Testing Treatment Effect Heterogeneity in Regression Discontinuity Designs

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: The authors would like to thank Arthur Lewbel, Vadim Marmer, Yuya Sasaki, Kevin Song, Liangjun Su, Ping Yu, and seminar participants at the department of economics in Havard/MIT, Singapore Management University, UC Irvine, University of British Columbia, University of Hong Kong, and the department of statistics in UC Davis, the 2016 Australasia Meeting of the Econometric Society, the 2016 All California Econometrics Conference, the 2016 Annual Meeting of Taiwan Econometric Society, the 2016 Cross-Strait Dialogue IV Workshop and the 2016 Midwest Econometrics Group for helpful comments. 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 the authors'.

Abstract

Treatment effect heterogeneity is frequently studied in regression discontinuity (RD) applications. This paper proposes, under the RD setup, the first set of formal tests for treatment effect heterogeneity among subpopulations with different characteristics. 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. Monte Carlo simulations show good small sample performance of the proposed tests. The empirical section applies the tests to study the impact of attending a better high school and discover interesting patterns of treatment effect heterogeneity neglected by previous studies.

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 values of 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.[2] However, when researchers are further interested in treatment effect heterogeneity, or how policy impact varies 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 that examine 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 3) has heterogeneous impact across all subpopulations. Both sharp and fuzzy RD designs are considered.

The proposed tests are useful because applied researchers are often interested in treatment effect heterogeneity. 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.[3] The common practice is to build a linear regression model

[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] Although controlling for additional covariates can potentially improve the estimation efficiency of the RD estimators. See, for example, Imbens and Lemieux (2008) and Calonico et al. (2016).

[3] The survey includes all 2015 issues of Quarterly Journal of Economics (QJE), Journal of Political Economy (JPE), Review of Economic Studies (RESTUD), American Economic Review (AER), American Economic Journal: Applied Economics (AEJ: AE) and American Economic Journal: Economic Policy (AEJ: EP) as well as 2016 issues of these journals published before April. A total of 17 papers (0 in QJE,

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

The interaction term method, which adds interaction terms between the dummy variable indicating whether an individual passes the cut-off value of the running variable and additional covariates of interest for the heterogeneity analysis, is parametric. The method 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 under certain mild kernel, bandwidth, and smoothness conditions of the underlying distribution.

The subsample regression method repeats the main RD analysis with different subsamples divided according to individual observed characteristics. The method is nonparametric. However, as the method typically involves running a number of subsample RD regressions at the same time, for correct inference it is essential to adjust the regressions for multiple testing (see, for example, Romano and Shaikh, 2010; Anderson, 2008). Unfortunately, none of the papers in our survey using the subsample regression method addressed on this issue. Besides, even if multiple testing is correctly accounted for, the subsample regression method is not ideal. First, it can produce over-rejected tests and under-covered confidence intervals under the fuzzy RD design if the sample size or the proportion of compliers is small for some subsamples (Feir et al., 2015). The main reason is that the subsample regression method compares estimated subsample local average treatment effects. But these estimators can have non-classical inference when the first stage is weak or when the subsample size is small. Second, the subsample regression method often requires categorizing continuous covariates into discrete groups, which often results in loss of information.

Our testing procedure applies the instrument function method developed in Andrews

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

and Shi (2013, 2015) to transform the hypotheses of interest, which are nonparametric conditional moment equalities/inequalities conditional on both the running variable of the RD model and additional covariates of interest, to (an 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 for covariates other than the running variable. Moreover, the proposed tests do not rely on plug-in estimators of the conditional average treatment effect, meaning that the proposed tests are robust to the weak identification problem discussed above. As we demonstrate in the Monte Carlo simulation section, the proposed tests have very good small sample performance, especially when compared to 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. 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 new testing procedures for conditional moment equality/inequalities that requires nonparametric boundary estimation. In addition, we propose a new multiplier bootstrap method for simulating critical values of such tests. For the uniform sign test, we discuss both critical values based on the least favorable configuration (LFC) and the generalized moment selection (GMS) method introduced by Andrews and Soares (2010) and Andrews and Shi (2013, 2015, 2017). The GMS method is similar to the recentering method in Hansen (2005) and Donald and Hsu (2016), and the contact set approach proposed in Linton et al. (2010), Lee et al. (2013), Aradillas-Lopez et al. (2016).

This paper is related to a large literature in regression discontinuity, especially some recent developments that also look at treatment effect heterogeneity. For example, Bertanha and Imbens (2014), Dong and Lewbel (2015) and Angrist and Rokkanen (2015) study treatment effect extrapolation away from the cut-off of the running variable, and Bertanha (2016) and Cattaneo et al. (2016) study treatment effect heterogeneity at different values of the running variable when the RD design has multiple cut-offs. In contrast, this paper focuses on the classic RD set-up with a single cut-off and examines treatment effect heterogeneity among marginal individuals around the cut-off of the running variable, rather than individuals away from the cut-off.

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) find 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 a heterogeneity analysis using the subsample regression method and find little evidence of treatment effect heterogeneity. In contrast, our tests detect a clear signal of treatment effect heterogeneity. We find that attending a more selective high school not only has a significant positive effect on the exam-taking rate for some subpopulations but also has some marginally significant negative effect for some other subpopulations. Our results suggest that the insignificant mean effect on the exam-taking rate found in Pop-Eleches and Urquiola (2013) should arise from the cancellation of opposite-signed effects among different parts of the population.

The paper is organized as follows. Section 2 sets up the model and identifies the conditional treatment effects of interest under both 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. Proofs and technical assumptions are provided in the appendix.

2 Model Framework

Let Y_i denote the outcome of interest and T_i the treatment decision of individual i. Use $Y_i(0)$ and $Y_i(1)$ to denote potential outcomes when $T_i = 0$ and $T_i = 1$, respectively; $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$. 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 last part of the next section, we will discuss how to implement the proposed tests when X_i contains discrete variables. For any $\delta > 0$, let $\mathcal{N}_{\delta,z}(c) = \{z : \ |z - c| \leq \delta\}$ denote a neighborhood of z around $Z = c$.

Assumption 2.1 *For a running variable Z_i continuously distributed in a neighborhood of the threshold value c, assume that for some $\delta > 0$*

(i) $E[Y_i(1)|X_i = x, Z_i = z]$ and $E[Y_i(0)|X_i = x, Z_i = z]$ are continuous in x and z on $\mathcal{X}_c \times \mathcal{N}_{\delta,z}(c)$.

(ii) The distribution function of $X_i|Z_i = z$ is continuous in z on $\mathcal{N}_{\delta,z}(c)$.

Assumption 2.1.(i) requires that the conditional means of the potential outcomes conditional on both the running variable and the additional controls of interest are continuous. It is stronger than the standard continuity assumption of $E[Y_i(t)|Z_i = z]$ used in the literature (c.f. Imbens and Lemieux, 2008) for the identification of the average treatment effect, or ATE. Assumption 2.1.(ii) requires that the conditional distribution of the additional controls conditional on the running variable is continuous. Although it is not required in the literature for the identification of ATE, either, Assumption 2.1.(ii) is in fact a direct implication of the "no precise control over the running variable" rule introduced in Lee and Lemieux (2010) and well-accepted in the applied RD literature.[4] We need both assumptions to identify the conditional average treatment effect, or CATE, that conditions on the value of X_i.

[4] According to 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, where V_i represents unobserved characteristics of individual i. By Bayes Rule, this condition 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 $z = c$.

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 ATE conditional on $Z_i = c$ only and the LATE conditional on both $Z_i = c$ and $X_i = x$ are identified as

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

More generally, when the treatment status T_i is a probabilistic function of Z_i, the RD model follows a *fuzzy* design. Suppose a policy intervention encourages an individual i to receive the treatment if the running variable Z_i is larger than or equal to c. Let $T_i(1)$ and $T_i(0)$ be the potential treatment decisions of individual i depending on whether he/she is encouraged or not; $T_i = T_i(1)1(Z_i \geq c) + T_i(0)1(Z_i < c)$. For identification in this general case we require the following assumptions in replacement of Assumption 2.1.(i).

Assumption 2.2 *Assume that for some $\delta > 0$,*

(i) For $t, t' \in \{0, 1\}$, $E[Y_i(t)|T_i(1) - T_i(0) = 1, X_i = x, Z_i = z]$ and $E[Y_i(t)|T_i(1) = T_i(0) = t', X_i = x, Z_i = z]$ are continuous in x and z on $\mathcal{X}_c \times \mathcal{N}_{\delta,z}(c)$;

(ii) For $t \in \{0, 1\}$, $P[T_i(1) - T_i(0) = 1|X_i = x, Z_i = z]$ and $P[T_i(1) = T_i(0) = t|X_i = x, Z_i = z]$ are continuous in x and z on $\mathcal{X}_c \times \mathcal{N}_{\delta,z}(c)$;

(iii) $T_i(1) \geq T_i(0)$;

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

Assumption 2.2.(i) requires the continuity of average potential outcomes for always-taker, compliers, and never-takes. Assumption 2.2.(ii) requires the continuity of the proportion of each group. Assumption 2.2.(iii) and (iv) require no defiers, and non-trivial presence of compliers, respectively. Assumption 2.2.(i), (ii) and (iv) are stronger than their counterparts that are unconditional on X_i (c.f. Dong and Lewbel, 2015). For example, (i) and (ii) imply that the conditional mean of potential outcomes and treatment decisions are continuous in both x and z on $\mathcal{X}_c \times \mathcal{N}_{\delta,z}(c)$. Assumption 2.2.(iii) is the monotonicity restriction that is commonly required in fuzzy RD models. It implies that $E[T_i(1) - T_i(0)|X_i = x, Z_i = c] \geq 0$ for all observational equivalent individuals with $x \in \mathcal{X}_c$. It is worth noting that this necessary condition of the monotonicity assumption could be tested by one of the tests proposed in this paper.

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

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

The identification of $LATE$ is standard. The identification of $CLATE(x)$ 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. All identified treatment effects, including ATE, $LATE$, $CATE$, and $CLATE$, can be estimated by standard local linear estimation methods.

3 Testing Under the Sharp RD Design

Researchers are often interested in knowing 1) whether a policy treatment is beneficial to at least some subpopulations defined by covariate values, 2) whether it has any impact on at least some subpopulations, and 3) 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 on average beneficial to at least some subpopulations defined by covariate values, the null and alternative hypotheses can be formulated as

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

The null and alternative hypotheses $H_{0,ate}^{neg}$ and $H_{1,ate}^{neg}$ are defined by conditional moment inequalities conditional on both the running variable Z and the additional control X. We apply the instrument function approach in Andrews and Shi (2013, 2015) to transform these inequalities to an infinite number of instrumented conditional moment inequalities conditional on only the running variable Z, without loss of information.

Let $\mathcal{G}$ be the set of the indicator functions of countable hypercubes C_ℓ such that

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

For each $\ell \in \mathcal{L}$, define the instrumented moment condition $\nu(\ell)$ by

$$
\nu(\ell) = E[g_\ell(X_i) CATE(X_i)|Z_i = z],
$$

which represents the average treatment effect for individuals with $X_i \in C_\ell$ multiplied by the proportion of such individuals in the population. The following lemma shows that hypotheses $H_{0,ate}^{neg}$ and $H_{1,ate}^{neg}$ can be characterized by the following instrumented conditional moment inequalities without loss of information.

$$
\begin{aligned}
&H_{0,ate}^{neg} : \ \nu(\ell) \leq 0, \ \forall \ \ell \in \mathcal{L}, \\
&H_{1,ate}^{neg} : \ \nu(\ell) > 0, \ \text{for some } \ell \in \mathcal{L}. \qquad (3.3)
\end{aligned}
$$

Lemma 3.1 *Under Assumption 2.1, hypotheses in (3.1) are equivalent to those in (3.3).*

Notice that when $q = 1$, $\ell = (\mathbf{0}, 1)$, $C_\ell = (0,1]^{d_x} = \mathcal{X}$, $\nu(\ell)$ reduces to $\nu(\mathbf{0}, 1) = E[CATE(X_i)|Z_i = z] = ATE$. When $q = 2$, the side length of the hypercubes is 1/2. Suppose that $d_x = 2$, then there are four possible values of ℓ's: $((0,0), 1/2)$, $((1/2, 1/2), 1/2)$, $((0, 1/2), 1/2)$ and $((1/2, 0), 1/2)$. They correspond to four hypercubes or C_ℓ's: $(0, 1/2]^2$, $(1/2, 1]^2$, $(0, 1/2] \times (1/2, 1]$ and $(1/2, 1] \times (0, 1/2]$. When q gets larger, the hypercubes get smaller. The null hypothesis (3.3) is violated if there exists at least one $\ell^* \in \mathcal{L}$ with $\nu(\ell^*) > 0$, which means that the average treatment effect of individuals with X_i value falling into hypercube C_{ℓ^*} is strictly positive. Therefore, the proposed testing procedure could also be used to identify hypercubes with positive average treatment effects.

Test Statistic and Asymptotic Results

By standard RD identification techniques, we know that $\nu(\ell)$ is identified by

$$\nu(\ell) = \lim_{z \searrow c} E[g_\ell(X_i)Y_i|Z_i = z] - \lim_{z \nearrow c} E[g_\ell(X_i)Y_i|Z_i = z], \tag{3.4}$$

for each $\ell \in \mathcal{L}$. See Appendix B for a proof. The equality implies that for any $\ell \in \mathcal{L}$, the instrumented moment condition $\nu(\ell)$ can be estimated by standard nonparametric RD regression methods.

To be specific, let $m_+(\ell) = \lim_{z \searrow c} E[g_\ell(X_i)Y_i|Z_i = z]$ and $m_-(\ell) = \lim_{z \nearrow c} E[g_\ell(X_i)Y_i|Z_i = z]$. Let $K(\cdot)$ be the kernel function and h the bandwidth. The estimators $\hat{m}_+(\ell)$ and $\hat{m}_-(\ell)$ for $m_+(\ell)$ and $m_-(\ell)$ are the constant terms in following regressions:

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

$$\min_{\hat{m}_-(\ell), \hat{b}_-(\ell)} \sum_{i=1}^{n} 1(Z_i < c) \cdot K\left(\frac{Z_i - c}{h}\right) \left[g_\ell(X_i)Y_i - \hat{m}_-(\ell) - \hat{b}_-(\ell)(Z_i - c)\right]^2.$$

In the Monte Carlo simulation and the empirical application sections of this paper, we follow the RD literature and use the triangular kernel (i.e. $K(u) = (1 - |u|) \cdot 1(|u| \leq 1)$) for all boundary local linear estimators.

Following Fan and Gijbels (1992), for $j = 0, 1, 2, \ldots$, define

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

and re-write the local linear estimators as

$$\hat{m}_+(\ell) = \frac{\sum_{i=1}^n 1(Z_i \geq c)K(\frac{Z_i-c}{h})[S_{n,2}^+ - S_{n,1}^+(Z_i-c)]g_\ell(X_i)Y_i}{S_{n,0}^+ S_{n,2}^+ - S_{n,1}^+ S_{n,1}^+} \equiv \sum_{i=1}^n w_{ni}^+ \cdot g_\ell(X_i)Y_i,$$
$$\hat{m}_-(\ell) = \frac{\sum_{i=1}^n 1(Z_i < c)K(\frac{Z_i-c}{h})[S_{n,2}^- - S_{n,1}^-(Z_i-c)]g_\ell(X_i)Y_i}{S_{n,0}^- S_{n,2}^- - S_{n,1}^- S_{n,1}^-} \equiv \sum_{i=1}^n w_{ni}^- \cdot g_\ell(X_i)Y_i,$$

where $w_{ni}^+ = \frac{1(Z_i \geq c)K(\frac{Z_i-c}{h})[S_{n,2}^+ - S_{n,1}^+(Z_i-c)]}{S_{n,0}^+ S_{n,2}^+ - S_{n,1}^+ S_{n,1}^+}$ and $w_{ni}^- = \frac{1(Z_i < c)K(\frac{Z_i-c}{h})[S_{n,2}^- - S_{n,1}^-(Z_i-c)]}{S_{n,0}^- S_{n,2}^- - S_{n,1}^- S_{n,1}^-}$. An estimator for $\nu(\ell)$ is then given by

$$\hat{\nu}(\ell) = \hat{m}_+(\ell) - \hat{m}_-(\ell).$$

Let $\vartheta_j = \int_0^\infty u^j K(u)du$ for $j = 0, 1, 2, \ldots$. Let $\sigma_+^2(\ell_1, \ell_2) = \lim_{z \searrow c} Cov[g_{\ell_1}(X)Y, g_{\ell_2}(X)Y|Z = z]$, and $\sigma_-^2(\ell_1, \ell_2) = \lim_{z \nearrow c} Cov[g_{\ell_1}(X)Y, g_{\ell_2}(X)Y|Z = z]$. We summarize the asymptotic properties of $\sqrt{nh}(\hat{\nu}(\cdot) - \nu(\cdot))$ in the following lemma.

Lemma 3.2 *Under Assumption 2.1, and Assumptions A.1 and A.2 stated 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,

$$\widehat{\Phi}_{\nu,n}(\cdot) \equiv \sqrt{nh}(\hat{\nu}(\cdot) - \nu(\cdot)) \Rightarrow \Phi_{h_{2,\nu}}(\cdot),$$

where $\Phi_{h_2}(\cdot)$ denotes 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 $\ell_1, \ell_2 \in \mathcal{L}$.

Assumptions A.1 and A.2 are standard kernel, bandwidth, and smoothness conditions of the underlying data distribution. They are stated in the appendix. We require undersmoothing so that $\sqrt{nh}\left(\hat{\nu}(\cdot) - \nu(\cdot)\right)$ weakly converges to a mean zero Gaussian process. The $\phi_{\nu,ni}(\ell)$ function defined in (3.5) is the influence function that contributes to the limiting distribution of $\sqrt{nh}\left(\hat{\nu}(\ell) - \nu(\ell)\right)$.

Let $\hat{\sigma}^2_{\nu,n}(\ell) = \sum_{i=1}^{n} \hat{\phi}_{\nu,ni}(\ell)^2$ where

$$\hat{\phi}_{\nu,ni}(\ell) = \sqrt{nh}\left(w^+_{ni} \cdot (g_\ell(X_i)Y_i - \hat{m}_+(\ell)) - w^-_{ni} \cdot (g_\ell(X_i)Y_i - \hat{m}_-(\ell))\right). \quad (3.6)$$

Call $\hat{\phi}_{\nu,ni}(\ell)$ the estimated influence function, which replaces $m_+(\ell)$ and $m_-(\ell)$ in $\phi_{\nu,ni}(\ell)$ by their nonparametric estimators. We will show that $\hat{\sigma}^2_{\nu,n}(\ell)$ is a consistent estimator for $\sigma^2_\nu(\ell) \equiv h_{2,\nu}(\ell,\ell)$ uniformly over $\ell \in \mathcal{L}$. Define $\hat{\sigma}^2_{\nu,\epsilon}(\ell) = \max\{\hat{\sigma}^2_{\nu,n}(\ell), \epsilon \cdot \hat{\sigma}^2_{\nu,n}(\mathbf{0},1)\}$ with some small positive ϵ that manually bounds the variance estimator away from zero. In the simulation and empirical sections of the paper, we set ϵ to 0.05 following the practice in Andrews and Shi (2013). Define the Kolmogorov-Smirnov (KS) type statistic for testing $H^{neg}_{0,ate}$ as

$$\widehat{S}^{neg}_{ate} = \sqrt{nh} \sup_{\ell \in \mathcal{L}} \frac{\hat{\nu}_n(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)}.$$

The test-statistic is scale-invariant.

Notice that we adopted the instrument function approach in Andrews and Shi (2013, 2015) to test the conditional moment inequalities stated in $H^{neg}_{0,ate}$. Other methods developed in the conditional moment inequality literature (e.g. Chernozhukov et al., 2013; Lee et al., 2013, 2015; Aradillas-Lopez et al., 2016; Chetverikov, forthcoming) could also be potentially applied to test the null. The simulation results in Aradillas-Lopez et al. (2016), suggest that no specific strategy is expected to outperform the rest in all data generating processes. We choose to use the instrument function approach because it transforms the null to a series of conditional moment inequalities that only condition on the running variable Z. The constructed test statistic then only involves one-dimensional local linear estimation, which is the same as the estimation strategy adopted in classic RD regression analysis and could therefore be attractive to practitioners.

Simulated Critical Value Based on the Least Favorable Configuration

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 simulation and empirical sections of the paper, the pseudo random variables are drawn from the standard normal distribution.

Let the simulated process $\widehat{\Phi}^u_{\nu,n}(\ell)$ be

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

with $\hat{\phi}_{\nu,ni}(\ell)$ defined in (3.6). The next lemma shows that the process $\widehat{\Phi}^u_{\nu,n}(\cdot)$ can approximate the empirical process $\widehat{\Phi}_{\nu,n}(\cdot)$ well. The regularity conditions and the proof are given in Appendix B.

Lemma 3.3 *Under Assumption 2.1 and Assumptions A.1, A.2 and A.3 stated in the appendix,* $\sup_{\ell\in\mathcal{L}} |\hat{\sigma}^2_{\nu,n}(\ell) - \sigma^2_\nu(\ell)| \xrightarrow{p} 0$ *and* $\widehat{\Phi}^u_n(\cdot) \overset{p}{\Rightarrow} \Phi_{h_{2,\nu}}(\cdot)$.[5]

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

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

i.e., $\hat{c}^{neg}_{n,ate}(\alpha)$ is the $(1-\alpha)$-th quantile of the simulated null distribution, $\sup_{\ell\in\mathcal{L}} \frac{\widehat{\Phi}^u_{\nu,n}(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)}$. Finally, define the *decision rule* of the test as: "Reject $H^{neg}_{0,ate}$ if $\widehat{S}^{neg}_{ate} > \hat{c}^{neg}_{n,ate}(\alpha)$."

Generalized Moment Selection

The above described testing procedure relies on the least favorable configuration, or LFC, and could be potentially conservative. In this section, we follow Andrews and Shi (2013) and apply the generalized moment selection, or GMS, method to improve the power of the proposed uniform sign test.

Assumption 3.1 *Let a_n and B_n be sequences of non-negative numbers.*

[5] 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 $\sup_{f\in BL_1} |E_u f(\Psi^u_n) - Ef(\Psi)| \xrightarrow{p} 0$ and $E_u f(\Psi^u_n)^* - E_u f(\Psi^u_n)_* \xrightarrow{p} 0$, 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 $\xrightarrow{p}$ requirements used in the definition for $\Psi^u_n \overset{p}{\Rightarrow} \Psi$ replaced by $\xrightarrow{a.s.}$. 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 $\Psi^u_{\ell_N} \overset{a.s.}{\Rightarrow} \Psi$.

1. a_n *satisfies that* $\lim_{n\to\infty} a_n = \infty$, *and* $\lim_{n\to\infty} a_n/\sqrt{nh} = 0$.

2. B_n *is non-decreasing and satisfies that* $\lim_{n\to\infty} B_n = \infty$, *and* $\lim_{n\to\infty} B_n/a_n = 0$.

For all $\ell \in \mathcal{L}$, define $\hat{\psi}_\nu(\ell)$ as

$$\hat{\psi}_\nu(\ell) = -B_n \cdot 1\left(\sqrt{nh}\cdot \frac{\hat{\nu}_n(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)} < -a_n\right), \tag{3.7}$$

and the simulated GMS critical value $\hat{c}^{\eta}_{n,ate}(\alpha)$ as

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

where $\eta > 0$ is some small positive number. Then $\hat{c}^{\eta}_{n,ate}(\alpha)$ is the $(1-\alpha+\eta)$-th quantile of the simulated distribution of $\sup_{\ell\in\mathcal{L}}\left(\frac{\widehat{\Phi}^u_{\nu,n}(\ell)}{\hat{\sigma}_{\mu,\epsilon}(\ell)} + \hat{\psi}_\nu(\ell)\right)$, plus η. Following Andrews and Shi (2013, 2015), in the simulation and empirical sections of the paper, we use $a_n = (0.3\ln(n))^{1/2}$, $B_n = (0.4\ln(n)/\ln\ln(n))^{1/2}$ and $\eta = 10^{-6}$.

Let the *decision rule* based on the GMS critical value be: "Reject $H^{neg}_{0,ate}$ if $\widehat{S}^{neg}_{ate} > \hat{c}^{\eta}_{n,ate}(\alpha)$." Since it is negative for ℓ vectors with relatively large and negative $\sqrt{nh}\frac{\hat{\nu}_n(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)}$ values and zero otherwise, the term $\hat{\psi}_\nu(\ell)$, intuitively, helps to suppress the influence of negative moment functions on the simulated critical value. Using the GMS critical value hence improves the power of the proposed inequality test.

Size and Power Properties

We summarize size and power properties of the proposed test in the following two theorems. The regularity conditions and the proof are given in the appendix.

Theorem 3.1 *Under Assumption 2.1 and Assumptions A.1, A.2 and A.3 described in the appendix, when* $\alpha < 1/2$, *we have*

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

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

Theorem 3.1 discusses the asymptotic property of the proposed test based on the LFC critical value. The test is consistent and its asymptotic size is less than or equal to the significance level α as a result of adopting the LFC critical value.

Theorem 3.2 *Under Assumption 2.1, Assumptions A.1, A.2 and A.3 described in the appendix, and the GMS Assumption 3.1, when* $\alpha < 1/2$*, we have*

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

(2) in addition, if $\mathcal{L}^o = \{\ell : \nu(\ell) = 0\}$ *is non-empty and there exists* $\ell^* \in \mathcal{L}^o$ *with* $\sigma_\nu^2(\ell^*) > 0$*, then under* $H_{0,ate}^{neg}$, $\lim_{\eta\to 0}\lim_{n\to\infty} P(\widehat{S}_{n,ate} > \hat{c}_{n,ate}^{\eta}(\alpha)) = \alpha$.

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

Theorem 3.2 shows the consistency and asymptotic size control of the proposed test based on the GMS critical value. Also, when the null hypothesis $H_{0,ate}^{neg}$ holds with equality for some ℓ vectors, using the GMS critical value can lead to exact asymptotic size control with small enough η value.

Note that although we focus on KS type tests in this paper, all of our testing results can be extended to Cramér-von Mises type tests fairly easily given the asymptotic results of $\hat{\nu}(\cdot)$ and the simulated process $\widehat{\Phi}_n^u(\cdot)$.

In addition, we would like to point out that the above described test for $H_{0,ate}^{neg}$ can be trivially extended to study the hypotheses

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

in any sharp RD design, or the first stage selection

$$H_{0,fs}^{pos} : E[T_i(1) - T_i(0)|X_i = x, Z_i = c] \geq 0, \ \forall\, x \in \mathcal{X}_c,$$
$$H_{0,fs}^{pos} : E[T_i(1) - T_i(0)|X_i = x, Z_i = c] < 0, \ \text{for some} \ \ x \in \mathcal{X}_c,$$

in any fuzzy RD design. As is discussed in the identification section, the second test above is a sufficient test for the monotonicity restriction commonly used in fuzzy RD models in the sense that if $H_{0,fs}^{pos}$ is rejected, the monotonicity assumption is rejected.

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, consider the case in which in addition to X_i, there is one binary variable, X_{di}, taking values in $\{0,1\}$. Let $\mathcal{G}_1 \equiv \{1(X_d = 1) \cdot g_\ell(\cdot) : \ \ell \in \mathcal{L}\}$ and $\mathcal{G}_0 \equiv \{1(X_d = 0) \cdot g_\ell(\cdot) : \ \ell \in \mathcal{L}\}$. Let $\widetilde{\mathcal{G}} = \mathcal{G} \cup \mathcal{G}_1 \cup \mathcal{G}_0$. It is straightforward to show that

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

are equivalent to

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

Then we can carry out the uniform sign test in the same way as is discussed earlier but with $\mathcal{G}$ replaced by $\widetilde{\mathcal{G}}$. All results discussed above 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, the null and alternative hypotheses can be formulated as

$$
\begin{aligned}
&H_{0,ate}^{zero} : CATE(x) = 0, \ \forall \ x \in \mathcal{X}_c, \\
&H_{1,ate}^{zero} : CATE(x) \neq 0, \ \text{for some } \ x \in \mathcal{X}_c. \qquad (3.8)
\end{aligned}
$$

Similar to the previous subsection, we can transform the hypotheses in (3.8) to

$$
\begin{aligned}
&H_{0,ate}^{zero} : \ \nu(\ell) = 0, \ \forall \ \ell \in \mathcal{L}, \\
&H_{1,ate}^{zero} : \ \nu(\ell) \neq 0, \ \text{for some } \ell \in \mathcal{L} \qquad (3.9)
\end{aligned}
$$

without loss of information, as is summarized in the following lemma.

Lemma 3.4 *Suppose that Assumption 2.1 holds. Then hypotheses in (3.8) are equivalent to those in (3.9).*

Define the KS type test statistic for this overall significance test as

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

Let the decision rule be: "Reject $H_{0,ate}^{zero}$ if $\widehat{S}_{ate}^{zero} > \hat{c}_{n,ate}^{zero}(\alpha)$", where α is the predetermined significance level and $\hat{c}_{n,ate}^{zero}(\alpha)$ the simulated critical value defined as

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

The following theorem summarizes the size and power property of the test.

Theorem 3.3 *Under Assumption 2.1 and Assumptions A.1, A.2 and A.3, when* $\alpha < 1/2$*, we have*

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

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

Since this test is two-sided, the simulated critical value is expected to have correct asymptotic size control. Again, we would like to point out that although we adopt the instrument function approach in Andrews and Shi (2013, 2015), other testing procedures developed in the moment equality literature (Bierens, 1982, 1990, Bierens and Ploberger, 1997, and Whang, 2000, 2001, etc.), could also be potentially modified to test the hypothesis of interest. As is discussed earlier, we adopt the instrument function approach because the resulted test statistic requires the same estimation strategy as classic RD regression.

3.3 Testing if the Treatment Effect is Heterogenous

To test for treatment effect heterogenous, we define the hypotheses as

$$\begin{aligned} H_{0,ate}^{hetero} &: CATE(x) = \gamma,\ \forall\ x \in \mathcal{X}_c \textit{ and some } \gamma \in R, \\ H_{1,ate}^{hetero} &:\ H_{0,ate}^{hetero} \text{ does not hold.} \end{aligned} \tag{3.10}$$

If $CATE(x) = \gamma$ for all $x \in \mathcal{X}_c$ for some $\gamma \in R$, then it would hold with $\gamma = ATE = \nu((\mathbf{0},1))$. Then $H_{0,ate}^{hetero}$ would imply that $\nu(\ell) = p(\ell)\cdot\nu((\mathbf{0},1))$, where $p(\ell) = E[g_\ell(X_i)|Z_i = c]$ is the conditional probability of $X_i \in C_\ell$. So the hypotheses in (3.10) are equivalent to

$$\begin{aligned} H_{0,ate}^{hetero} &:\ \nu_{hetero,ate}(\ell) = \nu(\ell) - \nu((\mathbf{0},1))\cdot p(\ell) = 0,\ \forall\ \ell\in\mathcal{L}, \\ H_{1,ate}^{hetero} &:\ \nu_{hetero,ate}(\ell) = \nu(\ell) - \nu((\mathbf{0},1))\cdot p(\ell) \neq 0,\ \text{for some } \ell\in\mathcal{L}. \end{aligned} \tag{3.11}$$

The following lemma formally summarizes the eqiuvalence result.

Lemma 3.5 *Suppose that Assumption 2.1 holds. Then hypotheses in (3.10) are equivalent to those in (3.11).*

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

$$\hat{p}(\ell) = \frac{\sum_{i=1}^{n} K(\frac{Z_i - c}{h})[S_{n,2} - S_{n,1}(Z_i - c)]g_\ell(X_i)}{S_{n,0}S_{n,2} - S_{n,1}^2} \equiv \sum_{i=1}^{n} w_{ni} \cdot g_\ell(X_i),$$

where

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

Let $\phi_{p,ni}(\ell) = \sqrt{nh}\Big(w_{ni}(g_\ell(X_i) - p(\ell))\Big)$. Similar to Lemma 3.2, we can show that

$$\left|\sqrt{nh}(\hat{p}(\ell) - p(\ell)) - \sum_{i=1}^{n} \phi_{p,ni}(\ell)\right| = o_p(1), \tag{3.12}$$

uniformly over $\ell \in \mathcal{L}$.

Let $\hat{\nu}_{hetero,ate}(\ell) = \hat{\nu}(\ell) - \hat{\nu}((\mathbf{0}, 1)) \cdot \hat{p}(\ell)$ be the estimator of $\nu_{hetero,ate}(\ell)$. Let $\phi_{ate,ni}^{hetero}(\ell) = \phi_{\nu,ni}(\ell) - p(\ell) \cdot \phi_{\nu,ni}((\mathbf{0}, 1)) - \nu((\mathbf{0}, 1)) \cdot \phi_{p,ni}(\ell)$, it is easy to show that

$$\left|\sqrt{nh}(\hat{\nu}_{hetero,ate}(\ell) - \nu_{hetero,ate}(\ell)) - \sum_{i=1}^{n} \phi_{ate,ni}^{hetero}(\ell)\right| = o_p(1) \tag{3.13}$$

uniformly over $\ell \in \mathcal{L}$. We give the proof in the appendix. Let $\hat{\phi}_{ate,ni}^{hetero}(\ell) = \hat{\phi}_{\nu,ni}(\ell) - \hat{p}(\ell) \cdot \hat{\phi}_{\nu,ni}((\mathbf{0}, 1)) - \hat{\nu}((\mathbf{0}, 1)) \cdot \hat{\phi}_{p,ni}(\ell)$ be the estimated influence function with $\hat{\phi}_{p,ni}(\ell) = \sqrt{nh}\Big(w_{ni}(g_\ell(X_i) - \hat{p}(\ell))\Big)$.

Let $\left(\hat{\sigma}_{ate,n}^{hetero}(\ell)\right)^2 = \sum_{i=1}^{n}\left(\hat{\phi}_{ate,ni}^{hetero}(\ell)\right)^2$. Define the KS type test statistic as

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

where $\hat{\sigma}_{ate,\epsilon}^{hetero}(\ell) = \sqrt{\max\left\{\left(\hat{\sigma}_{ate,n}^{hetero}(\ell)\right)^2, \epsilon \cdot \hat{\sigma}_{\nu,n}^2((\mathbf{0}, 1))\right\}}$ for some small positive ϵ. Again, $\sigma_{\nu,n}^2((\mathbf{0}, 1))$ is used in the definition of $\hat{\sigma}_{ate,\epsilon}^{hetero}(\ell)$ to obtain a scale invariant test statistic. Note that we use $\sigma_{\nu,n}^2((\mathbf{0}, 1))$ because when $\ell = (\mathbf{0}, 1)$, both $\nu_{hetero,ate}(\ell)$ and $\hat{\nu}_{hetero,ate}(\ell)$ reduce to zero and that estimated moment condition does not contribute to the test statistic $\widehat{S}_{ate}^{hetero}$. With smaller cubes, e.g. $\ell = (\mathbf{0}, 1/2)$, $\nu_{hetero,ate}(\ell)$ examines whether the ATE among individuals with characteristic values belonging to C_ℓ is equal to the population ATE multiplied by the proportion of such individuals. We follow Andrews and Shi (2013) and use $\epsilon = 0.05$. Let the simulated process $\widehat{\Phi}_{n,ate}^{hetero,u}(\ell)$ be

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

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

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

Let the decision rule be: "Reject $H_{0,ate}^{hetero}$ if $\widehat{S}_{ate}^{hetero} > \hat{c}_{n,ate}^{hetero}(\alpha)$." The following theorem summarizes the asymptotic propoerties of the proposed heterogeneity test.

Theorem 3.4 *Under Assumption 2.1 and Assumptions A.1, A.2 and A.3, when $\alpha < 1/2$, we have*

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

Note that this heterogeneity 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.

4 Testing in Fuzzy RD Design

In this section, we extend the proposed tests to the fuzzy RD design. Similar to the sharp RD case, we are interested in testing the following three null hypotheses:

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

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

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

Recall that

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

Since Assumption 2.2.(iv) requires that $E[T_i(1)-T_i(0)|X_i = x, Z_i = c] > 0$ for all $x \in \mathcal{X}_c$, 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, for all $x \in \mathcal{X}_c$. In other words, 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 $x \in \mathcal{X}_c$. Both $CLATE(x)$ and $LATE$ could be estimated using local linear regression methods. However, as is discussed in Feir et al. (2015), when the sample size is small or the proportion of compliers is small, the nonparametric RD estimator of $LATE$ can have Cauchy-type finite sample distribution analogous to the concerns raised in the weak IV literature (see Staiger and Stock, 1997, Stock and Yogo, 2005, among others). The problem of weak identification is worse with $CLATE$, as the effect conditions not only on the running variable Z_i but also on the additional covariate X_i. Further, in many empirical applications, the first stage is heterogeneous resulting low complying rates for some x values. Therefore, heterogeneity tests relying on plug-in estimators of $LATE$ and $CLATE(x)$ such as the subsample regression method are not ideal. In the next, we look at null transformations that can avoid the use of plug-in estimators of $LATE$ and $CLATE(x)$.

Let

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

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 individuals with $X_i \in C_\ell$. In the appendix, we show that the null hypothesis in (4.3) is equivalent to

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

Let $\hat{\mu}(\ell)$ be the estimator for $\mu(\ell)$ that is defined in the same way as $\hat{\nu}(\ell)$ except that Y_i is replaced by T_i. Let $\hat{\nu}_{hetero,late}(\ell) = \hat{\nu}(\ell) \cdot \hat{\mu}((\mathbf{0}, 1)) - \hat{\nu}((\mathbf{0}, 1)) \cdot \hat{\mu}(\ell)$ be the estimator for $\nu_{hetero,late}(\ell)$. Let $\phi_{\mu,ni}(\ell)$ be the influence function for $\sqrt{nh}(\hat{\mu}(\ell) - \mu(\ell))$ that is defined in the same way as $\phi_{\nu,ni}(\ell)$ except that Y_i is replaced by T_i, and let $\hat{\phi}_{\mu,ni}(\ell)$ be its estimator. Let $\phi_{late,ni}^{hetero}(\ell) = \mu((\mathbf{0}, 1)) \cdot \phi_{\nu,ni}(\ell) + \nu(\ell) \cdot \phi_{\mu,ni}((\mathbf{0}, 1)) - \nu((\mathbf{0}, 1)) \cdot \phi_{\mu,ni}(\ell) - \mu(\ell) \cdot \phi_{\nu,ni}((\mathbf{0}, 1))$

and $\hat{\phi}^{hetero}_{late,ni}(\ell) = \hat{\mu}((\mathbf{0},1))\cdot\hat{\phi}_{\nu,ni}(\ell)+\hat{\nu}(\ell)\cdot\hat{\phi}_{\mu,ni}((\mathbf{0},1))-\hat{\nu}((\mathbf{0},1))\cdot\hat{\phi}_{\mu,ni}(\ell)-\hat{\mu}(\ell)\cdot\hat{\phi}_{\nu,ni}((\mathbf{0},1))$ be its estimator. Let $\left(\hat{\sigma}^{hetero}_{late,n}(\ell)\right)^2 = \sum_{i=1}^{n}\left(\hat{\phi}^{hetero}_{late,ni}(\ell)\right)^2$. Define the test statistic for $H^{hetero}_{0,late}$ as

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

where $\hat{\sigma}^{hetero}_{late,\epsilon}(\ell) = \sqrt{\max\left\{\left(\hat{\sigma}^{hetero}_{late,n}(\ell)\right)^2, \epsilon\cdot\hat{\sigma}^2_{\nu,n}((\mathbf{0},1))\right\}}$ for some smal positive ϵ. Define the simulated process $\widehat{\Phi}^{hetero,u}_{n,late}(\ell)$ as

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

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

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

Finally, the decision rule would be: "Reject $H^{hetero}_{0,late}$ if $\widehat{S}^{hetero}_{late} > \hat{c}^{hetero}_{n,late}(\alpha)$." Again, the proposed test for $H^{hetero}_{0,late}$ controls size asymptotically and is consistent. We omit the details of the size and power properties in the interest of space.

5 Simulations

In this section, we carry out Monte Carlo simulations. First, we investigate the small sample size and power performance of the proposed tests using data generating processes (DGPs) estimated from the dataset in the empirical section. Second, we modify the data-driven DGPs to 1) compare the size and power performances of the proposed uniform sign test based on the LFC and the GMS critical values, 2) demonstrate the size distortion of the interaction term method and the subsample regression method which are popular in the applied RD literature for heterogeneity analysis, and 3) study the small sample performance of proposed tests when the DGP introduces larger finite sample bias in local linear estimation.

For all DGPs, the running variable Z, the additional control X, and the error term u in the outcome equation are generated following

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

The outcome Y and the treatment decision T are DGP specific. With each DGP, 5,000 simulation samples are drawn unless otherwise noted. In each test, the bootstrap critical value is calculated using 1,000 bootstrap simulations.

All tests carried out in this section use the triangular kernel and bandwidths selected according to the formula $h_{CCT} \times n^{1/5-1/k}$, where h_{CCT} is the robust bandwidth following Calonico et al. (2014) (CCT) and k is the under-smoothing parameter. In all simulation tables, we report results with $k = 4.25, 4.5$ and 4.75. The cubes defined in equation (3.2) have side-lengths $1/q$ for $q = 1, ..., Q$. We use benchmark $Q = 10$ which includes a total of 55 overlapping intervals. In DGPs 1-4, for example, when $n = 1,000$, since the average bandwidth ranges from 0.22 to 0.29, the expected effective sample size of the smallest cubes ranges from 16 to 21[6] (for each local linear regression on one side of the RD cut-off) with the benchmark Q choice. In Tables 1 and 2 we also report robustness checks with $Q = 7$ and 13.

Next, we give models of Y and T for the first set of four data-driven DGPs. The left and middle graphs in Figure 1 visualize the outcome equation in these DGPs.

DGP 1: Sharp RD, Homogeneous Zero Effect

$$Y = -0.555 + 0.581X - 0.553 + 0.060XZ - 0.058Z^2 + 1.074X^2 + 0.1u;$$
$$T = 1(Z > 0).$$

DGP 2: Sharp RD, Heterogeneous Effects

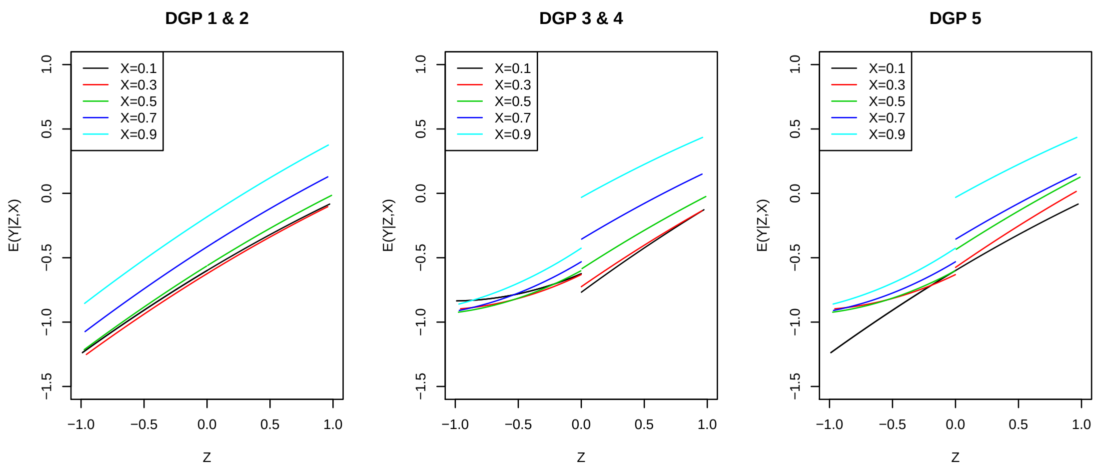
$$Y = \begin{cases} -0.755 - 0.254X + 0.742Z - 0.219XZ - 0.063Z^2 + 1.175X^2 + 0.1u & \text{if } Z \geq 0 \\ -0.607 - 0.220X + 0.386Z + 0.288XZ + 0.204Z^2 + 0.469X^2 + 0.1u & \text{if } Z < 0 \end{cases};$$
$$T = 1(Z > 0).$$

DGP 3: Fuzzy RD, Homogeneous Zero Effect

$$Y \sim \ DGP\ 1;$$
$$T = \begin{cases} 1(0.596 - 2.103X + 0.128Z + 0.352XZ + 0.013Z^2 + 2.454X^2 + u > 0) & \text{if } Z \geq 0 \\ 0 & \text{if } Z < 0 \end{cases}.$$

[6] Andrews and Shi (2013) suggests choosing Q such that the smallest cubes have expected sample size around 10 to 20.

Figure 1: The Data Generating Processes: DGPs 1-5



Note: DGPs 1-4 are estimated from the dataset in the empirical section. To obtain the outcome equation in DGPs 1 and 3, we first normalize the additional control of interest (i.e. peer transition score) in that dataset to $[-1, 1]$, and then regress the outcome (i.e. Baccalaureate exam score) on the running variable (i.e. transition score), the additional control of interest, as well as their interaction term and second order polynomial terms. To obtain the outcome equation in DGPs 2 and 4, we fit the same regression model described above separately for the subsamples to the left and the right of the cutoff (i.e. zero) of the running variable. The first stage equation in DGPs 3 and 4 are estimated through running two separate Probit regressions of the treatment decision (i.e. dummy for attending a more selective high school) on the running variable, the additional control, their interaction term and second order polynomial terms using data to the left and the right of the cutoff. DGP 5 is a combination of DGP 1 and 2 as is described in the paper.

DGP 4: Fuzzy RD, Heterogeneous Effects

$$
\begin{aligned}
&Y \sim \; DGP\ 2; \\
&T = \begin{cases} 1(0.596 - 2.103X + 0.128Z + 0.352XZ + 0.013Z^2 + 2.454X^2 + u > 0) & \text{if } Z \geq 0 \\ 0 & \text{if } Z < 0 \end{cases}.
\end{aligned}
$$

Table 1 report the simulation results 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 $H_0 : CATE(x) = ATE,\ \forall x \in [0, 1]$. It also reports the results of the standard mean test $H_0 : ATE = 0$ for comparison. Simulation results of the uniform significance test are omitted as they are highly similar to the results of the uniform sign test. DGPs 1 & 3 are constructed to show the small sample size property of the three tests. DGPs 2 & 4 are constructed to illustrate the power property. Recall from the discussions in Section 4 that the mean test and the uniform sign test have exactly the same testing results under DGPs 1 & 3 and DGPs 2 & 4, so they are only reported once in Table 1.

The simulation results show that the proposed tests control size well. Even when the sample size is small with $n = 1000$, the rejection rate is controlled under 5.5%, comparing to the 5% significance level. The proposed tests also have good power performance with the rejection rate going to one as the sample size increases. The reported rejection rate is somewhat dependent on the bandwidth choice, which is common to all kernel based tests. In the empirical application, we also report testing results with the same three under-smoothing parameters. We find that our empirical results are very robust to the bandwidth choice. Last but not least, simulation results in Table 1 show that the size and power performance of proposed tests are robust to the Q choice.

In simulations conducted above, we have been using the LFC critical value in the uniform sign test. Next, we illustrate the size and power performances of the uniform sign test with GMS critical values. The right graph in Figure 1 visualizes the outcome equation in DGP 5. We test both the null hypothesis of uniform non-positive effect ($H_0 : CATE(x) \leq 0,\ \forall x \in [0, 1]$) and the null of uniform non-negative effect ($H_0 : CATE(x) \geq 0,\ \forall x \in [0, 1]$).

DGP 5: Sharp RD, Mixture of Zero and Positive Effects

When $X < 0.3, Y \sim$ *DGP* 1; when $X \geq 0.6, Y \sim$ *DGP* 1; when $0.3 \leq X < 0.6$;

$$Y = \begin{cases} -0.905 + 0.742X - 0.254Z - 0.219XZ - 0.063Z^2 + 1.175X^2 + 0.1u & \text{if } Z \geq 0 \\ -0.607 + 0.386X - 0.220Z + 0.288XZ + 0.204Z^2 + 0.469X^2 + 0.1u & \text{if } Z < 0 \end{cases}.$$

With DGP 5, the null hypothesis of non-negative effect is true but the LFC only holds when $X < 0.3$. Therefore, using the LFC critical value should result in rejection rates lower than the significance level, or 5%, while using the GMS critical value should bring the rejection rate closer to 5%. This is exactly what is shown in Table 2. On the other hand, the null hypothesis of non-positive effect is false, and Table 2 shows that using the GMS critical value can slightly improve the power performance of the uniform sign test.

DGPs 6 and 7 are used to compare the proposed heterogeneity test (Hetero) with the interaction term method (Hetero-INT) and the subsample regression method (Hetero-SUB) that are popular in the applied RD literature. Details of the Hetero-INT and

Table 1: Small Sample Performance of Proposed Tests

	$H_0 : ATE = 0$			$CATE(\cdot) \leq 0$			$CATE(\cdot) = ATE$			$CLATE(\cdot) = LATE$		
	Sharp & Fuzzy RD			Sharp & Fuzzy RD			Sharp RD			Fuzzy RD		
	(DGPs 1-4)			(DGPs 1-4)			(DGPs 1 & 3)			(DGPs 2 & 4)		
	k=4.25	4.5	4.75	4.25	4.5	4.75	4.25	4.5	4.75	4.25	4.5	4.75
Panel A: $Q = 10$												
DGP 1 & 3: Homogeneous Zero Effect												
n=1000	0.053	0.056	0.059	0.054	0.054	0.053	0.053	0.052	0.048	0.045	0.045	0.044
n=2000	0.057	0.059	0.060	0.055	0.053	0.054	0.051	0.054	0.054	0.050	0.052	0.052
n=4000	0.050	0.050	0.053	0.047	0.049	0.049	0.047	0.048	0.046	0.046	0.047	0.045
n=8000	0.052	0.056	0.057	0.048	0.052	0.051	0.049	0.048	0.051	0.049	0.047	0.052
DGP 2 & 4: Heterogeneous Treatment Effect												
n=1000	0.279	0.301	0.324	0.279	0.318	0.357	0.178	0.191	0.205	0.145	0.167	0.177
n=2000	0.497	0.527	0.561	0.616	0.672	0.725	0.323	0.362	0.395	0.272	0.303	0.330
n=4000	0.745	0.779	0.809	0.935	0.958	0.972	0.630	0.686	0.736	0.529	0.583	0.635
n=8000	0.924	0.946	0.958	0.999	1.000	1.000	0.931	0.956	0.974	0.864	0.902	0.930
Panel B: $Q = 7$												
DGP 1 & 3: Homogeneous Zero Effect												
n=1000	0.053	0.056	0.059	0.054	0.054	0.054	0.055	0.055	0.050	0.047	0.047	0.046
n=2000	0.057	0.059	0.060	0.055	0.053	0.054	0.051	0.055	0.053	0.049	0.052	0.050
n=4000	0.050	0.050	0.053	0.048	0.048	0.049	0.048	0.048	0.046	0.047	0.047	0.045
n=8000	0.052	0.056	0.057	0.048	0.052	0.051	0.050	0.049	0.052	0.049	0.049	0.051
DGP 2 & 4: Heterogeneous Treatment Effect												
n=1000	0.279	0.301	0.324	0.286	0.325	0.365	0.184	0.196	0.212	0.151	0.169	0.182
n=2000	0.497	0.527	0.561	0.623	0.679	0.731	0.329	0.369	0.403	0.278	0.310	0.337
n=4000	0.745	0.779	0.809	0.937	0.960	0.974	0.634	0.694	0.742	0.534	0.589	0.644
n=8000	0.924	0.946	0.958	0.999	1.000	1.000	0.935	0.957	0.976	0.867	0.906	0.932
Panel C: $Q = 13$												
DGP 1 & 3: Homogeneous Zero Effect												
n=1000	0.053	0.056	0.059	0.054	0.054	0.053	0.052	0.051	0.047	0.045	0.045	0.043
n=2000	0.057	0.059	0.060	0.054	0.053	0.053	0.051	0.053	0.054	0.049	0.052	0.051
n=4000	0.050	0.050	0.053	0.047	0.048	0.048	0.048	0.048	0.046	0.046	0.047	0.044
n=8000	0.052	0.056	0.057	0.048	0.052	0.051	0.049	0.048	0.051	0.049	0.048	0.052
DGP 2 & 4: Heterogeneous Treatment Effect												
n=1000	0.279	0.301	0.324	0.277	0.316	0.354	0.177	0.190	0.204	0.146	0.165	0.175
n=2000	0.497	0.527	0.561	0.613	0.670	0.723	0.321	0.360	0.393	0.271	0.301	0.329
n=4000	0.745	0.779	0.809	0.934	0.958	0.972	0.627	0.685	0.735	0.527	0.581	0.634
n=8000	0.924	0.946	0.958	0.999	1.000	1.000	0.928	0.956	0.974	0.862	0.901	0.930

Note: Reported are rejection proportions among 5,000 simulations, where all tests are carried out using the 5% significance level. The uniform sign tests are based on the LFC critical value. For each test, the simulated critical value is calucalted with 1,000 bootstrap repetitions.

Table 2: Uniform Sign Test Using the LFC and the GMS Critical Values

	$H_0 : CATE(\cdot) \geq 0$ (LFC)			$CATE(\cdot) \leq 0$ (LFC)			$CATE(\cdot) \geq 0$ (GMS)			$CATE(\cdot) \leq 0$ (GMS)		
	k=4.25	4.5	4.75	4.25	4.5	4.75	4.25	4.5	4.75	4.25	4.5	4.75
					Panel A: $Q = 10$							
n=1000	0.040	0.041	0.040	0.800	0.841	0.880	0.048	0.049	0.049	0.813	0.854	0.888
n=2000	0.040	0.040	0.039	0.982	0.988	0.992	0.047	0.048	0.045	0.983	0.989	0.993
n=4000	0.033	0.033	0.033	1.000	1.000	1.000	0.039	0.040	0.041	1.000	1.000	1.000
n=8000	0.031	0.030	0.029	1.000	1.000	1.000	0.038	0.038	0.035	1.000	1.000	1.000
					Panel B: $Q = 7$							
n=1000	0.029	0.031	0.032	0.659	0.716	0.761	0.038	0.038	0.040	0.672	0.729	0.769
n=2000	0.032	0.030	0.030	0.933	0.957	0.971	0.039	0.039	0.038	0.934	0.959	0.972
n=4000	0.023	0.024	0.022	0.998	0.999	1.000	0.031	0.030	0.030	0.999	0.999	1.000
n=8000	0.023	0.022	0.023	1.000	1.000	1.000	0.029	0.029	0.031	1.000	1.000	1.000
					Panel C: $Q = 13$							
n=1000	0.040	0.041	0.040	0.798	0.839	0.877	0.047	0.048	0.048	0.811	0.852	0.886
n=2000	0.039	0.038	0.039	0.981	0.988	0.992	0.047	0.047	0.045	0.982	0.989	0.993
n=4000	0.033	0.033	0.033	1.000	1.000	1.000	0.039	0.040	0.040	1.000	1.000	1.000
n=8000	0.033	0.033	0.032	1.000	1.000	1.000	0.038	0.038	0.039	1.000	1.000	1.000

Note: Reported are rejection proportions among 5,000 simulations, where all tests are carried out using the 5% significance level. For each test, the simulated critical value is calucalted with 1,000 bootstrap repetitions.

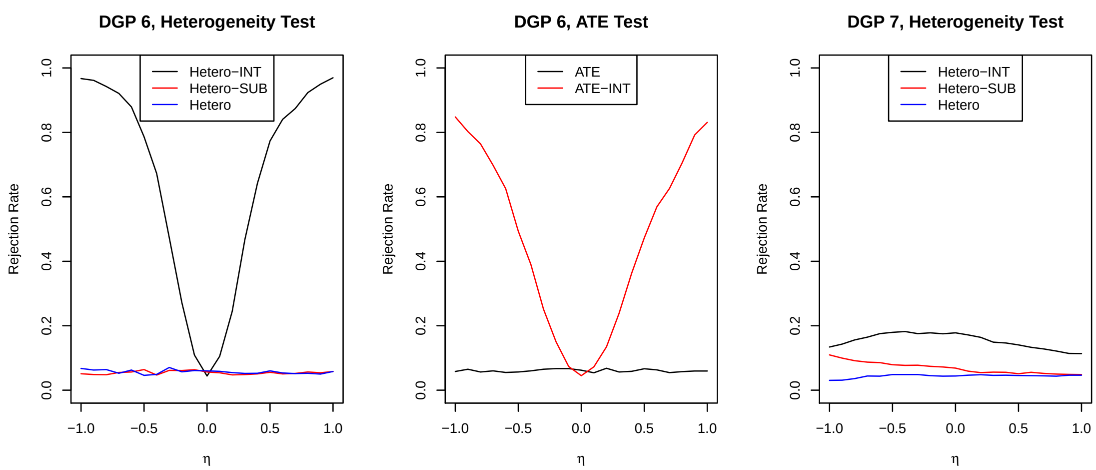
Hetero-SUB tests are described in the footnote of Figure 2.

DGP 6: Sharp RD, Homogeneous Zero Effect

$$
\begin{aligned}
Y =& -0.483 - 1.376X + 0.301Z \\
&+ \eta(0.112XZ + 0.194Z^2 + 3.234X^2 - 0.295XZ^2 + 0.469XZ^2 - 1.548X^3 - 0.021Z^3) \\
&+ 0.1u; \\
T =& 1(Z > 0).
\end{aligned}
$$

In DGP 6, the treatment effect is homogenous and zero, and the control parameter η determines the degree of model misspecification. The left graph of Figure 2 summarizes the size property of the three tests. We notice that the parametric Hetero-INT test controls size at 5% only when $\eta = 0$, or when the linear regression model is correctly specified. In contrast, the Hetero and Hetero-SUB tests control size well irrespective of η because both tests are nonparametric. Besides the three tests for treatment effect heterogeneity, we also report in the middle graph of Figure 2 rejection rates of the sig-

Figure 2: Performance of Naive and Proposed Testing Methods



Note: Reported are rejection proportions among 1,000 simulations. All tests are carried out using the 5% significance level and the simulated critical values calucalted with 1,000 bootstrap repetitions. The sample size is 1000. The Hetero test is the proposed heterogeneity test with $k = 4.5$ and $Q = 10$. 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 subsample ATE estimates of 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 ATE. The Hetero-SUB adjusts for multiple testing using the Bonferroni method. In DGP 7, the first-stage take-up rate ranges from 0.15 to 0.3 when $\eta = -1$ and ranges from 0.55 to 0.75 when $\eta = 1$.

nificance test of ATE from the classic RD regression (Mean RD test) and the interaction term method (Mean RD-INT test). We see that the Mean RD-INT test also over-rejects severely unless $\eta = 0$, which further supports our recommendation against the interaction term method in RD heterogeneity analysis.

DGP 7: Fuzzy RD, Homogeneous Positive Effect

$$Y = \begin{cases} -0.755 + 0.742Z - 0.063Z^2 + 0.1u & \text{if } Z \geq 0 \\ -0.607 + 0.386Z + 0.204Z^2 + 0.1u & \text{if } Z < 0 \end{cases};$$

$$T = \begin{cases} 1(\eta \cdot 0.596 - 2.103X + 0.128Z + 0.352XZ + 0.013Z^2 + 2.454X^2 + u > 0) & \text{if } Z \geq 0 \\ 0 & \text{if } Z < 0 \end{cases}.$$

DGP 7 is modified from DGP 2 by suppressing the role of the additional covariate X in the outcome equation so that the model has a homogeneous positive effect. The control parameter η in the selection equation determines the first stage take-up rate. The smaller the value of η, the weaker the first stage. As is shown in the right graph of Figure 2, not surprisingly, the proposed heterogeneity test has excellent size control while the Hetero-SUB test has sizable over-rejection when the first-stage is weak. The parametric Hetero-INT test again overrejects because of model misspecification.

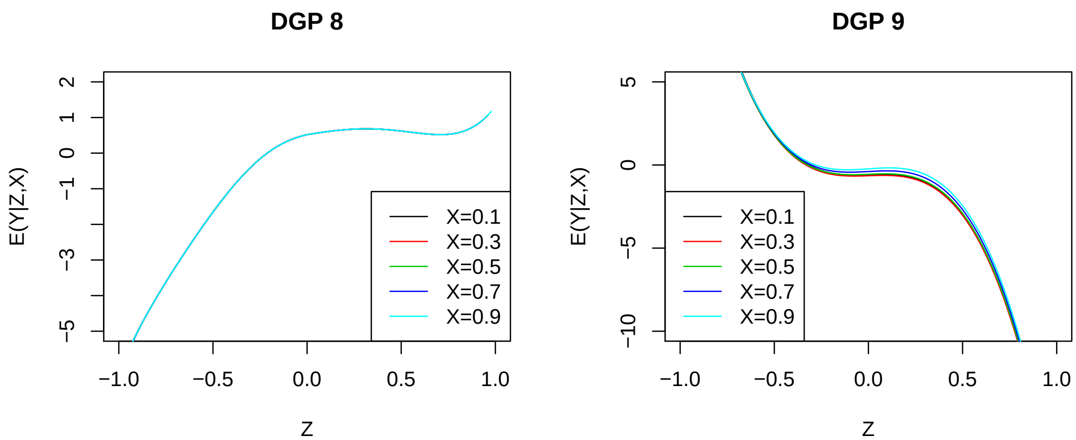
Last but not least, we examine the small sample performance of the proposed tests when DGP models have asymmetric and exaggerated curvature patterns around the cut-off of the running variable. DGP 8 is taken from Calonico et al. (2014), which is specifically designed to show the importance of bias correction. In this paper, we do not use the bias correction technique but instead require under-smoothed bandwidth to avoid having nuisance bias terms in the limiting distributions of local linear estimators. Therefore, this DGP from Calonico et al. (2014) should serve as a good example to examine the small sample performance of our testing methods when the DGP does not favor the under-smoothing technique that we employ.

DGP 8: Sharp RD, Homogeneous Zero Effect, Exaggerated Curvature

$$Y = \begin{cases} 0.52 + 0.84Z - 0.3Z^2 - 2.4Z^3 - 0.9Z^4 + 3.56Z^5 + 0.1u & \text{if } Z \geq 0 \\ 0.52 + 1.27Z - 3.59Z^2 + 14.15Z^3 + 23.69Z^4 + 11.36Z^5 + 0.1u & \text{if } Z < 0 \end{cases};$$

$$T = 1(Z > 0).$$

Figure 3: The Data Generating Processes: DGPs 8-9



Note: DGP 8 is taken from Calonico et al. (2014). DGP 9 is modified from a data-driven model estimated from the dataset of the empirical section.

The left graph in Figure 3 illustrates the model in DGP 8 and panel A of Table 3 reports the simulation results. For results reported in the first nine columns, we only see slight over-rejection with rejection rates always under 7%. The rejection rate also gets quite close to the 5% significance level when the sample size gets larger. However, when the GMS critical value is used for the uniform sign test, the rejection rate gets higher to close to 9% when $n = 1,000$, although it then steadily decreases as the sample size gets larger.

DGP 9: Sharp RD, Homogeneously Zero Effect, Exaggerated Curvature

$$Y = \begin{cases} -0.905 + 0.742X - 0.254Z - 0.219XZ - 0.063Z^2 + 1.175X^2 + 0.1u & \text{if } Z \geq 0 \\ -0.607 + 0.386X - 0.220Z + 0.288XZ + 0.204Z^2 + 0.469X^2 + 0.1u & \text{if } Z < 0 \end{cases};$$

$T = 1(Z > 0)$.

To verify the findings of DGP 8, we complement it with DGP 9, which is modified from a higher-order polynomial model estimated using our empirical dataset. DGP 9 also has exaggerated curvature that is asymmetric around the cut-off of the running variable, as is shown in the right graph of Figure 3. Test esults are reported in Panel B of Table 3.

Table 3: Uniform Sign Test Using the LFC and the GMS Critical Values

	$H_0 : ATE = 0$			$CATE(\cdot) \leq 0$ (LFC)			$CATE(\cdot) = LATE$			$CATE(\cdot) \leq 0$ (GMS)		
	k=4.25	4.5	4.75	4.25	4.5	4.75	4.25	4.5	4.75	4.25	4.5	4.75
Panel A: DGP 8												
n=1000	0.058	0.058	0.061	0.062	0.063	0.066	0.059	0.058	0.057	0.081	0.085	0.086
n=2000	0.057	0.055	0.058	0.057	0.056	0.055	0.052	0.055	0.052	0.074	0.073	0.071
n=4000	0.054	0.055	0.059	0.057	0.059	0.060	0.054	0.057	0.056	0.077	0.075	0.075
n=8000	0.061	0.063	0.062	0.058	0.061	0.059	0.055	0.057	0.056	0.076	0.073	0.072
Panel B: DGP 9												
n=1000	0.070	0.069	0.071	0.062	0.062	0.063	0.058	0.057	0.055	0.075	0.074	0.073
n=2000	0.060	0.063	0.063	0.061	0.057	0.057	0.052	0.052	0.050	0.073	0.069	0.070
n=4000	0.056	0.057	0.059	0.051	0.052	0.053	0.049	0.049	0.047	0.063	0.061	0.062
n=8000	0.055	0.057	0.061	0.053	0.053	0.053	0.051	0.053	0.056	0.062	0.061	0.062

Note: Reported are rejection proportions among 5,000 simulations. All tests are carried out using the 5% significance level and the simulated critical values calucalted with 1,000 bootstrap repetitions. All tests in this table uses $Q = 10$.

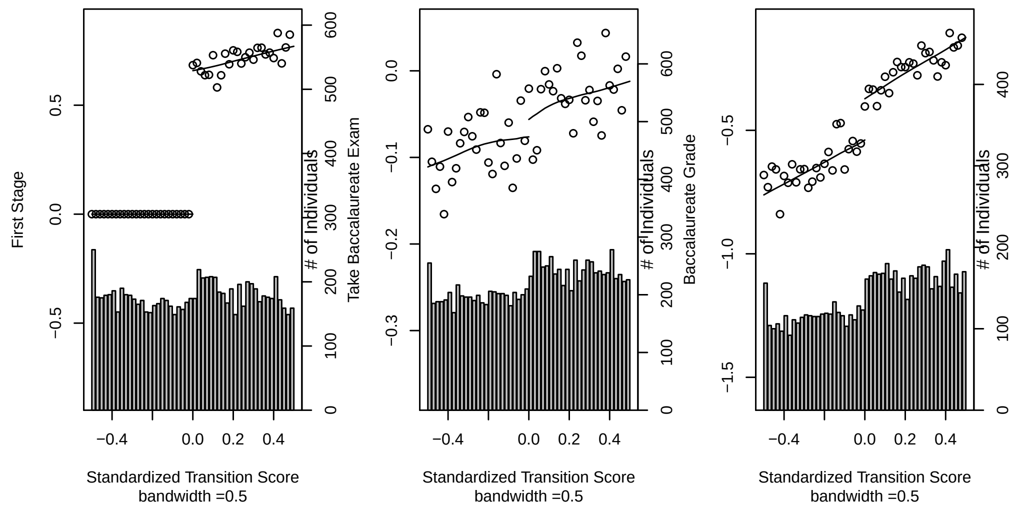
Similar to DGP 8, we only observe very mild over-rejection except for the uniform sign test with the GMS critical value. And the over-rejection problem always improves with the sample size.

In summary, we conclude that our proposed tests have very good small sample performance. When the underlying RD model has excess curvature and the sample size is small, using the GMS critical value for the uniform sign test might lead to moderate over-rejection. In such cases, the LFC critical value is recommended.

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 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 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

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 outcome within each bin and the local linear estimates. The bandwidth is set to 0.5 for all graphs for the purpose of data illustration and cross-comparison.

selective high school. They find that attending a better school significantly improves a student's performance on the Baccalaureate exam, but does not affect the exam take-up rate. Further, they find that a marginal student attending a more selective high school is also more likely to face negative peer interactions and perceive himself as weak.

In this section, we investigate the treatment effect heterogeneity among schools with different peer quality, where peer quality is defined as average admission score of the more selective school in town. In contrast to Pop-Eleches and Urquiola (2013), who find qualitatively similar results across schools in three terciles of the admission score cut-off, we find a clear signal that attending a better high school has a heterogeneous effect on whether a marginal student takes the Baccalaureate exam.

Figure 4 summarizes the classic mean RD results. Note that we restrict our attention to two-school towns because score cutoffs within a town are often quite close and estimation bias might be introduced as a result.[7] In all three graphs, the x-axis represents

[7]In fact, it is easy to prove that if the potential outcome monotonically increases with the running variable and also 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

Table 4: Uniform Sign and Heterogeneity Tests - P-values

$LATE = 0$			$CLATE(\cdot) \leq 0$			$CLATE(\cdot) \geq 0$			$CLATE(\cdot) = LATE$		
k=4.25	4.5	4.75	4.25	4.5	4.75	4.25	4.5	4.75	4.25	4.5	4.75
Q=100											
Exam-taking Rate (LFC for Uniform Sign Tests)											
0.482	0.350	0.285	0.012	0.007	0.005	0.104	0.116	0.111	0.004	0.001	0.001
Exam Grade (LFC for Uniform Sign Tests)											
0.002	0.003	0.002	0.002	0.002	0.001	0.652	0.687	0.749	0.074	0.103	0.135
Exam-taking Rate (GMS for Uniform Sign Tests)											
-	-	-	0.012	0.007	0.005	0.092	0.099	0.097	-	-	-
Exam Grade (GMS for Uniform Sign Tests)											
-	-	-	0.002	0.002	0.001	0.545	0.579	0.662	-	-	-
First-stage Take-up Rate (LFC for Uniform Sign Tests)											
0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.003	0.003	0.006
First-stage Take-up Rate (GMS for Uniform Sign Tests)											
-	-	-	0.000	0.000	0.000	1.000	1.000	1.000	-	-	-

Notes: Data are from Pop-Eleches and Urquiola (2013). Nonparametric local linear estimations are conducted using the triangular kernel and the undersmoothed CCT bandwidth defined in the simulation section. All simulated critical values are calculated with 1,000 bootstrap repetitions.

the running variable, which is a student's standardized transition score subtracting the school admission cut-off. The y-axis in the left graph represents the first stage take-up rate, or the proportion of eligible students attending a more selective school. The y-axis in the middle and right graphs represent two different outcomes, the demeaned probability of a student taking the Baccalaureate exam and the demeaned Baccalaureate exam grade among exam-takers, respectively. Both outcomes are demeaned by subtracting the school fixed effects following Pop-Eleches and Urquiola (2013). Both the middle and the right graphs see a jump in the average outcome at the discontinuity point, although the jump in the exam-taking rate is quite noisy.

Table 4 reports the testing results for the heterogeneity analysis. All tests use the triangular kernel, the undersmoothed CCT bandwidth defined in the simulation section and the cubes defined in (3.2). In Table 4, the tests are conducted with $Q = 100$.[8]

estimator.

[8]Since this empirical sample has a very large sample size, if the recommendation in (Andrews and Shi, 2013), also see footnote 6 of the paper, is followed, the Q value needs to be 200-400, which is computationally highly demanding. Instead, we report results with $Q = 75$, 100, and 125. We find our empirical results insensitive to the Q choice.

The testing results are very interesting. As is shown in Figure 4 and by the first two numbers of Table 4, the average effect of attending a better school on the probability of a marginal student taking the Baccalaureate exam is noisy and statistically insignificant. But the testing results in Table 4 reveal that 1) we can reject the null of non-positive effect at the 1% significance level, 2) we can reject the null that the effect is non-negative at the 10% significance level when the GMS critical value is used, and 3) we can reject the null that the effect does not vary with peer quality in the more selective school at the 1% significance level. Adding up the three pieces of information we conclude that the insignificant LATE on exam-taking rate results from the cancellation of negative and positive effects among different groups of the population.

Besides, testing results in Table 4 confirm the positive effect of attending a better school on the Baccalaureate exam grade. Testing results also reveal strong evidence of first stage heterogeneity, which is intuitive as selective schools with higher peer quality are expected to have higher attendance rate among qualified students.

Table 5 conducts robustness checks with two alternative Q values. The results suggest that the above discussed empirical results are not sensitive to the Q choice.

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 identification. Monte Carlo simulations show that our tests have very good small sample performance. We apply our methods to a dataset from Romania and reveal strong evidence of treatment effect heterogeneity previously neglected by the literature.

There are several interesting directions for future research. First, one could extend the proposed testing procedure to examine heterogeneity in distributional treatment effects (see previous researches in the treatment effect or RD literature by Bitler et al., 2008; Shen and Zhang, 2016; Shen, forthcoming, among others), rank shuffling (Dong and Shen, forthcoming), or the marginal threshold treatment effect (Dong and Lewbel, 2015). Second, as is pointed out by a referee, it is of practical interest to identify the set of

Table 5: Uniform Sign and Heterogeneity Tests - Alternative Q Values

$LATE = 0$			$CLATE(\cdot) \leq 0$			$CLATE(\cdot) \geq 0$			$CLATE(\cdot) = LATE$		
k=4.25	4.5	4.75	4.25	4.5	4.75	4.25	4.5	4.75	4.25	4.5	4.75
Q=75											
Exam-taking Rate (LFC for Uniform Sign Tests)											
0.482	0.350	0.285	0.012	0.007	0.005	0.104	0.116	0.111	0.004	0.001	0.001
Exam Grade (LFC for Uniform Sign Tests)											
0.002	0.003	0.002	0.002	0.002	0.001	0.651	0.686	0.747	0.074	0.103	0.135
First-stage Take-up Rate (LFC for Uniform Sign Tests)											
0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.003	0.003	0.006
Exam-taking Rate (GMS for Uniform Sign Tests)											
-	-	-	0.012	0.007	0.005	0.092	0.099	0.097	-	-	-
Exam Grade (GMS for Uniform Sign Tests)											
-	-	-	0.002	0.002	0.001	0.544	0.578	0.659	-	-	-
First-stage Take-up Rate (GMS for Uniform Sign Tests)											
-	-	-	0.000	0.000	0.000	1.000	1.000	1.000	-	-	-
Q=125											
Exam-taking Rate (LFC for Uniform Sign Tests)											
0.482	0.350	0.285	0.012	0.007	0.005	0.104	0.116	0.111	0.004	0.001	0.001
Exam Grade (LFC for Uniform Sign Tests)											
0.002	0.003	0.002	0.002	0.002	0.001	0.652	0.687	0.749	0.074	0.103	0.135
First-stage Take-up Rate (LFC for Uniform Sign Tests)											
0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.003	0.003	0.006
Exam-taking Rate (GMS for Uniform Sign Tests)											
-	-	-	0.012	0.007	0.005	0.092	0.099	0.097	-	-	-
Exam Grade (GMS for Uniform Sign Tests)											
-	-	-	0.002	0.002	0.001	0.545	0.579	0.663	-	-	-
First-stage Take-up Rate (GMS for Uniform Sign Tests)											
-	-	-	0.000	0.000	0.000	1.000	1.000	1.000	-	-	-

Notes: Data are from Pop-Eleches and Urquiola (2013). Nonparametric local linear estimations are conducted using the triangular kernel and the undersmoothed CCT bandwidth defined in the simulation section. All simulated critical values are calculated with 1,000 bootstrap repetitions.

covariate values with strictly positive CATE, i.e., $\mathcal{X}_c^+ \equiv \{x \in \mathcal{X}_c : \; CATE(x) > 0\}$. The testing procedure proposed in this paper is capable of identifying hypercubes with positive treatment effects on average (see discussions following introducing hypotheses (3.3)), but could not be directly used to uncover the $\mathcal{X}_c^+$ set. To study the $\mathcal{X}_c^+$ set, one might want to adapt the testing procedure in Chernozhukov et al. (2013), which involves estimating $CATE$ using higher order local linear regressions (see, for example, Armstrong and Shen, 2012).

APPENDIX

Appendix A gives Assumptions A.1, A.2 and A.3 that are required to show the asymptotic properties of the proposed tests. Appendix B collects proofs of Lemmas, Theorems, and some other equivalence results stated in the paper.

A Regularity Conditions

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 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 all the proofs will still go through, but with much more complicated notations. 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 (2017).

Assumption A.2 *Assume that*

(i) The function $K(\cdot)$ is a non-negative symmetric bounded kernel with a compact support.

(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 term converge to zero even after we multiply it with $\sqrt{nh}$.

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 $M > 0$. $\{U_i : 1 \leq i \leq n\}$ is independent of the sample path $\{(Y_i, X_i, Z_i, T_i) : 1 \leq i \leq n\}$.*

Assumption A.3 is required for the validity of the multiplier bootstrap.

B Proofs

B.1 Identification

Proof of Equation (2.1):

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

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 assumptions in Assumptions 2.2.(i) and (ii). The rest of the equalities hold from standard derivations.

Further,

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

by the definition of T_i and the continuity assumptions in Assumption 2.2.(ii). Collecting the two results proves the identification result stated in Equation (2.1).

Also, notice that when T_i is deterministic and $E[T_i(1)-T_i(0)|X_i=x,Z_i=c]=1$, the above proof for Equation (2.1) also gives the identification result for $CATE(x)$ under the sharp RD design. □

B.2 Lemmas

Proof of Lemma 3.1: Without loss of generality, we assume that X is a scalar, so $d_x = 1$. To show the equivalence of hypotheses in (3.1) and (3.3), we just need to show the equivalence of the null hypotheses. The direction from (3.1) to (3.3) is straightforward, so we omit the details. To show the other direction, suppose that there is an $x^* \in \mathcal{X}_c$ such that $CATE(x^*) > 0$. Then by the continuity of $E[Y_i(t)|X_i = x, Z_i = c]$ in Assumption 2.1, we know that $CATE(x)$ is continuous in x as well. The continuity of $CATE(x)$ then implies that there exists $x_\ell < x_u$ such that $CATE(x) > 0$ for all $x \in [x_\ell, x_u]$. Then there exists $\ell^* = (x^*, r^*) \in \mathcal{L}$ with $[x^*, x^* + r^*] \subseteq [x_\ell, x_u]$ so that $E[g_{\ell^*}(X_i)CATE(X_i)|Z_i = c] > 0$. Therefore, the null in (3.3) implies the null in (3.1). This then completes the proof of the lemma. $\square$

Proof of Lemma 3.2: Define

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

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

By the independence of data, 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

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

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 (FCLT), 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 in Pollard (1990). To show condition (ii), note that

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

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) of Theorem 10.2 in Pollard (1990), note that for any $\epsilon > 0$,

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

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) of Theorem 10.2 in Pollard (1990), 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 the 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 $\sqrt{nh}(\hat{\nu}(\ell)-\nu(\ell))=\sqrt{nh}(\hat{m}_+(\ell)-m_+(\ell))-\sqrt{nh}(\hat{m}_-(\ell)-m_-(\ell))\Rightarrow\Phi_{h_{2,\nu}}(\ell)$. □

Proof of Lemma 3.3: 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

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

Note that this result implies that $\sup_{\ell\in\mathcal{L}} |\hat{\sigma}^2_{\mu,n}(\ell) - \sigma^2_{\mu}(\ell)| \xrightarrow{p} 0$.

Also, we can show that

$$nh \sum_{i=1}^{n} \left(|w^+_{ni}| \cdot (|Y_i| + \overline{|Y|}^+_n)\right)^2 \xrightarrow{p} M_1 < \infty,$$
$$n^3h^3 \sum_{i=1}^{n} \left(|w^+_{ni}| \cdot (|Y_i| + \overline{|Y|}^+_n)\right)^4 \xrightarrow{p} M_2 < \infty,$$

for some positive M_1 and M_2.

Then by the same proof of Theorem 2.1 of Hsu (2016), we can show that $\widehat{\Phi}^{+,u}_n(\ell) \overset{p}{\Rightarrow} \Phi_{h_{2,m+}}(\ell)$. These results complete the proof for Lemma 3.3. □

Proof of Lemma 3.4: Similar to Lemma 3.1, we assume that X is a scalar. It is straightforward to show that the null hypothesis in (3.8) implies the null hypothesis in (3.9), so we omit the details. To show the other direction, suppose that there is $x^* \in \mathcal{X}$ such that $CATE(x^*) \neq 0$ and without loss of generality, we assume that $CATE(x^*) > 0$. Then then continuity of $CATE(x)$ implies that there exist $x_\ell < x_u$ such that $CATE(x) > 0$ for all $x \in [x_\ell, x_u]$. Then there exists $\ell^* = (x^*, r^*) \in \mathcal{L}$ with $[x^*, x^* + r^*] \subseteq [x_\ell, x_u]$ so that $\nu(\ell^*) = E[g_{\ell^*}(X_i)CATE(X_i)|Z_i = c] > 0$. That is, $\nu(\ell^*) \neq 0$. This completes the proof. □

Proof of Lemma 3.5: Similar to Lemmas 3.1 and 3.4, we assume that X is a scalar. It is straightforward to show the null hypothesis in (3.10) implies the null hypothesis in (3.11), so we omit the details. To show the other direction, suppose that there is $x^* \in \mathcal{X}_c$ such that $CATE(x^*) \neq ATE = \nu((0,1))$ and without loss of generality, we assume that $CATE(x^*) > \nu((0,1))$. Then continuity of $CATE(x)$ implies that there exist $x_\ell < x_u$ such that $CATE(x) > \nu((0,1))$ for all $x \in [x_\ell, x_u]$. Then there exists $\ell^* = (x^*, r^*) \in \mathcal{L}$ with $[x^*, x^* + r^*] \subseteq [x_\ell, x_u]$ so that $E[g_{\ell^*}(X_i)CATE(X_i)|Z_i = c] > E[g_{\ell^*}(X_i)\nu((0,1))|Z_i = c] = \nu((0,1)) \cdot p(\ell)$. That is, $\nu_{hetero,ate}(\ell^*) \neq 0$. This completes the proof. □

B.3 Theorems

Proof of Theorem 3.1: Given the results of Lemma 3.2 and Lemma 3.3 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 3.3 and 3.4: Note that the process results and simulated process results for Theorem 3.3 and 3.4 are similar to Lemma 3.2 and Lemma 3.3, so we omit the details for brevity. Then, the proofs for Theorem 3.3 and 3.4 are similar to that for Theorem 3.1. □

B.4 Other Results

Proof of Equation (3.4):

For any $\ell \in \mathcal{L}$, we have

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

The first equality is due to the identification of $CATE$. The second equality is from the continuity of $X_i | Z_i = z$ around the neighborhood of c. The last equality is from the Law of Iterated Expectation. □

Proof of Equation (3.13): Note that

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

The $o_p(1)$ result in the second to the last equality is due to the uniform consistency of $\hat{p}(\cdot)$, the $o_p(1)$ result in the last equation is due to the inference function representations in Equations (3.5) and (3.12). This completes the proof. $\square$

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

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

It is straightforward to show that the above equality implies the null hypothesis in 3.11, so we omit the details. Next, we show the equality in the other direction.

Denote $\lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z]$ as $CARE(x)$, or conditional average reduced-from effect, and $\lim_{z \searrow c} E[T_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[T_i|X_i = x, Z_i = z])$ as $CAFE(x)$, or conditional average first-stage effect. Suppose there is an $x^* \in \mathcal{X}_c$ such that $CARE(x^*) \cdot \mu((\mathbf{0}, 1)) - CAFE(x^*) \cdot \nu((\mathbf{0}, 1)) \neq 0$ and without loss of generality, we assume that $CARE(x^*) \cdot \mu((\mathbf{0}, 1)) - CAFE(x^*) \cdot \nu((\mathbf{0}, 1)) > 0$. Assumption 2.2.(i) and (ii) imply that both $CARE(x)$ and $CAFE(x)$ are continuous in x. Therefore, $CARE(x) \cdot \mu((\mathbf{0}, 1)) - CAFE(x) \cdot \nu((\mathbf{0}, 1))$ is continuous in x and there exists $x_l < x_u$ such that for all $x \in [x_l, x_u]$, $CARE(x) \cdot \mu((\mathbf{0}, 1)) - CAFE(x) \cdot \nu((\mathbf{0}, 1)) > 0$. Then there exists $\ell^* = (x^*, r^*) \in \mathcal{L}$ with $[x^*, x^* + r^*] \subseteq [x_\ell, x_u]$ so that $E[g_{\ell^*}(X_i)\,(CARE(X_i) \cdot \mu((\mathbf{0}, 1)) - CAFE(X_i) \cdot \nu((\mathbf{0}, 1)))\,|Z_i = c] = \nu(\ell) \cdot \mu((\mathbf{0}, 1)) - \mu(\ell) \cdot \nu((\mathbf{0}, 1)) > 0$. This completes the proof. $\square$

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. and X. Shi (2017): "Inference Based on Many Conditional Moment Inequalities," *Journal of Econometrics*, 196(2), 275–287.

Andrews, D. and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78, 119–157.

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

——— (2015): "Nonparametric Inference Based on Conditional Moment Inequalities," *Journal of Econometrics*, 179, 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, North Holland, vol. 4, 2111–3155.

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, 533–575.

Angrist, J. D. and M. Rokkanen (2015): "Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff," *Journal of the American Statistical Association*, 110(512), 1331–1344.

Aradillas-Lopez, A., A. Gandhi, and D. Quint (2016): "A simple test for moment inequality models with an application to English auctions," *Journal of Econometrics*, 194.1, 96–115.

Armstrong, T. and S. Shen (2012): "Inference on Optimal Treatment Assignment," Working paper, UC Davis.

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

Bertanha, M. (2016): "Regression discontinuity design with many thresholds," *work.*

Bertanha, M. and G. W. Imbens (2014): "External validity in fuzzy regression discontinuity designs," *working paper, National Bureau of Economic Research*, No. w20773.

Bierens, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20.1, 105–134.

——— (1990): "A consistent conditional moment test of functional form," *Econometrica*, 58.6, 1443–1458.

Bierens, H. J. and W. Ploberger (1997): "Asymptotic theory of integrated conditional moment tests," *Econometrica: Journal of the Econometric Society*, 1129–1151.

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

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

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2016): "Regression discontinuity designs using covariates," *working paper, University of Michigan.*

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

Cattaneo, M. D., R. Titiunik, G. Vazquez-Bare, and L. Keele (2016): "Interpreting regression discontinuity designs with multiple cutoffs," *The Journal of Politics*, 78(4), 1229–1248.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737.

Chetverikov, D. (forthcoming): "Adaptive test of conditional moment inequalities," *Econometric Theory*.

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

Dong, Y. and A. Lewbel (2015): "Identifying the effect of changing the policy threshold in regression discontinuity models," *Review of Economics and Statistics*, 97(5), 1081–1092.

Dong, Y. and S. Shen (forthcoming): "Testing for Rank Invariance or Similarity in Program Evaluation." *Review of Economics and Statistics*.

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.

Hansen, P. R. (2005): "A test for superior predictive ability," *Journal of Business & Economic Statistics*, 23, 365–380.

Hsu, Y.-C. (2016): "Multiplier Bootstrap," Tech. rep., Academia Sinica.

——— (2017): "Consistent Tests for Conditional Treatment Effects," *Econometric Journal*, 20.1, 1–22.

Imbens, G. W. and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 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, 281–355.

Lee, S., K. Song, and Y.-J. Whang (2013): "Testing Functional Inequalities," *Journal of Econometrics*, 172.1, 14–32.

——— (2015): "Testing for a General Class of Functional Inequalities," *working paper*, arXiv:1311.1595.

Linton, O., K. Song, and Y.-J. Whang (2010): "An improved bootstrap test of stochastic dominance," *Journal of Econometrics*, 154, 186–202.

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, 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. (forthcoming): "Estimation and Inference of Distributional Partial Effects: Theory and Application," *Journal of Business and Economic Statistics*.

Shen, S. and X. Zhang (2016): "Distributional Tests for Regression Discontinuity: Theory and Empirical Examples," *Review of Economics and Statistics*, 98.4, 685–700.

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

Stock, J. and M. Yogo (2005): *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Cambridge University Press, chap. 6, Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments.

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.

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

Whang, Y.-J. (2000): "Consistent bootstrap tests of parametric regression functions," *Journal of Econometrics*, 98.1, 27–46.

——— (2001): "Consistent specification testing for conditional moment restrictions," *Economics Letters*, 71.3, 299–306.