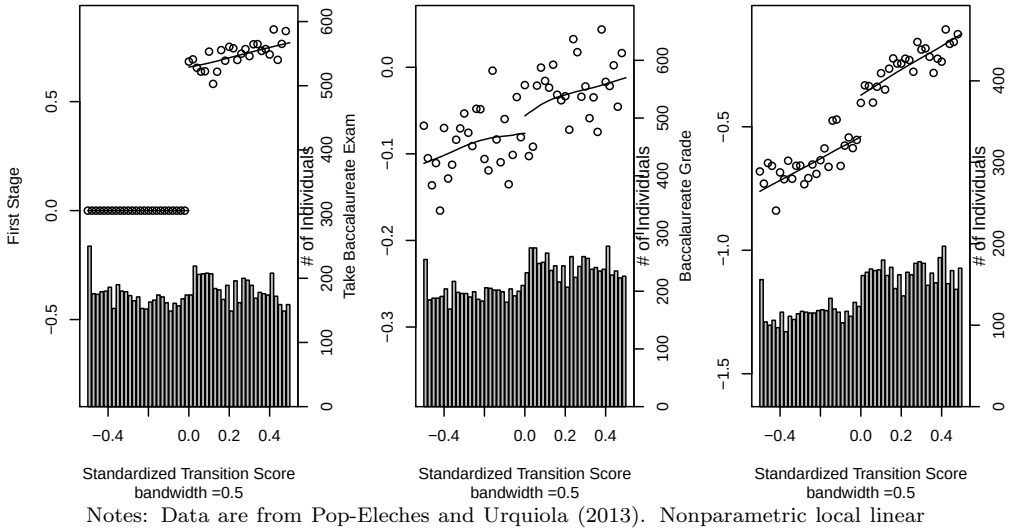
often quite close and we are concerned about introducing estimation bias[5] from having more than one discontinuity within the estimation window. Figure 4 summarizes the the RD regression with this dataset. In all three graphs, the x-axis represents the running variable (i.e. Z_i), which is a student's standardized transition score subtracting the school admission cut-off. The y-axis in the left graph represents the treatment dummy (i.e. T_i), or whether a student attends a more selective school. The y-axis in the middle and right graphs represent two different outcome variables (i.e. Y_i), the demeaned probability of a student taking the Baccalaureate exam and the demeaned Baccalaureate exam grade among exam-takers, respectively. Both outcome variables are demeaned by subtracting the school fixed effects following Pop-Eleches and Urquiola (2013). The left graph shows that the proportion of compliers is around 65%. Since there are no always-takers in this analysis, the reduced-form RD regressions reported in the middle and right graphs represent the intent-to-treat effect of going to a better high school. The middle and the right graphs both reveal a jump in the average outcome at the discontinuity point, with the jump for the average exam taking rate being far noisier than that for the average exam grade among exam takers.

Following Pop-Eleches and Urquiola (2013), we investigate the treatment effect heterogeneity among schools with different admission score cut-offs. But instead of grouping schools by terciles of score cut-offs, we apply the our proposed (uniform) tests which do not require arbitrary discretization of the continuous control variable. In contrast to the results in Pop-Eleches and Urquiola (2013), we find a clear signal that attending a more selective high school has a significant effect on the exam-taking rate for at least some subpopulations, as well as strong evidence supporting treatment effect heterogeneity.

Figure 5 reports the testing results for the two fuzzy RD regressions. The test uses the triangular kernel, the undersmoothed IK bandwidth defined in the simulation section with undersmoothing constant $c = 4.5$ and the cubes defined in Equation 3.2 with $q = 1, ..., Q = 10, .$ Critical values are calculated using the multiplier bootstrap with

[5]In fact, it is easy to prove that if both potential outcomes monotonically increase with the running variable and jumps positively at all discontinuity points (a proper assumption with this application), having extra discontinuity points within the estimation window can severely downward bias the ATE estimator.

Figure 4: Pooled Regression Discontinuity Analysis



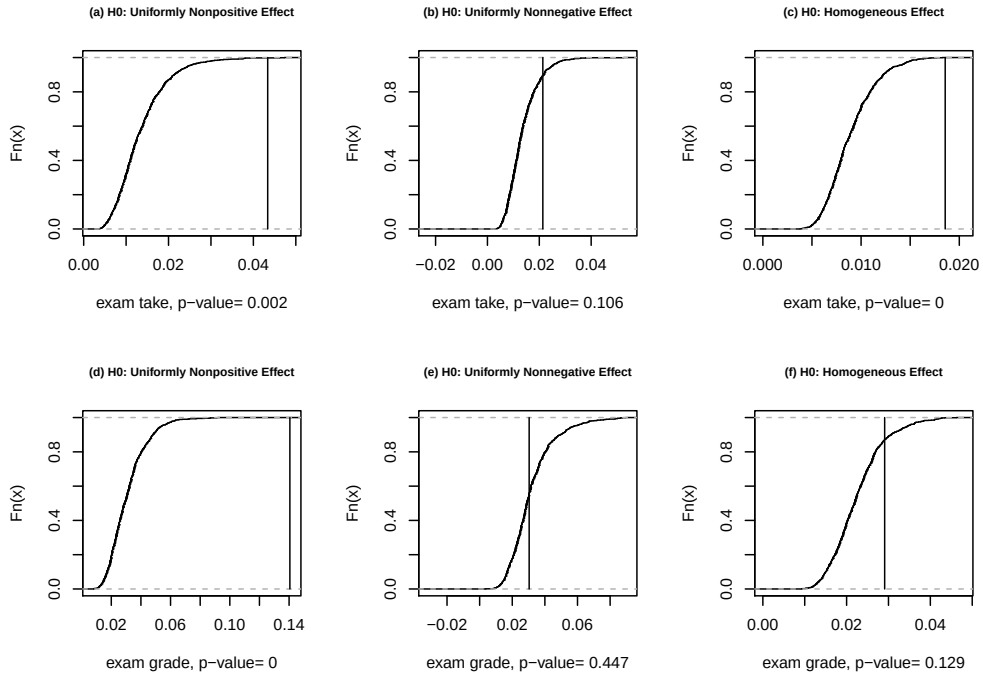
Notes: Data are from Pop-Eleches and Urquiola (2013). Nonparametric local linear estimations are conducted using a triangular kernel. The bar chart reports the histogram of the standardized running variable, while the circles and lines report the average score outcome within each bin and the results of the local linear regression conducted using a triangular kernel. The bandwidth is set to 0.5 for all three graphs for the purpose of data illustration and cross-comparison.

1000 bootstrap simulations. The top panel reports results for the exam-taking outcome. The three graphs correspond to results for the uniform negative, uniform positive, and heterogeneity tests, respectively. The lower panel reports results for the exam grade outcome. For all graphs, the step function represents the distribution of the simulated test statistics under the null, while the vertical line represents the test statistic obtained from the data. The p-value of each test is reported underneath the x-axis.

The test results are interesting. As shown in Figure 4 and by the first two numbers of Table 3, the average effect of attending a better school on the probability of a student taking the Baccalaureate exam is noisy and, in fact, insignificant. However, the top left graph shows that attending a better school certainly increases the probability of a student taking the Baccalaureate exam for some subpopulations. The top middle graph fails to reject the uniform nonnegative sign, but the p-value is very close to 10%. This indicates that the insignificant LATE effect may come from the cancelation of negative and positive effects among different groups of the population. Still for the effect on the exam-taking rate, the top right graph clearly rejects the null hypothesis of treatment effect homogeneity, leading to the conclusion that the effect depends on the admission score cut-off, or how selective a school is. In conclusion, our proposed heterogeneity tests reveal substantial heterogeneity in the effect of attending a better school on the probability of a student taking the Baccalaureate exam that was not picked up by the classic mean RD regression approach. On the other hand, the bottom panel of Figure 5 confirms the positive effect of attending a better school on the Baccalaureate exam grade and there is not enough evidence to conclude that there is treatment effect heterogeneity.

Table 3 reports the above testing results, as well as testing results for the first stage. Table 4 reports robustness checks of the above results with different bandwidth and cube choices. Panel A again uses undersmoothed bandwidth based on Imbens and Kalyanaraman (2012) while Panel B uses undersmoothed bandwidth with the formula $h_{CCT} \times n^{1/5-1/c}$ where h_{CCT} is the robust bandwidth proposed in Calonico, Cattaneo, and Titiunik (2014). The empirical findings by the proposed tests are consistent irrespective of bandwidth and cube choices.

Figure 5: Testing For Treatment Effect Heterogeneity



Notes: Data are from Pop-Eleches and Urquiola (2013). Nonparametric local linear estimations are conducted using a triangular kernel and the undersmoothed IK bandwidth as is described in the simulation section.

Table 3: Benchmark Testing Results

H_0	$ATE = 0$		$CATE(x) \leq 0$		$CATE(x) \geq 0$		$CATE(x) = ATE$	
h	c=4.5	c=4.75	c=4.5	c=4.75	c=4.5	c=4.75	c=4.5	c=4.75
	Treatment Effect							
Took Exam	0.195	0.158	0.002	0.002	0.106	0.117	0.000	0.001
Exam Grade	0.000	0.000	0.000	0.000	0.447	0.425	0.140	0.120
	First Stage							
Full Sample	0.000	0.000	0.000	0.000	1.000	1.000	0.064	0.036
Exam-takers	0.000	0.000	0.000	0.000	1.000	1.000	0.128	0.103

Table 4: Robustness Checks

H_0	$ATE = 0$		$CATE(x) \leq 0$		$CATE(x) \geq 0$		$CATE(x) = ATE$	
h	c=4.5	c=4.75	c=4.5	c=4.75	c=4.5	c=4.75	c=4.5	c=4.75
Panle A	First Stage							
Full Sample (Q=10)	0.000	0.000	0.000	0.000	1.000	1.000	0.064	0.036
Full Sample (Q=20)	0.000	0.000	0.000	0.000	1.000	1.000	0.064	0.036
Exam-takers (Q=10)	0.000	0.000	0.000	0.000	1.000	1.000	0.128	0.103
Exam-takers (Q=20)	0.000	0.000	0.000	0.000	1.000	1.000	0.129	0.103
	Treatment Effect							
Took Exam (Q=10)	0.195	0.158	0.002	0.002	0.106	0.117	0.000	0.001
Took Exam (Q=20)	0.195	0.158	0.002	0.002	0.106	0.117	0.000	0.001
Exam Grade (Q=10)	0.000	0.000	0.000	0.000	0.447	0.425	0.140	0.120
Exam Grade (Q=20)	0.000	0.000	0.000	0.000	0.469	0.446	0.141	0.121
Panel B	First Stage							
Full Sample (Q=10)	0.000	0.000	0.000	0.000	1.000	1.000	0.046	0.069
Full Sample (Q=20)	0.000	0.000	0.000	0.000	1.000	1.000	0.046	0.069
Exam-takers (Q=10)	0.000	0.000	0.000	0.000	1.000	1.000	0.202	0.172
Exam-takers (Q=20)	0.000	0.000	0.000	0.000	1.000	1.000	0.206	0.177
	Treatment Effect							
Took Exam (Q=10)	0.382	0.305	0.002	0.001	0.065	0.078	0.001	0.001
Took Exam (Q=20)	0.382	0.305	0.002	0.001	0.065	0.078	0.001	0.001
Exam Grade (Q=10)	0.000	0.000	0.001	0.000	0.598	0.540	0.080	0.143
Exam Grade (Q=20)	0.000	0.000	0.001	0.000	0.628	0.566	0.082	0.146

7 Conclusion

In this paper, we propose (uniform) tests for treatment effect heterogeneity under both sharp and fuzzy RD designs. Compared with other methods currently adopted in applied RD studies, our tests have the advantage of being both fully nonparametric and robust to weak inference. Monte Carlo simulations show that our tests have very good small sample performance. We apply our methods to a dataset from Romania and discover that the treatment effect of attending a better school on the probability of a student taking the Baccalaurate exam is heterogenous, but that the effect on the Baccalaurate exam grade is homogenous. One interesting question for future study is to extend the proposed testing procedure to examine the heterogeneity in the distributional treatment effect among subpopulations defined by covariate values, as researchers in the treatment effect literature are often interested in analyzing treatment effect heterogeneity along outcome distributions (Bitler, Gelbach, and Hoynes, 2008; Hsu, 2015; Shen and Zhang, 2015; Bitler, Hoynes, and Domina, 2016, etc.).

APPENDIX

In the Appendix, we give Assumptions A.1, A.2 and A.3, and the proofs of Lemmas 3.1 and 3.2 in Section A. The proofs of Theorems 3.1, 3.2 and 3.3, and the proof of the equivalence of Equations (3.1) and (3.3) are given in Section B.

A Regularity Conditions and Proofs of Lemmas

We introduce more notation. Let $f_z(z)$ denote the probability density function (pdf) of Z, $f_{xz}(x,z)$ denote the conditional pdf of X and $Z=z$ and $\mu_d(x,z)=E[Y(d)|X=x,Z=z]$. Let f' and f" denote the first and second derivatives of function f. Let for any $\delta>0$, $\mathcal{N}_{\delta,z}(c)=\{z||z-c|\leq\delta\}$ denote a neighborhood of z around $Z=c$. Let $\sigma_d^2(x,z)=Var(Y(d)|X=x,Z=z)$ and $\mathcal{X}_z$ denote the support of X conditioning on $Z=z$. We make the following assumptions.

Assumption A.1 *Assume that there exists $\delta>0$ such that*

(i) $\mathcal{X}_z=\mathcal{X}_c$ *for all* $z\in\mathcal{N}_{\delta,z}(c)$,

(ii) $f_z(z)$ *is twice continuously differentiable in* z *on* $\mathcal{N}_{\delta,z}(c)$,

(iii) $f_z(z)$ *is bounded away from zero on* $\mathcal{N}_{\delta,z}(c)$,

(iv) for each $x\in\mathcal{X}_c$, $f_{xz}(x,z)$ *is twice continuously differentiable in* z *on* $\mathcal{N}_{\delta,z}(c)$,

(v) $|\partial^2 f_{xz}(x,z)/\partial z\partial z|$ *is uniformly bounded on* $x\in\mathcal{X}_c$ *and* $z\in\mathcal{N}_{\delta,z}(c)$,

(vi) for $d=0$ *and* 1 *and for each* $x\in\mathcal{X}_c$, $\mu_d(x,z)=E[Y(d)|X=x,Z=z]$ *is twice continuously differentiable in* z *on* $\mathcal{N}_{\delta,z}(c)$,

(vii) for $d=0$ *and* 1, $|\partial^2\mu_d(x,z)/\partial z\partial z|$ *is uniformly bounded on* $x\in\mathcal{X}_c$ *and* $z\in\mathcal{N}_{\delta,z}(c)$,

(viii) for $d=0$ *and* 1, $E[Y^4|Z=z]\leq M$ *for some* $M>0$ *for all* $z\in\mathcal{N}_{\delta,z}(c)$, *and*

(ix) for $d=0$ *and* 1, $\sigma_d^2(x,z)$ *is uniformly bounded on* $x\in\mathcal{X}_c$ *and* $z\in\mathcal{N}_{\delta,z}(c)$.

Assumption A.1(i) is assumed for notational simplicity. We can allow $\mathcal{X}_z$ to depend on z and the theory will be the same, but it is more tedious in terms of notation. Assumption A.1(ii)-(vi) are standard in nonparametric estimation. Assumptions A.1(vii) is needed to show that the bias terms of the $\hat{\nu}(\ell)$ are asymptotically negligible uniformly over $\ell\in\mathcal{L}$. Assumption A.1$(viii)$ and (ix) are assumed so the covariance kernel estimator of the limiting process is uniformly consistent which is needed to show the validity of the multiplier bootstrap. Such conditions are also assumed in Andrews and Shi (2015) and Hsu (2016).

Assumption A.2 *Assume that*

(i) The $K(\cdot)$ is a non-negative symmetric bounded kernel with a compact support in R (say $[-1,1]$).

(ii) $\int K(u)du = 1$,

(iii) $h \to 0$, $nh \to \infty$ and $nh^5 \to 0$ as $n \to \infty$.

Assumption A.2 is standard for nonparametric estimation. Note that $nh^5 \to 0$ as $n \to \infty$ implies undersmoothing so that the bias terms converge to zero even after we multiply it with $\sqrt{nh}$ and that this condition is standard if one wants to obtain the asymptotic normality of the estimators.

Assumption A.3 *Let $\{U_i : 1 \leq i \leq n\}$ be a sequence of i.i.d. random variables $E[U] = 0$, $E[U^2] = 1$, and $E[|U|^4] < M$ for some $\delta > 0$ and $M > 0$, and $\{U_i : 1 \leq i \leq n\}$ is independent of the sample path $\{(Y_i, X_i, Z_i, T_i) : 1 \leq i \leq n\}$.*

Assumption A.3 is standard for the multiplier bootstrap as in Hsu (2016) and $E[|U|^4] < M$ is needed for the multiplier bootstrap for nonparametric method.

Proof of Lemma 3.1: Define

$$h_{2,m+}(\ell_1,\ell_2) = \frac{\int_0^\infty (\vartheta_2 - u\vartheta_1)^2 K^2(u)du}{(\vartheta_2\vartheta_0 - \vartheta_1^2)^2} \frac{\sigma_+^2(\ell_1,\ell_2)}{f_z(c)}$$
$$h_{2,m-}(\ell_1,\ell_2) = \frac{\int_0^\infty (\vartheta_2 - u\vartheta_1)^2 K^2(u)du}{(\vartheta_2\vartheta_0 - \vartheta_1^2)^2} \frac{\sigma_-^2(\ell_1,\ell_2)}{f_z(c)},$$

then it is easy to see that $h_{2,\nu}(\ell_1,\ell_2) = h_{2,m+}(\ell_1,\ell_2) + h_{2,m-}(\ell_1,\ell_2)$.

Recall that

$$\hat{m}_+(\ell) = \frac{\sum_{i=1}^n 1(Z_i \geq c) \cdot K(\frac{Z_i - c}{h})[S_{n,2}^+ - S_{n,1}^+(Z_i - c)]g_\ell(X_i)Y_i}{\sum_{i=1}^n 1(Z_i \geq c) \cdot K(\frac{Z_i - c}{h})[S_{n,2}^+ - S_{n,1}^+(Z_i - c)]} = \frac{1}{nh}\sum_{i=1}^n w_{ni}^+ g_\ell(X_i)Y_i,$$

and it is true that

$$\sqrt{nh}(\hat{m}_+(\ell) - m_+(\ell)) = \sqrt{nh}\,(\hat{m}_+(\ell) - E_Z[\hat{m}_+(\ell)]) + \sqrt{nh}\,(E_Z[\hat{m}_+(\ell)] - m_+(\ell))$$

in which E_Z denotes the conditional expectation conditional on sample path $\{Z_1, Z_2, \ldots\}$. By Theorem 4 of Fan and Gijbels (1992), we know that

$$E_Z[\hat{m}_+(\ell) - m_+(\ell)] = O_p(\sqrt{nh^5}) = o_p(1).$$

The first equality holds because the magnitude is proportional to $m''_+(\ell)$ which is equal to $E_Z[g_\ell(X)\cdot(\partial^2\mu_1(x,z)/\partial z\partial z)]$ and $|\partial^2\mu_1(x,z)/\partial z\partial z|$ is assumed to be uniformly bounded on $x\in\mathcal{X}_c$ and $z\in\mathcal{N}_{\delta,z}(c)$. Therefore,

$$\begin{aligned}&\sqrt{nh}(\hat{m}_+(\ell)-m_+(\ell))\equiv\sqrt{nh}\,(\hat{m}_+(\ell)-E_Z[\hat{m}_+(\ell)])+o_p(1),\\&=\sqrt{nh}\sum_{i=1}^n w_{ni}^+\big(g_\ell(X_i)Y_i-E_Z[g_\ell(X_i)Y_i]\big)+o_p(1).\end{aligned}$$

We use the functional central limit theorem, Theorem 10.6 of Pollard (1990), to show that

$$\sqrt{nh}\sum_{i=1}^n w_{ni}^+\big(g_\ell(X_i)Y_i-E_Z[g_\ell(X_i)Y_i]\big)\Rightarrow\Phi_{h_{2,m+}}(\ell).$$

Our arguments condition on the sample path of Z_i's and in other words, w_{ni}^+ can be treated as constants. Define our triangular array as $\{f_{ni}(\ell):\ \ell\in\mathcal{L},\ i\le n,n\ge 1\}$ and $f_{ni}(\ell)=\sqrt{nh}w_{ni}^+\big(g_\ell(X_i)Y_i-E_Z[g_\ell(X_i)\cdot Y_i]\big)$. Let the envelope functions be $\{F_{ni}:\ i\le n,n\ge 1\}$ with $F_{ni}=\sqrt{nh}|w_{ni}^+|\cdot(|Y_i|+E_Z[|Y_i|])$. Define our empirical process as $\widehat{\Phi}_n^+(\ell)=\sum_{i=1}^n f_{ni}(\ell)$. First, $\{g_\ell(X):\ \ell\in\mathcal{L}\}$ is a Type I class of functions in Andrews (1994) and by Lemma E1 of Andrews and Shi (2013), $\{f_{ni}(\ell):\ \ell\in\mathcal{L},\ i\le n,n\ge 1\}$ satisfies condition (i) of Theorem 10.2 of Pollard (1990). To show condition (ii), note that

$$\begin{aligned}&\hat{h}_{2,m+}(\ell_1,\ell_2)=E_Z[\widehat{\Phi}_n^+(\ell_1)\widehat{\Phi}_n^+(\ell_2)]=E[f_{ni}(\ell_1)f_{ni}(\ell_2)]\\=&nh\sum_{i=1}^n(w_{ni}^+)^2\Big(E_Z[g_{\ell_1}(X_i)g_{\ell_2}(X_i)Y_i^2]\\&\qquad-E_Z[g_{\ell_1}(X_i)\cdot Y_i]E_Z[g_{\ell_2}(X_i)\cdot Y_i]\Big)\to h_{2,m+}(\ell_1,\ell_2),\end{aligned}$$

where the third equality holds because $f_{ni}(\ell_1)$ and $f_{nj}(\ell_2)$ are mutually independent for $i\neq j$. Then by the arguments of the second part of Theorem 4 of Fan and Gijbels (1992), we can show that $E_Z[\widehat{\Phi}_n(\ell_1)\widehat{\Phi}_n(\ell_2)]$ converges to $h_{2,m+}(\ell_1,\ell_2)$. Furthermore, it is true that the convergence result holds uniformly over $\ell_1,\ell_2\in\mathcal{L}$. Condition (iii) can be shown by the same arguments for condition (ii). To show condition (iv), note that for any $\epsilon>0$,

$$\begin{aligned}&\sum_{i=1}^n E_Z[F_{ni}^2\cdot 1(F_{ni}>\epsilon)]\le\sum_{i=1}^n E_Z\Big[\frac{F_{ni}^4}{\epsilon^2}\Big]\\=&\epsilon^{-2}(nh)^2\sum_{i=1}^n(w_{ni}^+)^4E_Z\big[(|Y_i|+E_Z[|Y_i|])^4\big].\end{aligned}$$

The first inequality holds because $1(F_{ni}>\epsilon)\le(F_{ni}/\epsilon)^\delta$ for any $\delta>0$ and we take $\delta=2$ here. By the same arguments from the second part of Theorem 4 of Fan and Gijbels (1992), we can show that

$$\epsilon^{-2}(nh)^2\sum_{i=1}^n(w_{ni}^+)^4E_Z\big[(|Y_i|+E_Z[|Y_i|])^4\big]=\epsilon^{-2}(nh)^2O_p((nh)^{-3})=O_p((nh)^{-1})=o_p(1),$$

and this implies that condition (iv) holds.

To show condition (v), note that

$$\begin{aligned}
&\hat{\rho}_{n,m+}(\ell_1,\ell_2) = \sum_{i=1}^{n}(f_{ni}(\ell_1) - f_{ni}(\ell_2))^2 \\
=&\sum_{i=1}^{n} f_{ni}^2(\ell_1) - 2\sum_{i=1}^{n} f_{ni}(\ell_1)f_{ni}(\ell_2) + \sum_{i=1}^{n} f_{ni}^2(\ell_2) \\
=&H_{1n}(\ell_1,\ell_1) - 2H_{1n}(\ell_1,\ell_2) + H_{1n}(\ell_2,\ell_2) \\
\to&h_{2,m+}(\ell_1,\ell_1) - 2h_{2,m+}(\ell_1,\ell_2) + h_{2,m+}(\ell_2,\ell_2) \equiv \rho_{m+}(\ell_1,\ell_2).
\end{aligned}$$

Note that similar to condition (ii), the convergence holds uniformly over $\ell_1, \ell_2 \in \mathcal{L}$. Then this is sufficient for condition (v). Then by FCLT of Pollard (1990), we can show that $\sqrt{nh}(\hat{m}_+(\ell) - m_+(\ell)) \Rightarrow \Phi_{h_{2,m+}}(\ell)$. By the same arguments, we can show that $\sqrt{nh}(\hat{m}_-(\ell) - m_-(\ell)) \Rightarrow \Phi_{h_{2,m-}}(\ell)$ and it follows that $\sqrt{nh}(\hat{\nu}(\ell) - \nu(\ell)) = \sqrt{nh}(\hat{m}_+(\ell) - m_+(\ell)) - \sqrt{nh}(\hat{m}_-(\ell) - m_-(\ell)) \Rightarrow \Phi_{h_{2,\nu}}(\ell)$. This completes the proof. □

Proof of Lemma 3.2: We use the same arguments of proof in Hsu (2016). Recall that $\widehat{\Phi}_n^u(\ell) = \sum_{i=1}^{n} U_i \cdot \hat{\phi}_{\nu,ni}(\ell)$ where

$$\hat{\phi}_{\nu,ni}(\ell) = \sqrt{nh}\Big(w_{ni}^+ \cdot (g_\ell(X_i)Y_i - \hat{m}_+(\ell)) - w_{ni}^- \cdot (g_\ell(X_i)Y_i - \hat{m}_-(\ell))\Big).$$

It is sufficient for us to show that $\widehat{\Phi}_n^{+,u}(\ell) = \sum_{i=1}^{n} U_i \cdot \hat{\phi}_{m+,ni}(\ell) \overset{p}{\Rightarrow} \Phi_{h_{2,m+}}(\ell)$ where

$$\hat{\phi}_{m+,ni}(\ell) = \sqrt{nh}\Big(w_{ni}^+(g_\ell(X_i)Y_i - \hat{m}_+(\ell))\Big).$$

First, it is straightforward to see that the triangular array $\{\hat{f}_{ni}(\ell) = U_i \cdot \hat{\phi}_{m+,ni}(\ell) : \ \ell \in \mathcal{L},\ i \le n,\ n \ge 1\}$ is manageable with respective to envelope functions $\{\widehat{F}_{ni} = \sqrt{nh}|U_i| \cdot \big(|w_{ni}^+| \cdot (|Y_i| + \overline{|Y|}_n^+)\big) : \ i \le n,\ n \ge 1\}$ in which $\overline{|Y|}_n^+ \equiv \sum_{i=1}^{n} |w_{ni}^+| \cdot |Y_i|$. Define $\hat{h}_{2,m+}(\ell_1,\ell_2) = \sum_{i=1}^{n} \hat{\phi}_{m+,ni}(\ell_1)\hat{\phi}_{m+,ni}(\ell_2)$. First, by the same argument in (12.24)-(12.26) of Andrews and Shi (2015) and the same argument from the second part of Theorem 4 of Fan and Gijbels (1992), we can show that

$$\sup_{\ell_1,\ell_2 \in \mathcal{L}} |\hat{h}_{2,m+}(\ell_1,\ell_2) - h_{2,m+}(\ell_1,\ell_2)| \overset{p}{\to} 0.$$

Also, we can show that

$$\begin{aligned}
nh\sum_{i=1}^{n} \big(|w_{ni}^+| \cdot (|Y_i| + \overline{|Y|}_n^+)\big)^2 &\overset{p}{\to} M_1 < \infty, \\
n^3h^3\sum_{i=1}^{n} \big(|w_{ni}^+| \cdot (|Y_i| + \overline{|Y|}_n^+)\big)^4 &\overset{p}{\to} M_2 < \infty,
\end{aligned}$$

for some positive M_1 and M_2.

Then by the same proof of Theorem 2.1 of Hsu (2016), we can show that $\widehat{\Phi}_n^{+,u}(\ell) \overset{p}{\Rightarrow} \Phi_{h_{2,m+}}(\ell)$.

B Proofs of Main Results

Proof of Theorem 3.1: Given the results of Lemma 3.1 and Lemma 3.2 hold, then by the same proof for Proposition 3 of Barrett and Donald (2003), Theorem 3.1 follows. We omit the details for brevity. □

Proofs of Theorem 3.2 and 3.3: Note that the process results and simulated process results for Theorem 3.2 and 3.3 are similar to Lemma 3.1 and Lemma 3.2, so we omit the details for brevity. Then, the proofs for Theorem 3.2 and 3.3 are similar to that for Theorem 3.1. □

Proof of Equation (2.1):

$$
\begin{aligned}
&\lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z] \\
=&\lim_{z \searrow c} E[Y_i(1)T_i + Y_i(0)(1 - T_i)|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i(1)T_i + Y_i(0)(1 - T_i)|X_i = x, Z_i = z] \\
=&\lim_{z \searrow c} E[Y_i(1)T_i(1) + Y_i(0)(1 - T_i(1))|X_i = x, Z_i = z] \\
&- \lim_{z \nearrow c} E[Y_i(1)T_i(0) + Y_i(0)(1 - T_i(0))|X_i = x, Z_i = z] \\
=&E[Y_i(1)T_i(1) + Y_i(0)(1 - T_i(1))|X_i = x, Z_i = z] - E[Y_i(1)T_i(0) + Y_i(0)(1 - T_i(0))|X_i = x, Z_i = z] \\
=&E[(Y_i(1) - Y_i(0))\,(T_i(1) - T_i(0))\,|X_i = x, Z_i = z] \\
=&E[Y_i(1) - Y_i(0)|X_i = x, Z_i = z, T_i(1) - T_i(0) = 1]P[T_i(1) - T_i(0) = 1|X_i = x, Z_i = z] \\
=&E[Y_i(1) - Y_i(0)|X_i = x, Z_i = z, T_i(1) - T_i(0) = 1]E[T_i(1) - T_i(0)|X_i = x, Z_i = z].
\end{aligned}
$$

The first equality holds by the definition of Y_i. The second holds by the definition of T_i. The third holds by the continuity assumption. The rest of the equalities hold from standard derivations.

Proof of Equation (3.4): Note that Andrews and Shi (2015) show that $E[m(W)|X] \leq 0$ a.s. in X conditional on $Z = z$ iff $E[m(W)g_\ell(X)|Z = z] \leq 0$ for all $\ell \in \mathcal{L}$. Therefore, $CATE(x) \leq 0,\ \forall\ x \in \mathcal{X}_c$ is equivalent to $CATE(X) \leq 0$ a.s. in X conditional on $Z = c$ and in turn, it is equivalent to $E[CATE(X)g_\ell(X)|Z = z] \leq 0$ for all $\ell \in \mathcal{L}$. Hence, it is sufficient to show that $E[CATE(X)g_\ell(X)|Z = z] = \lim_{z \searrow c} E[g_\ell(X_i)Y_i|Z_i = z] - \lim_{z \nearrow c} E[g_\ell(X_i)Y_i|Z_i = z]$ for all $\ell \in \mathcal{L}$. We show $E[CATE(X)|Z = z] = \lim_{z \searrow c} E[Y_i|Z_i = z] - \lim_{z \nearrow c} E[Y_i|Z_i = z]$ and the argument for general $g_\ell(X)$ is similar.

Let $\mu_1(x, z) = E[Y(1)|X = x, Z = z]$ and it is true that $\mu_1(x, z) = E[Y|X = x, Z = z]$ whenever $z \geq c$. Let $\mu_1(x, c+) = \lim_{z \searrow c} E[Y(1)|X = x, Z = z]$. First, note that $\lim_{z \searrow c} E_{X|Z=z}[\mu_1(X, c+)] = E_{X|Z=c}[\mu_1(X, c+)]$ because the distribution of X conditional on $Z = z$ is continuous at $Z = c$. Also, by the continuity of $\mu_1(x, z)$, we have that $E_{X|Z=c}[\mu_1(X, c+)] = \lim_{z \searrow c} E_{X|Z=z}[\mu_1(X, c+)] =$

$\lim_{z\searrow c} E_{X|Z=z}[\mu_1(X,z)]$. Given that $\mu_1(X,z) = E[Y|X, Z=c]$, we have that $E_{X|Z=z}[\mu_1(X,z)] = E_{X|Z=z}[E[Y|X,Z=z]]$ and by the law of iterated expectations, we have that $E_{X|Z=z}[E[Y|X,Z=z]] = E[Y|Z=z]$ for $z \geq c$. Last, we have that $E_{X|Z=c}[\mu_1(X,c+)] = \lim_{z\searrow c} E_{X|Z=z}[\mu_1(X,z)] = \lim_{z\searrow c} E[Y|Z=z]$. Similarly, we have that $E_{X|Z=c}[\mu_0(X,c-)] = \lim_{z\nearrow c} E[Y|Z=z]$. These complete the proof. □

Proof of Equation (3.10): Note that

$$
\begin{aligned}
&\sqrt{nh}(\hat{\nu}_{hetero,ate}(\ell) - \nu_{hetero,ate}(\ell)) \\
=&\sqrt{nh}(\hat{\nu}(\ell) - \hat{\nu}((\mathbf{0},1)) \cdot \hat{p}(\ell) - \nu(\ell) + \nu((\mathbf{0},1)) \cdot p(\ell)) \\
=&\sqrt{nh}(\hat{\nu}(\ell) - \nu(\ell)) - \sqrt{nh}(\hat{\nu}((\mathbf{0},1)) \cdot \hat{p}(\ell) - \nu((\mathbf{0},1)) \cdot p(\ell)) \\
=&\sqrt{nh}(\hat{\nu}(\ell) - \nu(\ell)) - \hat{p}(\ell) \cdot \sqrt{nh}(\hat{\nu}((\mathbf{0},1)) - \nu((\mathbf{0},1))) - \nu((\mathbf{0},1))\sqrt{nh}(\hat{p}(\ell) - p(\ell)) \\
=&\sqrt{nh}(\hat{\nu}(\ell) - \nu(\ell)) - p(\ell) \cdot \sqrt{nh}(\hat{\nu}((\mathbf{0},1)) - \nu((\mathbf{0},1))) + o_p(1) - \nu((\mathbf{0},1))\sqrt{nh}(\hat{p}(\ell) - p(\ell)) \\
=&\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \phi_{\nu,ni}(\ell) - p(\ell)\phi_{\nu,ni}((\mathbf{0},1)) - \nu((\mathbf{0},1)) \cdot \phi_{p,ni}(\ell) + o_p(1).
\end{aligned}
$$

This completes the proof. □

Proof of the Equivalence of (4.3) and (4.4): Recall that $H_{0,late}^{hetero}$ in (4.3) is equivalent to: $H_{0,late}^{hetero}: \ CLATE(x) = LATE$ for all $x \in \mathcal{X}_c$ in which

$$
\begin{aligned}
CLATE(x) &= \frac{\lim_{z\searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z\nearrow c} E[Y_i|X_i = x, Z_i = z]}{E[T_i(1) - T_i(0)|X_i = x, Z_i = c]} \\
&= \frac{\lim_{z\searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z\nearrow c} E[Y_i|X_i = x, Z_i = z]}{\lim_{z\searrow c} E[T_i|X_i = x, Z_i = z] - \lim_{z\nearrow c} E[T_i|X_i = x, Z_i = z]} \\
LATE &= \nu((\mathbf{0},1))/\mu((\mathbf{0},1)).
\end{aligned}
$$

Therefore, $H_{0,late}^{hetero}$ is equivalent to

$$
\begin{aligned}
H_{0,late}^{hetero}: &(\lim_{z\searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z\nearrow c} E[Y_i|X_i = x, Z_i = z]) \cdot \mu((\mathbf{0},1)) \\
&- (\lim_{z\searrow c} E[T_i|X_i = x, Z_i = z] - \lim_{z\nearrow c} E[T_i|X_i = x, Z_i = z]) \cdot \nu((\mathbf{0},1)) = 0 \ \text{ for all } x \in \mathcal{X}_c.
\end{aligned}
$$

Then by the instrument function method and the proof for (3.4), $H_{0,late}^{hetero}$ is equivalent to

$$
H_{0,late}^{hetero}: \nu(\ell) \cdot \mu((\mathbf{0},1)) - \mu(\ell) \cdot \nu((\mathbf{0},1)) = 0 \ \text{ for all } \ell \in \mathcal{L}.
$$

This completes the proof. □

References

Anderson, M. L. (2008): "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103, 1481–1495.

Andrews, D. W., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81(2), 609–666.

——— (2015): "Nonparametric Inference Based on Conditional Moment Inequalities," *Journal of Econometrics*, 179(1), 31–45.

Andrews, D. W. K. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4, pp. 2111–3155. North Holland.

Angrist, J. D., and A. B. Krueger (1992): "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, 87(418), 328–336.

——— (1995): "Split-sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, 13(2), 225–235.

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.

Barrett, G. F., and S. G. Donald (2003): "Consistent Tests for Stochastic Dominance," *Econometrica*, 71(1), 71–104.

Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2008): "Distributional Impacts of the Self-Sufficiency Project," *Journal of Public Economics*, 92(3), 748–765.

Bitler, M. P., H. W. Hoynes, and T. Domina (2016): "Experimental Evidence on Distributional Effects of Head Start," *working paper*.

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

Bugni, F. A., I. A. Canay, and X. Shi (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185(1), 259–282.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression? Discontinuity Designs," *Econometrica*, 82(6), 2295–2326.

Donald, S. G., and Y.-C. Hsu (2016): "Improving the Power of Tests of Stochastic Dominance," *Econometric Review*, 35(4), 553–585.

Fan, J., and I. Gijbels (1992): "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 20(4), 2008–2036.

Feir, D., T. Lemieux, and V. Marmer (2015): "Weak Identification in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, forthcoming.

He, C. (2016): "How Does Adverse Adolescent Experience Affect Subjective Long-term Well-being? A Study on the Send-Down Movement in China," *working paper*.

Hsu, Y.-C. (2015): "Consistent Tests for Conditional Treatment Effects," Discussion paper, Academia Sinica.

——— (2016): "Multiplier Bootstrap," Discussion paper, Academia Sinica.

Imbens, G., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of Economic Studies*, 79(3), 933–959.

Imbens, G. W., and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142(2), 615 – 635.

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*. Springer: New York.

Lee, D. S., and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48(2), 281–355.

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Sructural Models," *Econometrica*, 71, 1027–1048.

Pollard, D. (1990): "Empirical processes: theory and applications," in *NSF-CBMS regional conference series in probability and statistics*.

Pop-Eleches, C., and M. Urquiola (2013): "Going to a Better School: Effects and Behavioral Responses," *American Economic Review*, 103(4), 1289–1324.

Romano, J. P., and A. M. Shaikh (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

Shen, S., and X. Zhang (2015): "Distributional Test for Regression Discontinuity: Theory and Applications," *Review of Economics and Statistics, forthcoming*.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 1997.

Stock, J. H., and M. Yogo (2005): "Testing for weak instruments in linear IV regression," *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg.*

van der Klaauw, W. (2002): "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression Discontinuity Approach," *International Economic Review*, 43(4)(97-10), 1249–1287.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer, New York.