Weak Identification in Fuzzy Regression Discontinuity Designs

Vadim Marmer†  Donna Feir‡  Thomas Lemieux†

January 21, 2014

Abstract

In fuzzy regression discontinuity (FRD) designs, the treatment effect is identified through a discontinuity in the conditional probability of treatment assignment. 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 as in a standard instrumental variables setting. This 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 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

---

*We thank our editor, Shakeeb Khan, an associate editor and two anonymous referees for very helpful comments. We also thank Chris Muris, Moshe Buchinsky, Karim Chalak, Jae-Young Kim, Sokbae Lee, Arthur Lewbel, Taisuke Otsu, Eric Renault, Yoon-Jae Whang for their comments on early drafts of the paper. Vadim Marmer gratefully acknowledges the financial support of the SSHRC under grants 410-2010-1394 and 435-2013-0331.

†Vancouver School of Economics, University of British Columbia, 997 - 1873 East Mall, Vancouver, BC, V6T 1Z1, Canada. E-mails: vadim.marmer@ubc.ca (Marmer), and thomas.lemieux@ubc.ca (Lemieux).

‡Department of Economics, University of Victoria, PO Box 1700 STN CSC, Victoria, BC, V8W 2Y2, Canada. Email: dfeir@uvic.ca.
also constructed using this null-restricted standard error. An extension to testing for constancy of the regression discontinuity effect across covariates is also discussed.

\textit{JEL Classification}: C12; C13; C14

\textit{Keywords}: Nonparametric inference; regression discontinuity design; treatment effect; weak identification; uniform asymptotic size

1 \textbf{Introduction}

Since the late 1990s regression discontinuity (RD) and fuzzy regression discontinuity (FRD) designs have been of growing importance in applied economics (Van der Klaauw, 2008; Lee and Lemieux, 2010).\footnote{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. There is extensive theoretical work on RD and FRD designs. A few examples include Hahn et al. (1999, 2001); Porter (2003); Buddelmeyer and Skoufias (2004); McCrary (2008); Frölich (2007); Frölich and Melly (2008); Otsu and Xu (2011); Imbens and Kalyanaraman (2012); Calonico et al. (2012); Arai and Ichimura (2013); Papay et al. (2011); Imbens and Zajonc (2011); Dong and Lewbel (2010); Fe (2012). See Van der Klaauw (2008) and Lee and Lemieux (2010) for a review of much of this literature.} Hundreds of recent applied papers have used RD, and in many cases FRD designs.\footnote{For example, as of July 18th, 2013 Imbens and Lemieux (2008) review of RD and FRD best practices was cited in 990 articles according to Google Scholar, with 372 of these articles explicitly considering FRD.} Around the same time, the seminal works of Bound et al. (1995) and Staiger and Stock (1997) made weak identification in an instrumental variables (IV) context an important consideration in applied work (see, Stock et al. (2002) and Andrews and Stock (2007) for surveys of the literature).

However, despite the close parallel between an IV setting and the FRD design (see Hahn et al. (2001)) there has been no theoretical or practical attempt to deal with
weak identification in the FRD design more broadly.\(^3\) This is surprising since it is now standard practice to account for the possibility of weak identification in an IV context.

Weak identification is likely an important problem in FRD. Although large data sets are often used in empirical applications, RD estimation only relies on data “right around” the discontinuity point, which means that effective estimation samples can be quite small. To get a sense of the practical importance of weak identification in the FRD design, we have examined a sample of influential applied papers that use the design. We then apply the \(F\)-statistic standards discussed below to see how many of these papers appear to suffer from a weak identification problem. We find that in about half of the papers where enough information is reported to compute the identification test, weak identification appears to be a problem in at least one of the empirical specifications reported.\(^4\) We take this as evidence that weak identification is a serious concern in the applied FRD design literature. Since it is a matter of practical importance, we examine weak identification in the context of the FRD design, demonstrate the problems that arise, and propose consistent testing procedures for both uniform and heterogeneous treatment effects.

In this paper, we show that the local-to-zero analytical framework common in the

\(^3\)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 assignment variable exceed the cutoff point. Weak identification in FRD corresponds to the situation where the discontinuity in the conditional probability function of treatment assignment is of a small magnitude.

\(^4\)We start with thirty applied papers that were cited by Lee and Lemieux (2010). Of the thirty papers, sixteen did not report enough information to perform the identification test. Of the remaining papers, more than half had specifications which would be suspect according to the test. We reach similar conclusions when only focusing on the ten most cited paper in the list (Pitt and Khandker (1998); Hoxby (2000); Angrist (1990); (Van der Klaauw, 2002); Thistlethwaite and Campbell (1960); Greenstone and Gallagher (2008); Jacob and Lefgren (2004); (Oreopoulos, 2006); Card et al. (2009); and (Kane, 2003)). These papers had between 203 and 888 Google Scholar citations. Four of the ten papers do not report enough information to compute the test, but four of the remaining six papers presented some specifications that failed the test.
weak instruments literature can be adapted to FRD, and 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. 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. We rely on novel techniques recently developed in the literature on uniform size properties of tests and confidence sets (Andrews et al., 2011) to formally justify our local-to-zero framework. Unlike the framework used in the weak IV literature, ours depends not only on the sample size but also on a smoothing parameter (the bandwidth).

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 modified $t$-statistic uses a null-restricted version of the standard error of the FRD estimator. The modified statistic can be used with standard normal critical values for two-sided testing. For two-sided testing, the proposed test is equivalent to the Anderson-Rubin test (Anderson and Rubin, 1949) adopted in the weak IV literature (Staiger and Stock, 1997). For one-sided testing, the modified $t$-statistic has to be used with non-standard critical values that must be simulated on a case-by-case basis following the approach of (Moreira, 2001, 2003).

We discuss how to evaluate the strength of identification and magnitude of potential size distortions in practice following the approach of Stock and Yogo (2005). The strength of identification is measured by the concentration parameter, which in the case of FRD depends on the magnitude of the discontinuity in the treatment variable and on the density of the assignment variable (the variable that determines treatment assignment). Identification can be tested by testing hypotheses about the concentration parameter with non-central $\chi^2_1$ critical values using the $F_n$-statistic, which is an
analogue of the first-stage $F$-statistic in IV regression. Surprisingly, we find that the critical value that ensures asymptotic validity of a 5% two-sided test based on the usual $t$-statistic for FRD is much higher than would be required in a simple IV setting. When $F_n$ is only around 10, which is often used as a threshold value for weak/strong identification in the IV literature, a two-sided test with nominal size of 5% is in fact a 13.4% test, and a 5% one-sided test is in fact a 16.6% test. For asymptotic validity of a 5% two-sided test, the $F_n$-statistic should be approximately 135.

Asymptotically valid confidence sets for the treatment effect can be obtained by inverting tests based on the modified $t$-statistic. Since the FRD is an exactly identified model, these confidence sets are easy to compute since their construction only involves solving a quadratic equation.\textsuperscript{5} These confidence sets are expected to be as informative as the standard ones, when identification is strong. However, unlike the usual confidence intervals, 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)).\textsuperscript{6}

We also discuss testing whether the RD effect is homogeneous over differing values of some covariates. The proposed testing approach is designed to remain asymptot-
ically 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 common RD effect is rejected when that confidence set is empty.

To illustrate how our proposed confidences sets may differ from the standard ones in practice, we compare the results of applying the standard confidence sets and the proposed confidence sets in two separate applications. The first application uses data from Israel (Angrist and Lavy, 1999) and the second, data from Chile (Urquiola and Verhoogen, 2009). Both applications use the FRD 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. Our main finding is that, as weak identification becomes more likely, the standard confidence sets and the weak identification robust confidence sets become increasingly divergent. Interestingly, in a number of cases the robust confidence sets provide more informative answers than the standard ones. More generally, the empirical applications, along with a Monte Carlo study reported in an online supplement (Marmer et al., 2014), 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, derive the uniform asymptotic size of usual $t$-tests for FRD, discuss size distortions and testing for weak identification, and describe weak-identification-robust inference for FRD. Section 3 discusses robust testing for constancy of the RD effect across covariates. We present our empirical applications in Section 4.
2 Theoretical results

2.1 The model, estimation, and standard inference approach

In RD designs, the observed outcome variable $y_i$ is modeled as

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

where $x_i$ is the treatment indicator variable, $y_{0i}$ 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, $z_i$ through $E(x_i|z_i = z)$. 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) $\lim_{z \downarrow z_0} E(y_{0i}|z_i = z) = \lim_{z \uparrow z_0} E(y_{0i}|z_i = z)$.

In the case where $x_i$ is binary, when $|\lim_{z \uparrow z_0} E(x_i|z_i = z) - \lim_{z \downarrow z_0} E(x_i|z_i = z)| = 1$ we have a sharp RD design, and a fuzzy design otherwise. When $x_i$ is a continuous treatment variable, the design is sharp when $x_i$ follows a deterministic rule, and fuzzy otherwise.

The focus of this paper is fuzzy designs, and the main object of interest is the RD effect:

$$ \beta = (y^+ - y^-)/(x^+ - x^-), \quad (1) $$

where

$$ y^+ = \lim_{z \downarrow z_0} E(y_i|z_i = z) \text{ and } y^- = \lim_{z \uparrow z_0} E(y_i|z_i = z), $$

If $x_i$ is binary, it takes on value one if the treatment is received and zero otherwise. When there are treatments of different intensity, $x_i$ may be non-binary.
and $x^+$ and $x^-$ are defined similarly with $y_i$ replaced by $x_i$. 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 et al. (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 $z_i = z_0$: $\beta = E(\beta_i|z_i = z_0)$. When $x_i$ is binary and under an alternative set of conditions, which allow for dependence between $x_i$ and $\beta_i$, Hahn et al. (2001) show that the RD effect captures the local ATE (LATE) or ATE for compliers at $z_0$, where compliers are observations for which $x_i$ switches its value from zero to one when $z_i$ changes from $z_0 - e$ to $z_0 + e$ for all small $e > 0$.

Regardless of its interpretation, the RD effect is estimated by replacing the unknown population objects in (1) with their estimates. Following Hahn et al. (2001), it is now a standard approach to estimate $y^+$, $y^-$, $x^+$, and $x^-$ using local linear kernel regression. Let $K(\cdot)$ and $h_n$ denote the kernel function and bandwidth respectively. For estimation of $y^+$, the local linear regression is

$$
\left(\hat{a}_n, \hat{b}_n\right) = \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - (z_i - z_0) b)^2 K\left(\frac{z_i - z_0}{h_n}\right) 1 \{z_i \geq z_0\}, \quad (2)
$$

and the local linear estimator of $y^+$ is given by

$$\hat{y}_n^+ = \hat{a}_n.$$ 

The local linear estimator for $y^-$ can be constructed analogously by replacing $1\{z_i \geq z_0\}$ with $1\{z_i < z_0\}$ in the above equation. Similarly, one can estimate $x^+$ and $x^-$ by replacing $y_i$ with $x_i$. Let $\hat{y}_n^-$, $\hat{x}_n^+$, and $\hat{x}_n^-$ denote the local linear estimators of $y^-$, $x^+$, and $x^-$, respectively.

---

8See the discussion on page 204 of their paper.
and \( x^- \) respectively. The corresponding estimator of \( \beta \) is given by

\[
\hat{\beta}_n = (\hat{y}_n^+ - \hat{y}_n^-)/(\hat{x}_n^+ - \hat{x}_n^-).
\]

The asymptotic properties of the local linear estimators and \( \hat{\beta}_n \) are discussed in Hahn et al. (1999) and Imbens and Lemieux (2008). We assume that the following conditions are satisfied.

**Assumption 2.**

(a) \( K(\cdot) \) is continuous, symmetric around zero, non-negative, and compactly supported second-order kernel.

(b) \( \{ (y_i, x_i, z_i) \}_{i=1}^n \) are iid; \( y_i, x_i, z_i \) have a joint distribution \( F \) such that:

(i) \( f_z(\cdot) \) (the marginal PDF of \( z_i \)) exists and is bounded from above, bounded away from zero, and twice continuously differentiable with bounded derivatives on \( \mathcal{N}_{z_0} \) (a small neighborhood of \( z_0 \)).

(ii) \( E(y_i|z_i) \) and \( E(x_i|z_i) \) are bounded on \( \mathcal{N}_{z_0} \) and twice continuously differentiable with bounded derivatives on \( \mathcal{N}_{z_0} \setminus \{z_0\} \); \( \lim_{\epsilon \downarrow 0} \frac{d}{dz} E(y_i|z_i = z_0 \pm \epsilon) \) and \( \lim_{\epsilon \downarrow 0} \frac{d}{dz} E(x_i|z_i = z_0 \pm \epsilon) \) exist for \( p = 0, 1, 2 \).

(iii) \( \sigma^2_y(z_i) = \text{Var}(y_i|z_i) \) and \( \sigma^2_x(z_i) = \text{Var}(x_i|z_i) \) are bounded from above and bounded away from zero on \( \mathcal{N}_{z_0} \); \( \lim_{\epsilon \downarrow 0} \sigma^2_y(z_0 \pm \epsilon) \), \( \lim_{\epsilon \downarrow 0} \sigma^2_x(z_0 \pm \epsilon) \), and \( \lim_{\epsilon \downarrow 0} \sigma_{xy}(z_0 \pm \epsilon) \) exist, where \( \sigma_{xy}(z_i) = \text{Cov}(x_i, y_i|z_i) \); \( |\rho_{xy}| \leq \bar{\rho} \) for some \( \bar{\rho} < 1 \), where \( \rho_{xy} = \sigma_{xy}/(\sigma_x \sigma_y) \), \( \sigma_{xy} = \lim_{\epsilon \downarrow 0} (\sigma_{xy}(z_0 + \epsilon) + \sigma_{xy}(z_0 - \epsilon)) \), and \( \sigma^2_x \) and \( \sigma^2_y \) defined similarly with the conditional covariance replaced by the conditional variances of \( x_i \) and \( y_i \) respectively.

(iv) For some \( \delta > 0 \), \( E \left( |y_i - E(y_i|z_i)|^{2+\delta} \big| z_i \right) \) and \( E \left( |x_i - E(x_i|z_i)|^{2+\delta} \big| z_i \right) \) are bounded on \( \mathcal{N}_{z_0} \).
As $n \to \infty$, $\sqrt{nh_n^2 h_n^2} \to 0$ and $nh_n^3 \to \infty$.

**Remark.** 1) The smoothness conditions imposed in Assumption 2(b) are standard for kernel estimation except for the left/right limit conditions in parts (ii) and (iii), which are due to the discontinuity design and have been used in Hahn et al. (1999). 2) Asymptotic normality of the local linear estimators is established using Lyapounov’s CLT, and part (iv) of Assumption 2(b) can be used to verify Lyapounov’s condition (see Davidson, 1994, Theorem 23.12, p. 373). 3) With twice differentiable functions, the bias of the local linear estimators is of order $h_n^2$ even near the boundaries. The condition $\sqrt{nh_n^2 h_n^2} \to 0$ in Assumption 2(c) is an under-smoothing condition, which makes the contribution of the bias term to the asymptotic distribution negligible. The condition $nh_n^3 \to \infty$ is imposed because of local linear estimation. Assumption 2(c) is satisfied if the bandwidth is chosen according to the rule $h_n = \text{constant} \times n^{-r}$ with $1/5 < r < 1/3$.

It is convenient for our purposes to present the asymptotic properties of the local linear estimators and the FRD estimator as follows. Define

$$\sigma^2(b) = \sigma_y^2 + b^2 \sigma_x^2 - 2b \sigma_{xy},$$

and

$$k = \frac{\int_0^\infty \left( \int_0^\infty s^2 K(s) ds - u \int_0^\infty s K(s) ds \right)^2 K^2(u) du}{\left( \int_0^\infty u^2 K(u) du \int_0^\infty K(u) du - \left( \int_0^\infty u K(u) du \right)^2 \right)^2}.$$

For

$$\Delta y = y^+ - y^- \quad \text{and} \quad \hat{\Delta} y_n = \hat{y}_n^+ - \hat{y}_n^-,$$

9The constant $k$ is a known as it depends only on the kernel function. In the case of asymmetric kernels, we will have two different constants for the left and right estimators, with the bounds of integration replaced by $(-\infty, 0]$ for the left estimators.
and similarly defined $\Delta x$ and $\widehat{\Delta x}_n$, by Assumption 2 and Lyapounov’s CLT we have:

$$
\sqrt{nh_n} \begin{pmatrix} \Delta y_n - \Delta y \\ \Delta x_n - \Delta x \end{pmatrix} \rightarrow d \sqrt{\frac{k}{f_z(z_0)}} \begin{pmatrix} \sigma_y \mathcal{Y} \\ \sigma_x \mathcal{X} \end{pmatrix},
$$

where

$$
\begin{pmatrix} \mathcal{Y} \\ \mathcal{X} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{pmatrix} \right),
$$

and

$$
\sqrt{nh_n} \left( \hat{\beta}_n - \beta \right) \rightarrow d N \left( 0, \frac{k}{f_z(z_0)} \sigma^2(\beta) \frac{1}{(\Delta x)^2} \right),
$$

where the last result holds by the delta method due to identification Assumption 1(a), i.e. when $\Delta x \neq 0$.

The asymptotic variance $\sigma_y^2$ can be consistently estimated by

$$
\hat{\sigma}_{y,n}^2 = \frac{1}{\hat{f}_{z,n}(z_0)} \frac{1}{nh_n} \sum_{i=1}^{n} (y_i - \hat{y}_n^+ 1\{z_i \geq z_0\} - \hat{y}_n^- 1\{z_i < z_0\})^2 K \left( \frac{z_i - z_0}{h_n} \right),
$$

where $\hat{f}_{z,n}(z_0)$ is the kernel estimator of $f_z(z_0)$: $\hat{f}_{z,n}(z_0) = (nh_n)^{-1} \sum_{i=1}^{n} K((z_i - z_0)/h_n)$. Consistent estimators of $\sigma_x^2$ and $\sigma_{xy}$ can be constructed similarly by replacing

$$(y_i - \hat{y}_n^+ 1\{z_i \geq z_0\} - \hat{y}_n^- 1\{z_i < z_0\})^2$$

with $(x_i - \hat{x}_n^+ 1\{z_i \geq z_0\} - \hat{x}_n^- 1\{z_i < z_0\})^2$ and

$$(x_i - \hat{x}_n^+ 1\{z_i \geq z_0\} - \hat{x}_n^- 1\{z_i < z_0\})(y_i - \hat{y}_n^+ 1\{z_i \geq z_0\} - \hat{y}_n^- 1\{z_i < z_0\})$$

respectively. Hence, a consistent estimator of $\sigma^2(b)$ can be constructed as

$$
\hat{\sigma}^2_n(b) = \hat{\sigma}_{y,n}^2 + b\hat{\sigma}_{x,n}^2 - 2b\hat{\sigma}_{xy,n}.
$$

A common inference approach for the FRD effect is based on the usual $t$-statistic.
Thus, when testing $H_0 : \beta = \beta_0$ one computes

$$T_n(\beta_0) = \frac{\sqrt{n}h_n (\hat{\beta}_n - \beta_0)}{\sqrt{k\hat{\sigma}_n^2(\hat{\beta}_n) / (f_{z,n}(z_0)(\hat{\Delta}x_n)^2)}}$$

and compares it with standard normal critical values, as $T_n(\beta) \to_d N(0,1)$ when $\Delta x \neq 0$. Confidence intervals for $\beta$ are constructed by collecting all values $\beta_0$ for which $H_0 : \beta = \beta_0$ cannot be rejected using a test based on $T_n(\beta_0)$.

### 2.2 Weak identification in FRD

Weak identification is a finite-sample problem, which occurs when the noise due to sampling errors is of the same magnitude or even dominates the signal in estimation of a model’s parameters. In such cases, the asymptotic normality result $T_n(\beta) \to_d N(0,1)$ provides a poor approximation to the actual distribution of the $t$-statistic, and as a result inference may be distorted.

Assuming that $H_0 : \beta = \beta_0$, we can re-write the $t$-statistic as

$$T_n(\beta) = \frac{\sqrt{n}h_n (\hat{\Delta}y_n - \beta\hat{\Delta}x_n)}{\sqrt{k\hat{\sigma}_n^2(\hat{\beta}_n) / f_{z,n}(z_0)}} \times \text{sign}(\hat{\Delta}x_n). \quad (5)$$

When testing $H_0$ against two-sided alternatives, one uses the absolute value of $T_n(\beta)$, which eliminates the sign term. Since $\sqrt{n}h_n (\hat{\Delta}y_n - \beta\hat{\Delta}x_n) \to_d N(0, k\sigma^2(\beta)/f_{z}(z_0))$, there will be no asymptotic size distortions as long as $\hat{\beta}_n$ is consistent and, therefore, $\hat{\sigma}_n^2(\hat{\beta}_n)$ approximates well $\sigma^2(\beta_0)$. Define $\Delta Y_n = (f_{z}(z_0)/k)^{1/2}(nh_n)^{1/2}(\hat{\Delta}y_n - \Delta y)$ and $\Delta X_n = (f_{z}(z_0)/k)^{1/2}(nh_n)^{1/2}(\hat{\Delta}x_n - \Delta x)$. We can now write

$$\hat{\beta}_n - \beta = \frac{\Delta Y_n - \beta \Delta X_n}{\Delta X_n + (f_{z}(z_0)/k)^{1/2}(nh_n)^{1/2}\Delta x}. \quad (6)$$
Note that in the above expression, estimation errors $\Delta Y_n$ and $\Delta X_n$ represent the noise components, while the signal component is given by $(nh_n)^{1/2}\Delta x$. Since the noise terms have bounded variances, the signal dominates the noise as long as $(nh_n)^{1/2}\Delta x \to \infty$. In this case, $\hat{\beta}_n \to_p \beta$. If, however, $\lim_{n \to \infty} |(nh_n)^{1/2}\Delta x| < \infty$, the signal and noise are of the same magnitude, which results in inconsistency of the FRD estimator and weak identification.

Thus, similarly to the weak IVs literature (Staiger and Stock, 1997), it is appropriate to model weak identification by assuming that $\Delta x$ is inversely related to the square root of the sample size. However, the kernel estimation framework and presence of the bandwidth, which is chosen by the econometrician, require some adjustments. Suppose one models weak identification as $\Delta x \sim 1/(ng_n)^{1/2}$, for some sequence $g_n \to 0$ as $n \to \infty$. In this case, the econometrician can obtain consistency of $\hat{\beta}_n$ and resolve weak identification simply by choosing $h_n$ so that $h_n/g_n \to \infty$. Hence, the worst case scenario, in which the econometrician cannot resolve weak identification by tweaking the bandwidth, occurs when $g_n = h_n$, i.e. $\Delta x \sim 1/(nh_n)^{1/2}$.

This idea can be formalized using the results obtained in the recent literature on uniform size properties of tests and confidence sets: Andrews and Guggenberger (2010), Andrews and Cheng (2012), and Andrews et al. (2011). The latter paper provides a general framework of establishing uniform size properties of tests and confidence sets. To describe this framework, let $S_n$ be a test statistic with exact finite-sample distribution (in a sample of size $n$) determined by $\lambda \in \Lambda$. Note that $\lambda$ may include infinite dimensional components such as distribution functions. Let $cr_n(\alpha)$ denote a possibly data-dependent critical region for nominal significance level $\alpha$. The test rejects a null hypothesis when $S_n \in cr_n(\alpha)$, and the rejection probability

---

10This situation resembles so-called nearly-weak or semi-strong identification, see Hahn and Kuestersteiner (2002), Caner (2009), Antoine and Renault (2009, 2012), and Antoine and Lavergne (forthcoming).
is given by

$$RP_n(\lambda) = P_\lambda(S_n \in cr_n(\alpha)),$$

where subscript $\lambda$ in $P_\lambda$ indicates that the probability is computed for a given value of $\lambda \in \Lambda$. The exact size is defined as $ExSz_n = \sup_{\lambda \in \Lambda} RP_n(\lambda)$. Note that $ExSz_n$ captures the maximum rejection probability for any combination of parameters $\lambda$ (the worst case scenario). In large samples, the exact size is approximated by asymptotic size:

$$AsySz = \limsup_{n \to \infty} \sup_{\lambda \in \Lambda} RP_n(\lambda).$$

Contrary to the usual point-wise asymptotic approach, $AsySz$ is determined by taking supremum over the parameter space before taking limit with respect to $n$. It has been argued in many papers that controlling $AsySz$ is crucial for ensuring reliable inference when test statistics have discontinuous asymptotic distribution, i.e. when point-wise asymptotic distribution changes with parametrization. In what follows, we rely on the following result of Andrews et al. (2011):

**Lemma 3** (Andrews et al. (2011)). Let $\{d_n(\lambda) : n \geq 1\}$ be a sequence of functions, where $d_n : \Lambda \to \mathbb{R}^J$. Define $D = \{d \in \mathbb{R} \cup \{\pm \infty\}^J : d_{p_n}(\lambda_{p_n}) \to d \text{ for some subsequence } \{p_n\} \text{ of } \{n\} \text{ and some sequence } \{\lambda_{p_n} \in \Lambda\}\}$. Suppose that for any subsequence $\{p_n\}$ of $\{n\}$ and any sequence $\{\lambda_{p_n} \in \Lambda\}$ for which $d_{p_n}(\lambda_{p_n}) \to d \in D$, we have that $RP_{p_n}(\lambda_{p_n}) \to RP(d)$ for some function $RP(d) \in [0, 1]$. Then, $AsySz = \sup_{d \in D} RP(d)$.

---

$^{11}$On the importance of uniform size, see for example Imbens and Manski (2004, p. 1848), Mikusheva (2007), and references in Andrews et al. (2011).

$^{12}$Lemma 3 combines Assumption B and Theorems 2.1 and 2.2 in Andrews et al. (2011).
To apply the above lemma, we define:

\[ \lambda_1 = \left( \frac{f_z(z_0)}{k} \right)^{1/2} \frac{|\Delta x|}{\sigma_x}, \quad \lambda_2 = \rho_{xy}, \quad \lambda_3 = \beta \sigma_x / \sigma_y. \tag{7} \]

We define \( \lambda_4 = F \), where \( F \) is the joint distribution of \( x_i, y_i, z_i \) and is such that, given \( \lambda_1 \in \mathbb{R}_+ \), \( \lambda_2 \in [-\bar{\rho}, \bar{\rho}] \), and \( \lambda_3 \in \mathbb{R} \), the three equations in (7) hold. Note that \( \lambda_4 \) is an infinite-dimensional parameter that depends on \( \lambda_1, \lambda_2, \) and \( \lambda_3 \). As explained in Andrews et al. (2011, pp. 8-9), \( d_n(\lambda) \) is chosen so that when \( d_n(\lambda_n) \) converges to \( d \in D \) for some sequence of parameters \( \{\lambda_n\} \), the test statistic converges to some limiting distribution, which might depend on \( d \). In view of (5) and (6), we therefore define:

\[ d_{n,1}(\lambda) = \sqrt{n\bar{h}_n} \lambda_1, \quad d_{n,2}(\lambda) = \lambda_2, \quad d_{n,3}(\lambda) = \lambda_3. \tag{8} \]

While \( \lambda_4 = F \) affects the finite-sample distribution of the test statistic, it does not enter its asymptotic distribution, and therefore can be dropped from \( d_n(\lambda) \) as discussed in Andrews et al. (2011, p. 8). Hence, \( D = \{\mathbb{R}_+ \cup \{+\infty\}\} \times [-\bar{\rho}, \bar{\rho}] \times \{\mathbb{R} \cup \{\pm\infty\}\} \).

Next, we describe the asymptotic size of tests for FRD based on the usual \( t \)-statistic and standard normal critical value. Let \( z_\nu \) denote the \( \nu \)-th quantile of the standard normal distribution.

**Theorem 4.** Let \((X, Y)'\) be defined as in (3), and suppose that Assumption 2 holds. Define

\[ T_{d_1,d_2,d_3} = \frac{Y - d_3X}{\sqrt{1 + \left( \frac{Y + d_3d_1}{X + d_1} \right)^2 - 2d_2 \frac{Y + d_3d_1}{X + d_1} \times \text{sign}(X + d_1)}}. \]

(a) For tests that reject \( H_0 : \beta = \beta_0 \) in favor of \( H_1 : \beta \neq \beta_0 \) when \( |T_n(\beta_0)| > z_{1-\alpha/2} \),

\[ \text{AsySz} = \sup_{d_1 \in \mathbb{R}_+ \cup \{+\infty\}, d_2 \in [0, \bar{\rho}], d_3 = \mathbb{R} \cup \{\pm \infty\}} P(|T_{d_1,d_2,d_3}| > z_{1-\alpha/2}). \]
(b) For tests that reject $H_0: \beta \leq \beta_0$ in favor of $H_1: \beta > \beta_0$ when $T_n(\beta_0) > z_{1-\alpha}$,

$$\text{AsySz} = \sup_{d_1 \in \mathbb{R}_+ \cup \{+\infty\}, d_2 \in [-\bar{\rho}, \bar{\rho}], d_3 = \mathbb{R}_+ \cup \{-\infty\}} P(T_{d_1,d_2,d_3} > z_{1-\alpha}).$$

Remark. A commonly used measure of identification strength is the so-called concentration parameter. In our framework, the concentration parameter is given by $d_{n,1}$, where $d_{n,1} \to \infty$ corresponds to strong (or semi-strong) identification, and identification is weak when the limit of $d_{n,1}$ is finite. As it is apparent from the expressions for $\lambda_1$ and $d_{n,1}$ in (7) and (8), the concentration parameter and, therefore, the strength of identification depend not only on the size of discontinuity in treatment assignment $\Delta x$, but also on $f_z(z_0)$, the PDF of the assignment variable at $z_0$. Hence, smaller values of $f_z(z_0)$ would correspond to a more severe weak identification problem.

For any permitted values of $d_2$ and $d_3$, when $d_1 = \infty$ we have $T_{\infty,d_2,d_3} \sim N(0,1)$. Thus, the asymptotic size of tests based on $T_n(\beta_0)$ is equal to nominal size $\alpha$ under strong or semi-strong identification. When $d_1 < \infty$, it is straightforward to compute $\text{AsySz}$ numerically. To compute asymptotic rejection probabilities given $d_1, d_2, d_3$, first using bivariate normal PDFs one integrates numerically $1(|T_{d_1,d_2,d_3}| > z_{1-\alpha/2})$ or $P(T_{d_1,d_2,d_3} > z_{1-\alpha})$ calculated for different realized values of $Y, X$. Rejection probabilities then can be numerically maximized over $d'$s.

Table 1 reports maximal rejection probabilities of one- and two-sided tests based on the usual $t$-statistic. It shows that $\text{AsySz}$ approaches one as the concentration parameter approaches zero. Size distortions decrease monotonically as the concentration parameter increases. To avoid size distortions, in the case of two-sided testing the concentration parameter has to be of order $d_1^2 \geq 10^2$ for asymptotic 5% tests, and

---

13 On the importance of the concentration parameter in IV estimation, see for example, Stock and Yogo (2005).

14 The rejection probabilities reported in Table 1 were computed by numerical integration using quad2d function in Matlab. Integration bounds for normal variables were set to $[-7, 7]$, and the rejection probabilities were maximized over the following grids of values: $d_2 \in \{-0.99, -0.9, -0.8, \ldots, 0.9, 0.99\}$ and $d_3 \in \{-1000, -100, -10, -9, \ldots, 10, 100, 1000\}$. 

16
Table 1: Maximal asymptotic rejection probabilities for different values of the concentration parameter \((d_1^2)\) of one- and two-sided \(t\)-tests for FRD with significance level \(\alpha\), and non-central \(\chi^2\) critical values for testing hypotheses about the concentration parameter at significance level \(\tau\).

<table>
<thead>
<tr>
<th>(d_1^2)</th>
<th>(\alpha = 0.05)</th>
<th>(\alpha = 0.01)</th>
<th>(\alpha = 0.05)</th>
<th>(\alpha = 0.01)</th>
<th>(\tau = 0.05)</th>
<th>(\tau = 0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10(^{-4})</td>
<td>0.901</td>
<td>0.883</td>
<td>0.892</td>
<td>0.877</td>
<td>3.84</td>
<td>6.64</td>
</tr>
<tr>
<td>0.01</td>
<td>0.690</td>
<td>0.635</td>
<td>0.664</td>
<td>0.612</td>
<td>3.88</td>
<td>6.70</td>
</tr>
<tr>
<td>0.25</td>
<td>0.363</td>
<td>0.283</td>
<td>0.322</td>
<td>0.260</td>
<td>4.76</td>
<td>8.08</td>
</tr>
<tr>
<td>1.0</td>
<td>0.221</td>
<td>0.152</td>
<td>0.185</td>
<td>0.133</td>
<td>7.00</td>
<td>11.06</td>
</tr>
<tr>
<td>4.0</td>
<td>0.143</td>
<td>0.086</td>
<td>0.112</td>
<td>0.070</td>
<td>13.28</td>
<td>18.72</td>
</tr>
<tr>
<td>9.0</td>
<td>0.119</td>
<td>0.062</td>
<td>0.088</td>
<td>0.049</td>
<td>21.57</td>
<td>28.37</td>
</tr>
<tr>
<td>16.0</td>
<td>0.105</td>
<td>0.050</td>
<td>0.075</td>
<td>0.037</td>
<td>31.87</td>
<td>40.03</td>
</tr>
<tr>
<td>25.0</td>
<td>0.096</td>
<td>0.042</td>
<td>0.066</td>
<td>0.031</td>
<td>44.15</td>
<td>53.67</td>
</tr>
<tr>
<td>36.0</td>
<td>0.090</td>
<td>0.037</td>
<td>0.060</td>
<td>0.026</td>
<td>58.45</td>
<td>69.34</td>
</tr>
<tr>
<td>49.0</td>
<td>0.090</td>
<td>0.033</td>
<td>0.055</td>
<td>0.023</td>
<td>74.73</td>
<td>86.98</td>
</tr>
<tr>
<td>64.0</td>
<td>0.081</td>
<td>0.030</td>
<td>0.052</td>
<td>0.020</td>
<td>93.03</td>
<td>106.63</td>
</tr>
<tr>
<td>81.0</td>
<td>0.078</td>
<td>0.028</td>
<td>0.051</td>
<td>0.018</td>
<td>113.31</td>
<td>128.28</td>
</tr>
<tr>
<td>10(^2)</td>
<td>0.075</td>
<td>0.026</td>
<td>0.050</td>
<td>0.017</td>
<td>135.60</td>
<td>151.94</td>
</tr>
<tr>
<td>25(^2)</td>
<td>0.061</td>
<td>0.016</td>
<td>0.050</td>
<td>0.011</td>
<td>709.96</td>
<td>746.72</td>
</tr>
<tr>
<td>50(^2)</td>
<td>0.056</td>
<td>0.013</td>
<td>0.050</td>
<td>0.010</td>
<td>2667.17</td>
<td>2738.06</td>
</tr>
</tbody>
</table>

\(d_1^2 \geq 50^2\) for asymptotic 1% tests. The table also shows that one-sided tests suffer from more substantial size distortions than two-sided tests, which is due to asymmetries in the distribution of \(T_{d_1,d_2,d_3}\). For example, with the concentration parameter exceeding 50\(^2\), both 5% and 1% one-sided tests still exhibit some (very small) size distortions.

### 2.3 Testing for weak identification

Table 1 can be used for testing the null hypothesis of weak identification against the alternative or strong identification following the approach of Stock and Yogo (2005). Suppose that the econometrician decides that identification is strong enough.
if, in the case of 1% one-sided testing, the maximal rejection probability does not exceed 5%. Thus, the econometrician effectively adopts tests with 5% significance level. According to the results in Table 1, the corresponding null hypothesis of weak identification and its alternative of strong identification in this case are $H_0^W : d_1^2 \leq 16$ and $H_1^S : d_1^2 > 16$ respectively. A test of $H_0^W$ can be based on the estimator of discontinuity $\Delta x$. Define

$$F_n = \frac{n h_n (\hat{\Delta} x_n)^2}{\hat{\sigma}_{x,n}^2 k / \hat{f}_{z,n}(z_0)} = ((\Delta X_n / \sigma_x) + d_{n,1})^2 + o_p(1).$$  (9)

As long as the concentration parameter is finite, $F_n \to_d \chi^2_1(d_1^2)$, a non-central $\chi^2_1$ distribution with non-centrality parameter $d_1^2$. Let $\chi^2_{1,1-\tau}(d_1^2)$ denote the $(1 - \tau)$-th quantile of the $\chi^2_1(d_1^2)$ distribution. Since size distortions are monotonically decreasing when the concentration parameter increases, an asymptotic size $\tau$ test of $H_0^W$ should reject it when $F_n > \chi^2_{1,1-\tau}(d_1^2)$.

Non-central $\chi^2_1$ critical values are reported in the last two columns of Table 1 for selected values of the concentration parameter and $\tau = 0.05, 0.01$. According to those values, $H_0^W : d_1^2 \leq 16$ should be rejected in favor of $H_1^S : d_1^2 > 16$ by a 5% test when $F_n > 31.87$. In the case of two-sided testing for FRD at 5% significance level, one needs the concentration parameter of at least $10^2$ to ensure that there are no size distortions. In that case, a 5% test should reject the null hypothesis of weak identification if $F_n > 135.6$.

Note that the critical values in Table 1 substantially exceed the rule-of-thumb of 10, which is often used in the literature as a threshold value for weak IVs. According to our calculations, with an $F$-statistic of only 10, one cannot reject $H_0^W : d_1^2 \leq 1.51^2$ at 5% significance level. However, a concentration parameter of $1.51^2$ corresponds to maximal rejection probabilities of 16.6% and 13.4% for 5% one-sided and two-sided
tests respectively.

2.4 Weak-identification-robust inference for FRD

A common approach adopted in the weak IVs literature is to use weak-identification-robust statistics to test hypotheses about structural parameters directly, instead of using their estimates and standard errors. The Anderson-Rubin (AR) statistic (Anderson and Rubin, 1949; Staiger and Stock, 1997) is often used for that purpose. In the context of IV regression, the AR statistic can be used to test $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ by testing whether the null-restricted residuals computed for $\beta = \beta_0$ are uncorrelated with the instruments.

In our case, the structural parameter is defined by (1). Hence, to test $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$, following the AR approach we can test instead

$$H_0 : \Delta y - \beta_0 \Delta x = 0 \text{ against } H_1 : \Delta y - \beta_0 \Delta x \neq 0.$$ 

A test, therefore, can be based on

$$\frac{nh_n \left( \tilde{\Delta} y_n - \beta_0 \tilde{\Delta} x_n \right)^2}{k \hat{\sigma}^2_n(\beta_0)/\hat{f}_{z,n}(z_0)} = \left| T_n^R(\beta_0) \right|^2,$$

where $T_n^R(\beta_0)$ denotes a modified or null-restricted version of the usual $t$-statistic:

$$T_n^R(\beta_0) = \frac{\sqrt{nh_n} (\hat{\beta}_n - \beta_0)}{\sqrt{k \hat{\sigma}^2_n(\beta_0)/\hat{f}_{z,n}(z_0)(\tilde{\Delta} x_n)^2}},$$

and the equality holds by (5). Unlike the usual $t$-statistic, $T_n^R(\beta_0)$ uses the null-restricted value $\beta_0$ instead of $\hat{\beta}_n$ when computing the standard error. In view of the discussion at the beginning of Section 2.2 and since the asymptotic distribution of
\(|T_n^R(\beta_0)|\) does not depend on the concentration parameter, replacing \(\hat{\sigma}_n^2(\hat{\beta}_n)\) by \(\hat{\sigma}_n^2(\beta_0)\) eliminates size distortions. Thus, under Assumption 2, tests that reject \(H_0 : \beta = \beta_0\) in favor of \(H_1 : \beta \neq \beta_0\) when \(|T_n^R(\beta_0)| > z_{1-\alpha/2}\) have asymptotic size equal to \(\alpha\).

Consider now a one-sided testing problem \(H_0 : \beta \leq \beta_0\) vs. \(H_1 : \beta > \beta_0\). Again, one can base a test on the null-restricted statistic. In this case under \(H_0\) when \(\beta = \beta_0\) we have

\[
T_n^R(\beta) = \frac{\Delta Y_n - \beta \Delta X_n}{\sigma(\beta)} \times \text{sign} (\Delta X_n \pm d_{n,1} + o_p(1)).
\]

When identification is strong or semi-strong, \(d_{n,1} \to \infty\), and the sign term is constant with probability one. Since the first term is asymptotically \(N(0,1)\), \(T_n^R(\beta)\) is also asymptotically \(N(0,1)\), and one could use standard normal critical values. On the other hand, when identification is weak and the concentration parameter is small, the sign term is random, and therefore, the null asymptotic distribution of the product differs standard normal. To obtain an asymptotically uniformly valid test, one can use data-dependent critical values that automatically adjust to the strength of identification. Such critical values can be generated using the approach of Moreira (2001, 2003) by conditioning on a statistic that is i) asymptotically independent of \(\Delta Y_n - \beta \Delta X_n\), and ii) summarizes the information on the strength of identification.\(^{15}\)

Define

\[
S_n = \frac{\Delta Y_n - \beta \Delta X_n}{\sigma(\beta)} \quad \text{and} \quad Q_n = \frac{\Delta X_n}{\sigma_x} - \frac{\sigma_{xy} - \beta \sigma_x^2}{\sigma_x \sigma(\beta)} S_n,
\]

so that, when \(\beta = \beta_0\),

\[
T_n^R(\beta) = S_n \times \text{sign} (Q_n \pm d_{n,1} + \sigma_{xy} - \beta \sigma_x^2 S_n) + o_p(1).
\]

\(^{15}\)See also Andrews et al. (2006) and Mills et al. (2013).
When identification is weak, $S_n$ and $Q_n$ are asymptotically independent by construction, while $S_n \xrightarrow{d} N(0,1)$. Therefore one can construct data-dependent critical values as follows. First, compute

$$
\dot{Q}_n(\beta_0) = \frac{\sqrt{nh_n} \Delta x_n}{\sqrt{k \hat{\sigma}_{x,n}^2 / \hat{f}_{z,n}(z_0) - \dot{\sigma}_{xy,n} - \beta_0 \hat{\sigma}_{x,n}^2}} \left( \frac{\sqrt{nh_n} \left( \Delta y_n - \beta_0 \Delta x_n \right)}{\sqrt{k \hat{\sigma}_n^2(\beta_0) / \hat{f}_{z,n}(z_0)}} \right).
$$

Second, simulate $M$ independent $N(0,1)$ random variables $\{S_1, \ldots, S_M\}$ for some large $M$. Third, for $m = 1, \ldots, M$ compute

$$
\dot{T}_{n,m}^R(\beta_0, \dot{Q}_n(\beta_0)) = S_m \times \text{sign} \left( \dot{Q}_n(\beta_0) + \frac{\dot{\sigma}_{xy,n} - \beta_0 \hat{\sigma}_{x,n}^2}{\hat{\sigma}_{x,n} \hat{\sigma}_n(\beta_0)} S_m \right).
$$

Let $\hat{c}v_{n,1-\alpha}(\beta_0, \dot{Q}_n(\beta_0))$ denote the $(1 - \alpha)$-th quantile of the sample distribution of $\{\dot{T}_{n,m}^R(\beta_0, \dot{Q}_n(\beta_0)) : m = 1, \ldots, M\}$. To obtain an asymptotically uniformly valid one-sided test, one can use $\hat{c}v_{n,1-\alpha}(\beta_0, \dot{Q}_n(\beta_0))$ as the critical value.

**Theorem 5.** Suppose that Assumption 2 holds. Tests that reject $H_0 : \beta \leq \beta_0$ in favor of $H_1 : \beta > \beta_0$ when $T_n^R(\beta_0) = \hat{c}v_{n,1-\alpha}(\beta_0, \dot{Q}_n(\beta_0))$ have asymptotic size equal to $\alpha$.

Weak-identification-robust confidence sets for $\beta$ can be constructed by inversion of the robust tests. For example, a confidence set for $\beta$ with asymptotic coverage probability $1 - \alpha$ can be constructed by collecting all values $\beta_0$ that cannot be rejected by the two-sided robust test:

$$
CS_{1-\alpha,n} = \{ \beta_0 \in \mathbb{R} : \| T_n^R(\beta_0) \| \leq z_{1-\alpha/2} \}.
$$

This confidence set can be easily computed analytically by solving for values of $\beta_0$.
that satisfy the inequality

$$(\hat{\beta}_n - \beta_0)^2 \hat{\sigma}^2 x,n F_n - z_{1-\alpha/2}^2 (\hat{\sigma}^2 y,n + \beta_0^2 \hat{\sigma}^2 x,n - 2\hat{\sigma} xy,n \beta_0) \leq 0,$$  

(11)

where $F_n$ is defined in (9).

Depending on the coefficients of the second-order polynomial (in $\beta_0$) in equation (11), $CS_{1-\alpha,n}$ can take one of the following forms: i) an interval, ii) a union of two disconnected half-lines $(-\infty, a_1] \cup [a_2, \infty)$, where $a_1 < a_2$, or iii) the entire real line. One will see cases ii) or iii) if the coefficient on $\beta_0^2$ in (11) is negative, which occurs when

$$F_n - z_{1-\alpha/2}^2 < 0.$$  

(12)

Thus, in practice one will see non-standard confidence sets if the null hypothesis $\Delta x = 0$ cannot be rejected using the $F_n$ statistic and central $\chi^2_{1,1-\alpha}$ critical values. Case iii) arises when the discriminant of the quadratic polynomial in (11) is negative, which occurs if

$$F_n \hat{\sigma}^2 n (\hat{\beta}_n) - z_{1-\alpha/2}^2 (\hat{\sigma}^2 y,n - \hat{\sigma}^2 xy,n / \hat{\sigma}^2 x,n) < 0.$$  

(13)

Positive definiteness of the variance-covariance matrix composed of $\hat{\sigma}^2 x,n$, $\hat{\sigma}^2 y,n$, and $\hat{\sigma} xy,n$ implies that (12) holds whenever (13) holds. Thus, negative discriminants implied by (13) are inconsistent with $F_n > z_{1-\alpha/2}^2$ or positive coefficients on $\beta_0^2$ in (11). This in turn implies that $CS_{1-\alpha,n}$ cannot be empty.

When identification is strong or semi-strong, the concentration parameter and, therefore, $F_n$ diverge to infinity. In such cases, both the discriminant and the coefficient on $\beta_0^2$ tend to be positive, and consequently, $CS_{1-\alpha,n}$ will be an interval with probability approaching one.

Furthermore, one can show that when identification is strong and under local
alternatives of the form $\beta = \beta_0 + \mu / (nh_n)^{1/2}$, tests based on $T_n(\beta_0)$ and $T_n^R(\beta_0)$ have the same asymptotic power. Thus, in practice there is no loss of asymptotic power from adopting the robust inference approach if identification is strong.

3 Testing for constancy of the RD effect across covariates

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 4 we use this test to argue that the effect of class sizes on educational achievements is different for secular and religious schools, and therefore it might be optimal to implement different rules concerning class sizes in those two categories of schools.\footnote{The problem is related to the classical ANOVA hypothesis of homogeneous populations (see, for example, Casella and Berger, 2002, Chapter 11).}

Similarly to Otsu and Xu (2011), we consider the RD effect conditional on some covariate $w_i$\footnote{See also Frölich (2007).}. Let $W$ denote the support of the distribution of $w_i$. Next, for $w \in W$ we define $y^+(w)$ using the conditional expectation given $z_i$ and $w_i = w$:

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

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

$$\beta(w) = (y^+(w) - y^-(w)) / (x^+(w) - x^-(w)).$$
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 the null hypothesis of constancy of the RD effect

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

against a general alternative $H_1 : \beta(w) \neq \beta(v)$ 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 $H_0$.\footnote{Such a test can be constructed similarly to the ANOVA $F$-test as in Casella and Berger (2002, Chapter 11), and is discussed in the supplement.} However, this approach can be unreliable if identification is weak. We therefore take an alternative approach. Suppose that $\mathcal{W} = \{\bar{w}^1, \ldots, \bar{w}^J\}$, i.e. the covariate is categorical and divides the population into $J$ 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 $j = 1, \ldots, J$, let $\hat{y}_n^+ (\bar{w}^j)$, $\hat{y}_n^- (\bar{w}^j)$, $\hat{x}_n^+ (\bar{w}^j)$, and $\hat{x}_n^- (\bar{w}^j)$ denote the local linear estimators of the corresponding population terms computed using only the observations with $w_i = \bar{w}^j$. Let $n_j$ be the number of such observations. $\sigma^2_y(\bar{w}^j)$, $\sigma^2_x(\bar{w}^j)$ and $\sigma_{xy}(\bar{w}^j)$ are defined as the conditional versions of the corresponding population terms, and $\hat{\sigma}^2_{y,n}(\bar{w}^j)$, $\hat{\sigma}^2_{x,n}(\bar{w}^j)$, and $\hat{\sigma}_{xy,n}(\bar{w}^j)$ denote the corresponding estimators.

Suppose that Assumption 2 holds for each of the $J$ categories, and none of the categories is redundant asymptotically: $n_j h_{n_j} / (n h_n) \to p_j > 0$ for $j = 1, \ldots, J$, where $n = \sum_{j=1}^J n_j$. If $H_0$ is true and the FRD effect is independent of $w$, one can construct
a robust confidence set for the common effect:

$$CS_{1-\alpha,n}^d = \{ \beta_0 \in \mathbb{R} : G_n(\beta_0) \leq \chi_{J,1-\alpha}^2 \},$$

where

$$G_n(\beta_0) = \sum_{j=1}^J \frac{n_j h_{n_j} \left( \hat{\beta}_n(\check{w}^j) - \beta_0 \right)^2}{k \hat{\sigma}^2_n(\beta_0, \check{w}^j)/(\hat{f}_{z,n}(z_0|\check{w}^j)(\Delta x_n(\check{w}^j))^2)},$$

where \( \hat{\beta}_n(\check{w}^j) = \Delta y_n(\check{w}^j)/\Delta x_n(\check{w}^j) \), \( \Delta x_n(\check{w}^j) = \hat{x}_n^+ - \hat{x}_n^- \); \( \hat{f}_n(\check{w}^j) \) is defined similarly to \( \hat{\sigma}^2_n(\beta_0, \check{w}^j) \) in (4) using the estimators conditional on \( w_i = \check{w}^j \); \( \hat{f}_{z,n}(z_0|\check{w}^j) = (n_j h_{n_j})^{-1} \sum_{i=1}^n K((z_i - z_0)/h_{n_j}) \{ w_i = \check{w}^j \} \) is the estimator for \( f_z(z_0|\check{w}^j) \), which denotes the conditional density of \( z_i \) at \( z_0 \) conditional on \( w_i = \check{w}^j \).

Under \( H_0 : \beta(w) = \beta \) for some \( \beta \in \mathbb{R} \), \( CS_{1-\alpha,n}^d \) is an asymptotically valid confidence set since \( G_n(\beta) \to_d \chi_J^2 \) under weak or strong identification. We consider the following size \( \alpha \) asymptotic test: Reject \( H_0 \) if \( CS_{1-\alpha,n}^d \) is empty. The test is asymptotically valid because under \( H_0 \), \( P(CS_{1-\alpha,n}^d = \emptyset) \leq P(\beta \notin CS_{1-\alpha,n}^d) = P(G_n(\beta) > \chi_{J,1-\alpha}^2) \to \alpha \), which again holds under weak or strong identification. Under the alternative, there is no common value \( \beta \) that will provide a proper re-centering for all \( J \) categories, and therefore, one can expect deviations from the asymptotic \( \chi_J^2 \) distribution.

We show below that the test is consistent if there is strong (or semi-strong) identification for at least two values \( \check{w}^{j_1} \) and \( \check{w}^{j_2} \) that satisfy \( \beta(\check{w}^{j_1}) \neq \beta(\check{w}^{j_2}) \). Let

$$d_{n,1}^2(\check{w}^j) = n_j h_{n_j} |x^+(\check{w}^j) - x^-(\check{w}^j)|^2 f_z(z_0|\check{w}^j)/(k \hat{\sigma}^2_x(\check{w}^j))$$

be the conditional version of the concentration parameter.

**Theorem 6.** Suppose that \( n_j h_{n_j}/(nh_n) \to p_j > 0 \) and Assumption 2 holds for each \( j = 1, \ldots, J \). Let \( \mathcal{W}^* = \{ \check{w}^1, \ldots, \check{w}^J \} \subset \mathcal{W} \) be such that \( d_{n,1}^2(\check{w}^j) \to \infty \) for \( \check{w}^j \in \mathcal{W}^* \) and \( \beta(\check{w}^{j_1}) \neq \beta(\check{w}^{j_2}) \) for some \( \check{w}^{j_1}, \check{w}^{j_2} \in \mathcal{W}^* \). Then, \( P(CS_{1-\alpha,n}^d = \emptyset) \to 1 \) as \( n \to \infty \).
4 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 standard method and our proposed method yield significantly different conclusions when weak identification is a problem, but similar results when it is not. We also show that the robust confidence sets can provide more informative answers than the standard confidence intervals in cases when the usual assumptions are violated. We also apply our weak identification robust constancy test.

We begin with a case where weak identification is not a serious issue. In an influential paper, Angrist and Lavy (1999) study the effect of class size on academic success in Israel. During their sample period, class size in Israeli public schools was capped at 40 students. This cap results in discontinuities in the relationship between class size and total school enrollment (for a given grade). In practice, school enrollment does not perfectly predict class size and thus the appropriate design is fuzzy rather than sharp.

The 1991 data set was compiled by Angrist and Lavy (1999) and consists of 4th and 5th grade class sizes, school enrollment and normalized class average verbal and mathematical achievement exam scores.\footnote{The data can be found at http://econ-www.mit.edu/faculty/angrist/data1/data/anglavy99.} We use the same sample selection rules as Angrist and Lavy (1999) and focus on language scores among 4th graders.\footnote{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). We use the uniform kernel in all cases.}

Figure 1 plots the observed values of class size as a function of enrollment. There is a clear discontinuity in the relationship between class size and enrollment at the cutoff value (40 students). Table 2 shows that the estimated discontinuity in the
Figure 1: Angrist and Lavy (1999): Empirical relationship between class size and school enrollment

Note: The solid line show the relationship when Maimonides’ rule (cap of 40 students) is strictly enforced.

treatment variable (the estimate of strength of identification) ranges from 8 to 14 students depending on the bandwidth chosen. The table also shows that, as expected, the F-statistic becomes smaller as the bandwidth gets smaller. Silverman’s normal rule of thumb and the optimal bandwidth procedure of Imbens and Kalyanaraman (2012) both suggest a bandwidth value of approximately 8, which corresponds to a relatively large value of the F-statistic (approximately 62). Applying the results from Table 1, we then conclude that weak identification is not a serious concern in this application. Using the 5% non-central \( \chi^2 \) critical value, we reject the null hypothesis that the concentration parameter is below 36, and therefore, the maximal size distortions of the 5% two-sided tests are expected to be under 1%. Note that even at the smallest bandwidth, the F-statistic is relatively large. This is consistent with Figure 2 which shows that the 95% standard and robust confidence sets for the class
size effect are very similar. The figure shows that the two sets of confidence intervals are essentially indistinguishable for larger bandwidths, and only differ slightly for smaller bandwidths.

Table 2: Angrist and Lavy (1999): Estimated discontinuity in the treatment variable 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>discont.estimates (standard errors)</th>
<th>F-Statistic</th>
<th>estimated effect of class size</th>
<th>standard CS</th>
<th>robust CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>-8.4040 (1.6028)</td>
<td>27.49</td>
<td>-0.0687</td>
<td>[-0.1440, 0.0066]</td>
<td>[-0.1702, -0.0003]</td>
</tr>
<tr>
<td>8</td>
<td>-9.9013 (1.2585)</td>
<td>61.90</td>
<td>-0.0722</td>
<td>[-0.1294, -0.0150]</td>
<td>[-0.1381, -0.0186]</td>
</tr>
<tr>
<td>10</td>
<td>-10.8283 (1.0314)</td>
<td>110.22</td>
<td>-0.056</td>
<td>[-0.0991, -0.0130]</td>
<td>[-0.1027, -0.0146]</td>
</tr>
<tr>
<td>12</td>
<td>-11.9974 (0.9149)</td>
<td>171.95</td>
<td>-0.0229</td>
<td>[-0.0562, 0.0105]</td>
<td>[-0.0581, 0.00970]</td>
</tr>
<tr>
<td>14</td>
<td>-12.6167 (0.7843)</td>
<td>258.78</td>
<td>-0.0301</td>
<td>[-0.0605, 0.0002]</td>
<td>[-0.0618, -0.0002]</td>
</tr>
<tr>
<td>16</td>
<td>-13.2053 (0.6864)</td>
<td>370.12</td>
<td>-0.0200</td>
<td>[-0.0475, 0.0075]</td>
<td>[-0.0486, 0.00710]</td>
</tr>
<tr>
<td>18</td>
<td>-13.8684 (0.6048)</td>
<td>525.81</td>
<td>-0.0212</td>
<td>[-0.0459, 0.0034]</td>
<td>[-0.0468, 0.00310]</td>
</tr>
<tr>
<td>20</td>
<td>-14.3463 (0.5552)</td>
<td>667.70</td>
<td>-0.0190</td>
<td>[-0.0424, 0.0045]</td>
<td>[-0.0434, 0.00420]</td>
</tr>
</tbody>
</table>

*Note: Silverman’s normal rule-of-thumb bandwidth is 7.84 and the optimal bandwidth suggested by Imbens and Kalyanaraman (2012) is 7.90. The scores are given in terms of standard deviations from the mean.*
Figure 2: Angrist and Lavy (1999): 95% confidence intervals for the effect of class size on verbal test scores for different values of the bandwidth

Note: This figure is for the enrollment cut-off of 40. The bandwidth according to Silverman’s normal rule-of-thumb is 7.94. The optimal bandwidth selected according to Imbens and Kalyanaraman (2012) is 7.90. The scores are given in terms of standard deviations from the mean.

In this application we also compare the results of the standard constancy test of the treatment effect across sub-groups to the results of our robust constancy test. The first set of results reported in Table 3 compare the treatment effect for secular and religious schools. The null hypothesis (the treatment effect is the same across subgroups) can never be rejected using a standard test. By contrast, the robust constancy test rejects the null hypothesis for the largest values of the bandwidth (18 and 20). We reach similar conclusions when comparing the treatment effect for schools with above and below median proportions of disadvantaged students. The null hypothesis is rejected by the robust test under the largest bandwidth (20). This suggests that our proposed test may have greater power against alternatives than the standard test in some contexts.
Table 3: 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>

The second application considers a similar policy in Chile originally studied by Urquiola and Verhoogen (2009).\textsuperscript{21} In this application, the discontinuity in the probability of being assigned to a smaller class is observed when enrollment in a given grade passes 45 students. Figure 3 shows the discontinuity in the empirical relationship be-

\textsuperscript{21}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 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.
tween class size and enrollment at the various multiples of 45. The figure shows that the appropriate design is fuzzy since the observed pattern in class size does not strictly follow the official rule and the discontinuity becomes smaller as enrollment increases. 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 (2009).\footnote{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 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. The unform kernel is used to compute all the results below.}

Figure 3: Urquiola and Verhoogen (2009): Empirical relationship between class size and enrollment

Note: The solid line show the relationship when the rule (cap of 45 students) is strictly enforced.

Table 4 shows that in the case of the first cutoff (45), the discontinuity in enrollment as a function of class size is large at larger bandwidths, but smaller and consistent with severe weak identification for bandwidths smaller than 10. This is
important since the bandwidth suggested by Silverman’s normal rule-of-thumb is only 8.59 and the optimal bandwidth suggested by Imbens and Kalyanaraman (2012) is 9.67.

Table 4: Urquiola and Verhoogen (2009): Estimated discontinuity in the treatment variable for the first cutoff and statistics for testing identification strength ($F_n$) for various values of the bandwidth

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Discontinuity Estimates</th>
<th>$F_n$-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1.388</td>
<td>0.8226</td>
</tr>
<tr>
<td>8</td>
<td>-0.387</td>
<td>0.0812</td>
</tr>
<tr>
<td>10</td>
<td>-3.107</td>
<td>6.8069</td>
</tr>
<tr>
<td>12</td>
<td>-4.779</td>
<td>20.684</td>
</tr>
<tr>
<td>14</td>
<td>-6.092</td>
<td>41.037</td>
</tr>
<tr>
<td>16</td>
<td>-7.870</td>
<td>84.236</td>
</tr>
<tr>
<td>18</td>
<td>-8.934</td>
<td>129.80</td>
</tr>
<tr>
<td>20</td>
<td>-9.968</td>
<td>188.43</td>
</tr>
</tbody>
</table>

Note: Silverman’s normal rule-of-thumb is only 8.59 and the optimal bandwidth suggested by Imbens and Kalyanaraman (2012) is 9.67. The scores are given in terms of standard deviations from the mean.

Table 5 reports the FRD estimates and the confidence sets for the different values of the bandwidth and cutoff points. As before, we set the size of the test at 5%. Starting with the first cutoff point, Table 5 shows that the robust and conventional confidence sets diverge dramatically as the bandwidth gets smaller. Interestingly, while the robust confidence interval is much wider than the conventional one, it nevertheless rejects the null hypothesis that the effect of class size is equal to zero while the conventional fails to reject the null.
Table 5: 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>
</tbody>
</table>

| second cutoff (90) | | | |
| 6 | 0.128 | \([-0.025, 0.281]\) | \([0.004, 3.093]\) |
| 8 | 0.261 | \([-0.061, 0.582]\) | \((-\infty, -0.587] \cup [0.085, \infty)\) |
| 10 | 0.227 | \([-0.111, 0.566]\) | \((-\infty, -0.241] \cup [0.046, \infty)\) |
| 12 | 0.306 | \([-0.296, 0.908]\) | \((-\infty, -0.118] \cup [0.053, \infty)\) |
| 14 | 0.486 | \([-1.092, 2.063]\) | \((-\infty, -0.056] \cup [0.068, \infty)\) |
| 16 | 1.636 | \([-18.745, 22.017]\) | \((-\infty, 0.002] \cup [0.065, \infty)\) |
| 18 | -1.056 | \([-10.968, 8.856]\) | \((-\infty, \infty)\) |
| 20 | -0.425 | \([-2.041, 1.190]\) | \((-\infty, 0.005] \cup [0.162, \infty)\) |

Silverman’s rule-of-thumb bandwidth is 8.59. The optimal bandwidth suggested by Imbens and Kalyanaraman (2012) for the cut-off of 45 is 9.67 and for the cut-off of 90, the suggested bandwidth is 11.60. The scores are given in terms of standard deviations from the mean.
<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>fourth 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>

Silverman’s rule-of-thumb bandwidth is 8.59. The optimal bandwidth suggested by Imbens and Kalyanaraman (2012) for the cut-off of 135 is 14.12 and for the cut-off of 180, the suggested bandwidth is 17.81. The scores are given in terms of standard deviations from the mean.

To help interpret the results, we also graphically illustrate the difference between standard and robust confidence sets in Figure 4. 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.\textsuperscript{23}

Figure 4: 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

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{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.}
\end{figure}

Note: This figure is for the first enrollment cut-off of 45. The bandwidth according to Silverman’s normal rule-of-thumb is 8.59. The optimal bandwidth selected according to Imbens and Kalyanaraman (2012) is 9.67. The scores are given in terms of standard deviations from the mean.

Identification is considerably weaker for the second cutoff point. At all band-

\textsuperscript{23}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.
widths, 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 do not include a zero effect. For example, for a bandwidth of 8 we cannot reject the null that class size is not related to grades when using the standard method, while the robust method suggests rejecting the null.

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 and the robust confidence sets now map the entire real line. This suggests that identification 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 of zero effects 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 discontinuity in the treatment variable is not large enough, the robust confidence sets are very different from those obtained using the standard method. We also demonstrate that our robust inference method provide more informative results than the standard method is used.

Appendix

**Proof of Theorem 4.** In what follows, population parameters should be viewed as drifting sequences indexed by \( n \). Let \( d_{n,1}^* = (f_z(z_0)/k)^{1/2}(nh_n)^{1/2}\Delta x/\sigma_x \), so that \( d_{n,1} = |d_{n,1}^*| \), and re-write (6) as

\[
\frac{\sigma_x}{\sigma_y} \hat{\beta}_n = \frac{(\Delta Y_n/\sigma_y) + d_{n,3} d_{n,1}^*}{(\Delta X_n/\sigma_x) + d_{n,1}^*}.
\]
Since $\Delta y = \beta \Delta x$, we can re-write (5) as

$$T_n(\beta) = \sqrt{\frac{\hat{f}_{z,n}(z_0)}{f_z(z_0)}} \frac{(\Delta Y_n/\sigma_y) - (\sigma_x/\sigma_y)\beta(\Delta X_n/\sigma_x)}{\sqrt{\frac{\hat{g}^2_{z,n}}{\sigma_y^2} + \frac{\hat{g}^2_{z,\hat{\beta}_n}}{\sigma_x^2} (\frac{\sigma_x}{\sigma_y})^2 - 2\frac{\hat{g}_{z,n}}{\sigma_x\sigma_y} (\frac{\sigma_x}{\sigma_y})\hat{\beta}_n}} \times \text{sign}\left((\Delta X_n/\sigma_x) + d_{n,1}^*\right).$$

In addition to $d_{n,1}, d_{n,2}, d_{n,3}$, the finite-sample distribution of $T_n(\beta)$ can also be indexed by $d_{n,4} = f_z(z_0)$, where $d_{n,4}$ takes values in a compact set by Assumption 2(b)(i). However, by usual results for kernel estimators and under Assumption 2(a) and part (i) of Assumption 2(b), $(E\hat{f}_{z,n}(z_0) - f_z(z_0))/f_z(z_0) \to 0$ and $\text{Var}(\hat{f}_{z,n}(z_0)/f_z(z_0)) \to 0$ (Li and Racine, 2007, Chapter 1). It follows that $\hat{f}_{z,n}(z_0)/f_z(z_0) \to_p 1$ as $n \to \infty$, and for any subsequence $\{p_n\}$ of $\{n\}$, $\hat{f}_{z,p_n}(z_0)/d_{p_n,4} \to_p 1$. Hence, $d_{n,4}$ does not affect AsySz of $T_n(\beta)$. Similarly, $\hat{\sigma}_x/\sigma_x \to_p 1$, $\hat{\sigma}_y/\sigma_y \to_p 1$, and $\hat{\sigma}_{xy}/(\sigma_x\sigma_y) \to_p 0$. Thus,

$$T_n(\beta) = \frac{((\Delta Y_n/\sigma_y) - d_{n,3}(\Delta X_n/\sigma_x)) \text{sign}\left((\Delta X_n/\sigma_x) + d_{n,1}^*\right)}{\sqrt{1 + \frac{(\Delta Y_n/\sigma_y + d_{n,3}d_{n,1}^*)}{(\Delta X_n/\sigma_x + d_{n,1}^*)^2} - 2d_{n,2} \frac{(\Delta Y_n/\sigma_y + d_{n,3}d_{n,1}^*)}{(\Delta X_n/\sigma_x + d_{n,1}^*)}}} + o_p(1).$$

Next, let $d_{n,5} = \sigma_y$ and $d_{n,6} = \sigma_x$. By Assumption 2(b)(iii), they take values in compact sets. Suppose that for any subsequence $\{p_n\}$ of $\{n\}$,

$$(\Delta Y_{p_n}/d_{p_n,5}, \Delta X_{p_n}/d_{p_n,6}) \to_d (\mathcal{Y}, \mathcal{X}).$$

(14)

Note again that $d_{n,5}$ and $d_{n,6}$ do not affect AsySz. Suppose further that $d_{p_n,1}^* \to d_1^*$ for some $|d_1^*| < \infty$, $d_{p_n,2} \to d_2 \in [-\tilde{\rho}, \tilde{\rho}]$, and $d_{p_n,3} \to d_3$ for some $|d_3| < \infty$. In this case,

$$T_{p_n}(\beta) \to_d \frac{\mathcal{Y} - d_3\mathcal{X}}{\sqrt{1 + \frac{(\mathcal{Y} + d_3\mathcal{X})^2}{\mathcal{X}^2 + d_1^2} - 2d_2 \frac{(\mathcal{Y} + d_3\mathcal{X})}{\mathcal{X} + d_1^2}}} \times \text{sign}(\mathcal{X} + d_1^*).$$

37
If \( d_1^* < 0 \), one can multiply \( \mathcal{Y} + d_3^* d_1^* \), \( \mathcal{X} + d_1^* \), and \( \mathcal{Y} - d_3^* \mathcal{X} \) each by \(-1\), and re-define \( \mathcal{Y} \) and \( \mathcal{X} \) as their negatives without changing the resulting asymptotic distribution.  

Hence, in this case \( T_{p_n}(\beta) \to_d \mathcal{T}_{d_1,d_2,d_3} \). Note also that the distribution of \( |\mathcal{T}_{d_1,d_2,d_3}| \) is the same as that of \( |\mathcal{T}_{d_1, -d_2, -d_3}| \), and therefore, without loss of generality, one can restrict \( d_2 \) to \([0, \overline{\rho}]\) for two-sided testing.

Suppose now that \( |d_1^*| < \infty, d_2 \in [-\overline{\rho}, \overline{\rho}] \), and \( d_3 = \pm \infty \). In this case,

\[
T_n(\beta) = \sqrt{\frac{1}{d_{n,3}^*}} \frac{((\Delta Y_n/d_{n,5}d_{n,3}) - (\Delta X_n/d_{n,6})) \text{sign} ((\Delta X_n/d_{n,6}) + d_{n,1}^*)}{(\Delta Y_n/d_{n,5}d_{n,3} + d_{n,1}^*)} + o_p(1).
\]  

Therefore, \( T_{p_n}(\beta) \to_d - \mathcal{X}(\mathcal{X} + d_1)/d_1 = \mathcal{T}_{d_1,d_2,\pm \infty} \) for any \( d_2 \in [-\overline{\rho}, \overline{\rho}] \).

Next, suppose that \( |d_1^*| = \infty, d_2 \in [-\overline{\rho}, \overline{\rho}] \), and \( |d_3| < \infty \). We have

\[
T_n(\beta) = \sqrt{\frac{1}{d_{n,3}^*}} \frac{((\Delta Y_n/d_{n,5}d_{n,3}) - (\Delta X_n/d_{n,6})) \text{sign} ((\Delta X_n/d_{n,6}) + d_{n,1}^*)}{(\Delta Y_n/d_{n,5}d_{n,3} + d_{n,1}^*)} + o_p(1),
\]

and, therefore, \( T_{p_n}(\beta) \) converges in distribution to \((\mathcal{Y} - d_3^* \mathcal{X})/(1 + d_2^2 - 2d_2d_3)^{1/2} \times \text{sign}(d_1^*)\) for any \( d_2 \in [-\overline{\rho}, \overline{\rho}] \). The case of \( |d_1^*| = \infty \) and \( |d_3| = \infty \) can be handled similarly to the previous to cases with \( T_{p_n}(\beta) \to_d \mathcal{T}_{\infty,d_2,\pm \infty} \sim N(0,1) \) for any \( d_2 \in [-\overline{\rho}, \overline{\rho}] \).

The results of the theorem now follow from Lemma 3 provided that (14) holds.

To show (14), consider \( \hat{y}_{i1}^+ \) first. As in Hahn et al. (1999, Lemma 2), write

\[
\sqrt{nh_n} \begin{pmatrix} \hat{y}_{i1}^+ - y^+ \\ h_n(\hat{y}_{i1}^{(1),+} - y^{(1),+}) \end{pmatrix} = \left( \frac{1}{nh_n} \sum_{i=1}^{n} Z_i Z_i' K_i \right)^{-1} \times \left( \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \xi_i + \sqrt{nh_n} E y_i^* Z_i K_i \right),
\]  

where \( E = \mathbb{E}(\epsilon_i | Z_i) \) is estimated by the empirical mean, \( \hat{y}_i^* = \sum_{i=1}^{n} \frac{Z_i y_i}{n} \).
where $y^{(1),+} = \lim_{\epsilon \to 0} dE(y_i|z_i = z_0 + \epsilon)/dz_i$, $\hat{y}_n^{(1),+}$ denotes the estimator of $y^{(1),+}$, $Z'_i = (1, (z_i - z_0)/h_n)$, $K_i = K((z_i - z_0)/h_n)1\{z_i \geq z_0\}$, and $\hat{y}_i = y_i - y^+ - y^{(1),+}(z_i - z_0)/h_n$, and

$$
\xi_{ni} = y_i^* Z_i K_i - E y_i^* Z_i K_i.
$$

Hahn et al. (1999) show that $E y_i^* Z_i K_i = h_n^2 f_z(z_0)(\lim_{\epsilon \to 0} d^2 E(y_i|z_i = z_0 + \epsilon)/dz_i^2) \times (k_1 + o(1))$, where $k_1$ is a vector of constants depending only on $K(\cdot)$, and the second derivative of $E(y_i|z_i = z)$ are bounded in the neighborhood of $z_0$ by Assumption 2(b)(i)-(ii), it follows from Assumption 2(c) that $\sqrt{p_n h_n} E y_i^* Z_i K_i \to 0$ for all subsequences $\{p_n\}$ of $\{n\}$. Similarly, since the variances are bounded from below by Assumption 2(b)(iii),

$$
\left( \text{Var} \left( \frac{1}{\sqrt{p_n h_n}} \sum_{i=1}^{n} \xi_{ni} \right) \right)^{-1/2} \sqrt{p_n h_n} E y_i^* Z_i K_i \to 0.
$$

By Lyapounov’s CLT (see for example, Lehmann and Romano, 2005, Corollary 11.2.1, p. 427) and the Cramér-Wold device (Davidson, 1994, Theorem 25.5, p. 405),

$$
\left( \text{Var} \left( \frac{1}{\sqrt{p_n h_n}} \sum_{i=1}^{n} \xi_{ni} \right) \right)^{-1/2} \frac{1}{\sqrt{p_n h_n}} \sum_{i=1}^{n} \xi_{ni} \to_d N(0, 1),
$$

where Lyapounov’s condition can be verified by Theorem 23.12 on p. 373 in Davidson (1994) using Assumption 2(b)(iv). Uniform positive definiteness of the variance-covariance matrix, which is needed to apply the Cramér-Wold device, holds because $\sigma^2_y(z_i)$ is bounded away from zero around $z_0$ by Assumption 2(b)(iii), and by Lemma 4 in Hahn et al. (1999). Let

$$
\Omega_n = \left( \frac{1}{nh_n} \sum_{i=1}^{n} Z_i Z'_i K_i \right)^{-1} \text{Var} \left( \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \xi_{ni} \right) \left( \frac{1}{nh_n} \sum_{i=1}^{n} Z_i Z'_i K_i \right)^{-1}.
$$

39
By (16)-(18),

\[ \Omega_{pn}^{-1/2} \sqrt{p_n h_{pn}} \begin{pmatrix} \hat{y}_{pn}^+ - y^+ \\ h_{pn} (\hat{y}_{pn}^{(1,+)} - y^{(1,+)} ) \end{pmatrix} \rightarrow_d N(0, 1). \]

Now, by Lemmas 1 and 4 in Hahn et al. (1999) and in view of Assumption 2, we conclude that

\[ \sqrt{p_n h_{pn}} (\hat{y}_{pn}^+ - y^+) / (\sigma_y^+ \sqrt{k/f_z(z_0)}) \rightarrow_d N(0, 1), \]

where \( \sigma_y^+ = \lim_{e \downarrow 0} \sigma_y(z_0 + e) \).

Let \( d_{n,7} = \sigma_y^+, d_{n,8} = \sigma_y^- = \lim_{e \downarrow 0} \sigma_y(z_0 - e) \), \( \Delta Y_n^+ = \sqrt{n h_{pn} / (k/f_z(z_0))}(\hat{y}_{pn}^+ - y^+) \), and let \( \Delta Y_n^- \) be defined similarly with the plus-terms replaced with the minus-terms.

Using the same arguments as above and applying the Cramér-Wold device, we can show that

\[ (\Delta Y_{pn}^+/d_{pn,7}, \Delta Y_{pn}^-/d_{pn,8}) \rightarrow_d (Y^+, Y^-) \]

where \( Y^+, Y^- \) are independent standard normal random variables. Next,

\[ \frac{\Delta Y_{pn}}{d_{pn,5}} = \frac{\Delta Y_{pn}^+}{d_{pn,7}} \frac{d_{pn,7}}{d_{pn,5}} + \frac{\Delta Y_{pn}^-}{d_{pn,8}} \frac{d_{pn,8}}{d_{pn,5}}. \]

Now, \( \Delta Y_{pn}/d_{pn,5} \rightarrow_d Y \) in (14) can be argued using (19), Assumption 2(b)(iii), and Lemma 3.

Lastly, the joint convergence in (14) can be shown using the same arguments as above in combination with the Cramér-Wold device applied to \( y \)- and \( x \)-terms. Note that since \( |\rho_{xy}| \) is bounded away from one by Assumption 2(b)(iii), the variance-covariance matrices will be positive definite, which ensures that the Cramér-Wold device can be applied. □

**Proof of Theorem 5.** Again, population parameters should be viewed as drifting sequences indexed by \( n \). First, note that under \( H_0 \), the rejection probability is the
largest when $\beta = \beta_0$. Next, as in the proof of Theorem 4, we can write

$$T_n^R(\beta) = \frac{((\Delta Y_n/d_{n,5}) - d_{n,3}(\Delta X_n/d_{n,6})) \text{sign} ((\Delta X_n/d_{n,6}) + d_{n,1}^*)}{\sqrt{1 + d_{n,3}^2 - 2d_{n,2}d_{n,3}}} + o_p(1). \quad (20)$$

Suppose that $d_{p_n,1}^* \to \pm \infty$, and $d_{p_n,3} \to d_3$, where $|d_3| < \infty$. By (14) we have that $T_{p_n}^R(\beta) \to_d N(0,1)$. Next, similarly to (20),

$$\hat{Q}_n(\beta) = \frac{\Delta X_n}{d_{n,6}} + d_{n,1}^* - \frac{(d_{n,2} - d_{n,3}) ((\Delta Y_n/d_{n,5}) - d_{n,3}(\Delta X_n/d_{n,6}))}{1 + d_{n,3}^2 - 2d_{n,2}d_{n,3}} + o_p(1),$$

$$\hat{T}_{n,m}(\beta, \hat{Q}_n(\beta)) = S_m \times \text{sign} \left( \hat{Q}_n(\beta) + \frac{d_{n,2} - d_{n,3}}{\sqrt{1 + d_{n,3}^2 - 2d_{n,2}d_{n,3}}} S_m \right) + o_p(1). \quad (21)$$

We have that $\hat{Q}_{p_n}(\beta)$ diverges to $\pm \infty$, and $\hat{T}_{p_n,m}(\beta, \hat{Q}_{p_n}(\beta)) \to_d N(0,1)$. Hence, it follows that for all subsequences $\{p_n\}$ of $\{n\}$, $\hat{c} \nu_{p_n,1-\alpha}(\beta_0, \hat{Q}_{p_n}(\beta)) \to_p z_{1-\alpha}$, and

$$P(T_{p_n}^R(\beta) > \hat{c} \nu_{p_n,1-\alpha}(\beta_0, \hat{Q}_{p_n}(\beta))) \to \alpha. \quad (22)$$

Suppose now that $d_{p_n,1}^* \to d_1^*$, where $|d_1^*| < \infty$, and $d_{p_n,3} \to d_3$, where $|d_3| < \infty$. We can re-write (20) as

$$T_n^R(\beta) = \frac{((\Delta Y_n/d_{n,5}) - d_{n,3}(\Delta X_n/d_{n,6}))}{\sqrt{1 + d_{n,3}^2 - 2d_{n,2}d_{n,3}}} \times \text{sign} \left( \hat{Q}_n(\beta) + \frac{(d_{n,2} - d_{n,3}) ((\Delta Y_n/d_{n,5}) - d_{n,3}(\Delta X_n/d_{n,6}))}{1 + d_{n,3}^2 - 2d_{n,2}d_{n,3}} \right) + o_p(1).$$

By (14),

$$\frac{((\Delta Y_{p_n}/d_{p_n,5}) - d_{p_n,3}(\Delta X_{p_n}/d_{p_n,6}))}{\sqrt{1 + d_{p_n,3}^2 - 2d_{p_n,2}d_{p_n,3}}} \to_d S = \frac{Y - d_3 X}{\sqrt{1 + d_3^2 - 2d_2}},$$

41
\[ \hat{Q}_{p_n}(\beta) \to_d \mathcal{Q} + d_1^*, \text{ where } \]
\[ \mathcal{Q} = \mathcal{X} - \frac{d_2 - d_3}{\sqrt{1 + d_3^2 - 2d_2}} S. \]

Note that the two limiting distributions represented by \( S \) and \( \mathcal{Q} \) are independent, and \( S \sim N(0, 1) \). Hence

\[ T_{p_n}^R(\beta) \to_d S \times \text{sign} \left( \mathcal{Q} + d_1^* + \frac{d_2 - d_3}{\sqrt{1 + d_3^2 - 2d_2}} S \right). \tag{23} \]

By (21), we have

\[ \hat{T}_{p_n,m}^R(\beta, \hat{Q}_{p_n}(\beta)) \to_d S_m \times \text{sign} \left( \mathcal{Q} + d_1^* + \frac{d_2 - d_3}{\sqrt{1 + d_3^2 - 2d_2d_3}} S_m \right), \tag{24} \]

where \( S_m \sim N(0, 1) \) and is independent from \( \mathcal{Q} \) by construction. The results in (23) and (24) imply that (22) holds also for \( |d_1^*| < \infty \).

Equation (22) remains true in the case of \( d_{p_n,3} \to \pm \infty \), which can be handled as in the proof of Theorem 4, equation (15). The result of the theorem now follows by Lemma 3. \( \square \)

**Proof of Theorem 6.** Let

\[ G_n^*(\beta_0) = \sum_{j=1}^{J^*} \frac{n_j b_{n_j} \left( \hat{\beta}_n(\bar{w}^j) - \beta_0 \right)^2}{k \hat{\sigma}_n^2(\beta_0, \bar{w}^j) / \left( \hat{f}_{z,n}(z_0|\bar{w}^j)(\Delta x_n(\bar{w}^j))^2 \right)} \leq G_n(\beta_0). \]

In the proof below, we allow for \( G_n^*(b) \) to be minimized at a set of points or infinity.

Let \( \Delta x(\bar{w}^j) = x^+(\bar{w}^j) - x^-(\bar{w}^j) \), \( \sigma^2(b, \bar{w}^j) = \sigma_y^2(\bar{w}^j) + b^2 \sigma_x^2(\bar{w}^j) - 2\sigma_{xy}(\bar{w}^j) \), and

\[ G_n^*(b) = \sum_{j=1}^{J^*} \frac{p_j}{\sigma^2(b, \bar{w}^j)} \left( \beta(\bar{w}^j) - b \right)^2 (\Delta x(\bar{w}^j))^2 / f_z(z_0|\bar{w}^j). \]

42
Since under the theorem’s assumptions inf_{b \in \mathbb{R}} G^*_n(b) > 0, it suffices to show that

$$\left| \inf_{b \in \mathbb{R}} G^*_n(b)/(nh_n) - \inf_{b \in \mathbb{R}} G^*(b) \right| \rightarrow_p 0, \quad (25)$$

as (25) implies that \( P(\inf_{b \in \mathbb{R}} G^*_n(b) > a) \rightarrow 1 \) for all \( a \in \mathbb{R} \) as \( n \to \infty \). However, the last equation can be shown to establishing that

$$\sup_{b \in \mathbb{R}} |G^*_n(b)/(nh_n) - G^*(b)| \rightarrow_p 0. \quad (26)$$

Since \( (\beta(\bar{w}^j) - b)^2 \) and \( \sigma^2(b, (\bar{w}^j)) \) are continuous for all \( b \in \mathbb{R} \), and the asymptotic variance-covariance matrix composed of \( \sigma^2_y(\bar{w}^j) \), \( \sigma^2_z(\bar{w}^j) \), and \( \sigma_{xy}(\bar{w}^j) \) is positive definite, it follows that the function \( (\beta(\bar{w}^j) - b)^2/\sigma^2(b, \bar{w}^j) \) is continuous for all \( b \in \mathbb{R} \) and bounded. By the same arguments,

$$\sup_{b \in \mathbb{R}} \frac{\left( \hat{\beta}_n(\bar{w}^j) - b \right)^2}{\hat{\sigma}^2_n(b, \bar{w}^j)} = O_p(1). \quad (27)$$

By triangle inequality,

$$|G^*_n(b)/(nh_n) - G^*(b)| \leq \sum_{j=1}^{J^*} p_j (\Delta x(\bar{w}^j))^2 \times \left| \frac{(\hat{\beta}_n(\bar{w}^j) - b)^2}{\hat{\sigma}^2_n(b, \bar{w}^j)k/f_z(z_0|\bar{w}^j)} - \frac{(\beta(\bar{w}^j) - b)^2}{\sigma^2(b, \bar{w}^j)k/f_z(z_0|\bar{w}^j)} \right| + \sum_{j=1}^{J^*} |R_{j,n}(b)|, \quad (28)$$

where \( |R_{j,n}(b)| \) is bounded by

$$\left| \frac{R_{j,n}(b)}{nh_n} \right| \leq \left| \frac{\left( n_jh_{n,j} - p_j \right) (\Delta x(\bar{w}^j))^2 + (\Delta x_n(\bar{w}^j))^2 - (\Delta x(\bar{w}^j))^2}{\hat{\sigma}^2_n(b, \bar{w}^j)k/f_z(z_0|\bar{w}^j)} \right| \left( \frac{(\hat{\beta}_n(\bar{w}^j) - b)^2}{\hat{\sigma}^2_n(b, \bar{w}^j)k/f_z(z_0|\bar{w}^j)} \right).$$

43
Since $n_j h_{n_j} / n h_n \to p_j$, it follows from (27) that for $j = 1, \ldots, J^*$, $\sup_{b \in \mathbb{R}} |R_{j,n}(b)| = o_p(1)$. Similarly, one can show that, for all $j = 1, \ldots, J^*$,

$$
\sup_{b \in \mathbb{R}} \left| \frac{\left( \hat{\beta}_n(\bar{w}^j) - b \right)^2}{\hat{\sigma}^2_n(b, \bar{w}^j) k/\hat{f}_z(z_0|\bar{w}^j)} - \frac{(\beta(\bar{w}^j) - b)^2}{\sigma^2(b, \bar{w}^j) k/f_z(z_0|\bar{w}^j)} \right| = o_p(1).
$$

(29)

The last result holds since it is assumed that there is strong or semi-strong identification for $j = 1, \ldots, J^*$. It also establishes (28), which now implies (26). □

References


Moreira, M. J. (2001), “Tests with Correct Size When Instruments Can Be Arbi-
trarily Weak,” Unpublished manuscript, Department of Economics, University of
California, Berkeley.

rica*, 71, 1027–1048.

Education When Compulsory Schooling Laws Really Matter,” *American Economic
Review*, 152–175.


Discontinuity Approach to Multiple Assignment Variables,” *Journal of Economet-
rics*, 161, 203–207.

grams on Poor Households in Bangladesh: Does the Gender of Participants Mat-

per, University of Wisconsin–Madison.


and Weak Identification in Generalized Method of Moments,” *Journal of Business


