Weak Identification in Fuzzy Regression
Discontinuity Designs*

Vadim Marmer†  Donna Feir†  Thomas Lemieux†

November 3, 2012

Abstract

In fuzzy regression discontinuity (FRD) designs, the treatment effect is identified through a discontinuity in the conditional probability of treatment assignment. As in a standard instrumental variables setting, we show that when identification is weak (i.e. when the discontinuity is of a small magnitude) the usual t-test based on the FRD estimator and its standard error suffers from asymptotic size distortions. This finite-sample problem can be especially severe in the FRD setting since only observations close to the discontinuity are useful for estimating the treatment effect. To eliminate those size distortions, we

*We thank Moshe Buchinsky, Karim Chalak, Jae-Young Kim, Sokbae Lee, Arthur Lewbel, Taisuke Otsu, Eric Renault, Yoon-Jae Whang, participants of the 2011 North American Winter and European Meetings of the Econometric Society, Conference in Honor of Halbert White, and econometrics seminar participants at Seoul National University, UC Davis, UC San Diego, and University of Montreal for helpful comments. Vadim Marmer gratefully acknowledges the financial support of the SSHRC under grant 410-2010-1394.
†Department of Economics, University of British Columbia, 997 - 1873 East Mall, Vancouver, BC, V6T 1Z1, Canada. E-mails: vadim.marmer@ubc.ca (Marmer), dfeir@interchange.ubc.ca (Feir), and thomas.lemieux@ubc.ca (Lemieux).
propose a modified \( t \)-statistic that uses a null-restricted version of the standard error of the FRD estimator. Simple and asymptotically valid confidence sets for the treatment effect can be also constructed using the FRD estimator and its null-restricted standard error. An extension to testing for constancy of the regression discontinuity effect across covariates is also discussed.

**JEL Classification:** C12; C13; C14

**Keywords:** Nonparametric inference; treatment effect; size distortions; Anderson-Rubin test; robust confidence set; class size effect

1 Introduction

In this paper, we discuss the problem of weak identification in the context of the fuzzy regression discontinuity (FRD) design. The regression discontinuity (RD) design has been studied recently by Hahn, Todd, and Van der Klaauw (2001) and Imbens and Lemieux (2008). The RD framework is concerned with evaluating the effects of interventions or treatments when assignment to treatment is determined completely or partly by the value of an observable assignment variable. In this framework, identification of the treatment effect comes from a discontinuity in the conditional probability of treatment assignment at some known cutoff value of the assignment variable. When assignment to the treatment is completely determined by the value of the assignment variable, the RD design is called sharp. When assignment to the treatment is only partly determined by the assignment variable, the RD design is called fuzzy. The later is the focus of this paper.

Hahn, Todd, and Van der Klaauw (2001) show there is a close parallel between the FRD design and an instrumental variable setting.\(^1\) However, since only the obs-

\(^1\)The FRD estimate of the treatment effect can be interpreted as an instrumental variable estimate, where the instrument for the treatment variable is a dummy variable indicating whether the
vations close to the cutoff point are useful for estimating the size of the discontinuity, the effective sample size available in the FRD design can be quite small even when the full sample is large. Thus, it is particularly important to study the finite sample properties of the estimated treatment effect, especially when identification is weak.

Weak identification in FRD corresponds to the situation where the discontinuity in the conditional probability function of treatment assignment is of a small magnitude. Similar to the weak instruments literature (see, for example, Andrews and Stock (2007) for a review), weak identification can be formally modeled using the local-to-zero framework. Specifically, we assume that the discontinuity in the conditional probability function of treatment assignment is local-to-zero.

When identification is weak, we show that the usual $t$-test based on the FRD estimator and its standard error suffers from asymptotic size distortions with an exception to a few specific situations. For example, one can still use the usual $t$-statistic when testing the hypothesis of zero treatment effect if the assignment to treatment and the outcome variables are asymptotically independent. However, in general the usual $t$-test is asymptotically invalid because it can over reject the null hypothesis when identification is weak. The usual confidence intervals constructed as estimate $\pm$ constant $\times$ standard error are also invalid because their asymptotic coverage probability can be below the assumed nominal coverage when identification is weak.

In this paper, we suggest a simple modification to the $t$-test that eliminates the asymptotic size distortions caused by weak identification. Unlike the usual $t$-statistic, the proposed modified $t$-statistic uses a null-restricted version of the standard error of the FRD estimator. Tests based on the $t$-statistic computed using the null-restricted standard errors do not suffer from asymptotic size distortions when identification assignment variable exceeds the cutoff point.
is weak and are asymptotically equivalent to the usual $t$-test when identification is strong.

Asymptotically valid confidence sets for the treatment effect can be obtained by inverting the test based on the $t$-statistic with the null-restricted standard error. Since the FRD is an exactly identified model, these confidence sets are easy to compute as their construction only involves solving a quadratic equation.\textsuperscript{2} These confidence sets are expected to be as informative as the standard ones, when identification is strong. However, unlike the usual confidence intervals constructed as estimate $\pm$ constant $\times$ standard error, the confidence sets we propose can be unbounded with positive probability. This property is expected from valid confidence sets in the situations with local identification failure and an unbounded parameter space (see Dufour (1997)).

In a recent paper, Otsu and Xu (2011), propose empirical likelihood based confidence sets for the RD effect. Their method does not involve variance estimation and for that reason is expected to be robust to weak identification. However, it requires computation of the empirical likelihood function numerically and is computationally more demanding than our approach. That being said, the empirical likelihood based confidence sets are expected to have better higher-order coverage properties.

We also discuss testing whether the RD effect is homogeneous over differing values of some covariates. The proposed testing approach is designed to remain asymptotically valid when identification is weak. This is achieved by building a robust confidence set for a common RD effect across covariates. The null hypothesis of the

\textsuperscript{2}Most of the literature on weak instruments deals with the case of over identified models (see, e.g., Andrews and Stock (2007)). In exactly identified models, the approach suggested by Anderson and Rubin (1949) results in efficient inference if instruments turn out to be strong and remains valid if instruments are weak. However, in over identified models, Anderson and Rubin’s tests are no longer efficient even when instruments are strong. Several papers (Kleibergen, 2002; Moreira, 2003; Andrews, Moreira, and Stock, 2006) proposed modifications to Anderson and Rubin’s basic procedure to gain back efficiency in over identified models. Since the FRD design is an exactly identified model, we can adapt Anderson and Rubin’s approach without any loss of power.
common RD effect is rejected when that confidence set is empty.

To demonstrate the empirical relevance of weak identification in fuzzy RD designs, we compare the results of both of these proposed robust tests to the standard ones in two separate applications for Israel (Angrist and Lavy (1999)) and Chile (Urquiola and Verhoogen (2009)). In both cases, we use the RD design to estimate the effect of class size on student achievement. The existence of caps in class size (40 in Israel, 45 in Chile) provides a discontinuity in the relationship between the number of students enrolled in the school (the assignment variable) and average class size (the treatment variable). In both cases, we have a FRD design because the caps are enforced imperfectly and can result in various class sizes. We revisit the Angrist and Lavy study by treating it explicitly as a FRD design (they used an instrumental variables approach instead). We show that weak identification is not an issue at the large discontinuity at the 40 students cutoff since the confidence sets obtained using our robust method are very close to those obtained using the standard method. We also use our proposed test for the homogeneity of the RD effect by comparing secular and religious schools, and schools with an above- and below-median fraction of disadvantaged students.

In the case of Chile, Urquiola and Verhoogen (2009) show that the discontinuity in class size gets progressively weaker at higher multiples of the 45 students cap. As weak identification becomes more of a problem, we find that the confidence sets obtained using standard methods and our robust procedure become more divergent. Interestingly, in a number of cases the robust confidence sets provides more informative answers than the standard method. More generally, the empirical applications, along with a Monte Carlo experiment, suggests that our simple and robust procedure for computing confidence sets performs well when identification is either strong or weak.

The rest of the paper proceeds as follows. In Section 2 we describe the FRD
model and present our analytical results. Section 3 discusses testing for constancy of the RD effect across covariates. In Section 4, we illustrate our results in a Monte Carlo experiment. We present our empirical applications in Section 5 and conclude in Section 6.

2 Theoretical results

2.1 Preliminaries

In the RD design, the observed outcome variable $y_i$ is written as

$$y_i = y_{0i} + x_i\beta_i,$$

where $x_i$ is the treatment indicator variable that takes on value one if the treatment is received and zero otherwise, $y_{0i}$ is the outcome without treatment, and $\beta_i$ is the random treatment effect for observation $i$. The treatment assignment depends on another observable assignment variable, $z_i$:

$$\Pr(x_i = 1|z_i = z) = E(x_i|z_i = z).$$

The main feature in this framework is that $E(x_i|z_i = z)$ is discontinuous at some known cutoff point $z_0$, while $E(y_{0i}|z_i)$ is assumed to be continuous at $z_0$.

**Assumption 1.** (a) $\lim_{z \downarrow z_0} E(x_i|z_i = z) \neq \lim_{z \uparrow z_0} E(x_i|z_i = z)$.

(b) $\lim_{z \downarrow z_0} E(y_{0i}|z_i = z) = \lim_{z \uparrow z_0} E(y_{0i}|z_i = z)$.

The RD design is called sharp if $|\lim_{z \uparrow z_0} E(x_i|z_i = z) - \lim_{z \downarrow z_0} E(x_i|z_i = z)| = 1$. In this case, the treatment assignment is completely determined by the value of $z_i$. 
The FRD design corresponds to the situation where

$$\left| \lim_{z \uparrow z_0} E (x_i | z_i = z) - \lim_{z \downarrow z_0} E (x_i | z_i = z) \right| < 1,$$

so either $0 < \lim_{z \uparrow z_0} E (x_i | z_i = z) < 1$ or $0 < \lim_{z \downarrow z_0} E (x_i | z_i = z) < 1$ or both, and therefore the treatment assignment is not a deterministic function of $z_i$.

The main object of interest is the RD effect

$$\beta = \frac{y^+ - y^-}{x^+ - x^-}, \quad (1)$$

where

$$y^+ = \lim_{z \downarrow z_0} E (y_i | z_i = z), \quad x^+ = \lim_{z \downarrow z_0} E (x_i | z_i = z),$$

$$y^- = \lim_{z \uparrow z_0} E (y_i | z_i = z), \quad x^- = \lim_{z \uparrow z_0} E (x_i | z_i = z). \quad (2)$$

The exact interpretation of $\beta$ depends on the assumptions that the econometrician is willing to make in addition to Assumption 1. As discussed in Hahn, Todd, and Van der Klaauw (2001), if $\beta_i$ and $x_i$ are assumed to be independent conditional on $z_i$, then $\beta$ captures the average treatment effect (ATE) at $z_i = z_0$: $\beta = E (\beta_i | z_i = z_0)$. This also covers a special case where the treatment effect is a deterministic function of $z_i$ in the neighborhood of $z_0$: $\beta_i = \beta (z_i)$. In this case, $\beta = \beta (z_0)$ and it is referred to in Hahn, Todd, and Van der Klaauw (2001) as a constant treatment effect.

Hahn, Todd, and Van der Klaauw (2001) show that another interpretation for $\beta$ can be obtained if one assumes that in the neighborhood of $z_0$ and with probability one, $x_i$ is a non-decreasing or non-increasing function of $z_i$, and $E (x_i \beta_i | z_i = z)$ is constant in the neighborhood of $z_0$. In this case, $\beta$ captures the local ATE or the
ATE for compliers, where compliers are observations $i$ for which $x_i$ switches its value from zero to one (or from one to zero) when $z_i$ changes from $z_0 - e$ to $z_0 + e$ for some small $e > 0$.

Regardless of its interpretation, it is now standard to estimate $\beta$ using the local linear approach. Define

$$
\left( \hat{y}^+, \hat{b}_y^+ \right) = \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - (z_i - z_0) b)^2 I_i^+ K \left( \frac{z_i - z_0}{h_n} \right), \tag{3}
$$

$$
\left( \hat{y}^-, \hat{b}_y^- \right) = \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - (z_i - z_0) b)^2 I_i^- K \left( \frac{z_i - z_0}{h_n} \right), \tag{4}
$$

$$
\left( \hat{x}^+, \hat{b}_x^+ \right) = \arg\min_{a,b} \sum_{i=1}^{n} (x_i - a - (z_i - z_0) b)^2 I_i^+ K \left( \frac{z_i - z_0}{h_n} \right), \tag{5}
$$

$$
\left( \hat{x}^-, \hat{b}_x^- \right) = \arg\min_{a,b} \sum_{i=1}^{n} (x_i - a - (z_i - z_0) b)^2 I_i^- K \left( \frac{z_i - z_0}{h_n} \right), \tag{6}
$$

where $K$ is the kernel function, $h_n$ is the bandwidth, and the indicator functions $I_i^-$ and $I_i^+$ are defined as

$$
I_i^- = 1 \{ z_i < z_0 \},
$$

$$
I_i^+ = 1 - 1 \{ z_i < z_0 \}.
$$

The local linear estimator of $\beta$ is given by

$$
\hat{\beta} = \frac{\hat{y}^+ - \hat{y}^-}{\hat{x}^+ - \hat{x}^-}.
$$

To describe its asymptotic behavior, consider the following high-level assumption.

**Assumption 2.** (a) The PDF of $z_i$ is continuous and bounded in the neighborhood of $z_0$; it is also bounded away from zero in the neighborhood of $z_0$. 

8
(b) The data \( \{(y_i, x_i, z_i)\}_{i=1}^n \), kernel function \( K \), and bandwidth \( h_n \) are such that

\[
\sqrt{n h_n} \begin{pmatrix}
\hat{y}^+ - y^+ \\
\hat{x}^+ - x^+ \\
\hat{y}^- - y^- \\
\hat{x}^- - x^-
\end{pmatrix} \rightarrow_d \begin{pmatrix}
Y^+ \\
X^+ \\
Y^- \\
X^-
\end{pmatrix} = d N\left(\begin{pmatrix}
0 \\
0 \\
0 \\
0
\end{pmatrix}, \begin{pmatrix}
\sigma_{yy}^+ & \sigma_{yx}^+ & 0 & 0 \\
\sigma_{yx}^+ & \sigma_{xx}^+ & 0 & 0 \\
0 & 0 & \sigma_{yy}^- & \sigma_{yx}^- \\
0 & 0 & \sigma_{yx}^- & \sigma_{xx}^-
\end{pmatrix}\right),
\]

where the asymptotic variance-covariance matrix is positive definite.

(c) There exist \( \hat{\sigma}_{fg}^s \), \( s \in \{+, -\} \) and \( f, g \in \{x, y\} \), such that \( \hat{\sigma}_{fg}^s \rightarrow_p \sigma_{fg}^s \) for all \( s \) and \( f, g \).

Remark. As discussed in Hahn, Todd, and Van der Klaauw (2001, pages 207-208) and Imbens and Lemieux (2008, page 630), Assumption 2 is satisfied when, for example, the data \( \{(y_i, x_i, z_i)\}_{i=1}^n \) are iid, \( K \) is a continuous, symmetric around zero, non-negative, and compactly supported second-order kernel function, \( h_n = cn^{-\delta} \) with \( c > 0 \) and \( 1/5 < \delta < 1 \), and provided that some additional technical conditions hold.\(^3\)

In this case,

\[
\begin{align*}
\sigma_{yy}^+ = \frac{k^+}{f_z(z_0)} \lim_{z_i \uparrow z_0} Var(y_i|z_i = z_0), & \quad \sigma_{yy}^- = \frac{k^-}{f_z(z_0)} \lim_{z_i \downarrow z_0} Var(y_i|z_i = z_0), \\
\sigma_{xx}^+ = \frac{k^+}{f_z(z_0)} \lim_{z_i \uparrow z_0} Var(x_i|z_i = z_0), & \quad \sigma_{xx}^- = \frac{k^-}{f_z(z_0)} \lim_{z_i \downarrow z_0} Var(x_i|z_i = z_0), \\
\sigma_{yx}^+ = \frac{k^+}{f_z(z_0)} \lim_{z_i \uparrow z_0} Cov(y_i, x_i|z_i = z_0), & \quad \sigma_{yx}^- = \frac{k^-}{f_z(z_0)} \lim_{z_i \downarrow z_0} Cov(y_i, x_i|z_i = z_0),
\end{align*}
\]

\(^3\) The requirement \( \delta > 1/5 \) corresponds to data under smoothing and is needed to eliminate the asymptotic bias of the kernel estimators.
where \( f_z(z_0) \) denotes the PDF of \( z_i \) at \( z = z_0 \),

\[
k^+ = \frac{\int_0^\infty \left( \int_0^\infty s^2 K(s) \, ds - u \int_0^\infty s K(s) \, ds \right)^2 K^2(u) \, du}{\left( \int_0^\infty u^2 K(u) \, du \int_0^\infty K(u) \, du - \left( \int_0^\infty u K(u) \, du \right)^2 \right)^2},
\]

and \( k^- \) is defined similarly to \( k^+ \) but with the integrals over \((-\infty, 0)\), see Theorem 4 in Hahn, Todd, and Van der Klaauw (2001).

The asymptotic variance \( \sigma_{yy}^+ \) can be consistently estimated by

\[
\hat{\sigma}_{yy}^+ = \frac{k^+}{\hat{f}_z^2(z_0)} \frac{1}{nh_n} \sum_{i=1}^n (y_i - \hat{y}^+)^2 I_i^+ K \left( \frac{z_i - z_0}{h} \right),
\]

where \( \hat{f}_z(z_0) \) is the kernel estimator of \( f_z(z_0) \): \( \hat{f}_z(z_0) = (nh_n)^{-1} \sum_{i=1}^n K((z_i - z_0)/h_n) \).

Consistent estimators of \( \sigma_{xx}^+, \sigma_{yx}^+, \sigma_{yy}, \sigma_{xx}^-, \sigma_{yx}^- \) can be constructed similarly.

When Assumption 2 holds and \( p \lim_{n \to \infty} (\hat{x}^+ - \hat{x}^-) \neq 0 \), by a standard application of the delta-method, the asymptotic distribution of the FRD estimator of \( \beta \) is given by

\[
\sqrt{nh_n} \left( \hat{\beta} - \beta \right) \to_d N (0, V(\beta)), \tag{7}
\]

where

\[
V(\beta) = \frac{\sigma^2(\beta)}{(\hat{x}^+ - \hat{x}^-)^2}, \tag{8}
\]

\[
\sigma^2(\beta) = \sigma_{yy}^+ + \sigma_{yy}^- + \beta^2 \left( \sigma_{xx}^+ + \sigma_{xx}^- \right) - 2\beta \left( \sigma_{yx}^+ + \sigma_{yx}^- \right). \tag{9}
\]

The asymptotic variance \( V_\hat{\beta}(\beta) \) can be consistently estimated by the plug-in method with

\[
\hat{V}(\hat{\beta}) = \frac{\hat{\sigma}^2(\hat{\beta})}{(\hat{x}^+ - \hat{x}^-)^2}, \tag{10}
\]
\[
\hat{\sigma}^2(\hat{\beta}) = \hat{\sigma}_{yy}^+ + \hat{\sigma}_{yy}^- + \hat{\beta}^2 \left( \hat{\sigma}_{xx}^+ + \hat{\sigma}_{xx}^- \right) - 2\hat{\beta} \left( \hat{\sigma}_{yx}^+ + \hat{\sigma}_{yx}^- \right).
\]

(11)

A test of \( H_0 : \beta = \beta_0 \) in practice is usually based on the \( t \)-statistic

\[
T(\beta_0) = \frac{\hat{\beta} - \beta_0}{\sqrt{V(\hat{\beta})}/(nh_n)}.
\]

and one rejects \( H_0 \) when \( T(\beta_0) \) exceeds a standard normal critical value.

### 2.2 Weak identification in FRD

The FRD effect \( \beta \) is not defined if Assumption 1(a) fails and \( x^+ - x^- = 0 \). Here we consider the situation where \( \beta \) is well-defined however only weakly identified. The issue of weak identification arises in the FRD model when the discontinuity in the conditional probability function of receiving the treatment is small. Similarly to the case of weak instruments in the IV regression model, this creates the problem of a nearly zero denominator in (1) and (8). The consequence of weak identification is that the asymptotic result in (7) provides a poor approximation to the actual behavior of the estimator in finite samples.

A useful device for analyzing the properties of estimators in the case of weak identification is local-to-zero asymptotics. We make the following assumption.

**Assumption 3 (Weak ID).** We assume that the bandwidth \( h \) in Assumption 2 is such that \( x^+ - x^- = \pi_n = \theta/\sqrt{nh_n} \) for some constant \( \theta \).

**Remarks.** (a) The weak ID condition assumes that the discontinuity in the function \( E(x_i|z_i = z) \) at \( z_0 \) is small. In the case of weak instruments, one usually assumes that in the first-stage equation, the coefficient of the IVs are local-to-zero: \( \theta/\sqrt{n} \). Such an assumption results in a non-trivial asymptotic distribution for the IV estimator.
In our case, since the effective sample size is $nh_n$ and due to nonparametric rates of convergence of the estimators, one has to consider the sequence $\pi_n$ that converges to zero at $1/\sqrt{nh_n}$ rate.

(b) Note that the bandwidth $h_n$ in the Weak ID condition is the bandwidth chosen by the econometrician for the estimation of $x^+, x^-, y^+, y^-$. Thus, formally the assumption Weak ID states that the model depends on the sample size and the choice of the bandwidth. Intuitively, the assumption implies that the econometrician cannot achieve identification by simply estimating $x^+, x^-, y^+, y^-$ using a different bandwidth, say $h_n^*/h_n \to \infty$, i.e. we assume that the weak identification problem persists regardless of the bandwidth choice.

**Theorem 1.** Under Assumptions 2 and 3, the following results hold jointly:

(a) $\hat{\beta} - \beta \to_d \xi_{\beta, \theta}$, where

$$\xi_{\beta, \theta} = \frac{Y^+ - Y^- - \beta (X^+ - X^-)}{(X^+ - X^-) + \theta}.$$

(b) $(nh_n)^{-1} \hat{V}(\hat{\beta}) \to_d \sigma^2 (\beta + \xi_{\beta, \theta}) / (X^+ - X^- + \theta)^2$, where the function $\sigma^2(\cdot)$ is defined in (9).

(c) Under $H_0: \beta = \beta_0$, $|T(\beta_0)| \to_d |Z_\beta| (\sigma(\beta) / \sigma(\beta + \xi_{\beta, \theta}))$, where

$$Z_\beta = \frac{Y^+ - Y^- - \beta (X^+ - X^-)}{\sigma(\beta)} \sim N(0, 1).$$

**Remarks.** (a) Part (a) of the theorem shows that, due to weak identification, the FRD estimator $\hat{\beta}$ is inconsistent.

(b) According to part (b) of the theorem, the estimator of the asymptotic variance of FRD $\hat{V}(\hat{\beta})$ diverges at rate $nh_n$ due to the presence of $(\hat{x}^+ - \hat{x}^-)^2$ in the
denominator. However the standard error $\sqrt{\hat{V}(\hat{\beta})/(nh_n)}$ is stochastically bounded in large samples.

(c) The asymptotic distribution of the $t$-statistic in part (c) is nonstandard. Although the marginal distribution of $Z_\beta$ is standard normal, the random variables $Z_\beta$ and $\sigma (\beta + \xi_{\beta,\theta})$ are not independent.

Consider a test of $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ with the nominal asymptotic size $\alpha$ based on the usual $t$-statistic. The econometrician rejects $H_0$ when $|T(\beta_0)| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the standard normal distribution. The true asymptotic size is given by

$$\lim_{n \to \infty} P \left( |T(\beta_0)| > z_{1-\alpha/2} \mid \beta = \beta_0 \right) = P \left( |Z_\beta| \frac{\sigma(\beta)}{\sigma (\beta + \xi_{\beta,\theta})} > z_{1-\alpha/2} \mid \beta = \beta_0 \right).$$

If $\sigma(\beta)/\sigma (\beta + \xi_{\beta,\theta}) \leq 1$ with probability one, the true asymptotic size is less or equal to $\alpha$. In such a case, the test based on $T(\beta_0)$ is asymptotically valid. For example, suppose that $\beta = 0$ and the assignment into treatment is asymptotically independent from the outcome variable, i.e. $\sigma_{yx}^+ = \sigma_{yx}^- = 0$. Since

$$\frac{\sigma^2(\beta)}{\sigma^2 (\beta + \xi_{\beta,\theta})} = \left( 1 + \left( \xi_{\beta,\theta}^2 + 2\beta \xi_{\beta,\theta} \right) \frac{\sigma_{xx}^+ + \sigma_{xx}^-}{\sigma^2 (\beta)} - 2\xi_{\beta,\theta} \frac{\sigma_{yx}^+ + \sigma_{yx}^-}{\sigma^2 (\beta)} \right)^{-1},$$

we obtain that in this case, $\sigma(\beta)/\sigma (\beta + \xi_{\beta,\theta}) \leq 1$ with probability one, and consequently the test based on $T(\beta_0)$ is conservative: $\lim_{n \to \infty} P \left( |T(\beta_0)| > z_{1-\alpha/2} \mid \beta = \beta_0 \right) \leq \alpha$.

If on the other hand $\sigma(\beta)/\sigma (\beta + \xi_{\beta,\theta}) > 1$ with high probability, one can expect asymptotic size distortions, i.e. $\lim_{n \to \infty} P \left( |T(\beta_0)| > z_{1-\alpha/2} \mid \beta = \beta_0 \right) > \alpha$, and that the usual confidence intervals constructed as $\hat{\beta} \pm z_{1-\alpha/2} \times \sqrt{\hat{V}(\hat{\beta})/(nh_n)}$ will have the asymptotic coverage probability less than their nominal coverage $1 - \alpha$. For example,
asymptotic size distortions are more likely to occur when the selection into treatment variable, $x_i$, and the outcome variable, $y_i$, are highly correlated. Note that for the IV regression model, substantial size distortions are reported when the instruments are weak and the correlation between endogenous regressors and errors is high (Staiger and Stock, 1997, page 577).

2.3 Weak identification robust inference for FRD

As it is apparent from Theorem 1, the failure of the standard $t$-test when identification is weak is due to the asymptotic behavior of $\hat{V}(\hat{\beta})$ which depends on the inconsistent estimator $\hat{\beta}$. Inconsistency of $\hat{\beta}$ leads to the appearance of the $\sigma(\beta)/\sigma(\beta + \xi_\beta, \theta)$ random factor in the asymptotic distribution of the $t$-statistic $T$.

Instead of $\hat{V}(\hat{\beta})$, we suggest using a null-restricted version of the estimator of the asymptotic variance. When testing $H_0 : \beta = \beta_0$, the null-restricted estimator of the asymptotic variance is given by $\hat{V}(\beta_0) = \hat{\sigma}^2(\beta_0)/(\hat{x}^+ - \hat{x}^-)^2$, where $\hat{\sigma}^2(\cdot)$ is defined in (11). Next, we consider a null-restricted version of the $t$-statistic based on $\hat{V}(\beta_0)$:

$$\tilde{T}(\beta_0) = \frac{\hat{\beta} - \beta_0}{\sqrt{\hat{V}(\beta_0) / (nh_n)}}.$$ 

**Theorem 2.** Let $Z \sim N(0, 1)$. Under Assumptions 2 and 3, and for a fixed constant $\delta = \beta - \beta_0$, $|\tilde{T}(\beta_0)| \rightarrow_d |Z + \frac{\theta \delta}{\sigma(\beta_0)}|$.

**Remarks.** (a) The $t$-statistic with a null-restricted variance estimator has a standard normal asymptotic distribution under $H_0 : \beta = \beta_0$. For fixed alternatives $\beta = \beta_0 + \delta$, the asymptotic distribution of $\tilde{T}(\hat{\beta}_0)$ is noncentral, and one can expect nontrivial power against such alternatives. As usual in the case of weak identification, there is

---

$^4$\(\tilde{T}(\beta_0)\) is the Anderson-Rubin statistic in our framework (Anderson and Rubin, 1949).
no power against local alternatives $\beta = \beta_0 + \delta / \sqrt{nh_n}$ since $\tilde{T}(\beta_0) \rightarrow_d Z$ for all values of $\delta$. Power of the test depends on the strength of identification $\theta$ and the distance from the null $\delta$.

(b) Thus, when identification is strong, a test based on $\tilde{T}(\beta_0)$ has nontrivial power against local alternatives.

While the usual $t$-test can have size distortions when identification is weak, a test based on the $t$-statistic with the null-restricted standard error remains asymptotically valid when identification is either weak or strong. The test consists of rejecting $H_0$ whenever $|\tilde{T}(\beta_0)| > z_{1-\alpha/2}$.

One can construct a confidence set for $\beta$ with asymptotic coverage probability $1 - \alpha$ by collecting the values $\beta_0$ that cannot be rejected by the $\tilde{T}(\beta_0)$ test:

$$CS_{1-\alpha} = \{\beta_0 \in \mathbb{R} : |\tilde{T}(\beta_0)| \leq z_{1-\alpha/2}\}. \quad (12)$$

The confidence set $CS_{1-\alpha}$ can be easily computed analytically by solving for the values of $\beta_0$ that satisfy the inequality

$$nh_n(\hat{\beta} - \beta_0)^2(\hat{x}^+ - \hat{x}^-)^2 - z_{1-\alpha}^2(\hat{\sigma}_{yy} + \beta_0^2\hat{\sigma}_{xx} - 2\hat{\sigma}_{yx}\beta_0) \leq 0, \quad (13)$$

where $\hat{\sigma}_{yy} = \hat{\sigma}_{yy}^+ + \hat{\sigma}_{yy}^-$, $\hat{\sigma}_{xx} = \hat{\sigma}_{xx}^+ + \hat{\sigma}_{xx}^-$, and $\hat{\sigma}_{yx} = \hat{\sigma}_{yx}^+ + \hat{\sigma}_{yx}^-$. The expression on the left-hand side in (13) is a second order polynomial in $\beta_0$. Depending on the coefficients of that polynomial, the confidence set $CS_{1-\alpha}$ can potentially take one of the following forms: (i) an interval, (ii) the entire real line, or (iii) a union of two disconnected half-lines $(-\infty, a_1] \cup [a_2, \infty)$, where $a_1 < a_2$.\footnote{We show in the appendix that the confidence set $CS_{1-\alpha}$ cannot be empty.}
polynomial in $\beta_0$ in (13) are both negative. The discriminant of that polynomial is negative when $n h_n (\hat{x}^+ - \hat{x}^-)^2 (\hat{\beta}^2 \hat{\sigma}_{xx} - 2 \hat{\beta} \hat{\sigma}_{yx} + \hat{\sigma}_{yy}) - z^2_{1 - \alpha / 2} (\hat{\sigma}_{xx} \hat{\sigma}_{yy} - \hat{\sigma}_{yx}^2) < 0$, and the coefficient on $\beta_0^2$ is negative when $n h_n (\hat{x}^+ - \hat{x}^-)^2 - z^2_{1 - \alpha / 2} \hat{\sigma}_{xx} < 0$. When identification is strong and as the sample size $n$ increases, both the discriminant and the coefficient on $\beta_0^2$ tend to be positive, and therefore, with probability approaching one, $CS_{1 - \alpha}$ will be an interval when identification is strong.

When identification is weak, however, $n h_n (\hat{x}^+ - \hat{x}^-)^2$ approaches a constant $\theta$ and $\hat{\beta} \xrightarrow{d} \beta + \xi_{\beta, \theta}$ as $n$ increases. In this case, the confidence set $CS_{1 - \alpha}$ can be unbounded with a positive probability. This probability depends on the strength of identification $\theta$ and is higher for smaller values of $|\theta|$. Thus, when $\theta = 0$, the confidence set $CS_{1 - \alpha}$ is equal to the entire real line with probability approaching one.

3 Testing for constancy of the RD effect across co-

variates

In this section, we develop a test of constancy of the RD effect across covariates which is robust to weak identification issues. Such a test can be useful in practice when the econometrician wants to argue that the treatment effect is different for different population sub-groups. For example, in Section 5 we use this test to argue that the effect of class sizes on educational achievements is different between secular and religious schools, and therefore it might be optimal to implement different rules concerning class sizes in those two categories of schools. Similarly, the rules concerning class sizes should be different for schools with small and large numbers of disadvantage students, as the data seem to suggest that the effect of class sizes on educational achievements is also different between schools with a greater and lower than median
percentage of disadvantage students.

Similarly to Otsu and Xu (2011), we consider the RD effect conditional on some covariate \( w_i \). Let \( \mathcal{W} \) denote the support of the distribution of \( w_i \). Next, for \( w \in \mathcal{W} \) we define \( y^+(w) \) similarly to \( y^+ \) in (2), except that we now use the conditional expectation given \( w_i = w \):

\[
y^+ = \lim_{z \downarrow z_0} E(y_i|z_i = z, w_i = w).
\]

Let \( y^-(w), x^+(w) \) and \( x^-(w) \) be defined similarly. The conditional RD effect given \( w_i = w \) is defined as

\[
\beta(w) = \frac{y^+(w) - y^-(w)}{x^+(w) - x^-(w)}.
\]

Similarly to the case without covariates, under an appropriate set of assumptions, \( \beta(w) \) captures the (local) ATE at \( z_0 \) conditional on \( w_i = w \). We are interested in testing

\[
H_0 : \beta(w) = \beta \text{ for some } \beta \in \mathbb{R} \text{ and all } w \in \mathcal{W},
\]

against

\[
H_1 : \beta(w) \neq \beta(v) \text{ for some } v, w \in \mathcal{W}.
\]

When identification is strong, the econometrician can estimate the conditional RD effect function consistently and then use it for testing of constancy of the RD effect. However, this approach can be unreliable if identification is weak. We therefore take an alternative approach.

Suppose that \( \mathcal{W} = \{ \bar{w}^1, \ldots, \bar{w}^Q \} \), i.e. the covariate is categorical and divides the population into \( Q \) groups. The assumption of a categorical covariate is plausible in many practical applications where the econometrician may be interested in the effect of gender, school type and etc. However, even when the covariate is continuous, in a
nonparametric framework it might be sensible to categorize it to have sufficient power (as is often done in practice). For \( q = 1, \ldots, Q \), let \( \hat{y}_q^+, \hat{y}_q^-, \hat{x}_q^+ \), and \( \hat{x}_q^- \) denote the local linear estimators as defined in (3)-(6) but computed using only the observations with \( w_i = \bar{w}^q \). Let \( n_q \) be the number of such observations. We assume that Assumption 2 holds for each of the \( Q \) categories with \( n \) replaced by \( n_q \) and the bandwidth \( h_{n_q} \). The asymptotic variances and covariances of \( \hat{y}_q^+, \hat{y}_q^-, \hat{x}_q^+ \), and \( \hat{x}_q^- \), denoted by \( \sigma_{yy,q}^+, \sigma_{yy,q}^-, \sigma_{xx,q}^+, \sigma_{xx,q}^- \), \( \sigma_{yx,q}^+ \) and \( \sigma_{yx,q}^- \), can vary across the categories. Let \( \hat{\sigma}_{yy,q}^+, \hat{\sigma}_{yy,q}^-, \hat{\sigma}_{xx,q}^+, \hat{\sigma}_{xx,q}^- \), \( \hat{\sigma}_{yx,q}^+ \) and \( \hat{\sigma}_{yx,q}^- \) be the corresponding estimators.

If \( H_0 \) is true and the FRD effect is independent of \( w \), one can construct a robust confidence set for the common effect:

\[
CS_{1-\alpha}^Q = \left\{ \beta_0 \in \mathbb{R} : \sum_{q=1}^Q n_q h_{n_q} \frac{(\hat{\beta}_q - \beta_0)^2}{\hat{V}_q(\beta_0)} \leq \chi^2_{Q,1-\alpha} \right\},
\]

where

\[
\hat{\beta}_q = \frac{\hat{y}_q^+ - \hat{y}_q^-}{\hat{x}_q^+ - \hat{x}_q^-},
\]

\( \chi^2_{Q,1-\alpha} \) is the \((1 - \alpha)\)-th quantile of the \( \chi^2_Q \) distribution, and \( \hat{V}_q(\beta_0) \) is defined as in equation (10) and (11), except that \( \hat{\sigma}_{yy,q}^+, \hat{\sigma}_{yy,q}^-, \hat{\sigma}_{xx,q}^+, \hat{\sigma}_{xx,q}^- \), \( \hat{\sigma}_{yx,q}^+ \) and \( \hat{\sigma}_{yx,q}^- \) are being replaced with \( \hat{\sigma}_{yy,q}^{+,q}, \hat{\sigma}_{yy,q}^{-,q}, \hat{\sigma}_{xx,q}^{+,q}, \hat{\sigma}_{xx,q}^{-,q}, \hat{\sigma}_{yx,q}^{+,q} \) and \( \hat{\sigma}_{yx,q}^{-,q} \) respectively. Under \( H_0 : \beta(w) = \beta \) for some \( \beta \in \mathbb{R} \), \( CS_{1-\alpha}^Q \) is an asymptotically valid confidence set as in this case,

\[
\sum_{q=1}^Q n_q h_{n_q} \frac{(\hat{\beta}_q - \beta)^2}{\hat{V}_q(\beta)} \xrightarrow{d} \chi^2_Q
\]

under weak or strong identification, where the convergence is as \( n \to \infty \) and under the assumption that \( n_q h_{n_q} / (n h_n) \to p_q > 0 \) for all \( q = 1, \ldots, Q \).
We consider the following size $\alpha$ asymptotic test:

Reject $H_0$ if $CS_{1-\alpha}^Q$ is empty.

The test is asymptotically valid because under $H_0$, $P(CS_{1-\alpha}^Q = \emptyset) \leq P(\beta \notin CS_{1-\alpha}^Q) \to \alpha$, which holds again under weak or strong identification. Under the alternative, there is no common value $\beta$ that will provide a proper recentering for all $Q$ categories, and therefore one can expect deviations from the asymptotic $\chi^2_Q$ distribution. We show below that the test is consistent in the case of strong identification.

**Theorem 3.** Suppose that Assumption 2 holds for each $q = 1, \ldots, Q$ with the sample size $n_q$, bandwidth $h_{n_q}$, and covariate-dependent asymptotic variances, covariances, and their estimators, where $n = \sum_{q=1}^Q n_q$ and $n_q h_{n_q} / (nh_n) \to p_q > 0$. Suppose further that identification is strong and $\hat{x}_q^+ - \hat{x}_q^- \to_p d_q \neq 0$ for all $q = 1, \ldots, Q$. Then, under $H_1$, $P(CS_{1-\alpha}^Q = \emptyset) \to 1$ as $n \to \infty$.

## 4 Monte Carlo experiment

In this section, we illustrate the problem of weak identification in the FRD model using a Monte Carlo experiment. In our experiment, the outcome variable $y_i$ is generated according to the following model:

\[
\begin{align*}
y_i &= y_{0i} + x_i \beta, \\
x_i &= \begin{cases} 
1 (u_i < 0), & z_i \leq 0, \\
1 (u_i < c), & z_i > 0,
\end{cases}
\end{align*}
\]
where the outcome without treatment variable $y_{0i}$ and the variable $u_i$ are bivariate normal:

$$
\begin{pmatrix}
y_{0i} \\
u_i
\end{pmatrix}
\sim
N
\left(0, \begin{pmatrix}
1 & \rho \\
\rho & 1
\end{pmatrix}\right).
$$

Note that in this setup, $u_i$ determines whether the treatment is received, and therefore the parameter $\rho$ captures the degree of endogeneity of the treatment. The assignment variable $z_i$ is generated to be either a standard normal or a $N(0, 10^2)$ random variable independent from $y_{0i}$ and $u_i$. The observations are simulated to be independent across $i$’s.

Let $\Phi(\cdot)$ denote the standard normal CDF. In this setting, $x^+ - x^- = \Phi(c) - \Phi(0)$, and weak identification corresponds to small values of $c$. We use the following values to generate the data: $\beta = 0$, $n = 1000$, $c = 2$ for strong identification and $c = 0.01$ for weak identification, and the values of $\rho = 0.50$ and 0.99.

For each Monte Carlo replication, we generate $n$ observations as described above. Using the bandwidth $h_n = n^{-1/5-1/100}$ and the uniform kernel, we compute $\hat{\beta}$, $\hat{V}_{\beta}$, and the confidence intervals $\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\beta})/(n h_n)}$ for $\alpha = 0.10, 0.05, 0.01$. We use 10,000 replications to compute the average coverage probabilities of those confidence intervals, as well as the bias and MSE of the FRD estimator.

The results are reported in Table 1. When identification is strong ($c = 2$), the usual confidence intervals have coverage probabilities very close to the nominal ones. This is the case regardless of the degree of endogeneity ($\rho = 0.50$ or $\rho = 0.99$). The FRD estimator is more biased when $\rho$ is large.

When identification is weak ($c = 0.01$) and the degree of endogeneity is small ($\rho = 0.50$) the usual confidence intervals include the true value of $\beta$ with higher than nominal probabilities. As a matter of fact, the obtained coverage probabilities are close to one.
The situation changes substantially when identification is weak \((c = 0.01)\) and endogeneity is strong \((\rho = 0.99)\). In this case we observe size distortions. For example, the actual coverage probabilities of the 90%, 95% and 99% confidence intervals are approximately 82%, 87%, and 95%, respectively. We also repeated the experiment with \(z_i\) generated as \(N(0, 10^2)\). In this case, more substantial size distortions were observed, and the actual coverages of the 90%, 95% and 99% confidence intervals were 76%, 82%, and 90%, respectively.

When identification is weak, the FRD estimator is more biased and has large MSE and standard errors. The distribution of the standard error \(\sqrt{\hat{V}(\hat{\beta})/(nh_n)}\) also exhibits heavy tails when identification is weak. This explains the large values for the average standard errors under weak identification reported in Table 1. For example, when \(c = 0.01\) and \(\rho = 0.50\), the median, 75th percentile, and maximum standard error are 3.98, 13.73, and \(6.97 \times 10^6\), respectively. The standard errors are well-behaved when identification is strong. For example, when \(c = 2\) and \(\rho = 0.50\), the median, 75th percentile, and maximum standard error are 0.63, 0.77, and 5.59, respectively.

Figure 1 shows the densities of the usual \(T\) statistic estimated by kernel smoothing (for a standard normal assignment variable). As a comparison, we also plot the standard normal density. For \(\rho = 0.50\) and strong identification, it is apparent that the standard normal distribution is a very good approximation to the distribution of \(T\). When \(\rho = 0.99\), the distribution of \(T\) is slightly skewed to the left, but the normal approximation still works reasonable well, because there is no substantial deviation of extreme values of the distribution of \(T\) from those of the standard normal distribution.

Figures 1(c) and (d) show that under weak identification, the distribution of \(T\) is very different from normal. It is strongly skewed to the left, though when \(\rho = 0.50\) it is also more concentrated around zero. As a result, we do not see size distortions when
identification is weak but the degree of endogeneity is small. The picture changes
drastically when $\rho = 0.99$. The distribution of $T$ is strongly skewed to the left and no
longer concentrated around zero. As a result, we observe size distortions in this case.

Note further that due to the skewness of the distribution of $T$ in the case of
weak identification and strong endogeneity (Figure 1(d)), larger size distortions than
those reported above are expected when considering one-sided hypothesis tests of
$H_0 : \beta \geq \beta_0$ against $H_1 : \beta < \beta_0$, or one-sided confidence intervals of the form
$[a, +\infty)$. For example, when identification is weak and $\rho = 0.99$, the actual coverage
of the 90% one-sided confidence intervals is 74% in the case of a standard normal
assignment variable and 68% in the case of $N(0, 10^2)$ assignment variable.\footnote{The coverage probability of the one-sided 95% confidence intervals is equal to that of the 90% two-sided intervals.} The results are summarized in Table 2.

We have also computed the simulated coverage probabilities of the weak identifi-
cation robust confidence set $CS_{1-\alpha}$ introduced in (12) in Section 2.3. We find that
regardless of the strength of identification $c$ and degree of endogeneity $\rho$, the simulated
coverage probabilities of $CS_{1-\alpha}$ are very close to the nominal coverage probabilities
(see Table 3). For example, in the case of weak identification, strong endogeneity and
standard normal assignment variable, the coverage probabilities of the 90%, 95%, and
99% confidence sets are 90.4%, 95.2%, and 99.2%, respectively. This supports our
claim that the inference based on the null-restricted statistic $\tilde{T}(\beta_0)$ does not suffer
from size distortions and is asymptotically valid unlike the testing procedures based
on the usual $t$-statistic.

By contrast, the shape (and the expected size) of the weak identification robust
certainty set $CS_{1-\alpha}$ does depend on the strength of identification. As reported in
Table 4, in the case of a strong FRD, the probabilities that the robust confidence sets are equal to the entire real line are very small. Regardless of the value of \( \rho \), they are below 1% for the 90% and 95% confidence sets, and approximately 2% for the 99% confidence set. The probabilities that the robust confidence sets are given by a union of two half lines are similarly small. In the case of a weak FRD, unbounded robust confidence sets are obtained with very high probabilities. Thus, the entire real line is obtained with probabilities 74%, 86% and 97% for the confidence sets with nominal coverage of 90%, 95% and 99% respectively.

5 Empirical Application

In this section we compare the results of standard and weak identification robust inference in two separate, but related, applications. We show that the two methods yield significantly different conclusions when weak identification appears to be a problem, but similar results when it is likely not. We also show that the robust confidence sets can actually provide more informative answers in some cases than the standard confidence intervals when the usual assumptions are violated.

We begin with a case where weak identification is not a serious issue. In an influential paper, Angrist and Lavy (1999) have studied the effect of class size on academic success in Israel.\(^7\) During the sample period, class sizes in Israeli public schools were capped at 40 students in accordance with the recommendations of the twelfth century Rabbinic scholar Maimonides. This rule, known as “Maimonides’ rule”, results in discontinuities in the relationship between class size and total school enrollment (for a given grade). In practice, class size is not perfectly predicted by enrollment and we have a fuzzy, as opposed to a sharp, RD design.

\(^7\)This application has also been used by Otsu and Xu (2011).
The data consists of 4th and 5th grade classes. Class size, enrollment and class average verbal and mathematical achievement exam scores are available at the school level. The exams were administered by the Israeli state in 1991 and compiled by Angrist and Lavy (1999).\(^8\) Scores are calculated on a 100 point scale in their study, but we have rescaled them to be in terms of standard deviations relative to the mean. We focus on class average language scores among 4th graders (similar results are obtained for math scores), but otherwise use the same sample selection rules as Angrist and Lavy (1999). There is a total of 2049 classes in 1013 schools with valid test results. Here we only look at the first discontinuity at the 40 students cutoff. The number of observations used in the estimation depends on the bandwidth. It ranges from 471 classes in 118 schools for the smallest bandwidth (6), to 722 observations in 484 schools for the widest bandwidth (20). Note that the bandwidth selected using a “rule-of-thumb” procedure is 7.84 students.

Figure 2 plots the observed values of class size as a function of enrollment. As discussed earlier, class size is not strictly set according to Maimonides’ rule (the solid line in the figure). There is, nonetheless, a clearly discontinuity in the relationship between class size and enrollment at the cutoff value (40 students). Table 5 shows that the size of the discontinuity (the first stage estimates \(x^+ - x^-\)) ranges from \(-8\) to \(-15\) depending on the bandwidth chosen, which is smaller than the 20 students drop predicted by Maimonides’ rule. The table also shows that, as expected, the standard errors get larger as the bandwidth gets smaller. Despite this, the first stage effect remains statistically significant (\(t\)-statistic above 5) even for the smallest bandwidth considered. This suggests that weak identification is not much of an issue in this particular application.

Table 5 also reports the FRD estimates of the class size effect on the class average

\(^8\)The data can be found at http://econ-www.mit.edu/faculty/angrist/data1/data/anglavy99.
verbal score, as well as the 95% standard and robust confidence sets for the class size effect. The FRD estimates are uniformly negative and only significant at smaller bandwidths. The robust confidence intervals are relatively close to their standard versions, except at small bandwidths where they are slightly asymmetric (lower limit of the confidence interval smaller for robust confidence intervals). This is also illustrated in Figure 3 which shows that the two sets of confidence intervals are essentially indistinguishable for larger bandwidths, and only slightly different from each other for smaller bandwidths. The close proximity of the two sets is consistent with the above reported evidence that the identification is strong in this particular example.

In this application we also compare the standard test of equality of the RD effect across subgroups to our robust test proposed in Section 3 using two examples. We look at both the difference between secular and religious schools, and the difference between schools with a greater and lower than median percentage of disadvantaged students. Table 6 reports the results for both these comparisons. In neither cases are any of the RD estimates individually significant using either method. However, in both cases we find that a higher bandwidths the robust tests rejects the null hypothesis that the RD effects are equal across groups, while the standard test fails to do so. Specifically, at the largest bandwidths (18 and 20), our test rejects the hypothesis that test scores in religious and secular schools respond in the same way to a change in class size. At a bandwidth of 20, our test again rejects the null of hypothesis of a common effect for schools with an above and below median proportion of disadvantaged students, while the standard test fails to do so. This may assuage the worry that our proposed

---

9 Comparable estimates reported in Angrist and Lavy are generally significant as they estimate the treatment effect by pooling data at all cutoff values (multiples of 40). For the sake of clarity, here we use a conventional FRD design by only focusing on the first cutoff value (40 students).
test has lower power against alternatives than the standard one.

The second application we consider is based on Urquiola and Verhoogen (2009) who look at a similar rule for Chile. In that country, public schools use a variant of Maimonides’ rule which stipulates that class size cannot exceed 45 students. As in the Israeli data, a discontinuity in the probability of being assigned to a smaller class is observed when enrollment in a given grade goes beyond the class size cutoff. Figure 4 shows the discontinuity in the empirical relationship between class size and enrollment at the various multiples of 45 (45, 90, 135 and 180 students). The figure again shows that we have a FRD design since the observed data does not strictly follow the relationship predicted under a strict application of the rule. While this is not immediately obvious from the graph, Table 3 in Urquiola and Verhoogen (2009) shows that the identification gets increasingly weaker at higher multiples of the 45 students cutoff rule. We, therefore, use this example to show how the null restricted confidence sets start diverging substantially from the conventional confidence sets as identification becomes progressively weaker.\(^\text{10}\) In this example, the outcome variable is average class scores on state standardized math exams and we restrict attention to 4th graders. We also strictly adhere to the sample selection rules used by Urquiola and Verhoogen.

The number of observations vary with the bandwidth and the enrollment cutoff of interest. At the first cutoff point (45) we use between 273 and 778 school level

\(^{10}\) It should be noted that Urquiola and Verhoogen (2009) are not attempting to provide causal estimates of the effect of class size on tests score. They instead show how the RD design can be invalid when there is some manipulation around the cutoff, which results in a violation of Assumption 1b (exogeneity of \(z_i\)). So while this particular application is useful for illustrating some pitfalls linked to weak identification in a FRD design, the results should be interpreted with caution.
observations, depending on the bandwidth. The range in the number of observations is 201 to 402, 45 to 95, and 17 to 34 at the 90, 135, and 180 enrollment cutoffs, respectively. As the sample size decreases, weak identification becomes generally more of a concern. For instance, Table 7 shows that in the case of the first cutoff (45), the first stage estimates are large and statistically significant for the larger bandwidths, but become smaller and insignificant for bandwidths smaller than 10. This is an important concern since the optimal bandwidth suggested by the rule-of-thumb procedure is only 8.59. The problem of weak identification becomes even more severe at the higher cutoffs for which the first stage is almost never significant for the range of bandwidths considered here.

Table 8 reports the FRD estimates and the confidence sets for the different values of the bandwidth and of the cutoff points. As before, we set the size of the test at the 5 percent level. In this application, there is now a substantial divergence between the confidence sets obtained using the two methods. As the (first stage) effect becomes weaker, the differences between the two methods become starker.

Starting with the first cutoff point, Table 8 shows that the robust and conventional confidence sets diverge dramatically as the bandwidth gets smaller and identification gets weaker (recall Table 7). The confidence sets are fairly similar for the largest bandwidth (20). By the time we get to a bandwidth of 12, however, the robust confidence set is very asymmetric (going from $-1.720$ to $-0.065$) around the FRD estimate of $-0.173$. Interestingly, while the robust confidence interval is much wider than the conventional one, it is sufficiently shifted to the left to reject the null hypothesis that the effect of class size is equal to zero. By contrast, a conventional test would fail to reject the null that the effect is zero.

As we move to smaller bandwidths, the number of observations decreases and the first stage gets quite weak. The consequences that the null restricted confidence sets
become two disjoint half-lines. While these confidence intervals are unbounded, we can nonetheless reject the null that the effect of class size on test scores is equal to zero. The problem is that we cannot tell whether the effect is positive or negative because of the weak first stage. Conventional confidence sets yield a very different conclusion as they suggest that that null hypotheses of no treatment effect cannot be rejected. For instance, with the smallest bandwidth (6) the conventional approach suggests that the treatment effect is relatively small (between −0.061 and 0.353) and that a zero effect cannot be ruled out. By contrast, the robust confidence sets suggest that the treatment effect can be potentially quite large, and that it is significantly different from zero.

To help interpret the results, we also graphically illustrate the difference between standard and conventional confidence sets in Figure 5. The first panel plots the standard confidence sets as a function of the bandwidth. The second panel does the same for the weak-identification robust method. The shaded area is the region covered by the confidence sets. As the bandwidth increases, the robust confidence sets evolve from two disjoint sections of the real line to a well defined interval.\footnote{Note that class size is a discrete rather than a strictly continuous variable, hence the break between bandwidths 11 and 12 when the robust confidence set switches from two disjoint half lines to a single interval.}

Identification is considerably weaker for the second cutoff point. At all bandwidths, the standard confidence intervals fail to reject the null that the effect of class size is zero. However, for most bandwidths, the robust confidence sets never contain zero. For example, at the rule-of-thumb bandwidth (about 8), the econometrician to would fail to reject that class sizes are not related to average class grades using the standard method. However, our method would allow the econometrician to con-
clude that, at the 5% level, the confidence interval does not include zero at most bandwidths.

Identification is even weaker at the third cutoff and, for most bandwidths, the robust confidence sets consists of two disjoint intervals. Finally, results get very imprecise at the fourth cutoff because of the small number of observations, and the robust confidence sets now map the entire real line. This suggests the first stage is very weak at these levels and the standard confidence sets are overly conservative, even if they do not lead the econometrician to reject the null hypothesis at conventional levels.

In summary, our results suggest that when weak identification is not a problem, the robust and standard confidence sets are similar, but when the assignment variable does not produce a large enough jump in the conditional probability of assignment, the robust confidence sets are very different from those obtained using the standard method. We also demonstrate that our robust inference method can actually provide more informative results than the one typically used.

6 Conclusion

In this paper, we propose a simple and asymptotically valid method for computing robust t-statistics and confidence sets for the treatment effect in the fuzzy regression discontinuity (FRD) design when identification is weak. We also discuss how to extend the method to test for the constancy of the regression discontinuity effect for different values of the covariates. Using a Monte Carlo experiment, we show that the simulated coverage probabilities of the robust intervals are very close to the nominal coverage probabilities regardless of whether identification is weak or strong. By contrast, conventional confidence intervals suffer from important size distortion
when identification is weak.

We illustrate how the method works in practice for two related empirical applications from Angrist and Lavy (1999) and Urquiola and Verhoogen (2009). As expected, robust and conventional confidence intervals are similar when identification is strong, but sometimes diverge substantially when identification is weak. Interestingly, in both applications the first stage relationship looks visually quite strong, as is often the case in a FRD design. The relationship tends to get substantially weaker, however, for the relatively small bandwidths suggested using a rule-of-thumb procedure. More generally, it is good empirical practice to show how RD estimates are robust to a wide range of bandwidths, including relatively small ones. As the number of observations and the precision of the estimates decline for smaller bandwidths, it becomes increasingly important to compute confidence intervals that are robust in the presence of weak identification. Therefore, we expect that the simple and robust method suggested here will be useful for a wide range of empirical applications.

Appendix A: Proofs of the theorems

Proof of Theorem 1. For part (a), using (1),

\[ \hat{\beta} - \beta = \frac{\sqrt{nh_n}((\hat{y}^- - \hat{y}^+) - (y^- - y^+)) - \beta \sqrt{nh_n}((\hat{x}^- - \hat{x}^+) - (x^- - x^+))}{\sqrt{nh_n}((\hat{x}^- - \hat{x}^+) - (x^- - x^+)) + \theta} \]

\[ \rightarrow_d \xi_{\beta, \theta}, \]

where the second equality is by Assumptions 3, and the result in the last line is by Assumption 2 and the Continuous Mapping Theorem.
For part (b), by Assumptions 3,

\[(nh_n)^{-1} \hat{V}(\hat{\beta}) = \frac{\hat{\sigma}_{yy}^2 + \hat{\sigma}_{yy}^2 + 2\hat{\beta}^2 (\hat{\sigma}_{xx}^2 + \hat{\sigma}_{xx}^2) - 2\hat{\beta} (\hat{\sigma}_{yx}^2 + \hat{\sigma}_{yx}^2)}{nh_n \left[ (\hat{x}^- - \hat{x}^+) - (x^+ - x^-) + \theta / \sqrt{nh_n} \right]^2}
\]

\[= \frac{\hat{\sigma}_{yy}^2 + \hat{\sigma}_{yy}^2 + 2\hat{\beta}^2 (\hat{\sigma}_{xx}^2 + \hat{\sigma}_{xx}^2) - 2\hat{\beta} (\hat{\sigma}_{yx}^2 + \hat{\sigma}_{yx}^2)}{\sqrt{nh_n} \left[ (\hat{x}^- - \hat{x}^+) - (x^+ - x^-) + \theta \right]^2}
\]

\[\rightarrow_d \frac{\sigma^2 (\beta + \xi_{\beta, \theta})}{(X^+ - X^- + \theta)^2}.
\]

For part (c), by imposing \(\beta = \beta_0\), collecting the results from (a) and (b), and since convergence in (a) and (b) is joint, we obtain

\[T(\beta_0) \rightarrow_d \text{sgn}(X^+ - X^-) \frac{(Y^+ - Y^- - \beta (X^+ - X^-))}{\sigma (\beta)} \frac{\sigma (\beta)}{\sigma (\beta + \xi_{\beta, \theta})}.
\]

where, \((Y^+ - Y^- - \beta (X^+ - X^-))/\sigma (\beta) \sim N(0, 1)\). □

**Proof of Theorem 2.** The absolute value of the null-restricted t-statistic \(|T(\beta_0)| = |\hat{\beta} - \beta_0| / \sqrt{\hat{V}(\beta_0) / (nh_n)}\) can be written as follows:

\[\sqrt{nh_n} |\hat{y}^+ - \hat{y}^- - \beta_0 (\hat{x}^+ - \hat{x}^-)| / \hat{\sigma} (\beta_0)
\]

\[\rightarrow_d \left| Z + \frac{\theta (\beta - \beta_0)}{\sigma (\beta_0)} \right|.
\]

□

**Proof of Theorem 3.** Define

\[S_n(b) = \sum_{q=1}^Q \frac{n_{hq_n}}{nh_n} \frac{(\hat{\beta}_q - b)^2 (\hat{x}_q^+ - \hat{x}_q^-)^2}{\hat{\sigma}_q^2 (b)}.
\]
It suffices to show that, under the alternative, \( P(\inf_{b \in R} n h_n S_n(b) > a) \to 1 \) for all \( a \in R \) as \( n \to \infty \). In the proof below, we allow for \( S_n(b) \) to be minimized at a set of points or infinity. Define \( y_q^+ = \lim_{z \downarrow z_0} E(y_i | z_i = z, w_i = \tilde{w}^q) \). Let \( y_q^-, x_q^+, \) and \( x_q^- \) be defined similarly, and let \( \beta_q = (y_q^+ - y_q^-)/d_q \), where \( d_q \)'s are defined in the statement of the theorem. Since Assumption 2 holds for each \( q = 1, \ldots, Q \), define \( \sigma_{yy, q}^+, \sigma_{yy, q}^-, \sigma_{xx, q}^+, \sigma_{xx, q}^-, \sigma_{yx, q}^+, \) and \( \sigma_{yx, q}^- \) as the corresponding asymptotic variances and covariances. Further, define \( \sigma_q^2(b) \) according to (9), however, with the category-dependent variances and covariances. Lastly, let

\[
S(b) = \sum_{q=1}^{Q} p_q \frac{(\beta_q - b)^2 d_q^2}{\sigma_q^2(b)}.
\]

We will show next that

\[
\sup_{b \in R} |S_n(b) - S(b)| \to_p 0. \tag{14}
\]

Since \( (\beta_q - b)^2 \) and \( \sigma_q^2(b) \) are continuous for all \( b \in R \), and the asymptotic variance-covariance matrix in Assumption 2(b) is positive definite, it follows that the function \( (\beta_q - b)^2/\sigma_q^2(b) \) is continuous for all \( b \in R \) (Khuri, 2003, part 3 of Theorem 3.4.1 on page 71). Next,

\[
\lim_{|b| \to \pm \infty} \frac{(\beta_q - b)^2 d_q^2}{\sigma_q^2(b)} = \lim_{|b| \to \pm \infty} \frac{((y_q^+ - y_q^-) - bd_q)^2}{\sigma_{yy, q}^+ + \sigma_{yy, q}^- + b^2(\sigma_{xx, q}^+ + \sigma_{xx, q}^-) - 2b(\sigma_{yx, q}^+ + \sigma_{yx, q}^-)}
\]

\[
= \frac{d_q^2}{\sigma_{xx, q}^+ + \sigma_{xx, q}^-}.
\]

Thus, for all \( \varepsilon > 0 \) there is \( M_\varepsilon > 0 \) such that

\[
\left| \frac{(\beta_q - b)^2 d_q^2}{\sigma_q^2(b)} - \frac{d_q^2}{\sigma_{xx, q}^+ + \sigma_{xx, q}^-} \right| < \varepsilon
\]

32
for all \( |b| \geq M_\varepsilon \). Hence, by the triangle inequality, the function \((\beta_q - b)^2/\sigma^2_q(b)\) is bounded for all \( |b| \geq M_\varepsilon \). This function is also bounded for \( b \in [-M_\varepsilon, M_\varepsilon] \) (Khuri, 2003, Theorem 3.4.5 on page 72). By a similar argument, one can show that

\[
\sup_{b \in \mathbb{R}} \frac{(\hat{\beta}_q - b)^2}{\hat{\sigma}^2_q(b)} = O_p(1).
\]

(15)

By the triangle inequality

\[
|S_n(b) - S(b)| \leq \sum_{q=1}^Q p_q d_q^2 \left( \frac{(\hat{\beta}_q - b)^2}{\hat{\sigma}^2_q(b)} - (\beta_q - b)^2 \right) + \sum_{q=1}^Q R_{q,n}(b),
\]

(16)

where

\[
|R_{q,n}(b)| \leq \left| \left( \frac{n_q h_{n_q}}{nh_n} - p_q \right) (\hat{x}_q^+ - \hat{x}_q^-)^2 + \left( (\hat{x}_q^+ - \hat{x}_q^-)^2 - d_q^2 \right) p_q \right| \left( \frac{(\hat{\beta}_q - b)^2}{\hat{\sigma}^2_q(b)} \right).
\]

Hence, since \( n_q h_{n_q}/nh_n \to p_q \) and \( \hat{x}_q^+ - \hat{x}_q^- \to_d d_q \) by the assumption of the theorem, it follows from (15) that for \( q = 1, \ldots, Q \),

\[
\sup_{b \in \mathbb{R}} |R_{q,n}(b)| = o_p(1).
\]

(17)

Lastly,

\[
\left| \frac{(\hat{\beta}_q - b)^2}{\hat{\sigma}^2_q(b)} - (\beta_q - b)^2 \right| \leq \frac{1}{\hat{\sigma}^2_q(b)} \left| \beta^2_q - \beta^2_q \right| + \frac{(\beta_q - b)^2}{\hat{\sigma}^2_q(b)\sigma^2_q(b)} \left| \hat{\sigma}^+_{yy,q} - \hat{\sigma}^-_{yy,q} - \sigma^+_{yy,q} + \sigma^-_{yy,q} \right|
\]

\[
+ \frac{(\beta_q - b)^2b^2}{\hat{\sigma}^2_q(b)\sigma^2_q(b)} \left| \hat{\sigma}^+_{xx,q} - \hat{\sigma}^-_{xx,q} - \sigma^+_{xx,q} + \sigma^-_{xx,q} \right|
\]

\[
+ \frac{2(\beta_q - b)^2|b|}{\hat{\sigma}^2_q(b)\sigma^2_q(b)} \left| \hat{\sigma}^+_{yx,q} - \hat{\sigma}^-_{yx,q} - \sigma^+_{yx,q} + \sigma^-_{yx,q} \right|.
\]
By the same argument as above,

\[
\sup_{b \in R} \frac{1}{\hat{\sigma}_q^2(b)} = O_p(1), \quad \sup_{b \in R} \frac{(\beta_q - b)^2}{\hat{\sigma}_q^2(b)\sigma_q^2(b)} = O_p(1), \quad \sup_{b \in R} \frac{(\beta_q - b)^2b^2}{\hat{\sigma}_q^2(b)\sigma_q^2(b)} = O_p(1), \quad \text{and}
\]

\[
\sup_{b \in R} \frac{(\beta_q - b)^2|b|}{\hat{\sigma}_q^2(b)\sigma_q^2(b)} = O_p(1).
\]

Hence, for all \(q = 1, \ldots, Q\),

\[
\sup_{b \in R} \left| \frac{(\hat{\beta}_q - b)^2}{\hat{\sigma}_q^2(b)} - \frac{(\beta_q - b)^2}{\sigma_q^2(b)} \right| = o_p(1).
\]

(18)

The result in (14) follows from (16), (17), and (18).

Next, we will show that

\[
\left| \inf_{b \in R} S_n(b) - \inf_{b \in R} S(b) \right| \rightarrow_p 0.
\]

(19)

For \(\epsilon > 0\), define an event \(E_{\epsilon,n} = \{\sup_{b \in R} |S_n(b) - S(b)| < \epsilon/2\}\). On that event, when \(\inf_{b \in R} S_n(b) \geq \inf_{b \in R} S(b)\),

\[
\left| \inf_{b \in R} S_n(b) - \inf_{b \in R} S(b) \right| = \inf_{b \in R} (S(b) + S_n(b) - S(b)) - \inf_{b \in R} S(b)
\]

\[
< \inf_{b \in R} (S(b) + \epsilon/2) - \inf_{b \in R} S(b)
\]

\[
= \epsilon/2.
\]

Similarly on \(E_{\epsilon,n}\), when \(\inf_{b \in R} S_n(b) \leq \inf_{b \in R} S(b)\),

\[
\left| \inf_{b \in R} S_n(b) - \inf_{b \in R} S(b) \right| = \inf_{b \in R} S(b) - \inf_{b \in R} (S(b) + S_n(b) - S(b))
\]

\[
< \inf_{b \in R} S(b) - \inf_{b \in R} (S(b) - \epsilon/2)
\]
= \varepsilon/2.

Hence,

\[
P \left( \left\| \inf_{b \in R} S_n(b) - \inf_{b \in R} S(b) \right\| \geq \varepsilon \right) \leq P \left( \left\| \inf_{b \in R} S_n(b) - \inf_{b \in R} S(b) \right\| \geq \varepsilon, E_{\varepsilon,n} \right) + P \left( E_{\varepsilon,n}^c \right)
\]

\[= P \left( E_{\varepsilon,n}^c \right) \rightarrow 0,
\]

where the convergence in the last line holds for all \( \varepsilon > 0 \) by (14). The result in (19) follows.

Lastly, since \( \inf_{b \in R} S(b) > 0 \) under \( H_1 \) (whether \( S(b) \) is minimized at a set of points or infinity), we have that \( P(\inf_{b \in R} n h_n S_n(b) > a) \rightarrow 1 \) for all \( a \in R \) as \( n \rightarrow \infty \).

\[\Box\]

**Appendix B**

Here we show that the robust confidence set \( CS_{1-\alpha} \) defined in (12) cannot be empty. From (13) it follows that for \( CS_{1-\alpha} \) to be empty, the following two conditions must be satisfied:

\[
n h_n (\hat{x}^+ - \hat{x}^-)^2 (\hat{\beta}^2 \hat{\sigma}_{xx} - 2 \hat{\beta} \hat{\sigma}_{yx} + \hat{\sigma}_{yy}) - z_{1-\alpha/2}^2 (\hat{\sigma}_{xx} \hat{\sigma}_{yy} - \hat{\sigma}_{yx}^2) < 0, \quad (20)
\]

\[
n h_n (\hat{x}^+ - \hat{x}^-)^2 - z_{1-\alpha/2}^2 \hat{\sigma}_{xx} > 0. \quad (21)
\]

Suppose that the first inequality holds. Since the variance-covariance matrix composed of \( \hat{\sigma}_{xx}, \hat{\sigma}_{yy}, \) and \( \hat{\sigma}_{yx} \) is positive definite, it follows that \( \hat{\beta}^2 \hat{\sigma}_{xx} - 2 \hat{\beta} \hat{\sigma}_{yx} + \hat{\sigma}_{yy} > 0 \)
and $\hat{\sigma}_{xx} \hat{\sigma}_{yy} - \hat{\sigma}_{yx}^2 > 0$. The inequality in (20) then can be re-written as

$$nh_n (\hat{x}^+ - \hat{x}^-)^2 < z_{1-\alpha/2}^2 \hat{\sigma}_{xx} \frac{\hat{\sigma}_{yy} - \hat{\sigma}_{yx}^2 / \hat{\sigma}_{xx}}{\hat{\sigma}_{xx}^2 - 2 \hat{\beta} \hat{\sigma}_{yx} + \hat{\sigma}_{yy}}$$

$$< z_{1-\alpha/2}^2 \hat{\sigma}_{xx} \left( 1 - \frac{(\hat{\sigma}_{yx} / \sqrt{\hat{\sigma}_{xx}} - \hat{\beta} \sqrt{\hat{\sigma}_{xx}})^2}{\hat{\sigma}_{xx}^2 - 2 \hat{\beta} \hat{\sigma}_{yx} + \hat{\sigma}_{yy}} \right)$$

$$< z_{1-\alpha/2}^2 \hat{\sigma}_{xx}.$$

It follows that the two inequalities (20)-(21) cannot be true together.

**References**


Figure 1: Kernel estimated density of the usual $T$ statistic (solid line) under strong ($c = 2$) and weak ($c = 0.01$) identification for different values of the endogeneity parameter $\rho$ against the standard normal PDF (dashed line).

(a) Strong identification, $\rho = 0.50$

(b) Strong identification, $\rho = 0.99$

(c) Weak identification, $\rho = 0.50$

(d) Weak identification, $\rho = 0.99$
Figure 2: Angrist and Lavy (1999): Empirical relationship between class size and school enrollment

Note: The solid line shows the relationship when Maimonides’ rule (cap of 40 students) is strictly enforced.
Figure 3: Angrist and Lavy (1999): 95% confidence intervals for the effect of class size on verbal test scores for different values of the bandwidth

Note: The rule-of-thumb bandwidth is 7.84. The scores are given in terms of standard deviations from the mean.
Figure 4: Urquiola and Verhoogen (2009): Empirical relationship between class size and enrollment

Note: The solid line shows the relationship when the rule (cap of 45 students) is strictly enforced.
Figure 5: Urquiola and Verhoogen (2009): 95% standard and robust confidence sets (CSs) for the effect of class size on class average math score for different values of the bandwidth

Note: The rule of thumb bandwidth is approximately 8, depending on the cutoffs. The scores are given in terms of standard deviations from the mean.
Table 1: Simulated coverage probabilities of the confidence intervals constructed as estimate $\pm$ constant $\times$ standard error, and bias, root MSE and average standard error of the FRD estimator

<table>
<thead>
<tr>
<th>enforcing variable</th>
<th>identification</th>
<th>endogeneity</th>
<th>nominal coverage</th>
<th>simulated coverage</th>
<th>bias</th>
<th>root MSE</th>
<th>average std.err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N(0, 1)$</td>
<td>strong</td>
<td>$\rho = 0.50$</td>
<td>0.90</td>
<td>0.9307</td>
<td>0.0557</td>
<td>0.7100</td>
<td>0.6987</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>0.9710</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.9951</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N(0, 1)$</td>
<td>strong</td>
<td>$\rho = 0.99$</td>
<td>0.90</td>
<td>0.9264</td>
<td>0.1114</td>
<td>0.8021</td>
<td>0.7360</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>0.9563</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.9858</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N(0, 1)$</td>
<td>weak</td>
<td>$\rho = 0.50$</td>
<td>0.90</td>
<td>0.9803</td>
<td>-1.0765</td>
<td>72.2496</td>
<td>$2.1235 \times 10^3$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>0.9927</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.9995</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N(0, 1)$</td>
<td>weak</td>
<td>$\rho = 0.99$</td>
<td>0.90</td>
<td>0.8219</td>
<td>-0.1356</td>
<td>133.3221</td>
<td>$1.3184 \times 10^4$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>0.8749</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.9459</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N(0, 10^2)$</td>
<td>weak</td>
<td>$\rho = 0.99$</td>
<td>0.90</td>
<td>0.7597</td>
<td>$1.3005 \times 10^{11}$</td>
<td>$1.3005 \times 10^{13}$</td>
<td>$2.1041 \times 10^{11}$</td>
</tr>
</tbody>
</table>
Table 2: Simulated coverage probabilities of the one-sided confidence intervals constructed as \( [\text{estimate} - \text{constant} \times \text{standard error}, \infty) \)

<table>
<thead>
<tr>
<th>enforcing variable</th>
<th>identification</th>
<th>endogeneity</th>
<th>nominal coverage</th>
<th>simulated coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N(0,1) )</td>
<td>weak</td>
<td>( \rho = 0.99 )</td>
<td>0.90</td>
<td>0.7405</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>0.8219</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.9220</td>
</tr>
<tr>
<td>( N(0,10^2) )</td>
<td>weak</td>
<td>( \rho = 0.99 )</td>
<td>0.90</td>
<td>0.6805</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>0.7597</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.8735</td>
</tr>
</tbody>
</table>
Table 3: Simulated coverage probabilities of the robust confidence sets

<table>
<thead>
<tr>
<th>enforcing variable</th>
<th>identification</th>
<th>endogeneity</th>
<th>nominal coverage</th>
<th>simulated coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N(0, 1) )</td>
<td>weak</td>
<td>( \rho = 0.50 )</td>
<td>0.90</td>
<td>0.9040</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>0.9522</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.9918</td>
</tr>
<tr>
<td>( N(0, 1) )</td>
<td>weak</td>
<td>( \rho = 0.99 )</td>
<td>0.90</td>
<td>0.9040</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>0.9522</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.9918</td>
</tr>
<tr>
<td>( N(0, 10^2) )</td>
<td>weak</td>
<td>( \rho = 0.50 )</td>
<td>0.90</td>
<td>0.9283</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>0.9757</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.9964</td>
</tr>
<tr>
<td>( N(0, 10^2) )</td>
<td>weak</td>
<td>( \rho = 0.99 )</td>
<td>0.90</td>
<td>0.9286</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>0.9760</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.9967</td>
</tr>
</tbody>
</table>
Table 4: Simulated probabilities for the weak identification robust confidence set $CS_{1-\alpha}$ to be the entire real line or a union of two disconnected half-lines

<table>
<thead>
<tr>
<th>identification</th>
<th>endogeneity</th>
<th>nominal coverage</th>
<th>entire real line</th>
<th>two half-lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>strong $\rho = 0.50$</td>
<td>0.90</td>
<td>0.0017</td>
<td>0.0033</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.0046</td>
<td>0.0080</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.0213</td>
<td>0.0271</td>
<td></td>
</tr>
<tr>
<td>strong $\rho = 0.99$</td>
<td>0.90</td>
<td>0</td>
<td>0.0061</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.0004</td>
<td>0.0122</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.0025</td>
<td>0.0455</td>
<td></td>
</tr>
<tr>
<td>weak $\rho = 0.50$</td>
<td>0.90</td>
<td>0.7441</td>
<td>0.1579</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.8581</td>
<td>0.0953</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.9719</td>
<td>0.0211</td>
<td></td>
</tr>
<tr>
<td>weak $\rho = 0.99$</td>
<td>0.90</td>
<td>0.7444</td>
<td>0.1578</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.8599</td>
<td>0.0963</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.9707</td>
<td>0.0227</td>
<td></td>
</tr>
</tbody>
</table>
Table 5: Angrist and Lavy (1999): First stage estimates for the first cutoff and their standard errors, estimated effect of class size on class average verbal score, and standard and robust 95% confidence sets (CSs) for the class size effect for different values of the bandwidth.

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>First Stage Estimates</th>
<th>Estimated Effect</th>
<th>Standard CS</th>
<th>Robust CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>-8.4040</td>
<td>-0.0687</td>
<td>[-0.1440, 0.0066]</td>
<td>[-0.1702, -0.0003]</td>
</tr>
<tr>
<td></td>
<td>(1.6028)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-9.9013</td>
<td>-0.0722</td>
<td>[-0.1294, -0.0150]</td>
<td>[-0.1381, -0.0186]</td>
</tr>
<tr>
<td></td>
<td>(1.2585)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-10.8283</td>
<td>-0.056</td>
<td>[-0.0991, -0.0130]</td>
<td>[-0.1027, -0.0146]</td>
</tr>
<tr>
<td></td>
<td>(1.0314)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-11.9974</td>
<td>-0.0229</td>
<td>[-0.0562, 0.0105]</td>
<td>[-0.0581, 0.00970]</td>
</tr>
<tr>
<td></td>
<td>(0.9149)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>-12.6167</td>
<td>-0.0301</td>
<td>[-0.0605, 0.0002]</td>
<td>[-0.0618, -0.0002]</td>
</tr>
<tr>
<td></td>
<td>(0.7843)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-13.2053</td>
<td>-0.0200</td>
<td>[-0.0475, 0.0075]</td>
<td>[-0.0486, 0.00710]</td>
</tr>
<tr>
<td></td>
<td>(0.6864)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-13.8684</td>
<td>-0.0212</td>
<td>[-0.0459, 0.0034]</td>
<td>[-0.0468, 0.00310]</td>
</tr>
<tr>
<td></td>
<td>(0.6048)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>-14.3463</td>
<td>-0.0190</td>
<td>[-0.0424, 0.0045]</td>
<td>[-0.0434, 0.00420]</td>
</tr>
<tr>
<td></td>
<td>(0.5552)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The rule-of-thumb bandwidth is 7.84. The scores are given in terms of standard deviations from the mean.
Table 6: Angrist and Lavy (1999): Test of equality of RD effect across groups at 5% significance level for different values of the bandwidth

<table>
<thead>
<tr>
<th>bandwidth</th>
<th>estimated effect</th>
<th>reject $H_0$ of equality?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>religious</td>
<td>secular</td>
</tr>
<tr>
<td>6</td>
<td>−0.0524</td>
<td>−0.1131</td>
</tr>
<tr>
<td>8</td>
<td>−0.0540</td>
<td>−0.0985</td>
</tr>
<tr>
<td>10</td>
<td>−0.0381</td>
<td>−0.0756</td>
</tr>
<tr>
<td>12</td>
<td>−0.0170</td>
<td>−0.0364</td>
</tr>
<tr>
<td>14</td>
<td>−0.0274</td>
<td>−0.0363</td>
</tr>
<tr>
<td>16</td>
<td>−0.0035</td>
<td>−0.0382</td>
</tr>
<tr>
<td>18</td>
<td>0.0052</td>
<td>−0.0505</td>
</tr>
<tr>
<td>20</td>
<td>0.0107</td>
<td>−0.0523</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>&lt;= 10% disadvantaged</th>
<th>&gt; 10% disadvantaged</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>−0.0390</td>
</tr>
<tr>
<td>8</td>
<td>−0.0626</td>
</tr>
<tr>
<td>10</td>
<td>−0.0387</td>
</tr>
<tr>
<td>12</td>
<td>−0.0259</td>
</tr>
<tr>
<td>14</td>
<td>−0.0343</td>
</tr>
<tr>
<td>16</td>
<td>−0.0290</td>
</tr>
<tr>
<td>18</td>
<td>−0.0368</td>
</tr>
<tr>
<td>20</td>
<td>−0.0360</td>
</tr>
</tbody>
</table>
Table 7: Urquiola and Verhoogen (2009): First stage estimates for the first cutoff with their standard errors and $t$-statistics for various values of the bandwidth

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>First-stage estimates</th>
<th>$t$-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>with standard errors</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.388</td>
<td>0.907</td>
</tr>
<tr>
<td></td>
<td>(1.532)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.3873</td>
<td>-0.285</td>
</tr>
<tr>
<td></td>
<td>(1.358)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-3.107</td>
<td>-2.609</td>
</tr>
<tr>
<td></td>
<td>(1.191)</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>-4.779</td>
<td>-4.548</td>
</tr>
<tr>
<td></td>
<td>(1.051)</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>-6.092</td>
<td>-6.406</td>
</tr>
<tr>
<td></td>
<td>(0.951)</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>-7.87</td>
<td>-9.178</td>
</tr>
<tr>
<td></td>
<td>(0.857)</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-8.934</td>
<td>-11.393</td>
</tr>
<tr>
<td></td>
<td>(0.784)</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>-9.968</td>
<td>-13.727</td>
</tr>
<tr>
<td></td>
<td>(0.726)</td>
<td></td>
</tr>
</tbody>
</table>
Table 8: Urquiola and Verhoogen (2009): The estimated effect of class size on the class average math score and its 95% standard and robust confidence sets (CSs) for different values of the bandwidth.

<table>
<thead>
<tr>
<th>bandwidth</th>
<th>estimated effect</th>
<th>standard CS</th>
<th>robust CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>first cutoff (45)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.146</td>
<td>$[-0.061, 0.353]$</td>
<td>$(-\infty, -0.433] \cup [0.043, \infty)$</td>
</tr>
<tr>
<td>8</td>
<td>3.378</td>
<td>$[-74.820, 81.576]$</td>
<td>$(-\infty, -0.120] \cup [0.129, \infty)$</td>
</tr>
<tr>
<td>10</td>
<td>-0.437</td>
<td>$[-1.867, 0.993]$</td>
<td>$(-\infty, -0.078] \cup [0.181, \infty)$</td>
</tr>
<tr>
<td>12</td>
<td>-0.173</td>
<td>$[-0.360, 0.014]$</td>
<td>$[-1.720, -0.065]$</td>
</tr>
<tr>
<td>14</td>
<td>-0.136</td>
<td>$[-0.246, -0.026]$</td>
<td>$[-0.376, -0.060]$</td>
</tr>
<tr>
<td>16</td>
<td>-0.091</td>
<td>$[-0.153, -0.029]$</td>
<td>$[-0.186, -0.042]$</td>
</tr>
<tr>
<td>18</td>
<td>-0.073</td>
<td>$[-0.115, -0.031]$</td>
<td>$[-0.127, -0.037]$</td>
</tr>
<tr>
<td>20</td>
<td>-0.063</td>
<td>$[-0.099, -0.027]$</td>
<td>$[-0.107, -0.032]$</td>
</tr>
<tr>
<td>second cutoff (90)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.128</td>
<td>$[-0.025, 0.281]$</td>
<td>$[0.004, 3.093]$</td>
</tr>
<tr>
<td>8</td>
<td>0.261</td>
<td>$[-0.061, 0.582]$</td>
<td>$(-\infty, -0.587] \cup [0.085, \infty)$</td>
</tr>
<tr>
<td>10</td>
<td>0.227</td>
<td>$[-0.111, 0.566]$</td>
<td>$(-\infty, -0.241] \cup [0.046, \infty)$</td>
</tr>
<tr>
<td>12</td>
<td>0.306</td>
<td>$[-0.296, 0.908]$</td>
<td>$(-\infty, -0.118] \cup [0.053, \infty)$</td>
</tr>
<tr>
<td>14</td>
<td>0.486</td>
<td>$[-1.092, 2.063]$</td>
<td>$(-\infty, -0.056] \cup [0.068, \infty)$</td>
</tr>
<tr>
<td>16</td>
<td>1.636</td>
<td>$[-18.745, 22.017]$</td>
<td>$(-\infty, 0.002] \cup [0.065, \infty)$</td>
</tr>
<tr>
<td>18</td>
<td>-1.056</td>
<td>$[-10.968, 8.856]$</td>
<td>$(-\infty, \infty)$</td>
</tr>
<tr>
<td>20</td>
<td>-0.425</td>
<td>$[-2.041, 1.190]$</td>
<td>$(-\infty, 0.005] \cup [0.162, \infty)$</td>
</tr>
</tbody>
</table>

The rule of thumb bandwidth is approximately 8. The scores are given in terms of standard deviations from the mean.
Table 8: (Continued)

<table>
<thead>
<tr>
<th>bandwidth</th>
<th>estimated effect</th>
<th>standard CS</th>
<th>robust CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>third cutoff (135)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>−2.145</td>
<td>[−15.627, 11.336]</td>
<td>(−∞, −0.076] ∪ [0.584, ∞)</td>
</tr>
<tr>
<td>8</td>
<td>−0.298</td>
<td>[−0.692, 0.097]</td>
<td>[−21.482, 0.007]</td>
</tr>
<tr>
<td>10</td>
<td>−0.307</td>
<td>[−0.850, 0.236]</td>
<td>(−∞, 0.027] ∪ [1.414, ∞)</td>
</tr>
<tr>
<td>12</td>
<td>−0.309</td>
<td>[−0.861, 0.243]</td>
<td>(−∞, 0.027] ∪ [1.550, ∞)</td>
</tr>
<tr>
<td>14</td>
<td>−0.328</td>
<td>[−0.885, 0.228]</td>
<td>(−∞, −0.001] ∪ [1.838, ∞)</td>
</tr>
<tr>
<td>16</td>
<td>−0.231</td>
<td>[−0.652, 0.190]</td>
<td>(−∞, 0.034] ∪ [1.604, ∞)</td>
</tr>
<tr>
<td>18</td>
<td>−0.181</td>
<td>[−0.500, 0.138]</td>
<td>(−∞, 0.041] ∪ [21.933, ∞)</td>
</tr>
<tr>
<td>20</td>
<td>−0.136</td>
<td>[−0.389, 0.117]</td>
<td>[−1.642, 0.063]</td>
</tr>
<tr>
<td>forth cutoff (180)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.048</td>
<td>[−0.119, 0.216]</td>
<td>(−∞, ∞)</td>
</tr>
<tr>
<td>12</td>
<td>0.035</td>
<td>[−0.130, 0.200]</td>
<td>(−∞, ∞)</td>
</tr>
<tr>
<td>14</td>
<td>−0.047</td>
<td>[−0.371, 0.278]</td>
<td>(−∞, ∞)</td>
</tr>
<tr>
<td>16</td>
<td>−0.045</td>
<td>[−0.343, 0.254]</td>
<td>(−∞, ∞)</td>
</tr>
<tr>
<td>18</td>
<td>−0.039</td>
<td>[−0.316, 0.238]</td>
<td>(−∞, ∞)</td>
</tr>
<tr>
<td>20</td>
<td>−0.029</td>
<td>[−0.299, 0.242]</td>
<td>(−∞, ∞)</td>
</tr>
</tbody>
</table>