Regression Discontinuity Designs with Unknown Discontinuity Points: Testing and Estimation

Jack Porter*  Ping Yu†
University of Wisconsin-Madison  University of Auckland

First Version: October 2010
This Version: March 2011
(Preliminary, Comments Welcome)

Abstract

The regression discontinuity design has become a common framework among applied economists for measuring treatment effects, while a key restriction in these works is to assume the discontinuity point to be known, which is not always possible in practice. This paper extends the applicability of the regression discontinuity design by allowing an unknown discontinuity point. First, we construct tests to test whether there is selection or treatment effect. Second, we estimate the treatment effect by estimating the nuisance discontinuity point first. In testing, we show that our tests are consistent. Also, a bootstrap method is proposed to find the critical values. In estimation, we show that the estimating of the discontinuity point will not affect the efficiency of the treatment effect estimator. Simulation studies confirm the usefulness of our procedures in finite samples.

Keywords: regression discontinuity design, unknown discontinuity point, specification testing, wild bootstrap, degenerate U-statistic, difference kernel estimator, superconsistency, compound Poisson process, asymptotic independence

JEL-Classification: C12, C13, C14, C21

*Email: jrporter@ssc.wisc.edu.
†Email: p.yu@auckland.ac.nz.
1 Introduction

Since its invention by Thistlethwaite and Campbell (1960), the Regression discontinuity design (RDD) has attracted much attention among econometricians; see Imbens and Lemieux (2008), van der Klaauw (2008) and Lee and Lemieux (2010) for excellent reviews on up-to-date theoretical developments and applications. In RDDs, an observable covariate is used to completely determine the treatment status, and is called the forcing (running or assignment) variable. When the value of the covariate for an individual is above a threshold or discontinuity point, the individual will be treated; otherwise, the individual will be put in the control group. Usually, the discontinuity point is set by the policy maker and is publicly known. But such information is not always available. Sometimes, the discontinuity point is only known to the policy maker and is unknown to the public including econometricians due to ethical reasons or privacy. In the classical application of RDDs on the effect of the scholarship offer by van der Klaauw (2002), the forcing variable is an underlying index of various individual characteristics. To avoid manipulation by individuals or competitions from other schools, the discontinuity point may not be disclosed.

To date, all existing literature on RDDs assumes the discontinuity point is known especially to econometricians. This paper studies the testing and estimation problem when the discontinuity point is unknown. In the testing, we want to know whether there is selection or treatment effect in our experiment. The testing of the presence of treatment effect is very close to the structural change problem in the nonparametric environment, and there are already at least three tests designed for this purpose in the statistical literature. Our test is inspired by the classical specification test started from Bierens (1982), and is novel in our context. A new testing problem is to check whether there is selection among individuals. Such a problem is not considered even among statisticians. We attack these two testing problems by a unifying test statistic, but adapt it to different problems by varying a smoothing parameter. In the estimation, our main interest lies on the treatment effect, but a primary input is the discontinuity point. So we estimate the discontinuity point first - show its superconsistency and find its asymptotic distribution, then we estimate the treatment effect as if the discontinuity point were known. It is surprising to find that the estimating of the discontinuity point does not affect the efficiency of the treatment effect estimator at all. The model we consider has many applications beyond RDDs; see Müller (1992), Wu and Chu (1993), Wang (1995), Müller and Stadtmüller (1999), and the reference therein for applications in statistics.

This paper is organized as follows. In Section 2, we define our framework and specify some regularity assumptions. Especially, we clarify what selection means and what its implication to observations. Section 3 presents our specification test statistic and develops its asymptotic distribution in different testing problems. A bootstrap method is used to find its critical values which have better finite-sample performance. Other alternative tests in the literature are also reviewed. Section 4 considers the estimation problem. We estimate the discontinuity point and the treatment effect, develop their asymptotic distributions and provide some intuitions to aid understanding. In Section 5, we extend the results in Section 3 and 4 to other settings and solve an important practical issue, the bandwidth selection, in both specification testing and estimation. Section 6 shows some simulation results and Section 7 concludes.

Throughout this paper, we concentrate on the sharp design and discuss the case of the fuzzy design only briefly. Such a structure allows us to focus on the main ideas of this paper. In the sharp design, we assume that only the response variable and the forcing variance are observable, while the treatment status is not. Otherwise, the problem of testing and estimation will degenerate to the case where the discontinuity point is known; see Section 2 of Yu (2007) for a detailed discussion on this point. We further concentrate on the sharp design with at most one discontinuity point; other generalizations are only discussed briefly.

A word on notation: the letter $C$ is used as a generic positive constant, which need not be the same in each
occurrence. WLOG means "without loss of generality". DGP means "data generating process". Throughout this paper, "discontinuity" and "jump" are used exchangeably. LLS means "local linear smoother" popularized by Fan (1992, 1993) and Fan and Gijbels (1996). \( \approx \) means that the higher-order terms are omitted or a constant term is omitted (depending on the context). w.p.a.1 means "with probability approaching one". For a nonnegative real \( s \), \( [s] \) is its integer part and \( \lfloor s \rfloor \) is its the lower integer part; that is, \( s - 1 < [s] \leq s \), and \( s - 1 \leq \lfloor s \rfloor < s \).

2 Framework and Assumptions

We first put RDDs in the usual treatment framework and discuss a key "selection" assumption. Such a framework can be treated as the structural form of RDDs. We then impose some smoothness assumptions on the usual reduced form formulation of RDDs. Such assumptions are most relevant for the development of this paper. At last, we provide some rough ideas for our specification testing and estimation.

2.1 Selection and Treatment Effects

Following Lee (2008), suppose the response \( y = y(x, U) \) where \( x \) is the one-dimensional forcing variable which is observable, and \( U \) is the unobservable component such as ability in the scholarship example of van der Klaauw (2002). We assume there can be any correlation between \( U \) and \( x \).

We assume \( y(x, U) \) satisfies the following smoothness assumption.

Assumption Y:

(a) If there is no treatment or there is treatment but are not treatment effects, \( y(x, U) = \bar{y}(x, U) \) is continuous in \((x, U)\) and is continuously differentiable in \( x \) for each \( U \).

(b) If there are treatment effects, \( y(x, U) = \bar{y}(x, U) + \alpha(U)1(x \geq \pi) \) with \( \alpha(U) \) being continuous.

Under Assumption Y, when there are not treatment effects, after controlling all specific characters (except \( x \)) of an individual, the response \( y \) is a smooth function of \( x \). In the special case where \( y \) takes the additively separable form, \( y = g(x) + U \), \( g(\cdot) \) is assumed to be continuously differentiable. This is understandable, because it is hard to believe \( y \) will change very differently when \( x \) changes to \( x + \Delta \) or \( x - \Delta \) for a small \( \Delta \). Even if there are treatment effects, \( y(x, U) \) only changes its size at \( x = \pi \), but the slope at the left and right side of \( \pi \) remains the same.\(^1\)

Under Assumption Y, using the notations of the conventional average treatment literature such as Heckman and Vytlacil (2007), we can express the responses of the control and treated group as follows:

\[ Y_0 = \mu_0(x, U_0), \quad Y_1 = \mu_1(x, U_1), \quad D = 1(x \geq \pi), \]

\(^1\)Of course, we can assume \( y = y(x, W, U) \) without affecting the following analysis except complicating the notations, where \( W \) includes other observable covariates; see Lee and Lemieux (2010) for some discussions on \( y(x, W, U) \).

\(^2\)From the Skorohod representation, \( y(x, U) \) can be expressed as \( Q(U|x) \), where \( Q(\cdot) \) is the conditional quantile function of \( y \) given \( x \), and \( U \sim U(0, 1) \) independent of \( x \). We consider the structural expression of \( y \) instead of the probabilistic expression in this paper, so \( U \) can be correlated with \( x \) in any way.

\(^3\)If \( \alpha(\cdot) \) also depends on \( x \) such that for a specific \( U \), \( \frac{\partial \alpha(x = \pi + U)}{\partial x} \neq 0 \), then a paradox will appear. Suppose \( \alpha(\pi, U) = 0 \). If the responses under both control and treated status are smooth about \( x \), then a natural expectation is that the treatment effect is not zero if the treatment point is deviated a little from \( \pi \). But human beings usually behave smoothly and such a scenario is unimaginable. If we assume \( \frac{\partial \alpha(x = \pi + U)}{\partial x} = 0 \) for all \( U \), then the analysis below won’t change much.
where
\[
\begin{align*}
\mu_0(x, U_0) &= y(x, U), U_0 = U, \\
\mu_1(x, U_1) &= y(x, U) + \alpha(U), U_1 = U.
\end{align*}
\]

In this special case, \(\mu_0(x, U_0)\) and \(\mu_1(x, U_1)\) do not take the additively separable form. The main difference of RDDs from the conventional average treatment framework is that the treatment status is determined by a single observable \(x\), so the treatment status can be sharply observed (or there is a discontinuity in the propensity score at \(\pi\), or the unconfoundedness condition is trivially satisfied). Such a benefit is not free. The usual overlap assumption is violated because given any \(x\), we can observe either \(Y_0\) or \(Y_1\) but not both. We must rely on the continuity of \(\mu_0(x, U_0)\) at the left neighborhood of \(\pi\) to predict its behavior in the right neighborhood of \(\pi\), and similarly for \(\mu_1(x, U_1)\). As a result, we only use the local information around \(\pi\) to identify the treatment effect, which makes the treatment effect estimator achieve only a nonparametric rate instead of the usual \(\sqrt{n}\) rate of average treatment effect estimators. Usually, we express the relationship between \(y\) and \(x\) in a reduced form,
\[
y = m(x) + \varepsilon = m_\pi(x) + \alpha_\pi d_\pi + \varepsilon, \quad \text{where } E[\varepsilon|x, d_\pi] = 0, \ d_\pi = D,
\]
where
\[
\begin{align*}
y &= Y_0(1-D) + Y_1 D, \ m(x) = \mu_0(x)(1-D) + \mu_1(x) D = E[y|x], \\
\mu_0(x) &= E[y(x,U)|x], \ \mu_1(x) = E[y(x,U) + \alpha(U)|x], \\
\varepsilon &= \varepsilon_0(1-D) + \varepsilon_1 D, \\
\varepsilon_0 &= y(x,U) - \mu_0(x), \ \varepsilon_1 = y(x,U) + \alpha(U) - \mu_1(x), \\
\alpha_\pi &= \mu_1(\pi+) - \mu_0(\pi-) = E[y|x = \pi+] - E[y|x = \pi-] = m_\pi(\pi) - m_\pi(-\pi), \\
m_\pi(x) &= m(x) - \alpha_\pi d_\pi,
\end{align*}
\]
and \(\pi \in \Pi = [\underline{\pi}, \overline{\pi}]\). Note that \(U\) is different from \(\varepsilon\) (or \(\varepsilon_0\)) in our setup since \(\varepsilon = y - E[y|x]\), which is not equal to \(U\) even if we assume an additively separable form of \(y\). Also, \(\varepsilon_0\) and \(\varepsilon_1\) generally follow different distributions.

The special structure of RDDs allows us to express \(E[y|x] = m_\pi(x) + \alpha_\pi d_\pi\). But when there is treatment, no matter there are treatment effects or not, individuals can respond to the treatment actively although they cannot control \(x\) precisely\(^4\). Such attempts to control \(x\) generate a cusp at \(\pi\) in \(f(x|U)\) and consequently \(m_\pi(x)\); see the figures in Lee (2008) for some intuitive impression. An example is useful to illustrate this point. Suppose \(x\) is related to \(U\) by \(x(U) = Z + e(U)\), where \(Z\) can be precisely controlled by individuals so its support is on \([\pi, \infty)\), but \(e\) is a random error independent of \(Z\). Because individuals exert different efforts to manage \(x\) greater than \(\pi\) from less than \(\pi\), the density of \(e\) is not continuous at 0. To be more specific, suppose \(\pi = 0\), the density of \(Z\), \(f_Z(z)\), is standard exponential, the density of \(U\), \(f_U(u)\), is uniform on \([0, 1]\), and \(e|U\) has a density as follows\(^5\).

\[
f(e|U) = \begin{cases} 
\frac{1}{2(2+U)} \exp \left\{ \frac{e}{2+U} \right\}, & \text{if } e < 0; \\
\frac{1}{2(1+U)} \exp \left\{ -\frac{e}{1+U} \right\}, & \text{if } e \geq 0.
\end{cases}
\]

\(^4\)See Lee (2008) and Lee and Lemieux (2010) for the discussions on the role of imprecise control of \(x\) by individuals in RDDs.

\(^5\)If \(U\) is interpreted as ability in the scholarship example of van der Klaauw (2002), then this density means that an individual with higher ability (smaller \(U\)) can exert a larger power to manage her score.
When there is selection, we impose the following smoothness assumption on $
abla$.

**Assumption S:** The marginal density of $U$, $f_U(u)$, is continuously differentiable. For each $U$, $f(U|x)$ is continuous for all $x$.\(^\text{6}\)

(a) When there is no selection, $f(U|x)$ is **continuously** differentiable in $x$ for each $U$.

(b) When there is selection, $f(U|x)$ is **continuously** differentiable for all $x$ except at $x = \pi$.

Assumption S(b) implies that $f(U|x)$ is a nontrivial function of $x$; that is, $U$ is not independent of $x$. Also, there is a cusp at $x = \pi$ in $f(U|x)$.

Under Assumptions Y(a) and S(a), $m(x) = E[y|x] = \int y(x,U)f(U|x)dU$ is continuously differentiable in $x$, so there is no cusp in $m(x)$. But different combinations of Y(b) and S(b) imply different behaviors of

---

\(^6\)Continuity of $f(x|U)$ is a basic assumption in RDDs. It implies the continuity of $f(U|x)$ and local randomization, and corresponds to imprecise control of $x$. 
Given that \( m(x) \). First, suppose Y(a) and S(b) hold, then \( m(x) \) is obviously continuous, but it is not smooth at \( x = \pi \). Note that

\[
m'_{+} (\pi) \equiv m' (\pi+) = \int \frac{\partial y(\pi+, U)}{\partial x} f(U|x = \pi+) dU + \int y(\pi+, U) \frac{\partial f(U|x = \pi+)}{\partial x} dU,
\]

\[
m'_{-} (\pi) \equiv m' (\pi-) = \int \frac{\partial y(\pi-, U)}{\partial x} f(U|x = \pi-) dU + \int y(\pi-, U) \frac{\partial f(U|x = \pi-)}{\partial x} dU.
\]

Given that \( \frac{\partial y(\pi+, U)}{\partial x} = \frac{\partial y(\pi-, U)}{\partial x} \) and \( f(U|x) \) is continuous at \( x = \pi \),

\[
m'_{+} (\pi) - m'_{-} (\pi) = \int \left( y(\pi+, U) \frac{\partial f(U|x = \pi+)}{\partial x} - y(\pi-, U) \frac{\partial f(U|x = \pi-)}{\partial x} \right) dU,
\]

which can be but generally is not zero. Of course, if \( U \) is independent of \( x \) (this is, there is no selection), then \( \frac{\partial f(U|x=\pi+)}{\partial x} = \frac{\partial f(U|x=\pi-)}{\partial x} = 0 \), and \( m'_{+} (\pi) = m'_{-} (\pi) \). Second, suppose Y(b) and S(a) hold, then

\[
m_{+} (\pi) - m_{-} (\pi) = \int \alpha(U)f(U|x = \pi)dU,
\]

which can be but generally is not zero; e.g., if \( \alpha(U) \) has the same sign for any \( U \), then \( m_{+} (\pi) - m_{-} (\pi) \) cannot be zero. But it is easy to show that \( m'_{+} (\pi) = m'_{-} (\pi) \). Finally, suppose Y(b) and S(b) hold, then neither \( m_{+} (\pi) - m_{-} (\pi) \) nor \( m'_{+} (\pi) - m'_{-} (\pi) \) is zero. The analysis above is summarized in the following Table 1 and intuitively shown in Figure 2.

<table>
<thead>
<tr>
<th>No Selection</th>
<th>Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Treatment Effect</td>
<td>( m_{+} (\pi) = m_{-} (\pi) )</td>
</tr>
<tr>
<td>( m'<em>{+} (\pi) = m'</em>{-} (\pi) )</td>
<td>( m'<em>{+} (\pi) \neq m'</em>{-} (\pi) )</td>
</tr>
<tr>
<td>Treatment Effect</td>
<td>( m_{+} (\pi) \neq m_{-} (\pi) )</td>
</tr>
<tr>
<td>( m'<em>{+} (\pi) = m'</em>{-} (\pi) )</td>
<td>( m'<em>{+} (\pi) \neq m'</em>{-} (\pi) )</td>
</tr>
</tbody>
</table>

Table 1: The Behavior of \( m(\cdot) \) and \( m'(\cdot) \) at \( \pi \) With and Without Selection or Treatment Effect

In a word, \( m'_{+} (\pi) \neq m'_{-} (\pi) \) is sufficient but not necessary for the existence of selection. Similarly, \( m_{+} (\pi) \neq m_{-} (\pi) \) is sufficient but not necessary for the existence of treatment effects. When some extra assumptions, such as \( \alpha(U) \) has the same sign for any \( U \), it is also necessary. Notice that \( f(x) = \int f(x|U = u) f_U(u) du \), so similar arguments show that \( f'_{+} (\pi) \neq f'_{-} (\pi) \) is sufficient but not necessary for the existence of selection and \( f_{+} (\pi) \neq f_{-} (\pi) \) is sufficient but not necessary for the existence of selection. When \( f_{+} (\pi) \neq f'_{-} (\pi) \), \( f_{+} (\pi) \) and \( f'_{-} (\pi) \) are similarly defined as \( m_{+} (\pi) \) and \( m'_{-} (\pi) \). If manipulation is monotonic, \( f_{+} (\pi) \neq f_{-} (\pi) \) is also necessary for the existence of precise control as argued in McCrary (2008).

At last, we briefly explain why we call Assumption S the selection assumption. In the usual treatment effect literature, selection means \( Cov(D|U_0|x) \neq 0 \); that is, control and treatment groups include individuals with different characteristics under the control treatment.\(^7\) In RDDs, selection is trivially avoided because given \( x, D \) is fixed at either 1 or 0. But if \( x \) is not observed, then \( Cov(D|U_0) = Cov(1(x \geq \pi), U) \neq 0 \)

\(^7\)In the model with additively separable form of \( \mu_0(x, U_0) \) and \( \mu_1(x, U_1) \), another endogeneity assumption is \( Cov(D, U_1 - U_0|x) \neq 0 \), which means sorting on the gain. The model associated with this assumption is called the model of essential heterogeneity. In RDDs, \( U_1 = U_0 \), so such a distinction is not necessary.
as discussed above. So Assumption S essentially corresponds to the selection assumption in the average treatment effect literature; that is, individuals are trying to benefit by managing their treatment status. Such an analysis also reveals the very fact that all specialties of RDDs are from the observability of the assignment variable $x$ which completely controls the treatment status.

### 2.2 Regularity Assumptions

For either specification testing or estimation, only the structures of $m(x)$ on $\Pi_\epsilon \equiv (\pi - \epsilon, \pi + \epsilon)$ for some $\epsilon > 0$ are relevant, so WLOG, we can assume that the support of $x$ is $\Pi_\epsilon$. The conditional distribution of $\epsilon$ given $x$ is assumed to satisfy the following conditions.

**Assumption E:**

(a) $f(\epsilon|x)$ is continuous in $(\epsilon, x)$ for $x \in [\pi, \pi + \epsilon)$ and $x \in (\pi - \epsilon, \pi)$, respectively.

(b) $E[|\epsilon|^4|x]$ is uniformly bounded on $\Pi_\epsilon$.

(a) implies that there is no point mass in the distribution of $\epsilon$. This assumption will be used to uniquely detect the location of $\pi$. Also, (a) implies that both $\sigma^2(x) = E[|\epsilon|^2|x]$ and $E[|\epsilon|^4|x]$ are continuous for $x \neq \pi$, $x \in \Pi_\epsilon$. Such continuity restrictions guarantee that the possible discontinuity at $\pi$ in $\sigma^2(x)$ and $E[|\epsilon|^4|x]$ is of the first kind.

---

8 We mean a version of the joint density satisfies this assumption because a density with the value on a set with Lebesgue measure zero changed is still valid.

9 This assumption is used in specification testing; in estimation, we need only $E[|\epsilon|^{2+\zeta}|x]$ bounded for some $\zeta > 0$. 

---

Figure 2: $m(\cdot)$, $m_\pi(\cdot)$ and $m(\cdot)$ in Models With and Without Selection or Treatment Effect: Solid Line for $m(\cdot)$, Dashed Line for $m_\pi(\cdot)$, and Dash-Dot Line for $m(\cdot)$
We must put some smoothness conditions on \( m(x) \) to continue our analysis. For \( s \in [0, 1) \), let the Hölder class \( C^+_\pi (L, s) \) be the set of maps \( m(\cdot) \) from \( \Pi^+ \) to \( \mathbb{R} \) with

\[
C^+_\pi (L, s) = \{ m(\cdot) : |m(x + \Delta) - m(x)| \leq L |\Delta|^s \text{ for all } x, x + \Delta \in \Pi^+_\pi \} \tag{10}
\]

where \( \Pi^+ = [\pi, \pi + \epsilon) \). For \( s \geq 1 \),

\[
C^+_{\pi, \epsilon} (L, s) = \left\{ m(\cdot) : \left| m(x + \Delta) - \sum_{j=0}^{[s]} \frac{m^{(j)}(x)}{j!} \Delta^j \right| \leq L |\Delta|^s \right\}
\]

for all \( x, x + \Delta \in \Pi^+_\pi \) and \( \sup_{j=0, \ldots, [s]} \sup_{x \in [\pi, \pi + \epsilon)} |m^{(j)}(x)| \leq L \),

where \( m^{(0)}(x) = m(x) \), \( m^{(j)}(x) \) is the \( j \)th-order derivative when \( x \neq \pi \), and \( m^{(j)}(\pi) \) is the \( j \)th-order right hand derivative at \( \pi \). Similarly, we can define \( C^-_{\pi, \epsilon} (L, s) \) with \( \Pi^- = (\pi - \epsilon, \pi) \) and \( m^{(j)}(\pi) \) being the \( j \)th-order left hand derivative at \( \pi \), and define \( C(L, s) \) with \( \Pi^+_\pi \) replaced by \( \Pi^- \). Further define

\[
C_{\pi, \epsilon} (L, s) \equiv C^+_\pi (L, s) \cap C^-_{\pi, \epsilon} (L, s), \quad M_{\Pi} (L, s) \equiv \bigcup_{\pi \in \Pi} C_{\pi, \epsilon} (L, s), \quad C_{\Pi} (L, s) \equiv M_{\Pi} (L, s) \cap C_0,
\]

where \( C_0 \) is the set of continuous functions on \( \Pi \). We put the following assumption on \( m(\cdot) \).

**Assumption M:** \( m(x) \in M_{\Pi} (L, s), m_{\pi}(x) \in C_{\Pi} (L, s), s \geq 1 \).

Note that \( C_{\Pi} (L, s) \) includes functions such that there is a cusp at \( \pi \). Such functions correspond to selection. Without selection, \( m_\pi (\cdot) \) is smooth. Figure 2 shows these two cases. Assumption M states that although there can be a cusp in \( m_\pi(\cdot) \), the cusp cannot be too sharp; see Wang (1995) for the case where the cusp is sharp. Such an assumption is consistent with the analysis in the last subsection. \( m'(\pi^+) - m'(\pi^-) = \int_{\pi}^{\pi + 0} \left( \frac{\partial f(U|x=\pi^+)}{\partial x} - \frac{\partial f(U|x=\pi^-)}{\partial x} \right) dU \) should not be very large if \( \frac{\partial f(U|x=\pi^+)}{\partial x} - \frac{\partial f(U|x=\pi^-)}{\partial x} \) is not large for each \( U \); see Figure 3 for the intuition. We next put some smoothness conditions on \( f(x) \):

**Assumption F:** \( f(x) \in C_0, f(x) \geq \underline{f} > 0 \) for \( x \in \Pi \).

(a) \( f(x) \in C(\overline{\pi}, \lambda), \lambda \geq 1 \).

(b) \( f(x) \in C_{\Pi} (\overline{\pi}, \lambda), \lambda \geq 1 \).

Assumption F corresponds to Assumption S. F(a) corresponds to the no selection case S(a), and F(b) corresponds to the selection case S(b); see Section 5.1 for tests on these two assumptions. Both \( C(\overline{\pi}, \lambda) \) and \( C_{\Pi} (\overline{\pi}, \lambda) \) include \( C_0 \). We repeat \( f(x) \in C_0 \) to emphasize that there is no precise control of \( x \).

---

10 When \( s = 0 \), this is just the class of bounded functions without any smoothness even continuity restriction.

11 A little stronger assumption on the smoothness of \( m(\cdot) \) in the literature is to assume \( |m^{(k)}(x) - m^{(k)}(y)| \leq L|x - y|^{s-[s]} \) for all \( x, y \in [\pi, \pi + \epsilon) \). Also, some literature defines the Hölder class using the lower integer part \([s]\) instead of the integer part \( s \). The former is a little broader than the latter, and the only difference lies in the case where \( s \) is an integer. For example, when \( s = 1 \), \( C^+_\pi (L, s) \) assumes that the right hand derivative at \( \pi \) exists, but if we use \([s]\) to define \( C^+_\pi (L, s) \), the derivative need not exist. We define \( C^+_\pi (L, s) \) in this way to ease the exposition below.
2.3 Sketch of The Paper

We first consider two kinds of specification tests. First, we test if there is selection. Specifically, the hypotheses are

\[ H_0^{(1)} : m(\cdot) \in \mathcal{C}(L, s), s \geq 1; \]
\[ H_1^{(1)} : m(\cdot) \in \mathcal{C}_1(L, s) \setminus \mathcal{C}(L, s), s \geq 1; \]

where \( H_0 \) corresponds to no effects (including selection and treatment effect), and \( H_1 \) corresponds to selection only. In Figure 2, we are testing the upper-left case against the upper-right case. Second, we test if there is treatment effect (no matter there is selection or not). Specifically, the hypotheses are

\[ H_0^{(2)} : m(\cdot) \in \mathcal{C}_1(L, s), s \geq 1; \]
\[ H_1^{(2)} : m(\cdot) \in \mathcal{M}_1(L, s), s \geq 1, \text{ and } \alpha_{\pi} \neq 0 \text{ for some } \pi \in \Pi; \]

where \( H_0 \) corresponds to two cases: no effects and selection only, and \( H_1 \) corresponds to the other two cases: treatment effect only and both selection and treatment effect. In Figure 2, we are testing the upper two cases against the bottom two cases. Both tests are relevant in practice, and can be sequentially conducted: first test \([3]\); if cannot reject \( H_0^{(2)} \), then test \([2]\). Denote the class of probability measures under \( H_0^{(\kappa)} \) as \( \mathcal{H}_0^{(\kappa)} \) and under \( H_1^{(\kappa)} \) as \( \mathcal{H}_1^{(\kappa)} \), \( \kappa = 1, 2 \). Both \( \mathcal{H}_0^{(\kappa)} \) and \( \mathcal{H}_1^{(\kappa)} \), \( \kappa = 1, 2 \), are characterized by \( m(\cdot) \). In what follows, we acknowledge the dependence of the distribution of \( y \) given \( x \) upon the regression function by denoting probabilities and expectations as \( P_m \) and \( E_m \), respectively.

We develop a unifying test statistic for both cases. We first find the best approximation of \( m(x) \), \( \bar{m}(\cdot) \), in \( \mathcal{C}(L, s) \) in both cases,

\[ \bar{m}(\cdot) = \arg \inf_{\tilde{m} \in \mathcal{C}(L, s)} E \left[ (m(x) - \tilde{m}(x))^2 1_{x}^{\Pi} \right] = \arg \inf_{\tilde{m} \in \mathcal{C}(L, s)} E \left[ (y - \tilde{m}(x))^2 1_{x}^{\Pi} \right]; \]

where \( 1_{x}^{\Pi} = 1(x \in \Pi) \), the second equality is from the fact that

\[ E \left[ (y - \tilde{m}(x))^2 1(x \in \Pi) \right] = E \left[ (\varepsilon + m(x) - \tilde{m}(x))^2 1_{x}^{\Pi} \right] = E \left[ \sigma^2(x) 1_{x}^{\Pi} \right] + E \left[ (m(x) - \tilde{m}(x))^2 1_{x}^{\Pi} \right], \]

and \( E \left[ \sigma^2(x) 1_{x}^{\Pi} \right] \) does not depend on \( \tilde{m}(\cdot) \). In short words, \( \bar{m}(\cdot) \) is the closest \( \tilde{m} \in \mathcal{C}(L, s) \) to \( m(\cdot) \) (or the observable \( y \)) under the \( L^2 \)-metric. It can be estimated by some standard nonparametric techniques. We then define a measure for the distance between \( \bar{m}(\cdot) \) and \( m(\cdot) \) to detect the deviation from the null. Our test is inspired by Fan and Li (1996) and Zheng (1996) although they consider different testing problems, and is detailed in the next section. The difference of our test statistic in testing \([2]\) and \([3]\) is to choose different smoothing parameters. This is intuitively illustrated in Figure 2. In testing \([2]\), we oversmooth \( m(x) \), then the bias in such smoothing will generate power. In contrast, in testing \([3]\), we undersmooth \( m(x) \). Now, the bias from the cusp will disappear, while the jump in \( m(\cdot) \) at \( \pi \) still generates power.

Of course, we need our test satisfy some basic properties. Abuse the notations a little here. \( H_0 \) represents either \( H_0^{(1)} \) or \( H_0^{(2)} \), and \( H_1 \) represents either \( H_1^{(1)} \) or \( H_1^{(2)} \). For a randomized test \( t_n \), its size (or level) is defined as

\[ \alpha(t_n) = \sup_{m(\cdot) \in H_0} P_m(t_n = 1). \]

t_n is consistent if its type II error, \( P_m(t_n = 0) \), converges to zero for any \( m(\cdot) \in H_1 \) under the Neyman-
Pearson restriction $\alpha(t_n) \leq \alpha + o(1)$ for some $\alpha > 0$. We show that our test is consistent in both cases.

If the specification test rejects $H_0^{(2)}$, we can believe that there is a discontinuity in $m(x)$. The question ensued is how to estimate the discontinuity location $\pi$ and size $\alpha_\pi$. In the RDD literature, the main interest lies on $\alpha_\pi$ instead of $\pi$ although an estimator of $\pi$ is a primary input to estimate $\alpha_\pi$. Suppose we have an estimator $\hat{\pi}$ in hand, then we use the local polynomial estimator as in Porter (2003) to estimate $\alpha_\pi$. Denote the estimator as $\hat{\alpha}(\hat{\pi})$ with

$$\hat{\alpha}(\pi) = \hat{m}_+(\pi) - \hat{m}_-(\pi),$$

where $\hat{m}_+(\pi)$ is an local polynomial estimator (LPE) which is determined by the minimizer $\hat{a}$ in the following minimization problem:

$$\min_{a,b_1,\ldots,b_p} \frac{1}{n} \sum_{j=1}^{n} k_h(x_j - \pi) d_j(\pi) [y_j - a - b_1(x_j - \pi) - \cdots - b_p(x_j - \pi)^p]^2,$$

where $p$ is the suitable order of local polynomials defined in Porter (2003), $d_j(\pi) = 1(x_j \geq \pi)$, $k_h(\cdot) = \frac{1}{h}k\left(\frac{\cdot}{h}\right)$, $k(\cdot)$ is a kernel function, and $h$ is the bandwidth. $\hat{m}_-(\pi)$ is similarly defined but using the data in the left neighborhood of $\pi$. Actually, $\hat{\alpha}(\hat{\pi})$ is a corollary of our estimation of $\pi$ because $\pi$ is estimated by maximizing $\hat{\alpha}_{}^2(\pi)$. We show that $\hat{\pi}$ is $n$-consistent, so it will not affect the asymptotic distribution of $\hat{\alpha}(\hat{\pi})$. As a result, our estimator of $\alpha_\pi$ achieves the optimal rate developed in Porter (2003).

3 Consistent Specification Testing

This section presents consistent specification testing for (2) and (3). It begins with consistent tests and their asymptotic distributions, followed by a bootstrap method to find their critical values, and concludes with a discussion about other alternative tests.

3.1 Consistent Tests and Their Asymptotic Distributions

First rewrite (1) as

$$y = \overline{m}(x) + e,$$

where $E[e|x] = E[e|x] = 0$ under $H_0^{(1)}$ for $x \in \Pi$ and $E[e|x] \neq 0$ under $H_1^{(1)}$ for some $x \in \Pi$. Observing that $E[eE[e|x]1_{h}] = E[E[e|x]2_{h}] = E\left(\overline{m}(x) - \overline{m}(x)\right)^2 1_{h} \geq 0$ and the equality holds if and only if $H_0^{(1)}$ is true, we can construct a consistent test for (2) based on $E[eE[e|x]1_{h}]$. $E[e|x]$ can be estimated by a kernel estimator. To overcome the random denominator problem in kernel estimation, we choose to estimate a density weighted version of $E[eE[e|x]1_{h}]$ given by $E[eE[e|x]f(x)1_{h}]$. If $e_i$ and $E[e_i|x_i]f(x_i)$ were available, then we could estimate $E[eE[e|x]f(x)1_{h}]$ by its sample analogue $n^{-1} \sum e_i\hat{E}[e_i|x_i]f(x_i)1_{h}$, where $1_{h} = 1(x_i \in \Pi)$. To get a feasible test statistic, we need to estimate $e_i$ by the corresponding residual from (5) and $E[e_i|x_i]f(x_i)$ by an appropriate kernel estimator. Specifically, we estimate $e_i$ by $\hat{e}_i = y_i - \hat{y}_i$, the nonparametric residual from (5), where $\hat{y}_i$ is a kernel estimator of $\overline{m}(x_i)$ defined as

$$\hat{y}_i = \frac{1}{n-1} \sum_{j \neq i} y_j L_{b,ij} \hat{f}_i,$$

and $\hat{f}_i$ is the corresponding kernel estimator of $f(x_i)$ given by

$$\hat{f}_i = \frac{1}{n-1} \sum_{j \neq i} L_{b,ij},$$
where \( L_{b,ij} = b_i(x_i - x_j) = \frac{1}{l} \left( \frac{x_i - x_j}{b} \right) \), \( b \) is the bandwidth, and \( l() \) is some kernel function. We estimate \( E[\varepsilon_i|x_i]f(x_i) \) by \( \frac{1}{n-1} \sum_{j \neq i} \hat{\varepsilon}_j K_{b,ij} \hat{y} \), where \( K_{b,ij} \) is similarly defined as \( L_{b,ij} \). Our test statistic is based on

\[
I_n = \frac{nh^{1/2}}{n(n-1)} \sum_i \sum_{j \neq i} \mathbb{1}_H^1 K_{b,ij} \hat{\varepsilon}_i \hat{\varepsilon}_j
\]

where

\[
\hat{\varepsilon}_i = (m_i - \hat{m}_i) + (\varepsilon_i - \hat{\varepsilon}_i),
\]

\( m_i = m(x_i), \hat{m}_i \) and \( \hat{\varepsilon}_i \) are defined in the same way as \( \hat{y}_i \) in \( 6 \) with \( y_j \) replaced by \( m(x_j) \) and \( \varepsilon_j \), respectively. Under \( H_0^{(1)} \), \( \hat{\varepsilon}_i \) is a good estimate of \( \varepsilon_i \). Under \( H_1^{(1)} \), \( \hat{\varepsilon}_i \) includes a bias term in the neighborhood of \( \pi \), so generates power. The indicator function \( \mathbb{1}_H^1 \) in \( I_n \) can improve the power properties by shrinking the asymptotic variance of \( I_n \).

To use \( I_n \) in testing \( 3 \), we must guarantee that the estimator of \( E[\varepsilon E[\varepsilon|x]f(x)]^1 \) is approximately zero under \( H_0^{(2)} \). Note that under \( H_0^{(2)} \), \( E[\varepsilon E[\varepsilon|x]f(x)]^1 = E[\varepsilon m(x) - \bar{m}(x)]^2 f(x) \mathbb{1}_H^1 \), so we must use a smaller \( b \) to make the bias in approximating \( m(x) \) by \( \bar{m}(x) \) disappear; see the upper-right panel of Figure 2 for some intuition. The following Assumption B imposes such kind of restrictions on \( b \) (and \( h \)) which are suitable in different cases.

**Assumption B:** \( b \to 0, nh \to \infty, h^{1/2}/b \to 0. \)

(a) \( nh^{1/2}b^{2\eta} \to 0, nh^{1/2}b^{3} \to \infty, \) where \( \eta = \min(\lambda + 1, s) > 1; \)

(b) \( nh^{1/2}(b^{2\eta} + b^{3}) \to 0, nh^{1/2}b \to \infty, \) where \( \eta = \min(\lambda + 1, s) \geq 1; \)

Assumption B implies that \( \max \{ h, b \} \to 0 \) and \( \min \{ h, b \} \to \infty \), which ensure that the kernel estimators involved are consistent. In (B-a), \( nh^{1/2}b^{2\eta} \) is the bias under \( H_0^{(1)} \), and the bias under \( H_1^{(1)} \) is \( nh^{1/2}(b^{2\eta} + b^{3}) \). The first term \( b^{2\eta} \) is the bias from the area out of a \( b \) neighborhood of \( \pi \), and the second term \( b^{3} \) is the bias from the \( b \) neighborhood of \( \pi \). Because \( \eta > 1.5 \), the second term dominates the first term. \( nh^{1/2}b^{2\eta} \to 0 \) ensures that \( I_n \) is centered correctly at zero under \( H_0^{(1)} \). Corresponding arguments can be applied to (B-b) for testing \( 3 \). We indeed need \( m() \) and \( f() \) to be a little smoother under \( H_0^{(1)} \) than \( H_0^{(2)} \) to control the bias, but it seems irrelevant in practice. \( h^{1/2}/b \to 0 \) implies that \( h \) is smaller than \( b \). This condition not only ensures the bias disappear under \( H_0^{(1)} \) and \( H_0^{(2)} \), but generates power under \( H_1^{(1)} \) and \( H_1^{(2)} \). Because \( \hat{\varepsilon}_j \) is estimated using the data in a \( b \) neighborhood of \( x_j \), and the distance between \( x_j \) and \( x_i \) is \( h \), we can treat \( x_i \) and \( x_j \) as the same point without changing the asymptotic results. In consequence, \( \hat{\varepsilon}_i \hat{\varepsilon}_j \) is like a squared term which generates power. Take \( b \sim n^{-\kappa} \) and \( h = n^{-(2\kappa+\delta)} \) for an arbitrarily small \( \delta \), then in (B-a), \( c \leq 1/4 \) and in (B-b), \( \frac{1}{2} < b \leq \frac{1}{2} \) when \( s \geq 1.5 \). This matches the intuition above that a larger bandwidth is required in testing \( 2 \) than in testing \( 3 \).

Next, we put some standard constraints on the kernel functions \( l() \) and \( k() \).

---

\(^{13}\)Fan and Li (1996) further eliminate the random denominators in \( \hat{\varepsilon}_i \) and \( \hat{\varepsilon}_j \) by considering

\[
\sum_{i,j \neq i} \mathbb{1}_H^1 K_{b,ij} \left( \hat{\varepsilon}_i \hat{\varepsilon}_j \right),
\]

but the form we use is more convenient for practitioners since it allows \( \hat{\varepsilon}_i \)'s to be estimated by other nonparametric methods. For example, although the local polynomial estimator is asymptotically equivalent to a higher-order kernel smoother in \( 6 \), its various forms, especially the LLS because of its minimax efficiency studied in Fan (1993), can improve the finite-sample performance of \( I_n \) and are more popular than \( 6 \). But it is not easy to define the random denominator for the LLS. Of course, our form of \( I_n \) is not new; Gu et al (2007) consider a similar test statistic in testing whether there are omitted variables.

\(^{14}\)\( nh^{1/2}b \to \infty \) is implied by \( nh \to \infty, h^{1/2}/b \to 0 \). We write it out explicitly to emphasize that it is the bias under \( H_1^{(2)} \).

\(^{15}\)In the classical specification tests such as Zheng (1996), \( \hat{\varepsilon}_i \) is obtained under the null parametric specification, which corresponds \( b = C \), a fixed number. Since \( b \to 0, h/b \to 0 \), and \( h \) is smaller than \( b \).
Assumption K: Both \( l(\cdot) \) and \( k(\cdot) \) are bounded, symmetric, Lipschitz function, zero outside a bounded set \([-M, M]\], \( \int u^i l(u) du = \delta_{i0}, \ i = 0, \ldots, [s] + [\lambda] \), where \( \delta_{i0} \) is Kronecker’s delta, and \( k(\cdot) \) is non-negative.

\( l(\cdot) \) can be a higher-order kernel to reduce the bias in estimating \( m(\cdot) \), but \( k(\cdot) \) is a usual second-order kernel. Such an assumption can simplify our proof. To further simplify our discussion, we assume \( M = 1 \) throughout this paper. Now, we state the asymptotic distribution of \( I_n \) and the consistency of the tests based on \( I_n \) in different cases.

**Theorem 1** Under Assumptions B(a), E, K and M,

(i) if \( F(a) \) holds, then

\[
I_n \xrightarrow{d} N(0, \Sigma)
\]

uniformly over \( \mathcal{H}^{(1)}_0 \), where

\[
\Sigma = 2E \left[ 1^I f(x) s^4(x) \right] \int k^2(u) du
\]

can be consistently estimated by

\[
v_n^2 = \frac{2h}{n(n-1)} \sum_i \sum_{j \neq i} 1^I 1^j K_{h,ij}^2 \varepsilon_i^2 \varepsilon_j^2.
\]

As a result, the test based on the studentized test statistic \( T_n = I_n/v_n \)

\[
t_n = 1(T_n > z_\alpha),
\]

has the significance level \( \alpha \), where \( z_\alpha \) is the \( 1 - \alpha \) quantile of the standard normal distribution.\(^\text{16}\)

(ii) if \( F(b) \) holds, then the test \( t_n \) is consistent; that is, \( P_m(T_n > z_\alpha) \rightarrow 1 \) for any \( m(\cdot) \in \mathcal{C}_1(L, s) \setminus \mathcal{C}(L, s) \).

Furthermore, \( z_\alpha \) can be replaced by any nonstochastic constant \( C_n = o(nh^{1/2}b^3) \), and the result still holds.

**Theorem 2** Under Assumptions B(b), F(b), E, K and M,

(i) \( I_n \xrightarrow{d} N(0, \Sigma) \)

uniformly over \( \mathcal{H}^{(2)}_0 \), where \( \Sigma \) is the same as in Theorem 1, and can be consistently estimated by

\[
v_n^2 = \frac{2h}{n(n-1)} \sum_i \sum_{j \neq i} 1^I 1^j K_{h,ij}^2 \varepsilon_i^2 \varepsilon_j^2.
\]

As a result, the test based on the studentized test statistic \( T_n = I_n/v_n \)

\[
t_n = 1(T_n > z_\alpha),
\]

has the significance level \( \alpha \), where \( z_\alpha \) is the \( 1 - \alpha \) quantile of the standard normal distribution.

(ii) The test \( t_n \) is consistent; that is, \( P_m(T_n > z_\alpha) \rightarrow 1 \) for any \( m(\cdot) \) such that \( \alpha_m \neq 0 \). Furthermore, \( z_\alpha \) can be replaced by any nonstochastic constant \( C_n = o(nh^{1/2}b^3) \), and the result still holds.

\(^\text{16}\)This is a one-side test because \( I_n \) is based on the \( L^2 \)-distance between \( m(\cdot) \) and \( \overline{m}(\cdot) \).
The asymptotic distributions of $I_n$ in Theorem 1 and Theorem 2 are the same, but under different assumptions. To provide a nondegenerate asymptotic distribution of $I_n$, we use the central limit theorem for fourth order degenerate $U$-statistics. This is standard in model specification tests where the null involves a nonparametric component; see, e.g., Fan and Li (1996). When the null takes a parametric form, then only the second order degenerate $U$-statistics emerges; see, e.g., Zheng (1996). $I_n$ is designed for the case where the number of cusps is at most one. It is easy to see that when the number of cusps is greater than 1 but finite, the specification tests based on $I_n$ remain valid.\(^\text{17}\)

### 3.2 A Bootstrap Method

From the proofs of Theorem 1 and 2, the convergence rate of $T_n$ to standard normal is very slow (how slow?)\(^\text{18}\). As argued in the literature of specification tests, see, e.g., Gu et al (2007), Härdle and Mammen (1993), Li and Wang (1998), and Stute et al (1998), a better way to approximate the finite-sample distribution is to use the wild bootstrap of Wu (1986) and Liu (1988)\(^\text{19}\). Specifically, the following procedure is used in testing both (2) and (3).

**Algorithm WB:**

**Step 1:** For $i = 1, \cdots, n$, generate the two-point wild bootstrap residual $\varepsilon_i^* = \hat{e}_i (1 - \sqrt{5})/2$ with probability $(1 + \sqrt{5})/(2\sqrt{5})$, and $\varepsilon_i^* = \hat{e}_i (1 + \sqrt{5})/2$ with probability $(\sqrt{5} - 1)/(2\sqrt{5})$, then $E^* [\varepsilon_i^*] = 0$, $E^* [\varepsilon_i^{*2}] = \hat{e}_i^2$ and $E^* [\varepsilon_i^{*3}] = \hat{e}_i^3$, where $E^* [\cdot] = E [\cdot | F_n]$ and $F_n = \{ (x_i, y_i) \}_{i=1}^n$.

**Step 2:** Generate the bootstrap resample $\{ y_i^*, x_i \}_{i=1}^n$ by\(^\text{21}\)
\[
y_i^* = \hat{y}_i + \varepsilon_i^*.
\]

Then obtain the bootstrap residuals $\hat{e}_i^* = y_i^* - \hat{y}_i^*$, where $\hat{y}_i^*$ is similarly defined as $\hat{y}_i$ with the only difference being that $y_j$ in (6) is replaced by $y_j^*$.

**Step 3:** Use the bootstrap samples to compute the test statistic
\[
I^*_n = \frac{nh^{1/2}}{n(n-1)} \sum_i \sum_{j \neq i} 1_i^{\bar{H}} 1_j^{\bar{H}} K_{h,i,j} \hat{e}_i^* \hat{e}_j^*
\]
and the estimated asymptotic variance
\[
\nu_n^{*2} = \frac{2h}{n(n-1)} \sum_i \sum_{j \neq i} 1_i^{\bar{H}} 1_j^{\bar{H}} K_{h,i,j}^2 \hat{e}_i^* \hat{e}_j^*.
\]

\(^{17}\) Most literature assumes the number of jumps is finite. The only exception coming to our attention is Müller and Stadtmüller (1999); see Section 3.3 for a discussion on their method. In RDDs, it is not imaginable that there are infinite design points.

\(^{18}\) In the classical specification tests, the bias in the test statistic of Zheng (1996) is $O(h^{1/2})$, which is $O(n^{-1/10})$ when $h = n^{-1/5}$, as shown in Li and Wang (1998).

\(^{19}\) The bootstrap naive bootstrap is not valid as argued in Härdle and Mammen (1993). This is because the regression function under the bootstrap distribution is not the conditional expectation of the observation: $E^* [y_i^* | x_i^*] = y_i^*$ which is typically different from the fitted value of (6) at $x_i^*$.

\(^{20}\) The bootstrap method is a standard way to find critical values in hypothesis testing. For example, when $m_n(\cdot)$ takes a parametric form, the test above is labeled as a test with nuisance parameter ($\pi$) unidentified under the null. In this case, Hansen (1996) finds that the bootstrap works very well.

\(^{21}\) To construct $I_n^*$, we need only the data with $x_i \in [x - b, x + b]$.}

12
The studentized bootstrap statistic is \( T^*_n = I^*_n / v^*_n \). Here, the same \( b \) and \( h \) are used as in \( I_n \) and \( v_n^2 \) in Theorem 1 and 2.

**Step 4:** Repeat step 1 through 3 \( B \) times, and use the empirical distribution of \( \{T^n_{nk}\}_{k=1}^{B} \) to approximate the null distribution of \( T_n \). We reject \( H_0 \) if \( T_n > T^*_{n(aB)} \), where \( T^*_{n(aB)} \) is the upper \( a \)-percentile of \( \{T^n_{nk}\}_{k=1}^{B} \).

In Algorithm WB, the only difference in testing \([2]\) and \([3]\) is to use different \( b \) and \( h \) as defined in Assumption B(a) and B(b), respectively. In Step 1, a more popular way to simulate \( \varepsilon^*_i \) is based on \( \hat{\varepsilon}_i \)'s centralized counterpart \( \bar{\varepsilon}_i = \hat{\varepsilon}_i - \bar{\varepsilon} \) instead of \( \hat{\varepsilon}_i \) itself, where \( \bar{\varepsilon} = \frac{\sum_{i=1}^{n} \varepsilon_i 1^*_{ib}}{\sum_{i=1}^{n} 1^*_{ib}} \), \( \Pi_b = (\varepsilon - b, \bar{\varepsilon} + b) \); see, e.g., Gijbels and Goderniaux (2004) and Su and Xiao (2008). Such a formulation can ensure \( \sum_{i=1}^{n} \bar{\varepsilon}_i 1^*_{ib} / \sum_{i=1}^{n} 1^*_{ib} = 0 \), which will not affect the asymptotic results, but may affect the finite-sample performance of Algorithm WB especially under the alternative.

The bootstrap sample is generated by imposing the null hypothesis. Therefore, the bootstrap statistic \( T^*_n \) will mimic the null distribution of \( T_n \) even when the null hypothesis is false. When the null is false, \( \hat{\varepsilon}_i \) is not a consistent estimate of \( \varepsilon_i \) in the neighborhood of \( \varepsilon \). Nevertheless, the following theorem shows that the above bootstrap procedure is valid. This is because our studentized test statistic \( T_n \) is invariant to the variance of \( \varepsilon \). In other words, the wild bootstrap procedure may not be valid if the test statistic \( I_n \) instead of \( T_n \) is used.

**Theorem 3** Under either \( H_{0}^{(1)} \) and Assumptions B(a), E, F(a), K and M or \( H_{0}^{(2)} \) and Assumptions B(b), E, F(b), K and M,

\[
\sup_{z \in \mathbb{R}} |P(T^*_n \leq z | \mathcal{F}_n) - \Phi(z)| = o_p(1),
\]

where \( \Phi(\cdot) \) is the cumulative distribution function of a standard normal random variable.

The above theorem only proves the first-order validity of the wild bootstrap procedure. The higher-order theory can follow the line of Li and Wang (1998) and Fan and Linton (2003). Also, the validity of the wild bootstrap using the bandwidth based on cross-validation can be justified by extending the technique developed in Hsiao et al (2007). But a formal development of these results is beyond the scope of this paper.

### 3.3 Other Specification Tests

There are no specific tests designed for testing \([2]\) in the statistical literature, while at least three alternative tests exists for testing \([3]\). All of them assume a fixed, essentially equally spaced design on \([0,1]\) and \( \varepsilon_i \) being i.i.d. sampled. Nevertheless, these results provide some insights for testing \([3]\).

A straightforward test is based on \( \int_{\pi}^{\pi} (\tilde{m}_+ (\pi) - \tilde{m}_- (\pi))^2 \, d\pi \), where \( \tilde{m}_\pm (\pi) \) are the local constant estimators of \( m_\pm (\pi) \).

To avoid the random denominator problem as in \( I_n \), we can use

\[
\tilde{I}_n = nh^{1/2} \int_{\pi}^{\pi} (\tilde{m}_+ (\pi) \tilde{f}_+ (\pi) - \tilde{m}_- (\pi) \tilde{f}_- (\pi))^2 \, d\pi
\]

\[22\] If we use a data adaptive bandwidth such as cross-validation based on each bootstrap sample, then the algorithm is extremely time-consuming. See Chapter 3 of Mammen (1992) for a related discussion.

\[23\] Throughout this paper, we mean this setup when we discuss the statistical literature if no further specification. Note that \( \Pi \) is a subinterval of \([0,1]\) now.

\[24\] In testing \([2]\), a similar test can be based on \( \int_{\pi}^{\pi} (\tilde{m}'_+ (\pi) - \tilde{m}'_+ (\pi))^2 \, d\pi \), where \( \tilde{m}'_\pm (\pi) \) are some estimators (such as the LLS) of \( m'_\pm (\pi) \). But it obviously suffers the drawbacks mentioned below.

\[25\] When deriving the asymptotic distribution of \( \tilde{I}_n \), we need only employ the central limit theorem for the second-order
where \( \hat{f}_+(\pi) = \frac{1}{n} \sum_{j=1}^{n} k_h (x_j - \pi) d_j (\pi) \), \( \hat{f}_-(\pi) = \frac{1}{n} \sum_{j=1}^{n} k_h (x_j - \pi) (1 - d_j (\pi)) \). Wu and Lai (1998) show that
\[
\bar{I}_n - \frac{\sigma^2}{\sqrt{n}} K_1 \xrightarrow{d} 2\sigma^2 N(0, K_2),
\]
where \( \sigma^2 = E[\varepsilon_i^2] \), and \( K_1 \) and \( K_2 \) are constants only related to \( k (\cdot) \). From the form of \( \bar{I}_n \) and \( I_n \), we can see their key difference. \( I_n \) is based on \( m(\cdot) \) under the null (e.g., \( \overline{m}(\cdot) \) is very close to \( m(\cdot) \) under the null, and \( \pi \) does not appear at all), while \( \bar{I}_n \) is based on \( m(\cdot) \) under the alternative (e.g., it measures the integrated jumping size at different possible jumping locations). The key problem of \( \bar{I}_n \) is that it does not have a mean zero because it takes a squared form. Also, the numerical integration is required for a practitioner, which is not attractive in practice. In our case, \( \sigma^2 \) will change to a functional of \( f(\cdot) \) and \( \sigma^2(\cdot) \); see, e.g., Härdle and Mammen (1993). Estimation of such nuisance parameters introduces extra troubles to practitioners.

Both \( I_n \) and \( \bar{I}_n \) measure the \( L^2 \) distance between \( m(\cdot) \) and \( \overline{m}(\cdot) \), so they are closely related; see Fan and Li (2000) for a clear argument for this point in the classical specification test. Such test statistics correspond to the average-statistic of Andrews and Ploberger (1994) in the parametric structural change environment. Another popular metric used in the structural change literature is the sup-norm; e.g., the sup-statistic in Andrews (1993) uses this metric. In our case, a natural test statistic under the sup-norm is \( \sup_{\pi \in \Pi} |\hat{g}(\pi)| \). To avoid the random denominator problem as in \( I_n \) and \( \bar{I}_n \), we use the following test statistic
\[
S_n = \sup_{\pi \in \Pi} \sqrt{n} h \left| \hat{m}_+(\pi) \hat{f}_+(\pi) - \hat{m}_-(\pi) \hat{f}_-(\pi) \right|
\]
\[
= \sup_{\pi \in \Pi} \left| \sqrt{n} h \sum_{j=1}^{n} k_h (x_j - \pi) y_j d_j (\pi) - \sqrt{n} h \sum_{j=1}^{n} k_h (x_j - \pi) y_j (1 - d_j (\pi)) \right|
\]
\[
= \sup_{\pi \in \Pi} |S_n|.
\]
By using the technique in Stadtmüller (1986), Wu and Chu (1993) study the asymptotic property of \( S_n \).\(^{26}\)

It turns out that the asymptotic distribution of \( S_n \) is related to the type-I extreme value distribution. Specifically,
\[
P(S_n < a_n + b_n x) \to \exp(-2 \exp(-x))
\]
where
\[
a_n = \sigma K_3 \left( C_\Pi + \frac{\log(3 \sqrt{4\pi})}{C_\Pi} \right), \quad b_n = \frac{\sigma K_3}{C_\Pi},
\]
\( K_3 \) is a constant only related to \( k (\cdot) \), and \( C_\Pi = \sqrt{C_1 + C_2 \log n} \) with \( C_1 \) related to the length of \( \Pi \) and \( C_2 \) related to the width of \( h \).

We provide some intuition for this result. First, by the central limit theory, the average form \( S_n \) in \( S_n \) will follow a normal distribution asymptotically. Second, because \( S_n \) only uses the data in a \( h \) neighborhood of \( \pi \), \( S_n_{\pi_1} \) and \( S_n_{\pi_2} \) are asymptotically independent when the distance between \( \pi_1 \) and \( \pi_2 \) is greater than \( 2h \). Third, since \( x_i \) is evenly sampled, and \( \varepsilon_i \) is homoskedastic, the normal distributions that \( S_n \) will follow are the same. In summary, \( S_n \) is asymptotically equivalent to the supremum of infinite i.i.d. normal random variables. By the extreme value theory (see, e.g., Example 21.16 of van der Vaart (1998)), for \( n \)

\(^{26}\)Strictly speaking, their test statistic uses the Gasser-Müller estimator instead of \( S_n \), but they are essentially the same. See also the local test in Grégoire and Hamrouni (2002a) and Horváth and Kokoszka (2002) for the case when the local polynomial estimator is used in estimating \( \hat{m}_+(\pi) \) and \( \hat{m}_-(\pi) \).
i.i.d. standard normal random variables $X_i$, $i = 1, \cdots, n$,

$$P \left( \max_{i} X_i < \sqrt{2 \log n} - \frac{1}{2} \frac{\log \log n + \log 4\pi}{\sqrt{2 \log n}} + \frac{1}{\sqrt{2 \log n}} \right) \to \exp \left( - \exp(-x) \right).$$

It is easy to see the similarity of the asymptotic distributions of $S_n$ and $\max_{i} X_i$. The "$2" in $\exp(-2\exp(-x))$ is because $S_n$ is the supremum of the absolute value of $S_{n,\pi}$ instead of $S_{n,\pi}$ itself. The simulation studies in Wu and Chu (1993) show that the type I error using $S_n$ to test $H_0$ is much higher than the nominal level, which is partially due to the low convergence rate of $S_n$ to its asymptotic distribution. It is unknown whether the wild bootstrap is valid or not for $S_n$.

It is noteworthy that we need normalize $S_n$ by a location constant $a_n$ to get a nondegenerate asymptotic distribution. When there is no normalizing constant $a_n$, the asymptotic distribution of $S_n$ will degenerate. To appreciate this result, consider $\max_{i} X_i$ again. It is shown in Gnedenko (1943) that for any $\varepsilon > 0$,

$$P \left( \left| \frac{\max_{i} X_i}{\sqrt{2 \log n}} - 1 \right| < \varepsilon \right) \to 1.$$

$\max_{i} X_i$ is called relatively stable in this case by Gnedenko (1943). We refer to Lepski and Tsybakov (2000) for the discussion about the degeneracy of $S_n$ in the classical specification test with the ideal white noise setup.

Although the argument above is elegant, it can not be applied in our case. The key point here is that $x_i$ is not evenly sampled and $\varepsilon_i$ may be heteroskedastic, but we need $S_{n,\pi}$ to be homoskedastic to apply the extreme value theory. Let us check the effect of either of the two conditions on $S_n$. If $\varepsilon_i$ is homoskedastic and $x_i$ is not evenly sampled, because $h$ is the same for all $S_{n,\pi}$, the asymptotic variance of $S_{n,\pi}$ will be smaller when $f(\pi)$ is larger since more data are used in $S_{n,\pi}$, and vice versa. To make the asymptotic variance same for all $S_{n,\pi}$, we must normalize $S_{n,\pi}$ by an estimator of $\sqrt{f(\pi)}$ or use variable bandwidths such as in Fan and Gijbels (1992). Second, if $\varepsilon_i$ is heteroskedastic and $x_i$ is evenly sampled, similarly, the asymptotic variance of $S_{n,\pi}$ will depend on $\pi$. We need divide $S_{n,\pi}$ by an estimator of $\sigma(\pi)$ to make it homoskedastic; See Eubank and Speckman (1993) for the discussion on this issue in confidence band construction in nonparametric regression. In summary, we need estimators of $f(\cdot)$ and $\sigma^2(\cdot)$ to make the extreme value theory apply. When $\varepsilon_i = \sigma(x_i) \varepsilon_i$ with $\varepsilon_i$ i.i.d. sampled, Hamrouni (1999) normalizes $S_{n,\pi}$ by $K_4 \sqrt{f(\pi)} / \sigma^2(\pi)$ to get the asymptotic distribution $\exp(-2\exp(-x))$, where $K_4$ is a constant only related to $k(\cdot)$. To our knowledge, there is no literature to consider normalizing $S_{n,\pi}$ by an estimator of $K_4 \sqrt{f(\pi)} / \sigma^2(\pi)$. Even there were, such a complicated procedure is not attractive to a practitioner.

The third test is proposed by Müller and Stadtmüller (1999); see also Dubowik and Stadtmüller (2000). This test is based on sums of squared differences of the data, formed with various span sizes:

$$Z_k = \sum_{j=1}^{n-L} \frac{(y_{j+k} - y_j)^2}{n-L}, 1 \leq k \leq L,$$

where $L = L(n) \geq 1$ is a sequence of integers depending on $n$. Müller and Stadtmüller (1999) show that the statistics $Z_k$ can be interpreted as dependent variables within the following two-parameter asymptotic linear model\footnote{This approximation is very clear when $k = 1$. Müller and Stadtmüller (1999) allow unknown number of jump points, and in this case, $\alpha^2_\pi$ in (8) should change to the sum of squared jump sizes.}

$$Z_k = 2\sigma^2 + \frac{k}{n-L} \alpha^2_\pi + \eta_k, 1 \leq k \leq L.$$ (8)
\( \alpha^2_\pi \) can be estimated by least squares and the asymptotic distribution of the estimator is as follows

\[
\sqrt{L} \left( \hat{\alpha}^2_\pi - \alpha^2_\pi \right) \xrightarrow{d} N \left( 0, \frac{12}{5} \left( E [\varepsilon^4] - \sigma^4 \right) \right)
\]

A usual \( t \) test can be conducted to test if \( \alpha_\pi = 0 \). Müller and Stadtmüller (1999) derive this test under the fixed design with i.i.d. errors, and it is unknown how to adjust this procedure to adapt to the case with random design and heterogeneous \( \varepsilon_i \)'s.

### 4 Efficient Estimation of the Treatment Effect

If the specification tests reject \( H^{(2)}_0 \) that \( \alpha_\pi = 0 \), we need to estimate \( \alpha_\pi \) with an unknown \( \pi \). Usually, we can use a two-step procedure: first estimate \( \pi \), then estimate \( \alpha_\pi \) as if \( \pi \) were known.

#### 4.1 Estimation of the Discontinuity Point

Because the weighted average nature of kernel smoothers, \( \hat{\alpha}(\pi) \) would be near to zero when there was no jump at a given point. Otherwise, the difference would be near the jump magnitude. So a natural estimator of \( \pi \) is as follows:

\[
\hat{\pi} = \arg \max_{\pi \in \Pi} \hat{\alpha}^2(\pi),
\]

Such an estimator is called The Difference Kernel Estimator (DKE) in Qiu et al (1991). In practice, we need only check if \( \pi = x_i, x_i \in \Pi \), maximizes \( \hat{\alpha}^2(\pi) \). Yu (2007) suggests to check the middle points of contiguous \( x_i \)'s to improve the efficiency of \( \hat{\pi} \). We expect either way will generate similar estimates of \( \alpha_\pi \) which is of main interest in RDDs. Before stating the asymptotic distributions of \( \hat{\pi} \), we impose the following standard restrictions on the bandwidth \( h \):

**Assumption H:** \( h \to 0, \, nh \to \infty, \) and \( \frac{nh}{\ln n} \to \infty \).

The following theorem provides the asymptotic distribution of \( \hat{\pi} \).

**Theorem 4** Suppose Assumptions E, \( F(b) \), H and M hold, and \( k(\cdot) \) satisfies the same conditions as specified in Assumption K, then \( \hat{\pi} - \pi_0 = O_p \left( n^{-1} \right) \), and

\[
n \left( \hat{\pi} - \pi_0 \right) \xrightarrow{d} \arg \max_{v \in \mathbb{R}} D(v),
\]

where

\[
D(v) = \begin{cases} 
\sum_{i=1}^{N_-(v)} \left( -\alpha^2_{\pi_0} + 2\alpha_{\pi_0} \varepsilon^-_i \right), & \text{if } v < 0; \\
\sum_{i=1}^{N_+(v)} \left( -\alpha^2_{\pi_0} + 2\alpha_{\pi_0} \varepsilon^+_i \right), & \text{if } v \geq 0;
\end{cases}
\]

\( N_-(\cdot) \) and \( N_+(\cdot) \) are Poisson processes with intensity \( f(\pi_0) \), \( \varepsilon^-_i \) has the density \( f(\varepsilon | x = \pi_0^{-}) \), \( \varepsilon^+_i \) has the density \( f(\varepsilon | x = \pi_0^{+}) \), and \( \{\varepsilon^+_i, \varepsilon^-_i\} \) are independent of each other. Furthermore, \( D(v) \) is caglad with \( D(0) = 0 \) almost surely, and the asymptotic distribution of \( \hat{\pi} \) is the same as that in the case when \( \alpha_{\pi_0} \) is known.

---

*28* See the introduction of Spokoiny (1998) for two different directions in the nonparametric regression analysis of a function with change-points. See also the introduction of Gréoire and Hamrouni (2002) for the other two different approaches in the estimation of nonparametric regression with discontinuities.
Note that there is an interval \((M_-, M_+)\) maximizing \(D(v)\), we must determine which point on the maximizing interval is taken as \(\arg \max_{v \in \mathbb{R}} D(v)\). If we check \(x_i \in \Pi\) when maximizing \(\hat{\alpha}^2(\pi)\) in (9), then \(\arg \max_{v \in \mathbb{R}} D(v) = M_+\); if we check middle points, then \(\arg \max_{v \in \mathbb{R}} D(v) = \frac{M_+ + M_-}{2}\). Also, when \(\varepsilon\) is independent of \(x, \varepsilon_i^-\) and \(\varepsilon_i^+\) have the same distribution as \(\varepsilon_i\). The explicit form of the asymptotic distribution of \(\hat{\pi}\) can be found in Appendix D of Yu (2007).

It is surprising that \(\hat{\pi}\) is \(n\)-consistent although \(m(\cdot)\) takes a nonparametric form; see the next subsection for some intuitions on this result. This result hinges on the assumption that \(k(0) > 0\); that is, we use some information in the neighborhood of \(\pi\) to estimate \(m_\pi(\pi)\). When \(k(0) = 0\) (and \(k'(0) > 0\), the convergence rate is \(\sqrt{\frac{nh}{k}} = n/\sqrt{\frac{nh}{k}}\) slower than \(n\); see Müller (1992), Wu and Chu (1993), and Delgado and Hidalgo (2000) for the details. Gréoire and Hamrouni (2002b) essentially use the objective function \(|\hat{\alpha}(\pi)|\), but their asymptotic distribution of \(\hat{\pi}\) is questionable. This is because \(\arg \max_{\pi \in \Pi} \hat{\alpha}^2(\pi) = \arg \max_{\pi \in \Pi} |\hat{\alpha}(\pi)|^\beta\) for any \(\beta \geq 1\), and their asymptotic distribution should be the same as in Theorem 4.

When there is only one discontinuity point, Korostelev (1987) shows that \(n^{-1}\) is the optimal minimax rate to estimate \(\pi\) in the ideal Gaussian white noise model. Actually, we can adapt the proof idea of Yu (2008b) to show that \(\pi\) can even be adaptively estimated. When the number of discontinuities is greater than 1, Spokoiny (1998) shows that the optimal minimax rate changes to \(n^{-1} \log n\).

### 4.2 Some Intuitions

It is useful to pause here and highlight some intuitions. We will first provide some intuitions for the \(n\)-consistency of \(\hat{\pi}\), then discuss two equivalence results on the asymptotic distribution of \(\hat{\pi}\) in Theorem 4.

#### 4.2.1 \(n\)-Consistency of \(\hat{\pi}\)

First, we discuss how the convergence rate is determined for a general estimator determined by maximizing an objective function. Suppose the parameter \(\theta \in \Theta\) is estimated by

\[
\hat{\theta} = \arg \max_{\theta \in \Theta} Q_n(\theta) = \arg \max_{\theta \in \Theta} [Q_n(\theta) - Q_n(\theta_0)],
\]

where \(Q_n(\cdot)\) is the objective function. Then because \(\hat{\theta}\) is the maximizer of \(Q_n(\theta) - Q_n(\theta_0)\) on \(\Theta\), and \(\theta_0 \in \Theta\),

\[
0 \leq Q_n(\hat{\theta}) - Q_n(\theta_0) = \left[Q(\hat{\theta}) - Q(\theta_0)\right] + \left[Q_n(\hat{\theta}) - Q(\hat{\theta}) - (Q_n(\theta_0) - Q(\theta_0))\right],
\]

where the first term on the right-hand side is the limit process and is less than zero since \(\theta_0 = \arg \max Q(\theta)\), and the second term is the modulus of continuity of the empirical process and is greater than zero. We must balance \(Q(\hat{\theta}) - Q(\theta_0)\) and

\[
\frac{\phi_n(\delta)}{\sqrt{n}} \equiv \sup_{|\theta - \theta_0| \leq \delta} [Q_n(\theta) - Q(\theta) - (Q_n(\theta_0) - Q(\theta_0))]
\]

\[29\] We agree with Hastie and Loader in their rejoinder to H.G. Müller in the 1993 review paper. It is hard to defend \(k(0) = 0\) given that we are estimating \(m_\pi(\pi)\) but do not use the data in the neighborhood of \(\pi\). Also, if we can achieve the \(n\)-consistency of \(\hat{\pi}\), why use an estimator with a lower convergence rate? One advantage of the estimator with \(k(0) = 0\) is that its asymptotic distribution is normal so that the inference is easy. But there are other methods to conduct inference on \(\pi\); see Section 6 of Yu (2008a) for a summary. What's more, we are interested in \(\alpha_\pi\) rather than \(\pi\) in RDDS, so inference on \(\pi\) is not very important, and we only need a precise estimator of \(\pi\) such that its impact on the estimation of \(\alpha_\pi\) is small.
such that their sum is greater than zero.

For a regular parameter $\mu$, say, the mean of a random variable, $Q(\mu)$ is smooth, and for $|\mu - \mu_0| \leq \delta$, $Q(\mu) - Q(\mu_0) = O(\delta^2)$, and $\hat{\Delta}(\delta) = O_p\left(\frac{\delta}{\sqrt{n}}\right)$. Set $\hat{\mu} - \mu_0 = O_p\left(r_n^{-1}\right)$, and let $\left(\frac{1}{r_n}\right)^2 \approx \frac{r_n^{-1}}{\sqrt{n}}$, we have $r_n = \sqrt{n}$.

The balancing in estimating $\pi$ is different. Consider a simple model with fixed design:

$$y_i = \alpha 1(x_i \geq \pi) + \varepsilon_i,$$

(10)

where $\varepsilon_i$’s are i.i.d. and follow $N(0, 1)$, $x_i = i/n$, $i = 1, \cdots, n$, and $\pi_0 = 1/2$. WLOG, suppose $\alpha > 0$. In the parametric estimation of $\pi$, $\hat{\alpha}$ maximizes $\hat{\alpha}(\pi) - \hat{\alpha}(\pi_0)$. For $|\pi - \pi_0| \leq \delta$, and $\pi > \pi_0$,

$$\hat{\alpha}(\pi) - \hat{\alpha}(\pi_0) = \left[\frac{1}{n(1-\pi)} \sum_{i=n\pi+1}^{n} y_i - \frac{1}{n\pi} \sum_{i=1}^{n\pi} y_i\right] - \left[\frac{1}{n(1-\pi_0)} \sum_{i=n\pi_0+1}^{n} y_i - \frac{1}{n\pi_0} \sum_{i=1}^{n\pi_0} y_i\right]$$

$$= \left[\alpha + \frac{1}{n(1-\pi)} \sum_{i=n\pi+1}^{n} \varepsilon_i - \frac{n(\pi - \pi_0)}{n\pi} \alpha - \frac{1}{n\pi} \sum_{i=1}^{n\pi} \varepsilon_i\right] - \left[\alpha + \frac{1}{n(1-\pi_0)} \sum_{i=n\pi_0+1}^{n} \varepsilon_i - \frac{n(\pi_0 - \pi_0)}{n\pi_0} \alpha - \frac{1}{n\pi_0} \sum_{i=1}^{n\pi_0} \varepsilon_i\right]$$

$$\approx -\alpha \left(\frac{\pi - \pi_0}{\pi}\right) - \left(\frac{1}{n\pi} + \frac{1}{n(1-\pi_0)}\right) \sum_{i=n\pi_0+1}^{n\pi} \varepsilon_i = O\left(\frac{\pi - \pi_0}{n}\right) + O\left(\frac{n-1}{r_n}\right) O_p\left(\sqrt{n}(\pi - \pi_0)\right)$$

$$= O\left(\delta\right) + O_p\left(\frac{\sqrt{\delta}}{\sqrt{n}}\right),$$

so $\phi_n(\delta) = \sqrt{\delta}$. Similar results hold for $\pi < \pi_0$. Set $\hat{\pi} - \pi_0 = O_p\left(r_n^{-1}\right)$, and let $\frac{1}{r_n} \approx \frac{1}{\sqrt{n}}$, we have $r_n = n$.

In the nonparametric estimation of $\pi$, the scale is changed to $h$. For $\pi = \pi_0 + ah$ with $a > 0$ small enough,

$$\hat{\alpha}(\pi) - \hat{\alpha}(\pi_0) = \left[\frac{1}{nh} \sum_{i=n\pi+1}^{n} y_i - \frac{1}{nh} \sum_{i=n(\pi-h)}^{n} y_i\right] - \left[\frac{1}{nh} \sum_{i=n\pi_0+h}^{n\pi_0} y_i - \frac{1}{nh} \sum_{i=n\pi_0-h}^{n\pi_0} y_i\right]$$

$$= \left[\alpha + \frac{1}{nh} \sum_{i=n\pi+1}^{n\pi} \varepsilon_i - \frac{n(\pi - \pi_0)}{nh} \alpha - \frac{1}{nh} \sum_{i=n(\pi-h)}^{n\pi} \varepsilon_i\right] - \left[\alpha + \frac{1}{nh} \sum_{i=n\pi_0+h}^{n\pi_0} \varepsilon_i - \frac{n(\pi_0 - \pi_0)}{nh} \alpha - \frac{1}{nh} \sum_{i=n\pi_0-h}^{n\pi_0} \varepsilon_i\right]$$

$$\approx -\alpha \left(\frac{\pi - \pi_0}{h}\right) - \frac{2}{nh} \sum_{i=n\pi_0+1}^{n\pi} \varepsilon_i = O\left(\frac{\pi - \pi_0}{h}\right) - O\left(\frac{1}{nh}\right) O_p\left(\sqrt{n}(\pi - \pi_0)\right)$$

$$= O\left(\frac{\delta}{h}\right) + O_p\left(\frac{1}{h}\frac{\sqrt{\delta}}{\sqrt{n}}\right).$$

So solve

$$\frac{r_n^{-1}}{h} \approx \frac{1}{h} \sqrt{\frac{1}{r_n}},$$

we have $r_n = n$.

Figure 3 illustrates the intuition above. Roughly speaking, in estimating $\pi$, $Q(\pi) - Q(\pi_0)$ is a non-smooth function of $\pi$ in the neighborhood of $\pi_0$ such that $\pi_0$ can be identified more easily than $\mu_0$. In the nonparametric case, we use a smaller scale $h$, and focus the discussion in a $h$ neighborhood of $\pi_0$. 

18
4.2.2 Two Equivalence Results

Given the \( n \)-consistency of \( \hat{\pi} \), we continue to find the asymptotic limit of the localized objective function. Here, we need the Lipschitz continuity of \( k(\cdot) \), and the uniform kernel above does not work. To simplify the discussion, let \( \pi_0 \) be the closest \( \frac{1}{n} \) to \( \frac{1}{2} \). Assume further that both \( v \) and \( nh \) are positive integers. Then for \( \hat{\alpha}(\pi) \) based on a kernel function \( k(\cdot) \),

\[
\begin{align*}
\hat{\alpha}(\pi_0 + \frac{v}{nh}) - \hat{\alpha}(\pi_0)
&= \sum_{i=1}^{nh} k\left(\frac{i}{nh}\right) y_{n\pi_0 + v + i} - \sum_{i=-nh}^{0} k\left(\frac{i}{nh}\right) y_{n\pi_0 + i} - \sum_{i=-nh}^{0} k\left(\frac{i}{nh}\right) y_{n\pi_0 + i} \\
&= \sum_{i=-nh}^{nh-v-1} \left[ k\left(\frac{i}{nh}\right) \varepsilon_{n\pi_0 + i} - \sum_{i=-nh}^{0} \left[ k\left(\frac{i-v}{nh}\right) - k\left(\frac{i}{nh}\right) \right] \varepsilon_{n\pi_0 + i} \\
&- \sum_{i=1}^{v} \left[ k\left(\frac{i}{nh}\right) \alpha + \left[ k\left(\frac{i}{nh}\right) + k\left(\frac{i-v}{nh}\right) \right] \varepsilon_{n\pi_0 + i} \right] \\
&+ \sum_{i=1}^{nh-v} \left[ k\left(\frac{i}{nh}\right) \varepsilon_{n\pi_0 + v + i} + \sum_{i=-nh-v+1}^{nh} k\left(\frac{i}{nh}\right) \varepsilon_{n\pi_0 + v + i} \right] \\
&= I + II + III + IV + V.
\end{align*}
\]

Figure 4 shows the effect of \( k(\cdot) \) on the five terms. When \( k(\cdot) \) is Lipschitz, all terms except III are \( o_p(1) \). Especially, \( k\left(\frac{1}{nh}\right) \) in I and V is close to zero. But if \( k(\cdot) \) is uniform, I and V will not disappear. The summands in III correspond to the jumps in \( D(v) \). Note that \( \frac{i-v}{nh} \) and \( \frac{i}{nh} \) are close to zero when \( i \) and \( v \) are finite, so the summand in III is approximately \(-k(0)(\alpha + 2\varepsilon_{n\pi_0 + i})\)\(^{30}\). Similar results hold for \( v < 0 \).

Now, we can state two equivalence results for the asymptotic distribution of \( \hat{\pi} \). To simplify our discussion, we continue using the model with fixed design. First, the asymptotic distribution of \( \hat{\pi} \) is the same as the parametric estimator based on \( \hat{\alpha}(\pi) \) in \( \hat{\alpha}(\pi) \). In other words, although RDDs are nonparametric, we can

\(^{30}\)Since our objective function is \( \hat{\alpha}^2(\pi) \), \( nh(\hat{\alpha}^2(\pi_0 + \frac{v}{nh}) - \hat{\alpha}^2(\pi_0)) \) = \( nh(\hat{\alpha}(\pi_0 + \frac{v}{nh}) - \hat{\alpha}(\pi_0)) \) (\( \hat{\alpha}(\pi_0 + \frac{v}{nh}) + \alpha \)) converges to \( 2\alpha \hat{\varepsilon}_\pi \), so the jump size in \( D(v) \) should be \(-k(0)(2\varepsilon_{n\pi_0 + i})\) \( (\alpha \hat{\varepsilon}_\pi + 2\varepsilon_{n\pi_0 + i}) \), which is different from the jumps of \( D(v) \) in Theorem 4 only by a constant \( 2k(0) \). Of course, a constant will not affect the argmax operator.
estimate \( \pi_0 \) as if the model is parametric as long as \( m(\cdot) \) is smooth. Given that \( \widehat{\pi} \) is \( n \)-consistent, only the data in a \( n^{-1} \) neighborhood of \( \pi_0 \) (that is, finite data points) are informative to \( \pi_0 \); see the derivation above. Because \( nh \to \infty \), there are infinite data points in the \( h \) neighborhood of the kernel smoother. Also, the kernel estimator treats \( m(\cdot) \) as a constant in any \( h \) neighborhood. So the kernel estimator is smooth enough to identify \( \pi_0 \) as if \( m(\cdot) \) were constant in big enough left and right neighborhoods of \( \pi_0 \). Following such an argument, it is not surprising that \( \pi_0 \) can be estimated as if in a parametric model.

Second, the asymptotic distribution of \( \widehat{\pi} \) is the same as that of the least squares estimator in (1). In other words, our objective function can be

\[
\sum_{i=1}^{n} \left( y_i - \alpha_{\pi} d_i(\pi) - \sum_{j=1}^{n} w_{ij}^2 (y_j - \alpha_{\pi} d_j(\pi)) \right)^2, \tag{12}
\]

where \( w_{ij} = \frac{K_{h,ij}}{\sum_{l=1}^{n} K_{h,il}} = \frac{K_{h,ij}}{nh} \), and \( \sum_{j=1}^{n} w_{ij}^2 (y_j - \alpha_{\pi} d_j(\pi)) \) can be treated as an estimator of \( m_{\pi}(x) \) at \( x_i \). This is exactly the objective function of partially linear estimation in Porter (2003). Note that (12) can be written as

\[
\| (I - W) Y - \alpha_{\pi} (I - W) D_{\pi} \|_2^2, \tag{13}
\]

where \( \| \cdot \| \) is the Euclidean norm in \( \mathbb{R}^n \), \( I \) is an \( n \times n \) identity matrix, \( W = (w_{ij})_{i,j=1,\ldots,n} \) is symmetric, and \( D_{\pi} = (d_i(\pi))_{i=1,\ldots,n} \). So the estimator of \( \alpha_{\pi} \) given \( \pi \) is

\[
\hat{\alpha}(\pi) = \frac{D'_{\pi} (I - W)^2 Y}{D'_{\pi} (I - W)^2 D_{\pi}},
\]

which is the coefficient in the least squares projection of \( (I - W) Y \) on \( (I - W) D_{\pi} \). Now, (13) becomes

\[
\| (I - W) Y \|_2^2 - \left\| (I - W) D_{\pi} \frac{D'_{\pi} (I - W)^2 Y}{D'_{\pi} (I - W)^2 D_{\pi}} \right\|^2 = \| (I - W) Y \|_2^2 - \| P_{\pi} (I - W) Y \|_2^2,
\]

\[\text{See Section 4.1 of Yu (2008b) for the asymptotic distribution of the least squares estimator in the parametric case, but as argued above, it should be the same as in the nonparametric case.}\]
where $P_\pi = \frac{(I - W)D_\pi D_\pi' (I - W)}{D_\pi' (I - W)^2 D_\pi}$ is a projection matrix. Because the first term $\|(I - W)Y\|^2$ does not involve $\pi$, minimizing the objective function with respect to $\pi$ is equivalent to maximizing $\|P_\pi (I - W)Y\|^2$ with respect to $\pi$. Note that $\|P_\pi (I - W)Y\|^2 = \hat{\alpha}^2(\pi) D_\pi' (I - W)^2 D_\pi$, and from Theorem 2 of Porter (2003),

$$D_\pi' (I - W)^2 D_\pi \overset{p}{\to} 2 \int_0^1 \left( \int_u^1 k(v) dv \right)^2 du$$

independent of $\pi$, so maximizing $\|P_\pi (I - W)Y\|^2$ is equivalent to maximizing $\hat{\alpha}^2(\pi)$.

### 4.3 Estimation of the Treatment Effect

Given $\hat{\pi}$, $\alpha_\pi$ is estimated by

$$\hat{\alpha}_\pi = \hat{\alpha}(\hat{\pi}),$$

so $\hat{\alpha}_\pi$ is a natural by-product of the estimation of $\pi$. The asymptotic distribution of $\hat{\alpha}_\pi$ is as follows.

**Theorem 5** Under the assumptions in Theorem 4,

$$\sqrt{nh} (\hat{\alpha}(\hat{\pi}) - \hat{\alpha}(\pi_0)) = o_p(1).$$

Furthermore, $\hat{\alpha}(\hat{\pi})$ is asymptotically independent of $\hat{\pi}$.

**Theorem 5** claims that the estimating of $\pi_0$ will not affect the efficiency of $\hat{\alpha}_\pi$. In other words, $\pi_0$ can be treated as known even if it is not in evaluating the treatment effect in RDDs. Combining this result with that in Theorem 4, we can treat $\alpha_\pi$ as if were known when estimating $\pi$, and treat $\pi$ as if were known when estimating $\alpha_\pi$. From Yu (2008b), $\pi_0$ is a "middle" boundary of $x$, and the information for the regular parameter $\alpha_\pi$ and the nonregular parameter $\alpha_\pi$ do not affect each other. So such an independence result is not surprising. Given this result, the asymptotic distribution of $\hat{\alpha}(\hat{\pi})$ is the same as that given in Theorem 3 of Porter (2003). Because the model with $\pi$ unknown is more difficult than that with $\pi$ known, and our estimator has the same efficiency as in the latter case, it achieves the optimal rate of convergence in the minimax sense.

When there are multiple, let’s say $q$, discontinuity points, $\pi_1, \ldots, \pi_q$ in ascending order (WLOG, suppose $\alpha_{\pi_1} > \alpha_{\pi_2} > \cdots > \alpha_{\pi_{q-1}} > \alpha_{\pi_q}$), a sequential procedure can be used to detect them. Specifically, we first estimate $\pi_1$ by maximizing $\hat{\alpha}^2(\pi)$ on $\Pi_1 = \Pi$. Then sequentially, $\pi_j$ is estimated by maximizing $\hat{\alpha}^2(\pi)$ on $\Pi_j = \Pi_{j-1} \backslash \bigcup_{k=1}^{j-1} [\pi_k - 2h, \pi_k + 2h]$, for $j = 2, \ldots, q$. Given $\hat{\pi}_j$’s, $\alpha_{\pi_j}$ can be estimated straightforward. Before stating the asymptotic distributions of $\hat{\pi}_j$ and $\hat{\alpha}_{\pi_j}$, we must assume $\pi_j$’s are separated apart. Rigorously,

**Assumption A:** $\min_{j=2, \ldots, q} (\pi_j - \pi_{j-1}) > \epsilon > 0$.

Assumption A does not exclude the possibility that $\pi_1 = \underline{\pi}$ and $\pi_q = \overline{\pi}$. This is because we assume the support of $x$ covers $\Pi_\epsilon$ so that $\underline{\pi}$ and $\overline{\pi}$ are not the boundaries of $x$’s support. The following corollary provides the joint asymptotic distribution of $\left\{\hat{\pi}_j, \hat{\alpha}_{\pi_j}\right\}_{j=1}^q$.

**Corollary 1** Suppose Assumptions A, E, F(b), H and M hold, and k(·) satisfies the same conditions as specified in Assumption K, 

---

32 A similar procedure can be found in Wu and Chu (1993) although they use a different estimator of $\alpha_\pi$; see also Yin (1988) and Delgado and Hidalgo (2000).
(i) \( \hat{\pi}_j - \pi_{j0} = O_p\left(n^{-1}\right), j = 1, \ldots, q, \) and \( n\left(\hat{\pi}_j - \pi_{j0}\right) \) has the same asymptotic distribution as \( n\left(\hat{\pi} - \pi_0\right) \) in Theorem 3 except that \( \pi_0 \) is replaced by \( \pi_{j0} \) in \( D(v) \).

(ii) \( \sqrt{n h} \left(\hat{\alpha}_{\pi_j} - \hat{\alpha}_{\pi_{j0}}\right) = o_p(1) \) for all \( 1 \leq j \leq q \).

(iii) All \( \hat{\pi}_j \)‘s and \( \hat{\alpha}_{\pi_j} \)’s are asymptotically independent of each other.

This corollary claims that each \( (\pi_j, \alpha_{\pi_j}) \) can be estimated independent of the others and as if the others were known. This is because each \( \hat{\pi}_j \) is \( n \)-consistent, and w.p.a.1. \( \pi_{j0} \) will stay in the \( h \) neighborhood of \( \hat{\pi}_j \). Given that we exclude a \( 2h \) neighborhood of \( \hat{\pi}_j \), we essentially use an independent data set to estimate \((\pi_{j+1}, \alpha_{\pi_{j+1}})\), and consequently, \((\hat{\pi}_{j+1}, \hat{\alpha}_{\pi_{j+1}})\) are asymptotically independent of \((\hat{\pi}_j, \hat{\alpha}_{\pi_j})\) and not affected by \((\hat{\pi}_{j}, \hat{\alpha}_{\pi_j})\).

There is also some literature discussing the \( L^p \)-consistency of the whole \( m(\cdot) \) function, where \( p \geq 1 \). Müller (1992) uses the boundary kernel developed in Gassar and Müller (1979) to deal with the estimation of \( m(\cdot) \) in the neighborhood of the discontinuity point, and derives the \( L^p \) convergence rate of his estimator. Oudshoorn (1998) uses the \( L^2 \) norm, and allows multiple discontinuity points (with unknown numbers and locations) in the ideal white noise model. He concentrates on the asymptotically minimax estimation of \( m(\cdot) \) based on series, and shows that the minimax rate is the same as that in Stone (1982) where no jump exists.

5 Extensions and A Practical Issue

In this section, we discuss three extensions of the results in Section 3 and 43. Also, we will deal with an important practical issue, the bandwidth selection, in both specification testing and estimation.

5.1 Two Auxiliary Tests

In this subsection, we will provide two auxiliary tests. The first one is to check Assumption F. There are two parallel tests as in Section 2.3. The first test is to check whether \( f(\cdot) \) is smooth at \( \pi \); that is, we want to know if there is control of \( x \) although the control is not precise. The second one is to check whether there is a jump in \( f(\cdot) \) at \( \pi \). McCrary (2008) interprets the second test as an indicator of whether there is precise manipulation of the running variable \( x \). He uses the local linear density estimator of Cheng (1994) to test this hypothesis in the spirit of the first or second alternative test in Section 3.3 but with \( \pi \) known. Our test statistic \( T_n \) can be extended to apply in both tests when \( \pi \) is unknown. The only adjustment is to redefine \( x_i \) and \( y_i \). For this purpose, we first divide \( \Pi_e \) into \( N = \left[ \frac{\pi_{e+1} - (\pi_e - 1)}{h} \right] + 1 \) subintervals, and denote the middle-points of these intervals as \( \{\pi_l\}_{l=1}^N \). Then define \( Y_l = \frac{1}{nh} \sum_{i=1}^n 1 \left( \pi_l - \frac{h}{2} \leq x_i < \pi_l + \frac{h}{2} \right) \). Now, substitute \((x_i, y_i)\) in \( I_n \) by \((\pi_l, y_i)\), and make sure that \( \bar{h} = o(h) \) and \( h = o(h) \), then the results in Theorem 1 and 2 apply. As argued in Section 3.2 of McCrary (2008), his procedure is robust to the choice of \( \bar{h} \), and a similar argument can be applied in our case. As to the choice of \( h \) and \( b \), see Section 5.4 below. As noted in McCrary (2008), "a running variable with a continuous density is neither necessary nor sufficient for identification except under auxiliary assumptions". Lee (2008) claims that the treatment effect can be identified even if there were manipulation as long as the treatment status cannot be precisely controlled

---

33 Müller (1992) uses the Gassar-Müller estimator whose properties are available only in the case where design points are \( O(n^{-1}) \) apart, which excludes random design, since in that circumstance the maximal spacing is of size \( n^{-1} \log n \).

34 The key assumption of this paper is that the smoothness index \( s \) is known beforehand. If \( s \) is unknown, we can extend Horowitz and Spokoiny (2001) in specification testing and Sun (2005) in estimation to adapt to the unknown \( s \). But such a development is beyond the scope of this paper.

35 They degenerate to the same test when \( \pi \) is known.
by individuals. A testable corollary of his framework is that the conditional mean of any pre-determined variable given \( x \) is continuous at \( \pi \). Of course, our testing procedure can be used to test such a hypothesis.

The second test is to check if there are enough compliers in a fuzzy design. To describe this test, first recall the identification structure in a fuzzy design. From Hahn et al (2001), under the local unconfoundedness condition, the treatment effect \( \Delta_\pi \) can be identified, and

\[
\Delta_\pi \equiv E[Y_1 - Y_0|x = \pi] = \frac{\alpha_\pi}{\beta_\pi},
\]

where \( \beta_\pi = p_+(\pi) - p_-(\pi) \), and \( p_\pm(\pi) = E[D|x = \pi \pm] \) are propensity scores in the right and left neighborhoods of \( \pi \) with \( D \) being the treatment status. Under the monotonicity assumption as used in Imbens and Angrist (1994), the denominator of \( \Delta_\pi \), \( \beta_\pi \), can be interpreted as the probability to be a complier. To check whether \( \Delta_\pi \neq 0 \) by testing whether \( \alpha_\pi = 0 \), we must make sure \( \beta_\pi \neq 0 \); that is, there are enough compliers. Our test statistic \( T_n \) can be applied directly except that \( D_i \) now plays the role of \( y_i \). Because \( |\beta_\pi| < 1 \), the significance test of the treatment effect based on \( \alpha_\pi \) is conservative.

### 5.2 Estimation in Fuzzy Design

We first discuss the estimation of \( \pi \). In a fuzzy design, the available data are \( \{y_i, x_i, D_i\}_{i=1}^n \). The extra data \( D_i \) can provide extra information about \( \pi \). To be specific, our estimator of \( \pi \) is

\[
\hat{\pi} = \arg\max_{\pi \in \Omega} \left[ \hat{\alpha}^2(\pi) + \hat{\beta}^2(\pi) \right],
\]

where \( \hat{\beta}(\pi) = \hat{p}_+(\pi) - \hat{p}_-(\pi) \), and \( \hat{p}_\pm(\pi) \) are nonparametric estimators of \( p_\pm(\pi) \), respectively. The asymptotic distribution of \( \hat{\pi} \) is the same as that of \( \hat{\pi} \) in Theorem 3 except that the \( D(v) \) process is redefined as

\[
D(v) = \begin{cases} 
\sum_{i=1}^{N^-([v])} \left(-\alpha_{\pi_0}^2 + 2\alpha_{\pi_0} \epsilon_i^- - \beta_{\pi_0}^2 \epsilon_i^- - 2\beta_{\pi_0} \eta_i^- \right), & \text{if } v < 0; \\
\sum_{i=1}^{N^+([v])} \left(-\alpha_{\pi_0}^2 - 2\alpha_{\pi_0} \epsilon_i^+ - \beta_{\pi_0}^2 \epsilon_i^+ - 2\beta_{\pi_0} \eta_i^+ \right), & \text{if } v \geq 0;
\end{cases}
\]

where \( \eta_i = D_i - E[D_i|x = x_i] \), and \( \epsilon_i^\pm \) are similarly defined as \( \epsilon_i^\pm \) \footnote{Note that \( \eta_i \) follows a discrete distribution given \( x_i \), so \( \epsilon_i^\pm \) have a discrete component. As a result, the maximizer of \( D(v) \) may not be uniquely defined. So we need a scheme to uniquely define the maximizer when there exist multiple ones; e.g., the average of all maximizers, or the average of the maximum and minimum of the maximizers. Note also that this is only an asymptotic result, and in finite samples, multiple maximizers hardly happen.} Because \( E\left[-\alpha_{\pi_0}^2 \pm 2\alpha_{\pi_0} \epsilon_i^+ - \beta_{\pi_0}^2 \pm 2\beta_{\pi_0} \eta_i^+ \right] = -\alpha_{\pi_0}^2 - \beta_{\pi_0}^2 \}, \), we expect \( \hat{\pi} \) is more efficient than \( \hat{\pi} \).

Given an estimator of \( \pi \), the treatment effect \( \Delta_\pi \) defined in [14] can be estimated as in Section 3.6 of Porter (2003). Of course, due to the superconsistency of \( \hat{\pi} \), the asymptotic distribution of the estimator of \( \Delta_\pi \) is the same as that in the case when \( \pi_0 \) is known. For this asymptotic distribution, we refer to Proposition 1 of Porter (2003).

### 5.3 The Number of Discontinuity Points is Unknown

Usually, the number of discontinuity points \( q \) is known in RDDs. When \( q \) is unknown, there are three procedures to get information about \( q \). We maintain Assumption A in the following discussion.

First, we can conduct tests about \( q \) sequentially; Wu and Chu (1993) belongs to this class. For this method, we use the notations in Section 4. Specifically, we first test \( H_0 : q = 0 \) versus \( H_1 : q > 0 \) using some...
test statistic $T_n$ on $\Pi_1$. If $H_0$ is rejected, then estimate $\pi_1$ using the procedure in Section 4. Generally, test $H_0 : q = j$ versus $H_1 : q > j$ using $T_n$ on $\Pi_{j+1}$. Keep on this procedure until $H_0$ cannot be rejected. Remark 4 of Wu and Chu (1993) uses the sup-statistic $S_n$ as $T_n$. Our test statistic $T_n$ can be used in a similar way. Because of the type-I error in testing, this method tends to overestimate $q$.

Second, we can estimate $q$. Yin (1988) and Qiu (1994) belong to this class. We briefly describe the procedure of Qiu (1994) here. Because the kernel estimator uses the data in a $2h$ range, we divide $\Pi$ into $N = \left\lceil \frac{\pi I}{nh} \right\rceil + 1$ (closed) subintervals. Denote the right endpoints of these intervals as $\pi_l$, where $\pi_l = \pi + (\pi - \pi)l/N$, $l = 1, \ldots, N$. Calculate $\hat{\alpha}(\pi)$ at all $\pi_l$’s, and find the $\pi_l$’s such that $|\hat{\alpha}(\pi_l)| > B_n$, where $B_n = O\left(\beta_n \sqrt{\log \frac{n}{mh}}\right)$ with $\lim_{n \to \infty} \beta_n = \infty$ such as $\beta_n = \log \log n$. Such a $B_n$ is understandable because when there is no jump, $\sup_{\pi \in \Pi} |\hat{\alpha}(\pi)| = O_p\left(\sqrt{\frac{\log n}{mh}}\right)$ from Section 3.3. Denote such $\pi_l$’s as $\{\pi_{l_1}, \ldots, \pi_{l_q}\}$.

In the neighborhood of a discontinuity point, there may be more than one $\pi_{l_q}$’s such that $|\hat{\alpha}(\pi_{l_q})| > B_n$, so the contiguous $\pi_{l_q}$’s are indicating the same discontinuity point. Combining the contiguous subintervals whose left endpoint is $\pi_{l_q}$, we get intervals $\mathcal{I}_1, \ldots, \mathcal{I}_k, \ldots, \mathcal{I}_q$, where $\hat{q}$ is the estimator of $q$, and the middle point of each $\mathcal{I}_k$ or the maximizer of $|\hat{\alpha}(\pi)|$ on each $\mathcal{I}_k$, $k = 1, \ldots, \hat{q}$, is the estimator of the location of jumps. Qiu (1994) shows the a.s. consistency and convergence rate of $\hat{q}$.

Third, we can select $q$ to obtain an optimal fitting. This approach is proposed in Braun and Müller (1998) and used in Müller and Stadtmüller (1999) and Gijbels and Goderniaux (2004), but no theoretical justification is provided yet. Specifically, we can choose $q$ and $h$ together when minimizing the objective function of cross-validation below, where $\hat{m}_{-i}(x_i)$ is a function of both $q$ and $h$.

### 5.4 Bandwidth Selection in Specification Testing and Estimation

Most statistical and econometric literature concentrates on the bandwidth selection in estimation instead of specification testing. We will first provide an algorithm for bandwidth selection in estimation, then adjust the algorithm for specification testing. Throughout this discussion, we assume there is at most one cusp or jump under the null and alternative.

For statistical literature on bandwidth selection when $m(\cdot)$ is discontinuous, see, e.g., Remark 2 of Wu and Chu (1993a), Wu and Chu (1993b), and Section 5 of Müller (1992). It seems that only Wu and Chu (1993b) give a rigorous analysis. There is also some econometric literature on bandwidth selection in estimation of $\alpha_\pi$ when $\pi$ is known. Ludwig and Miller (2005) use the conventional CV approach to select the bandwidth. To avoid the boundary problem, they use the one-side kernel in estimating $m(\cdot)$ on both sides of $\pi$. Such a scheme induces some efficiency loss in data using. DeJardins and McCall (2008) and Imbens and Kalyanaraman (2009) use the plug-in method, but as argued in Loader (1999), plug-in approaches are not as elegant as our algorithm below when $\pi$ is either known or unknown.
tuned by arbitrary specification of pilot estimators and prone to oversmoothing when presented with difficult
smoothing problems." See Imbens and Kalyanaraman (2009) for detailed description on these methods.

Cross-validation remains the dominating approach in bandwidth selection for practitioners. Our approach
is based on and extend the CV method of Ludwig and Miller (2005), Wu and Chu (1993b)\(^\text{43}\) Different from
the CV method of Ludwig and Miller (2005), Wu and Chu (1993b) eliminate the boundary effect by projecting or reflecting the data
at the discontinuity point. Such an idea is used in Wu and Chu (1993c) who borrow from Hall and Wehrly
(1991). Given that the local linear smoother (LLS) is dominating in practice, we adapt the procedures in
Hall and Wehrly (1991) and Wu and Chu (1993b) by using the LLS. Because this procedure is not popular
in the RDD literature, we give a detailed description here.

Algorithm CV:

**Step 1:** Fix \( h \in H_n \equiv [An^{-1+p}, Bn^{-p}] \) for arbitrary small \( p \) and constants \( A \) and \( B \).

**Step 2:** Estimate \( \pi \) using \( \text{44} \) with the bandwidth \( c_h \), where \( c \) is determined by

\[
c = \frac{C(k^*_+)}{C(k^*)},
\]

where \( C(K) = \left\{ \int K^2(u)du \right\}^{1/5} \) for a kernel function \( K(u) \),

\[
k^*(u) = (1 0) \left( \int_{-1}^{1} k(t)dt \int_{-1}^{1} tk(t)dt \right)^{-1} \begin{pmatrix} k(u) \\ uk(u) \end{pmatrix} = k(u)
\]

is the equivalent kernel of \( k(u) \), and

\[
k^*_+(u) = \frac{\left( \int_{0}^{1} t^2k(t)dt \right) k(u) - \left( \int_{0}^{1} tk(t)du \right) uk(u)}{\left( \int_{0}^{1} k(t)dt \int_{0}^{1} t^2k(t)dt - \left( \int_{0}^{1} tk(t)dt \right)^2 \right)}
\]

is the equivalent kernel of \( k_+(u) \equiv k(u)1(0 \leq u \leq 1) \) in the local linear regression\(^\text{45}\) To avoid the
difficulty in deciding \( x_i \) should be used in estimating \( m_+(x_i) \) or \( m_-(x_i) \) and improve the efficiency of
\( \hat{\pi} \), check the middle points of contiguous \( x_i \)'s on \( \Omega \) in maximizing \( \hat{\sigma}^2(\pi)\]^\text{46}\)

**Step 3:** Given \( \hat{\pi}, \hat{m}_+(\hat{\pi}) \) and \( \hat{m}_-(\hat{\pi}) \) in step 2, generate pseudo-data \( (\bar{x}_{i+}, \bar{y}_{i+}) \equiv (2\hat{\pi} - x_{i-}, 2\hat{\pi} - \hat{\pi} - y_{i-}) \),

\( i = 1, \cdots, L_h \), in the right \( h \) neighborhood and \( (\bar{x}_{i-}, \bar{y}_{i-}) \equiv (2\hat{\pi} - x_{i+}, 2\hat{\pi} - y_{i+}) \), \( i = 1, \cdots, R_h \),

in the left \( h \) neighborhood of \( \hat{\pi} \), where \( \{x_{i-}, y_{i-}\}_{i=1}^{L_h} \) are the data points such that \( \hat{\pi} - x_{i-} \leq h \), and

\( \{x_{i+}, y_{i+}\}_{i=1}^{R_h} \) are the data points such that \( x_{i+} - \hat{\pi} \leq h \).

**Step 4:** For \( x_i \in [\hat{\pi}, \bar{\pi}] \), use \( \{x_{i-}, y_{i-}\}_{i=1}^{L_h} \cup \{x_{i+}, y_{i+}\}_{i=1}^{R_h} \) with \( x_{i-} \in [\bar{\pi} - h, \hat{\pi}] \) and for \( x_i \in [\bar{\pi}, \pi] \), use

\( \{x_{i+}, y_{i+}\}_{i=1}^{R_h} \) with \( x_{i+} \in [\bar{\pi}, \pi + h] \) to get the "leave-one-out" version of \( \hat{m}_i(x_i) \). Here, the
bandwidth of the LLS is \( h \), and the resulting estimates are denoted as \( \hat{m}_i(x_i) \)

\(^{43}\) Wu and Chu (1993b) assume that \( m_a(\cdot) \) is twice continuously differentiable, and the i.i.d errors \( (\varepsilon_i) \) have infinite moments.
Even though such assumptions seem too strong, we expect their procedure to work better than the conventional one.

\(^{44}\) The equivalent kernel \( c \) is first introduced by Lejeune (1985).

\(^{45}\) \( c \) is determined as follows. It is well known that the optimal bandwidth minimizing the MSE is \( n^{-1/5}C(K)C(m, f, \sigma^2) \)
for both the interior point and the boundary point, where \( C(m, f, \sigma^2) \) is a constant related to \( m(\cdot), f(\cdot) \) and \( \sigma^2(\cdot) \) but not to \( K(\cdot) \), so the optimal bandwidths for these two kinds of points are different only by a constant ratio \( c \) if \( m(\cdot), f(\cdot) \) and \( \sigma^2(\cdot) \) are stable on the interested area.

\(^{46}\) Also, in step 3 below, we do not need to consider the reflection of \( (x_i, y_i) \) itself.
Step 5: Get the objective function of CV:

\[ CV(h) = \sum_{i=1}^{n} (\hat{m}_{\pi}(x_i) - y_i)^2 1_i^\Pi \]  

(15)

for each \( h \), and minimize \( CV(h) \) for \( h \in H_n \) to find the CV bandwidth \( \hat{h}_{CV} \).

This algorithm is designed for the case where \( \pi \) is unknown. When \( \pi \) is known, it can be simplified in an obvious way. Note that \( \hat{\pi} \) and \( \hat{\sigma}_\pi \) are by-products of Algorithm CV. In Step 2, we use a delicate bandwidth in estimating \( \pi \) to achieve an optimal fitting in (15). But \( ch \) may not be a good choice in estimating \( \pi \). For example, suppose \( k(u) = 3(1 - u^2)/4 \), the Epanechnikov kernel, \(^{47}\) it can be shown that \( c = 1.554 \) which is greater than one. This is reasonable because we need more data in estimating \( m(\cdot) \) at the boundary to improve the efficiency. But for our purpose, we want to precisely estimate \( \pi \) rather than precisely estimate \( m(\cdot) \) around \( \pi \). Because \( \pi \) is identified by the jump of \( m(\cdot) \), small bias in estimating \( m(\cdot) \) is more important than small variance. So we suggest to use a smaller bandwidth than \( h \) in estimating \( \pi \). \(^{56}\) This is also a common sense in simulation studies of the literature.

Now, we describe how to determine \( H_n \), and minimize \( CV(h) \). Suppose a vector \( (x_{(1)}, \cdots, x_{(n)}) \) includes the sorted \( x_i \)'s in \( \Pi \). Define the lower bound of \( H_n \) as \( h \equiv \max \{ x_{(i+1)} - x_{(i)} : i = 1, \cdots, n-1 \} \). This \( h \) guarantees that there are at least 2 data points in a \( h \) neighborhood of any \( x \in \Pi \), so the multicollinearity problem in the LLS can be avoided. The upper bound \( \bar{h} \) of \( H_n \) can be set as \( \frac{1}{2} (\pi - \bar{\pi}) \), which roughly corresponds to the parametric formulation of \( m(\cdot) \) when \( \pi \) is in the middle of \( \Pi \). When minimizing \( CV(h) \), we need only search over a discretized subset \( D_h \) of \( H_n \); e.g., \( D_h = \left\{ h + i \cdot \text{step} : i = 0, 1, \cdots, \left\lfloor \frac{\bar{h} - h}{\text{step}} \right\rfloor \right\} \), where step = \( \frac{1}{2} \min \{ x_{(i+1)} - x_{(i)} : i = 1, \cdots, n-1 \} \). The step size in \( D_h \) ensures that at most one more data point is covered by the \( h + (i+1) \cdot \text{step} \) ball than the \( h + i \cdot \text{step} \) ball centered at any \( x \in \Pi \), so essentially all possible LLS estimates are covered for a fixed data set. If such a step size is too small, we may set step = \( \frac{1}{2} \frac{\pi - \bar{\pi}}{n} \). In this case, \( CV(h) \) is evaluated at roughly \( n \) points. Such a specification of \( H_n \) and \( D_h \) is not rigid. In practice, we must make sure the minimizer is not obtained at the boundary points, \( \hat{h} \) and \( \bar{h} \), of \( H_n \), and hope \( CV(h) \) is relatively flat near its minimizer so that the estimation of \( \alpha_\pi \) is relatively robust to the bandwidth selection. \(^{49}\)

Wu and Chu (1993b) use a different approach, which results in a smaller bandwidth in estimating \( \pi \), in step 2 because their estimating procedure assumes \( k(0) = 0 \) as mentioned in Section 4. A smaller bandwidth can make the uncertainty of \( \hat{\pi} \) in evaluating \( CV(h) \) disappear. Also, they generate a different pseudo-data set in Step 3. In their case, \( (\bar{x}_{i+}, \bar{y}_{i+}) = (2\bar{\pi} - x_{i-}, y_{i-} + 2\hat{m}_{\pi}(\hat{\pi})(\bar{\pi} - x_{i-}), i = 1, \cdots, L_h \), and \( (\bar{x}_{i-}, \bar{y}_{i-}) = (2\bar{\pi} - x_{i+}, y_{i+} + 2\hat{m}_{\pi}(\hat{\pi})(x_{i+} - \bar{\pi}), i = 1, \cdots, R_h \). Our procedure in Step 3 is based on Hall and Wehrly (1991). For comparison, we check the following Figure 5. From Figure 5, when reflecting the data to the right (left) neighborhood of \( \hat{\pi} \), Hall and Wehrly (1991) make the ray of light through the fixed point \( (\bar{\pi}, \hat{m}_{\pi}(\hat{\pi}))(\bar{\pi}, \hat{m}_{\pi}(\bar{\pi}))) \), while Wu and Chu (1993b) project the ray of light parallelly with the slope \( \hat{m}_{\pi}(\bar{\pi}) \). Intuitively, both methods maintain the linear component of \( m(\cdot) \) in the neighborhood of \( \hat{\pi} \), so the bias term will be \( O(h^2) \) which is the same order at the interior point. Because we only reflect the data in a \( h \) neighborhood of \( \hat{\pi} \), the resulting contribution to the MISE in \( CV(h) \) is \( O \left( h^3 \right) \). \(^{50}\) This contribution is negligible because the bias component of the MISE from other area in \( CV(h) \) is \( O \left( h^5 \right) \).

\(^{47}\) This kernel is optimal in minimizing MSE and MISE at interior points and is nearly optimal at the most boundary point. Cheng et al (1997) shows that the optimal kernel for the boundary point is the triangular or Bartlett kernel max \( \{ 0, (1 - |u|) \} \).

\(^{48}\) It is easy to see that smaller bandwidth in estimating \( \pi \) will not affect the asymptotic argument in Theorem 5.

\(^{49}\) See the figures in Lee and Lemieux (2010) for some intuitions on this fact. This is also why the CV bandwidth converges to the optimal value in a very slow rate \( n^{-1/10} \) as shown in Wu and Chu (1993b).

\(^{50}\) Without reflection or projection, the contribution would be \( O \left( h^2 h \right) = O \left( h^3 \right) \).
Figure 5: Comparison of Reflection in Hall and Wehrly (1991) and Projection in Wu and Chu (1993b): HW for Hall and Wehrly (1991) and WC for Wu and Chu (1993b), Red Dot for Real Data, and Blue Circle for Reflected or Projected Data.

rate Hall and Wehrly (1991)'s method over Wu and Chu (1993b)'s because levels $\hat{m}_{\pi} (\pi)$ can be estimated more precisely than slopes $\hat{m}'_{\pi} (\pi)$. 

Now, we give our bandwidth selection of $b$ (and $h$) in specification testing. The key difference between specification testing and estimation is that there may or may not be a cusp or discontinuity in the data, so we do not need step 2 and 3 in Algorithm CV. In step 4, we use the original data to estimate $\hat{m}_{-i} (x_i)$. Such an algorithm is to choose the best smooth fit of the original data, so matches the idea in the construction of $I_n$. If the DGP is from $H_0^{(1)}$, then balancing the integrated bias squared ($O(h^4)$) and variance ($O (\frac{1}{nh})$), we get $\hat{b}_{CV} = O (n^{-1/5})$, the most popular rate of optimal bandwidth. If the DGP is from $H_1^{(1)}$, then balancing the integrated bias squared ($O(h^3)$) and variance ($O (\frac{1}{nh})$), we get $\hat{b}_{CV} = O (n^{-1/4})$. At last, if the DGP is from $H_1^{(2)}$, then balancing the integrated bias squared ($O(h)$) and variance ($O (\frac{1}{nh})$), we get $\hat{b}_{CV} = O (n^{-1/2})$. So when $m(\cdot)$ gets rougher, our bandwidth gets smaller. But this is exactly what we want in Assumption B. Given $\hat{b}_{CV}$, $\hat{h}$ can be set as $\hat{b}^{CV}$. 

As mentioned above, we recommend to use a small bandwidth in estimating $\pi$ when $\alpha_\pi > 0$. The bandwidth selection method in specification testing can serve this purpose. In this case, the bandwidth is $O (n^{-1/2})$, much smaller than the optimal bandwidth (which is $O(n^{-1/5})$) in estimating $\alpha_\pi$. Actually, this choice is also suggested in Wu and Chu (1993b).

We do not suggest to use the bandwidth above sti$dly$. Instead, our procedure of bandwidth selection better serves as a method to strengthen and rigorize the intuition in a practical question. Before conducting any analysis discussed in this paper, we suggest to draw graphs of local moving averages for a range of bandwidths or globally polynomial fitting for different orders to get some senses about where $\pi$ is and whether there is selection or treatment effect. Even in a rigorous analysis, we suggest to check a range of bandwidths to explore the sensitivity of our estimator and testing procedure to the bandwidth selection. For example, in specification testing, we can draw a graph of decision (0 for accepting and 1 for rejecting) against bandwidth and check how the decision changes when the bandwidth is smaller or larger than $\hat{b}_{CV}$.
6 Monte Carlo Results

The goal of this Monte Carlo study is to check how our tests and estimators depend on the bandwidth selection. We will use the four setups of Figure 2 in our simulations.

\[ DGP1 : y = x^2 (1 - 2 \leq x \leq 3) + \varepsilon, \]
\[ DGP2 : y = x^2 (1 - 2 \leq x < 1) + [(x - 3)^2 - 3] 1 (1 \leq x \leq 3) + \varepsilon, \]
\[ DGP3 : y = x^2 (1 - 2 \leq x < 1) + (x^2 + 1) 1 (1 \leq x \leq 3) + \varepsilon, \]
\[ DGP4 : y = x^2 (1 - 2 \leq x < 1) + [(x - 3)^2 - 3] 1 (1 \leq x \leq 3) + \varepsilon. \]

In all setups, \( \varepsilon \)'s are i.i.d. sampled and follow \( N(0, 0.2^2) \) and \( x \) is uniformly distributed on \([-2, 3] \). \( \Pi = [0.5, 1.5] \), and the number of discontinuity points is known as 1. In specification testing, we will study size under the null and also power under the alternative. The rejection probability under \( DGP1 \) is the size for testing both (2) and (3). The rejection probability under \( DGP2 \) is the power in testing (2) and the size in testing (3). The rejection probabilities under \( DGP3 \) and \( DGP4 \) are the power in testing (3). The significance level is set at 5%. In estimation, we will only consider \( DGP3 \) and \( DGP4 \) and study the bias and variance properties of the estimators of \( \pi \) and \( \alpha \).

Because we will check the performance of our procedures for each bandwidth in a reasonable range, the computation burden is tremendous. To save simulation time, we let \( n = 500 \) which guarantees that about 100 data points fall in \( \Pi \). The number of replications is set as 100. In the bootstrap method in Section 3.2, \( B \) is set as 199. Even for this small simulation study, many of the results in the main text are confirmed. Throughout the simulations, all estimators are based on the LLS with the kernel function \( k(u) = 3 (1 - u^2) / 4 \).

6.1 Specification Testing

In this simulation, \( h \) is set as \( b^{2.1} \). The construction of the test statistic and the bootstrap method to find the critical values can be found in Section 3.1 and 3.2. The simulation results for our specification testing are summarized in the following Figure 3. A few results from Figure 3 are as follows. (i) For the same bandwidth, when \( m(\cdot) \) gets rougher, the probability of rejection gets higher. This matches the construction of \( T_n \), because the bias will increase when \( m(\cdot) \) gets rough. Similarly, given a DGP, the larger the bandwidth, the higher the probability of rejection. (ii) In both tests, the bootstrap method has a better finite-sample approximation to the distribution of \( T_n \) than the asymptotic method under the null. In testing (2), the bootstrap method has a almost correct size for a wide range of bandwidths, while the asymptotic method is either over-sized or under-sized so that it is hard to pick the right bandwidth. In consideration of the power properties, it seems that \([0.25, 0.35]\) is a suitable range for \( b \). In testing (3), similarly, the size property of the asymptotic method is not satisfying. For the bootstrap method, \([0.1, 0.15]\) seems to be a suitable range of \( b \) because under the null which covers both \( DGP1 \) and \( DGP2 \), the sizes match the target 5% for \( b \) in this range. These ranges of bandwidths also justify Assumption B; that is, a smaller bandwidth should be used in testing (3) than in testing (2). Third, the bootstrap method has a better power in testing (3) than the asymptotic method for \( b \in [0.1, 0.15] \). The powers under \( DGP3 \) and \( DGP4 \) are similar. This is because the power is mostly provided by the jump instead of the cusp in this case. The lesson here is that the

\[ \text{The variance of } \varepsilon \text{ is set as } 0.2^2 \text{ to make the simulation nontrivial. If the variance is too large, such as 1, then the null and alternative cannot be distinguished visually.} \]
\[ \text{It is hard to compare the power property in testing (2) because the associated size property is not comparable as mentioned in the last point.} \]
Figure 6: Probability of Rejection Based on $T_n$ Under Different DGPs

The bootstrap method is indeed preferable than the asymptotic method in both size and power properties and the robustness to the bandwidth selection in specification testing.

6.2 Estimation

Similar as in the specification testing, we check how our estimators of $\pi$ and $\alpha_\pi$ depend on the bandwidth selection. Meanwhile, we illustrate the efficiency loss in estimating $\alpha_\pi$ with an unknown $\pi$. Our simulation results are summarized in Figure [7] and Figure [8] for DGP3 and DGP4, respectively. The optimal $h$ for $\pi$ selected in the two figures are based on the RMSE risk of $\tilde{\pi}$.

Some common structures in these two figures are as follows. (i) The risk of $\tilde{\pi}$ is much smaller than $\hat{\alpha}_\pi$, which confirms the superconsistency of $\tilde{\pi}$. (ii) When the bias of $\tilde{\pi}$ is small, the risk of $\hat{\alpha}_\pi$ is close to the case where $\pi$ is known, which confirms the result that $\tilde{\pi}$ will not affect the efficiency of $\hat{\alpha}_\pi$. Especially, when the optimal bandwidth in estimating $\pi$ is used, the risk of $\hat{\alpha}_\pi$ is almost not affected by the estimating of $\pi$. But when the bias of $\tilde{\pi}$ is large, there are indeed considerable efficiency losses in $\hat{\alpha}_\pi$ compared with the case with $\pi$ known. (iii) The bandwidth suitable for $\tilde{\pi}$ should be much smaller than that for $\hat{\alpha}_\pi$. Due to the smoothness of $m(\cdot)$ in our specification, the risk of $\hat{\alpha}_\pi$ is decreasing even for $h = 1$.

There are indeed some differences between Figure [7] and Figure [8]. Under DGP3, the slope and curvature of $m(\cdot)$ in the left and right neighborhoods of $\pi$ are the same, so even we use the same bandwidth to estimate $\pi$ and $\alpha_\pi$, there is no much efficiency loss in estimating $\alpha_\pi$. While under DGP4, where the structures of $m(\cdot)$ are very different in the left and right neighborhoods of $\pi$, choosing a different bandwidth in estimating $\alpha_\pi$ is very important to get good performance of $\hat{\alpha}_\pi$. The lesson here is that when we suspect there is selection, bandwidth selection should be very careful in estimating $\pi$ and $\alpha_\pi$. 
Figure 7: Comparison of Estimators of $\pi$ and $\alpha_\pi$ in Bias and RMSE in $DGP3$

Figure 8: Comparison of Estimators of $\pi$ and $\alpha_\pi$ in Bias and RMSE in $DGP4$
7 Conclusion and Future Research

This paper is designed friendly for practitioners, and many theoretical problems are solved. Nevertheless, there are still many theoretical and practical questions left behind, some of which are already mentioned in the main text. We summarize these questions here for future research. First, the Edgeworth expansion of the wild bootstrap test statistic is attractable for a theoretical justification of our testing procedure. Second, when the smoothness of the conditional mean of the response variable given the forcing variable is unknown, how to adapt our procedures to the unknown smoothness index? Third, consistency is only a basic requirement for a test; a more challenging problem is to find the minimax optimal test in our testing environment. Relevant literature in the classical specification test includes Ingster (1993) and Guerre and Lavergne (2002). Fourth, as mentioned in the introduction of Outshoorn (1998), the limit experiment of the model we considered should be the ideal Gaussian white noise model with discontinuities. A rigorous development of this result is theoretically intriguing. The counterpart in the nonparametric regression without discontinuities can be found in Brown and Low (1996). Fifth, we test whether there is selection after excluding the possibility of treatment effect. If we want to test the presence of selection no matter there is treatment effect or not, we must exclude the influence of treatment effect (if it is present) on the test statistic. Construction of such a test is challenging. Sixth, our asymptotic arguments in both testing and estimation are based on a fixed sequence of bandwidths; a rigorous analysis on the asymptotic behaviors based on the data-driven bandwidth such as the CV bandwidth in Section 5.4 is preferable. Seventh, more simulation studies should be carried out to provide more practical suggestions on the bandwidth selection in both specification testing and estimation.
References


Hansen, B.E., 1996, Inference when a Nuisance Parameter is not Identified under the Null Hypothesis, *Econometrica*, 413-430.


Yu, P., 2008a, The Bootstrap in Threshold Regression, unpublished manuscript, Department of Economics, University of Auckland.

Yu, P., 2008b, Adaptive Estimation of the Threshold Point in Threshold Regression, unpublished manuscript, Department of Economics, University of Auckland.