Weak Identification in Fuzzy Regression Discontinuity Designs*

Donna Feir† Thomas Lemieux‡ Vadim Marmer‡

April 17, 2016

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

*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.

†Deparment of Economics, University of Victoria, PO Box 1700 STN CSC, Victoria, BC, V8W 2Y2, Canada. Email: dfeir@uvic.ca.

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

treatment effect can be also constructed using this 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; regression discontinuity design; treatment effect; weak identification; uniform asymptotic size

1 Introduction

Since the late 1990s regression discontinuity (RD) and fuzzy regression discontinuity (FRD) designs have been of growing importance in applied economics.[1] Hundreds of recent applied papers have used RD, and in many cases FRD designs.[2] 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.

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

[1]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 et al. (forthcoming); Imbens and Kalyanaraman (2012); Calonico et al. (2014); 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.

[2]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.

papers may suffer from a weak identification problem. We find that in about half of the papers where enough information is reported to compute the F-statistic, weak identification appears to be a problem in at least one of the empirical specifications.[3] 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 uniformly valid testing procedures for treatment (RD) effects.

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

[3]For the procedure followed to obtain the sample of papers, see the online Supplement, Section 1.

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 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). The magnitude of potential size distortions can be tested by testing hypotheses about the concentration parameter with non-central χ^2_1 critical values using the F-statistic, which is an analogue of the first-stage F-statistic in IV regression. Surprisingly, we find critical values that are much higher than would be required in a simple IV setting. When the F-statistic 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 13.6% test, and a 5% one-sided test is in fact a 16.9% test. Nearly zero (under 0.5%) size distortions of a 5% two-sided test correspond to the values of the F-statistic above 93.

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, as their construction only involves solving a quadratic equation.[4] These confidence sets are expected to be as

[4]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 et al., 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.

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)).[5]

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 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 that use the FRD design to estimate the effect of class size on student achievement. 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, suggest 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,

[5]In a recent paper, Otsu et al. (forthcoming), propose empirical likelihood based inference for the RD effect. Using the profile empirical likelihood function, they propose confidence sets for the RD effect, which are expected to be robust against weak identification. However, they do not provide a formal analysis of the weak identification. While their method does not involve variances estimation and for that reason can enjoy better higher-order properties than our approach, it requires computation of the empirical likelihood function numerically and is computationally more demanding.

derive the uniform asymptotic size of usual t-tests for FRD, discuss size distortions and testing for potential size distortions, 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. The online Supplement (Feir et al., 2015) contains additional materials including the proofs and the Monte Carlo results.

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 β_i is the random treatment effect for observation i.[6] 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) $\lim_{z\downarrow z_0} E(y_{0i}|z_i = z) = \lim_{z\uparrow z_0} E(y_{0i}|z_i = z)$.

For binary x_i, 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 if x_i is a deterministic function of z_i, and fuzzy otherwise.

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

[6] 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.

effect:

$$\beta = (y^+ - y^-)/(x^+ - x^-), \tag{1}$$

where $y^+ = \lim_{z \downarrow z_0} E(y_i | z_i = z)$, $y^- = \lim_{z \uparrow z_0} E(y_i | z_i = z)$, and x^+ and x^- are defined similarly with y_i replaced by x_i. The exact interpretation of β depends on the assumptions that the econometrician is willing to make in addition to Assumption 1. As discussed in Hahn et al. (2001), if β_i and x_i are assumed to be independent conditional on z_i, then β 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 β_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 $e > 0$.[7]

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

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

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 (2). 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. The corresponding

[7] See the discussion on page 204 of their paper.

estimator of β 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 $\mathcal{N}_{z_0}\backslash\{z_0\}$; $\lim_{e\downarrow 0} \frac{d^p}{dz^p} E(y_i|z_i = z_0 \pm e)$ and $\lim_{e\downarrow 0} \frac{d^p}{dz^p} E(x_i|z_i = z_0 \pm e)$ exist for $p = 0, 1, 2$.*

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

(iv) *For some $\delta > 0$, $E\left(|y_i - E(y_i|z_i)|^{2+\delta} \middle| z_i\right)$ and $E\left(|x_i - E(x_i|z_i)|^{2+\delta} \middle| z_i\right)$ are bounded on $\mathcal{N}_{z_0}$.*

(c) *As* $n \to \infty$, $\sqrt{nh_n}h_n^2 \to 0$ *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 $\sqrt{nh_n}h_n^2 \to 0$ in Assumption 2(c) is an under-smoothing condition, which makes the contribution of the bias term to the asymptotic distribution negligible. The condition $nh_n^3 \to \infty$ ensures that the variance of the local linear estimator tends to zero. 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[8]

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

For $\Delta y = y^+ - y^-$, $\widehat{\Delta y}_n = \hat{y}_n^+ - \hat{y}_n^-$, and similarly defined Δx and $\widehat{\Delta x}_n$, by Assumption 2 and Lyapounov's CLT we have:

$$\sqrt{nh_n}\begin{pmatrix} \widehat{\Delta y}_n - \Delta y \\ \widehat{\Delta x}_n - \Delta x \end{pmatrix} \to_d \sqrt{\frac{k}{f_z(z_0)}}\begin{pmatrix} \sigma_y \mathcal{Y} \\ \sigma_x \mathcal{X} \end{pmatrix},$$

[8]The constant k is 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.

where $\mathcal{Y}$ and $\mathcal{X}$ are two bivariate normal variables with zero means, unit variances and correlation coefficient ρ_{xy}. This in turn implies that under standard asymptotics, $\sqrt{nh_n}(\hat{\beta}_n - \beta) \to_d N\left(0, k\sigma^2(\beta)/(f_z(z_0)(\Delta x)^2)\right)$, where $\sigma^2(b) = \sigma_y^2 + b^2\sigma_x^2 - 2b\sigma_{xy}$. The last result holds due to identification Assumption 1(a), i.e. only when $\Delta x \neq 0$ and is fixed.

The asymptotic variance σ_y^2 can be consistently estimated by

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

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 σ_x^2 and σ_{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

$$\hat{\sigma}_n^2(b) = \hat{\sigma}_{y,n}^2 + b\hat{\sigma}_{x,n}^2 - 2b\hat{\sigma}_{xy,n}. \tag{3}$$

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

$$T_n(\beta_0) = \sqrt{nh_n}\left(\hat{\beta}_n - \beta_0\right) / \sqrt{k\hat{\sigma}_n^2(\hat{\beta}_n)/(\hat{f}_{z,n}(z_0)(\widehat{\Delta x}_n)^2)}$$

and compares it with standard normal critical values, as $T_n(\beta) \to_d N(0,1)$, when $\Delta x \neq 0$ and is fixed. Confidence intervals for β are constructed by collecting all values β_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

$$T_n(\beta) = \frac{\sqrt{nh_n}\left(\widehat{\Delta y}_n - \beta\widehat{\Delta x}_n\right)}{\sqrt{k\hat{\sigma}_n^2(\hat{\beta}_n)/\hat{f}_{z,n}(z_0)}} \times \text{sign}\left(\widehat{\Delta x}_n\right). \tag{4}$$

When testing H_0 against two-sided alternatives, one uses the absolute value of $T_n(\beta)$, which eliminates the sign term. Since under standard (fixed distribution) asymptotics $\sqrt{nh_n}\left(\widehat{\Delta y}_n - \beta\widehat{\Delta x}_n\right) \to_d N(0, k\sigma^2(\beta)/f_z(z_0))$, the usual t-test has no size distortions as long as $\hat{\beta}_n$ is consistent and $\hat{\sigma}_n^2(\hat{\beta}_n)$ approximates $\sigma^2(\beta_0)$ very well. Define $\Delta Y_n = (f_z(z_0)/k)^{1/2}(nh_n)^{1/2}(\widehat{\Delta y}_n - \Delta y)$ and $\Delta X_n = (f_z(z_0)/k)^{1/2}(nh_n)^{1/2}(\widehat{\Delta x}_n - \Delta x)$. We can now write

$$\hat{\beta}_n - \beta = \frac{\Delta Y_n - \beta\Delta X_n}{\Delta X_n + (f_z(z_0)/k)^{1/2}(nh_n)^{1/2}\Delta x}. \tag{5}$$

Note that in the above expression, estimation errors ΔY_n and ΔX_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 Δx 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 $h_n/g_n \to \infty$.[9] 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 λ may include infinite dimensional components such as distribution functions. Let $cr_n(\alpha)$ denote a possibly data-dependent critical region for nominal significance level α. The test rejects a null hypothesis when $S_n \in cr_n(\alpha)$, and the rejection probability is given by $RP_n(\lambda) = P_\lambda(S_n \in cr_n(\alpha))$, where subscript λ in P_λ 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 λ (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

[9] This situation resembles so-called nearly-weak or semi-strong identification, see Hahn and Kuersteiner (2002), Caner (2009), Antoine and Renault (2009, 2012), and Antoine and Lavergne (forthcoming).

when test statistics have discontinuous asymptotic distribution, i.e. when point-wise asymptotic distribution is discontinuous in a parameter.[10] In what follows, we rely on the following result of Andrews et al. (2011):[11]

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* $D = \{d \in \{\mathbb{R} \cup \{\pm\infty\}\}^J : d_{p_n}(\lambda_{p_n}) \to d$ *for some subsequence* $\{p_n\}$ *of* $\{n\}$ *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,* $AsySz = \sup_{d \in D} RP(d)$*.*

To apply Lemma 3, we define:

$$\lambda_1 = \left(\frac{f_z(z_0)}{k}\right)^{1/2} \frac{|\Delta x|}{\sigma_x}, \quad \lambda_2 = \rho_{xy}, \quad \lambda_3 = \beta\sigma_x/\sigma_y. \tag{6}$$

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 (6) hold. Note that λ_4 is an infinite-dimensional parameter that depends on λ_1, λ_2, and λ_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 $\{\lambda_n \in \lambda : n \geq 1\}$, the test statistic converges to some limiting distribution, which might depend on d. In view of (4) and (5), we therefore define:

$$d_{n,1}(\lambda) = \sqrt{nh_n}\lambda_1, \quad d_{n,2}(\lambda) = \lambda_2, \quad d_{n,3}(\lambda) = \lambda_3. \tag{7}$$

While $\lambda_4 = F$ affects the finite-sample distribution of the test statistic, it does not en-

[10] On the importance of uniform size, see for example Imbens and Manski (2004, p. 1848), Mikusheva (2007), and references in Andrews et al. (2011).

[11] Lemma 3 combines Assumption B and Theorems 2.1 and 2.2 in Andrews et al. (2011).

ter 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_ν denote the ν-th quantile of the standard normal distribution.

Theorem 4. *Suppose that Assumption 2 holds. Let $\mathcal{X}, \mathcal{Y}$ be two bivariate normal variables with zero means, unit variances, and correlation d_2. Define*

$$\mathcal{T}_{d_1,d_2,d_3} = \frac{\mathcal{Y} - d_3\mathcal{X}}{\sqrt{1 + \left(\frac{\mathcal{Y}+d_3 d_1}{\mathcal{X}+d_1}\right)^2 - 2d_2\frac{\mathcal{Y}+d_3 d_1}{\mathcal{X}+d_1}}} \times \text{sign}(\mathcal{X} + d_1).$$

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

$$AsySz = \sup_{d_1 \in \mathbb{R}_+ \cup \{+\infty\}, d_2 \in [0,\bar{\rho}], d_3 = \mathbb{R} \cup \{\pm\infty\}} P(|\mathcal{T}_{d_1,d_2,d_3}| > z_{1-\alpha/2}).$$

(b) *For tests that reject $H_0 : \beta \leq \beta_0$ in favor of $H_1 : \beta > \beta_0$ when $T_n(\beta_0) > z_{1-\alpha}$,*

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

Remark. A commonly used measure of identification strength is the so-called *concentration parameter.*[12] In our framework, the concentration parameter is given by $d_{n,1}^2$, where $d_{n,1}^2 \to \infty$ corresponds to strong (or semi-strong) identification, and identification is weak when the limit of $d_{n,1}^2$ is finite. As it is apparent from the expressions for λ_1 and $d_{n,1}$ in (6) and (7), the concentration parameter and, therefore, the strength of identification depend not only on the size of discontinuity in treatment assignment Δx, 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.

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

For any permitted values of d_2 and d_3, when $d_1 = \infty$ we have $\mathcal{T}_{\infty,d_2,d_3} \sim N(0,1)$. Thus, the asymptotic size of tests based on $T_n(\beta_0)$ is equal to nominal size α under strong or semi-strong identification. When $d_1 < \infty$, it is straightforward to compute $AsySz$ numerically. To compute asymptotic rejection probabilities given d_1, d_2, d_3, first using bivariate normal PDFs one integrates numerically $1(|\mathcal{T}_{d_1,d_2,d_3}| > z_{1-\alpha/2})$ or $1(\mathcal{T}_{d_1,d_2,d_3} > z_{1-\alpha})$ calculated for different realized values of $\mathcal{Y}, \mathcal{X}$. Rejection probabilities then can be numerically maximized over d's.

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 α, and non-central χ_1^2 critical values for testing hypotheses about the concentration parameter at significance level τ.

	maximal rejection prob. for FRD					
	one-sided		two-sided		non-central $\chi_1^2(d_1^2)$ critical values	
d_1^2	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\tau = 0.05$	$\tau = 0.01$
10^{-4}	0.906	0.885	0.893	0.877	3.84	6.64
0.01	0.691	0.636	0.664	0.622	3.88	6.70
0.25	0.363	0.294	0.322	0.261	4.76	8.08
1.0	0.221	0.153	0.187	0.134	7.00	11.06
4.0	0.144	0.086	0.113	0.070	13.28	18.72
9.0	0.119	0.062	0.099	0.050	21.57	28.37
16.0	0.106	0.051	0.076	0.038	31.87	40.03
25.0	0.097	0.045	0.067	0.031	44.15	53.67
36.0	0.091	0.037	0.060	0.029	58.45	69.34
49.0	0.086	0.033	0.056	0.023	74.73	86.98
64.0	0.081	0.032	0.053	0.022	93.03	106.63
81.0	0.078	0.029	0.052	0.022	113.31	128.28
10^2	0.076	0.029	0.052	0.020	135.60	151.94
25^2	0.061	0.020	0.051	0.015	709.96	746.72
50^2	0.056	0.014	0.051	0.012	2667.17	2738.06

Table 1 reports maximal rejection probabilities of one- and two-sided tests based on the usual t-statistic.[13] It shows that $AsySz$ approaches one as the concentration

[13]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

parameter approaches zero. Size distortions decrease monotonically as the concentration parameter increases. In the case of two-sided testing, nearly zero size distortions (under 0.5%) correspond to the concentration parameter of order $d_1^2 \geq 64$ for asymptotic 5% tests, and $d_1^2 \geq 50^2$ 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 $\mathcal{T}_{d_1,d_2,d_3}$.

2.3 Testing for potential size distortions

Following the approach of Stock and Yogo (2005), Table 1 can be used for testing a null hypothesis about the largest potential size distortion against an alternative hypothesis under which the largest potential size distortion does not exceed a certain pre-specified level. Suppose that the econometrician decides that identification is strong enough if, in the case of 1% two-sided testing, the maximal rejection probability does not exceed 5%. Thus, the econometrician effectively adopts tests with 5% significance level, however uses the 1% standard normal critical value. According to the results in Table 1, the corresponding null hypothesis and its alternative in this case can be stated in terms of the concentration parameter d_1^2 as $H_0^W : d_1^2 \leq 9$ and $H_1^S : d_1^2 > 9$ respectively. A test of H_0^W can be based on the estimator of discontinuity Δx. Define

$$F_n = \frac{nh_n(\widehat{\Delta x}_n)^2}{\hat{\sigma}_{x,n}^2 k / \hat{f}_{z,n}(z_0)} = ((\Delta X_n/\sigma_x) + d_{n,1})^2 + o_p(1). \tag{8}$$

As long as the concentration parameter is finite, $F_n \to_d \chi_1^2(d_1^2)$, a non-central χ_1^2 distribution with non-centrality parameter d_1^2. Let $\chi_{1,1-\tau}^2(d_1^2)$ denote the $(1-\tau)$-th quantile of the $\chi_1^2(d_1^2)$ distribution. Since size distortions are monotonically decreasing

rejection probabilities were maximized over the following grids of values: from -0.99 to 0.99 at 0.01 intervals for d_2, and from -1000 to 1000 at 0.5 intervals for d_3.

when the concentration parameter increases, an asymptotic size τ test of H_0^W should reject it when $F_n > \chi^2_{1,1-\tau}(d_1^2)$.

Non-central χ^2_1 critical values are reported in the last two columns of Table 1 for selected values of the concentration parameter and $\tau = 0.05, 0.01$. For example, $H_0^W : d_1^2 \leq 9$ should be rejected in favor of $H_1^S : d_1^2 > 9$ by a 5% test when $F_n > 21.57$. In the case of 5% two-sided testing of β, one needs the concentration parameter of at least 64 to ensure that size distortions are under 0.5%. In that case, a 5% test should reject the null hypothesis of weak identification if $F_n > 93.03$.

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 16.9% and 13.6% for 5% one-sided and two-sided tests respectively.

The results from Table 1 can also be used for designing valid tests (for the FRD effect β) based on usual t-statistics in combination with somewhat larger than usual critical values. For example, suppose one is interested in a 5% two-sided test about β, and rejects the null hypothesis when $F_n > 21.57$ and $|T_n(\beta_0)|$ exceeds the 1% standard normal critical value. According to Table 1, if the concentration parameter $d_1^2 \geq 9$, the asymptotic size does not exceed 5%. On the other hand, if $d_1^2 \leq 9$, $\lim_{n\to\infty} P(F_n > 21.75) \leq 0.05$. Hence, overall this test has an asymptotic 5% significance level. Intuitively, such a test is valid because the null-hypothesis for the F-pre-test assumes size distortions, and one proceeds using the t-statistic only if it is rejected, i.e. if the concentration parameter is found to be large enough. Note, however, that the procedure is conservative. Furthermore, passing the F-test does not completely safeguard against size distortions, and the usual t-statistic must be

used with somewhat larger critical values.

Although the F-test provides useful guidance on the potential magnitude of size distortions, practitioners should not solely rely on this test to decide whether it is worth proceeding with the estimation. With this in mind, we present a robust inference approach in the next section that always yields valid confidence intervals regardless of the strength of identification and does not rely on any pre-tests.

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 $H_0 : \Delta y - \beta_0 \Delta x = 0$ against $H_1 : \Delta y - \beta_0 \Delta x \neq 0$. A test, therefore, can be based on

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

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

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

and the equality holds by (4). Unlike the usual t-statistic, $T_n^R(\beta_0)$ uses the null-

restricted value β_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.

Theorem 5. *Suppose that Assumption 2 holds. 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 AsySz equal to α.*

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 $T_n^R(\beta) = (\Delta Y_n - \beta \Delta X_n) \times \text{sign}\left(\Delta X_n \pm d_{n,1}\right)/\sigma(\beta) + o_p(1)$. 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 from the 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.[14]

Define $S_n = (\Delta Y_n - \beta \Delta X_n)/\sigma(\beta)$ and $Q = \Delta X_n/\sigma_x - (\sigma_{xy} - \beta\sigma_x^2)S_n/(\sigma_x \sigma(\beta))$, so that, when $\beta = \beta_0$, $T_n^R(\beta) = S_n \times \text{sign}[Q_n \pm d_{n,1} + (\sigma_{xy} - \beta\sigma_x^2)S_n/(\sigma_x\sigma(\beta))] + o_p(1)$. When identification is weak, S_n and Q_n are asymptotically independent by construction, while $S_n \to_d N(0,1)$. Therefore one can construct data-dependent

[14] See also Andrews et al. (2006) and Mills et al. (2014).

critical values as follows. First, compute

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

Second, simulate M independent $N(0,1)$ random variables $\{\mathcal{S}_1, \ldots, \mathcal{S}_M\}$ for some large M. Third, for $m = 1, \ldots M$ compute

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

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

Theorem 6. *Suppose that Assumption 2 holds. Tests that reject $H_0 : \beta \leq \beta_0$ in favor of $H_1 : \beta > \beta_0$ when $T^R_n(\beta_0) > \hat{cv}_{n,1-\alpha}(\beta_0, \hat{Q}_n(\beta_0))$ have AsySz equal to α.*

Weak-identification-robust confidence sets for β can be constructed by inversion of the robust tests. For example, a confidence set for β with asymptotic coverage probability $1-\alpha$ can be constructed by collecting all values β_0 that cannot be rejected by the two-sided robust test:

$$CS_{1-\alpha,n} = \left\{ \beta_0 \in \mathbb{R} : \left| T^R_n(\beta_0) \right| \leq z_{1-\alpha/2} \right\}. \tag{9}$$

This confidence set can be easily computed analytically by solving for values of β_0 that satisfy the inequality

$$(\hat{\beta}_n - \beta_0)^2\hat{\sigma}^2_{x,n}F_n - z^2_{1-\alpha/2}(\hat{\sigma}^2_{y,n} + \beta_0^2\hat{\sigma}^2_{x,n} - 2\hat{\sigma}_{xy,n}\beta_0) \leq 0, \tag{10}$$

where F_n is defined in (8).

Depending on the coefficients of the second-order polynomial (in β_0) in equation (10), $CS_{1-\alpha,n}$ 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 β_0^2 in (10) is negative, which occurs when

$$F_n - z_{1-\alpha/2}^2 < 0. \tag{11}$$

Thus, in practice one will see non-standard confidence sets if the null hypothesis $\Delta x = 0$ cannot be rejected using the F-statistic and central $\chi_{1,1-\alpha}^2$ critical values. Case iii) arises when the discriminant of the quadratic polynomial in (10) is negative, which occurs if

$$F_n \hat{\sigma}_n^2(\hat{\beta}_n) - z_{1-\alpha/2}^2 \left(\hat{\sigma}_{y,n}^2 - \hat{\sigma}_{xy,n}^2/\hat{\sigma}_{x,n}^2\right) < 0. \tag{12}$$

Positive definiteness of the variance-covariance matrix composed of $\hat{\sigma}_{x,n}^2$, $\hat{\sigma}_{y,n}^2$, and $\hat{\sigma}_{xy,n}$ implies that (11) holds whenever (12) holds. Thus, negative discriminants implied by (12) are inconsistent with $F_n > z_{1-\alpha/2}^2$ or positive coefficients on β_0^2 in (10). 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 β_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.[15]

Similarly to Otsu et al. (forthcoming), we consider the RD effect conditional on some covariate w_i.[16] Let $\mathcal{W}$ denote the support of the distribution of w_i. Next, for $w \in \mathcal{W}$ we define $y^+(w)$ using the conditional expectation given z_i and $w_i = w$: $y^+(w) = \lim_{z \downarrow z_0} E\left(y_i | z_i = z, w_i = w\right)$. Let $y^-(w), x^+(w)$ and $x^-(w)$ be defined similarly. The conditional RD effect given $w_i = w$ is defined a s$\beta(w) = (y^+(w) - y^-(w))/(x^+(w) - x^-(w))$. 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

$$H_0 : \beta(w) = \beta \text{ for some } \beta \in \mathbb{R} \text{ and all } w \in \mathcal{W}, \tag{13}$$

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.[17] However, this approach can be unre-

[15] The problem is related to the classical ANOVA hypothesis of homogeneous populations (see, for example, Casella and Berger, 2002, Chapter 11).

[16] See also Frölich (2007).

[17] Such a test can be constructed similarly to the ANOVA F-test as in Casella and Berger (2002,

liable 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, 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 $\hat{x}_{j,n}^-(\bar{w}^j)$ 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. $\sigma_y^2(\bar{w}^j)$, $\sigma_x^2(\bar{w}^j)$ and $\sigma_{xy}(\bar{w}^j)$ are defined as the conditional versions of the corresponding population terms, and $\hat{\sigma}_{y,n}^2(\bar{w}^j)$, $\hat{\sigma}_{x,n}^2(\bar{w}^j)$, 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: $n_j h_{n_j}/(nh_n) \to p_j > 0$ 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: $CS_{1-\alpha,n}^J = \left\{\beta_0 \in \mathbb{R} : G_n(\beta_0) \leq \chi_{J,1-\alpha}^2\right\}$, where

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

$\hat{\beta}_n(\bar{w}^j) = \widehat{\Delta y}_n(\bar{w}^j)/\widehat{\Delta x}_n(\bar{w}^j)$, $\widehat{\Delta x}_n(\bar{w}^j) = \hat{x}_n^+(\bar{w}^j) - \hat{x}_n^-(\bar{w}^j)$; $\hat{\sigma}_n^2(\beta_0, \bar{w}^j)$ is defined similarly to $\hat{\sigma}_n^2(\beta_0)$ in (3) using the estimators conditional on $w_i = \bar{w}^j$; and $\hat{f}_{z,n}(z_0|\bar{w}^j) = (n_j h_{n_j})^{-1} \sum_{i=1}^n K((z_i - z_0)/h_{n_j}) 1\{w_i = \bar{w}^j\}$ is the estimator for $f_z(z_0|\bar{w}^j)$, which denotes the conditional density of z_i at z_0 conditional on $w_i = \bar{w}^j$.

Under $H_0 : \beta(w) = \beta$ for some $\beta \in \mathbb{R}$, $CS_{1-\alpha,n}^J$ is an asymptotically valid confi-

Chapter 11) and is discussed in the supplement.

dence set since $G_n(\beta) \to_d \chi^2_J$ under weak or strong identification. We consider the following size α asymptotic test: Reject H_0 if $CS^J_{1-\alpha,n}$ is empty. The test is asymptotically valid because under H_0, $P(CS^J_{1-\alpha,n} = \varnothing) \leq P(\beta \notin CS^J_{1-\alpha,n}) = P(G_n(\beta) > \chi^2_{J,1-\alpha}) \to \alpha$, which again holds under weak or strong identification. Under the alternative, there is no common value β that will provide a proper re-centering for all J categories, and therefore, one can expect deviations from the asymptotic χ^2_J distribution.

We show below that the test is consistent if there is strong (or semi-strong) identification for at least two values $\bar{w}^{j_1}$ and $\bar{w}^{j_2}$ that satisfy $\beta(\bar{w}^{j_1}) \neq \beta(\bar{w}^{j_2})$. Let $d^2_{n,1}(\bar{w}^j) = n_j h_{n_j} |x^+(\bar{w}^j) - x^-(\bar{w}^j)|^2 f_z(z_0|\bar{w}^j)/(k\sigma^2_x(\bar{w}^j))$ be the conditional version of the concentration parameter.

Theorem 7. *Suppose that $n_j h_{n_j}/(nh_n) \to p_j > 0$ and Assumption 2 holds for each $j = 1, \ldots, J$.*

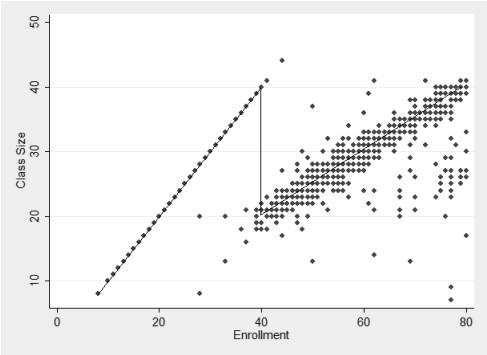
(a) *Tests that reject H_0 of constancy in (13) when $CS^J_{1-\alpha,n} = \varnothing$ have AsySz less or equal to α.*

(b) *Let $\mathcal{W}^* = \{\bar{w}^1, \ldots, \bar{w}^{J^*}\} \subset \mathcal{W}$ be such that $d^2_{n,1}(\bar{w}^j) \to \infty$ for $\bar{w}^j \in \mathcal{W}^*$ and $\beta(\bar{w}^{j_1}) \neq \beta(\bar{w}^{j_2})$ for some $\bar{w}^{j_1}, \bar{w}^{j_2} \in \mathcal{W}^*$. Then, $P(CS^J_{1-\alpha,n} = \varnothing) \to 1$ as $n \to \infty$.*

4 Empirical Applications

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

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.

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.

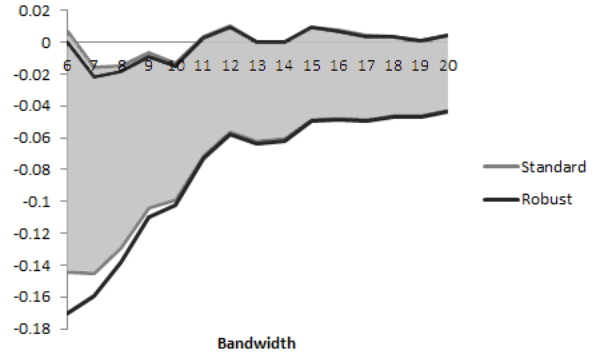
We begin with a case where weak identification is not a serious issue. In an influential paper, Angrist and Lavy (1999) study the effect of class size on academic success in Israel using the fact class size in Israeli public schools was capped at 40 students during their sample period. As demonstrated in Figure 1, 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. We use the same sample selection rules as Angrist and Lavy (1999) and focus on language scores among 4th graders.[18]

Table 2 shows that the estimated discontinuity in the treatment variable (the

[18]The data can be found at http://econ-www.mit.edu/faculty/angrist/data1/data/anglavy99. 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.

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 standards of Table 1, we then conclude that weak identification is not a serious concern in this application. Using the 5% non-central χ^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.

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.

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

bandwidth	discont.	std errors	F-stat	effect	standard CS	robust CS
6	−8.40	1.60	27.5	−0.07	$[-0.145, 0.007]$	$[-0.170, -0.000]$
8	-9.90	1.26	61.9	−0.07	$[-0.129, -0.015]$	$[-0.138, -0.019]$
10	-10.83	1.03	110.2	−0.06	$[-0.103, -0.015]$	$[-0.103, -0.015]$
12	-12.00	0.92	172.0	−0.02	$[-0.056, 0.011]$	$[-0.058, 0.010]$
14	-12.62	0.78	258.8	−0.03	$[-0.061, 0.000]$	$[-0.062, -0.000]$
16	-13.21	0.69	370.1	−0.02	$[-0.048, 0.008]$	$[-0.049, 0.007]$
18	-13.87	0.61	525.8	−0.02	$[-0.046, 0.003]$	$[-0.047, 0.003]$
20	-14.35	0.56	667.7	−0.02	$[-0.042, 0.005]$	$[-0.043, 0.004]$

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.

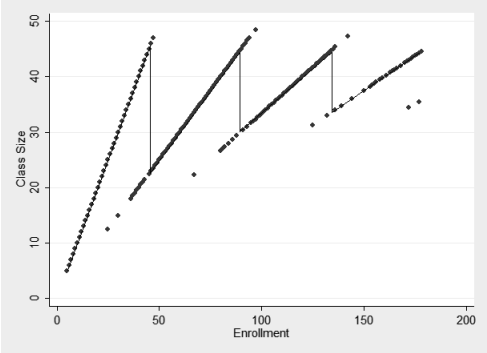
In this application we also compare the results of the standard constancy test of the treatment effect across sub-groups to the results of our robust constancy test. The first set of results reported in Section 5 of the online Supplement compare the treatment effect for secular and religious schools. 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.

The second application considers a similar policy in Chile originally studied by Urquiola and Verhoogen (2009).[19] In this application, the class sizes are capped at 45

[19]It should be noted that Urquiola and Verhoogen (2009) are not attempting to provide causal

students. Figure 3 shows the fuzzy discontinuity in the empirical relationship between class size and enrollment at the various multiples of 45. The figure also shows that 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).[20]

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 3 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 3 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

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.

[20]The total number of observations is 1,636. The effective number of observations varies 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 uniform kernel is used to compute all the results below.

nevertheless rejects the null hypothesis that the effect of class size is equal to zero while the conventional fails to reject the null.

Table 3: 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

bandwidth	estimated effect	standard CS	robust CS
first cutoff (45)			
6	0.146	$[-0.061, 0.353]$	$(-\infty, -0.433] \cup [0.043, \infty)$
8	3.378	$[-74.820, 81.576]$	$(-\infty, -0.120] \cup [0.129, \infty)$
10	-0.437	$[-1.867, 0.993]$	$(-\infty, -0.078] \cup [0.181, \infty)$
12	-0.173	$[-0.360, 0.014]$	$[-1.720, -0.065]$
14	-0.136	$[-0.246, -0.026]$	$[-0.376, -0.060]$
16	-0.091	$[-0.153, -0.029]$	$[-0.186, -0.042]$
18	-0.073	$[-0.115, -0.031]$	$[-0.127, -0.037]$
20	-0.063	$[-0.099, -0.027]$	$[-0.107, -0.032]$
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 3: (Continued)

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

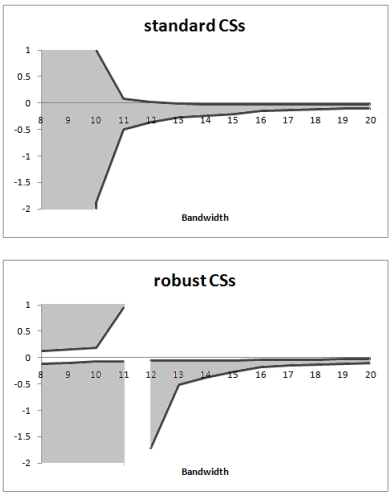
Silverman's rule-of-thumb bandwidth is 8.59 . The optimal bandwidth suggested by Imbens and Kalyanaraman (2012) for the cut-off of 135 is 14.12 and for the cut-off of 180, the suggested bandwidth is 17.81. The scores are given in terms of standard deviations from the mean.

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.[21] This is consistent with the size of the discontinuity in class size as a function of enrollment estimated at different bandwidths and the corresponding F-statistic. At bandwidths below 10, the estimated discontinuity is small and the F-statistic is below 7. However

[21]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.

at bandwidths higher than 12, the estimated discontinuity is progressively closer to 10 students and the F-statistic ranges from just over 40 to just over 188. 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. See Section 5 in the online supplement for a complete listing of the F-statistic and discontinuity estimates at different bandwidths.

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



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 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 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 liberal, 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 provides more informative results than the standard method.

References

Anderson, T. W. and H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D. W. K. and X. Cheng (2012): "Estimation and Inference With Weak, Semi-Strong, and Strong Identification," *Econometrica*, 80, 2153–2211.

Andrews, D. W. K., X. Cheng, and P. Guggenberger (2011): "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," Cowles Foundation Discussion Paper 1813.

Andrews, D. W. K. and P. Guggenberger (2010): "Asymptotic Size and a Problem with Subsampling and with the m out of n Bootstrap," *Econometric Theory*, 26, 426–468.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006): "Optimal Invariant Similar Tests For Instrumental Variables Regression," *Econometrica*, 74, 715–752.

Andrews, D. W. K. and J. H. Stock (2007): "Inference with Weak Instruments," in *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, ed. by R. Blundell, W. K. Newey, and T. Persson, Cambridge, UK: Cambridge University Press, vol. III.

Angrist, J. D. and V. Lavy (1999): "Using Maimonides' Rule to Estimate The Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533–575.

Antoine, B. and P. Lavergne (forthcoming): "Conditional Moment Models Under Semi-Strong Identification," *Journal of Econometrics*.

Antoine, B. and E. Renault (2009): "Efficient GMM with Nearly-Weak Instruments," *Econometrics Journal*, 12, S135–S171.

——— (2012): "Efficient Minimum Distance Estimation with Multiple Rates of Convergence," *Journal of Econometrics*, 170, 350–367.

Arai, Y. and H. Ichimura (2013): "Optimal Bandwidth Selection for Differences of Nonparametric Estimators with an Application to the Sharp Regression Discontinuity Design," Working Paper, National Graduate Institute for Policy Studies.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Buddelmeyer, H. and E. Skoufias (2004): "An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA," World Bank Policy Research Working Paper 3386.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Bias-Corrected Inference in the Regression Discontinuity Design," *Econometrica*, 82, 2295–2326.

Caner, M. (2009): "Testing, Estimation in GMM and CUE with Nearly-Weak Identification," *Econometric Reviews*, 29, 330–363.

Casella, G. and R. L. Berger (2002): *Statistical Inference*, Duxbury Press, Pacific Grove, CA, second ed.

Davidson, J. (1994): *Stochastic Limit Theory*, New York: Oxford University Press.

Dong, Y. and A. Lewbel (2010): "Regression Discontinuity Marginal Threshold Treatment Effects," Working Paper.

Dufour, J.-M. (1997): "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1387.

Fe, E. (2012): "Efficient Estimation in Regression Discontinuity Designs via Asymmetric Kernels," Working Paper.

Feir, D., T. Lemieux, and V. Marmer (2015): "Supplement to "Weak Identification in Fuzzy Regression Discontinuity Designs"," UBC Working paper.

Frölich, M. (2007): "Regression Discontinuity Design with Covariates," Working Paper 2007-32, University of St. Gallen.

Frölich, M. and B. Melly (2008): "Quantile Treatment Effects in the Regression Discontinuity Design," IZA Discussion Paper 3638.

Hahn, J. and G. Kuersteiner (2002): "Discontinuities of Weak Instrument Limiting Distributions," *Economics Letters*, 75, 325–331.

Hahn, J., P. Todd, and W. Van der Klaauw (1999): "Evaluating the Effect of an Antidiscrimination Law Using a Regresion-Discontinuity Design," NBER Working Paper 7131.

——— (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

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

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

Imbens, G. W. and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

Imbens, G. W. and T. Zajonc (2011): "Regression Discontinuity Design with Multiple Forcing Variables," Working Paper.

Kleibergen, F. (2002): "Pivotal Statistics For Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803.

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

McCrary, J. (2008): "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142, 698–714.

Mikusheva, A. (2007): "Uniform Inference in Autoregressive Models," *Econometrica*, 75, 1411–1452.

Mills, B., M. J. Moreira, and L. P. Vilela (2014): "Tests Based on t-Statistics for IV Regression with Weak Instruments," *Journal of Econometrics*, 182, 351–363.

Moreira, M. J. (2001): "Tests with Correct Size When Instruments Can Be Arbitrarily Weak," Unpublished manuscript, Department of Economics, University of California, Berkeley.

——— (2003): "A Conditional Likelihood Ratio Test For Structural Models," *Econometrica*, 71, 1027–1048.

Otsu, T., K.-L. Xu, and Y. Matsushita (forthcoming): "Empirical Likelihood for Regression Discontinuity Design," *Journal of Econometrics*.

Papay, J. P., J. B. Willett, and R. J. Murnane (2011): "Extending the Regression-Discontinuity Approach to Multiple Assignment Variables," *Journal of Econometrics*, 161, 203–207.

Porter, J. (2003): "Estimation in the Regression Discontinuity Model," Working Paper, University of Wisconsin–Madison.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression With Weak Instruments," *Econometrica*, 65, 557–586.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20.

Stock, J. H. and M. Yogo (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. W. K. Andrews and J. H. Stock, New York: Cambridge University Press, chap. 6, 80–108.

Urquiola, M. and E. Verhoogen (2009): "Class-size caps, sorting, and the regression-discontinuity design," *American Economic Review*, 99, 179–215.

Van der Klaauw, W. (2008): "Regression–Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, 22, 219–245.

Supplement To "Weak Identification in Fuzzy Regression Discontinuity Designs"

Donna Feir* Thomas Lemieux† Vadim Marmer†

February 19, 2015

Abstract

This paper contains supplemental materials for Feir et al. (2015). It discusses the problem of weak identification in applied Fuzzy Regression Discontinuity (FRD) literature, provides the proofs of the analytical results in the main paper, reports Monte Carlo results, and provides additional tables for the empirical application.

1 Influential applied papers sample procedure

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 F-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

*Deparment of Economics, University of Victoria, PO Box 1700 STN CSC, Victoria, BC, V8W 2Y2, Canada. Email: dfeir@uvic.ca.

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

enough information to compute the test, but four of the remaining six papers presented some specifications that failed the test.

2 Proofs of Theorem 4, 5 and 6

Proof of Theorem 4. In what follows, the 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

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

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 $Var(\hat{f}_{z,n}(z_0)/f_z(z_0)) \to 0$ (Li and Racine, 2007, Chapter 1), since f_z is bounded away from zero around z_0. It follows that $\hat{f}_{z,n}(z_0)/d_{n,4} \to_p 1$ as $n \to \infty$, and therefore 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 $T_n(\beta)$. Next, let $d_{n,5} = \sigma_y$ and $d_{n,6} = \sigma_x$. By Assumption 2(b)(iii), they take values in compact sets, and $\hat{\sigma}_{x,n}/d_{n,6} \to_p 1$, $\hat{\sigma}_{y,n}/d_{n,5} \to_p 1$, and $\hat{\sigma}_{xy,n}/(d_{n,5}d_{n,6}) - d_{n,2} \to_p 0$. Thus,

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

Suppose now that for any subsequence $\{p_n\}$ of $\{n\}$,

$$(\Delta Y_{p_n}/d_{p_n,5}, \Delta X_{p_n}/d_{p_n,6}) \to_d (\mathcal{Y}, \mathcal{X}). \tag{S.1}$$

Note that $d_{n,5}$ and $d_{n,6}$ do not affect $AsySz$. Suppose further that $d^*_{p_n,1} \to d^*_1$ for some $|d^*_1| < \infty$, $d_{p_n,2} \to d_2 \in [-\bar{\rho}, \bar{\rho}]$, and $d_{p_n,3} \to d_3$ for some $|d_3| < \infty$. In this case,

$$T_{p_n}(\beta) \to_d \frac{\mathcal{Y} - d_3\mathcal{X}}{\sqrt{1 + \left(\frac{\mathcal{Y}+d_3 d^*_1}{\mathcal{X}+d^*_1}\right)^2 - 2d_2\frac{\mathcal{Y}+d_3 d^*_1}{\mathcal{X}+d^*_1}}} \times \text{sign}(\mathcal{X} + d^*_1).$$

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 $[0, \bar{\rho}]$ for two-sided testing.

Suppose now that $|d^*_1| < \infty$, $d_2 \in [-\bar{\rho}, \bar{\rho}]$, and $d_3 = \pm\infty$. In this case,

$$T_n(\beta) = \frac{((\Delta Y_n/d_{n,5}d_{n,3}) - (\Delta X_n/d_{n,6}))\,\text{sign}\left((\Delta X_n/d_{n,6}) + d^*_{n,1}\right)}{\sqrt{\frac{1}{d^2_{n,3}} + \left(\frac{(\Delta Y_n/d_{n,5}d_{n,3})+d^*_{n,1}}{(\Delta X_n/d_{n,6})+d^*_{n,1}}\right)^2 - \frac{2d_{n,2}}{d_{n,3}}\frac{(\Delta Y_n/d_{n,5}d_{n,3})+d^*_{n,1}}{(\Delta X_n/d_{n,6})+d^*_{n,1}}}} + o_p(1). \tag{S.2}$$

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

Next, suppose that $|d^*_1| = \infty$, $d_2 \in [-\bar{\rho}, \bar{\rho}]$, and $|d_3| < \infty$. We have

$$T_n(\beta) = \frac{((\Delta Y_n/d_{n,5}) - d_{n,3}(\Delta X_n/d_{n,6}))\,\text{sign}\left((\Delta X_n/d_{n,6}) + d^*_{n,1}\right)}{\sqrt{1 + \left(\frac{(\Delta Y_n/d_{n,5}d^*_{n,1})+d_{n,3}}{(\Delta X_n/d_{n,6}d^*_{n,1})+1}\right)^2 - 2d_{n,2}\frac{(\Delta Y_n/d_{n,5}d^*_{n,1})+d_{n,3}}{(\Delta X_n/d_{n,6}d^*_{n,1})+1}}} + o_p(1),$$

and, therefore, $T_{p_n}(\beta)$ converges in distribution to $(\mathcal{Y}-d_3\mathcal{X})/(1+d_3^2-2d_2d_3)^{1/2} \times \text{sign}(d^*_1) =^d \mathcal{T}_{\infty,d_2,d_3} \sim N(0,1)$ for any $d_2 \in [-\bar{\rho}, \bar{\rho}]$. 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 $d_2 \in [-\bar{\rho}, \bar{\rho}]$.

The results of the theorem now follow from Lemma 3 provided that (S.1) holds. To show (S.1), consider $\hat{y}^+_n$ first. As in Hahn et al. (1999, Lemma 2), write

$$\sqrt{nh_n}\begin{pmatrix} \hat{y}^+_n - y^+ \\ h_n(\hat{y}^{(1),+}_n - y^{(1),+}) \end{pmatrix} = \left(\frac{1}{nh_n}\sum_{i=1}^n Z_i Z_i' K_i\right)^{-1}$$

$$\times \left(\frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \xi_{ni} + \sqrt{nh_n} E y_i^* Z_i K_i \right), \quad \text{(S.3)}$$

where $y^{(1),+} = \lim_{e \downarrow 0} dE(y_i | z_i = z_0 + e)/dz_i$, $\hat{y}_n^{(1),+}$ 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 $y_i^* = y_i - y^+ - y^{(1),+}(z_i - z_0)/h_n$, and

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

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

$$\left(Var \left(\frac{1}{\sqrt{p_n h_{p_n}}} \sum_{i=1}^{n} \xi_{p_n i} \right) \right)^{-1/2} \sqrt{p_n h_{p_n}} E y_i^* Z_i K_i \to 0. \quad \text{(S.4)}$$

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),

$$\left(Var \left(\frac{1}{\sqrt{p_n h_{p_n}}} \sum_{i=1}^{n} \xi_{p_n i} \right) \right)^{-1/2} \frac{1}{\sqrt{p_n h_n}} \sum_{i=1}^{n} \xi_{p_n i} \to_d N(0, 1), \quad \text{(S.5)}$$

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 $\sigma_y^2(z_i)$ is bounded away from zero around z_0 by Assumption 2(b)(iii), and by Lemma 4 in Hahn et al. (1999). Let

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

By (S.3)-(S.5),

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

Now, by Lemmas 1 and 4 in Hahn et al. (1999) and in view of Assumption 2, we conclude that $\sqrt{p_n h_{p_n}}(\hat{y}_{p_n}^{+} - y^{+})/(\sigma_y^{+}\sqrt{k/f_z(z_0)}) \to_d N(0,1)$, where $\sigma_y^{+} = \lim_{e\downarrow 0}\sigma_y(z_0+e)$.

Let $d_{n,7} = \sigma_y^{+}$, $d_{n,8} = \sigma_y^{-} = \lim_{e\downarrow 0}\sigma_y(z_0-e)$, $\Delta Y_n^{+} = \sqrt{nh_n/(k/f_z(z_0))}(\hat{y}_n^{+} - y^{+})$, and let ΔY_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

$$(\Delta Y_{p_n}^{+}/d_{p_n,7}, \Delta Y_{p_n}^{-}/d_{p_n,8}) \to_d \left(\mathcal{Y}^{+}, \mathcal{Y}^{-}\right). \tag{S.6}$$

where $\mathcal{Y}^{+}, \mathcal{Y}^{-}$ are independent standard normal random variables. Next,

$$\frac{\Delta Y_{p_n}}{d_{p_n,5}} = \frac{\Delta Y_{p_n}^{+}}{d_{p_n,7}}\frac{d_{p_n,7}}{d_{p_n,5}} + \frac{\Delta Y_{p_n}^{-}}{d_{p_n,8}}\frac{d_{p_n,8}}{d_{p_n,5}}.$$

Now, $\Delta Y_{p_n}/d_{p_n,5} \to_d \mathcal{Y}$ in (S.1) can be argued using (S.6), Assumption 2(b)(iii), and Lemma 3.

Lastly, the joint convergence in (S.1) 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, the population parameters should be viewed as drifting sequences indexed by n. First, note that under H_0, the rejection probability is largest when $\beta = \beta_0$. Next, as in the proof of Theorem 4, we can write

$$T_n^R(\beta) = \frac{((\Delta Y_n/d_{n,5}) - d_{n,3}(\Delta X_n/d_{n,6}))\,\text{sign}\left((\Delta X_n/d_{n,6}) + d_{n,1}^{*}\right)}{\sqrt{1 + d_{n,3}^2 - 2d_{n,2}d_{n,3}}} + o_p(1). \tag{S.7}$$

Suppose that $d_{p_n,1}^{*} \to \pm\infty$, and $d_{p_n,3} \to d_3$, where $|d_3| < \infty$. By (S.1) we have that

$T^R_{p_n}(\beta) \to_d N(0,1)$. Next, similarly to (S.7),

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

$$\hat{\mathcal{T}}^R_{n,m}(\beta, \hat{Q}_n(\beta)) = \mathcal{S}_m \times \text{sign}\left(\hat{Q}_n(\beta) + \frac{d_{n,2} - d_{n,3}}{\sqrt{1 + d^2_{n,3} - 2d_{n,2}d_{n,3}}}\mathcal{S}_m\right) + o_p(1). \tag{S.8}$$

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

$$P(T^R_{p_n}(\beta) > \hat{cv}_{p_n,1-\alpha}(\beta_0, \hat{Q}_{p_n}(\beta))) \to \alpha. \tag{S.9}$$

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 (S.7) as

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

By (S.1),

$$\frac{(\Delta Y_{p_n}/d_{p_n,5}) - d_{p_n,3}(\Delta X_{p_n}/d_{p_n,6})}{\sqrt{1 + d^2_{p_n,3} - 2d_{p_n,2}d_{p_n,3}}} \to_d \mathcal{S} = \frac{\mathcal{Y} - d_3\mathcal{X}}{\sqrt{1 + d^2_3 - 2d_2}},$$

$$\hat{Q}_{p_n}(\beta) \to_d \mathcal{Q} + d^*_1, \text{ where}$$

$$\mathcal{Q} = \mathcal{X} - \frac{d_2 - d_3}{\sqrt{1 + d^2_3 - 2d_2}}\mathcal{S}.$$

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

$$T^R_{p_n}(\beta) \to_d \mathcal{S} \times \text{sign}\left(\mathcal{Q} + d^*_1 + \frac{d_2 - d_3}{\sqrt{1 + d^2_3 - 2d_2}}\mathcal{S}\right). \tag{S.10}$$

By (S.8), we have

$$\hat{\mathcal{T}}^R_{p_n,m}(\beta, \hat{Q}_{p_n}(\beta)) \to_d \mathcal{S}_m \times \text{sign}\left(\mathcal{Q} + d_1^* + \frac{d_2 - d_3}{\sqrt{1 + d_3^2 - 2d_2 d_3}}\mathcal{S}_m\right), \tag{S.11}$$

where $\mathcal{S}_m \sim N(0,1)$ and is independent from $\mathcal{Q}$ by construction. The results in (S.10) and (S.11) imply that (S.9) holds also for $|d_1^*| < \infty$.

Equation (S.9) 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 (S.2). The result of the theorem now follows by Lemma 3. □

Proof of Theorem 6. The result in part (a) holds since the concentration parameter does not affect the asymptotic distribution of $G_n(\beta)$. To prove part (b), let

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

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

$$G^*(b) = \sum_{j=1}^{J^*} p_j \frac{(\beta(\bar{w}^j) - b)^2(\Delta x(\bar{w}^j))^2}{\sigma^2(b, \bar{w}^j)} \frac{f_z(z_0|\bar{w}^j)}{k}.$$

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

$$|\inf_{b\in\mathbb{R}} G_n^*(b)/(nh_n) - \inf_{b\in\mathbb{R}} G^*(b)| \to_p 0, \tag{S.12}$$

as (S.12) implies that $P(\inf_{b\in\mathbb{R}} G_n^*(b) > a) \to 1$ for all $a \in \mathbb{R}$ as $n \to \infty$. However, the last equation can be shown to establishing that

$$\sup_{b\in\mathbb{R}} |G_n^*(b)/(nh_n) - G^*(b)| \to_p 0. \tag{S.13}$$

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 $\sigma_y^2(\bar{w}^j)$, $\sigma_x^2(\bar{w}^j)$, 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,

$$\sup_{b\in\mathbb{R}} \frac{\left(\hat{\beta}_n(\bar{w}^j) - b\right)^2}{\hat{\sigma}_n^2(b, \bar{w}^j)} = O_p(1). \tag{S.14}$$

By triangle inequality,

$$|G_n^*(b)/(nh_n) - G^*(b)| \leq \sum_{j=1}^{J^*} p_j (\Delta x(\bar{w}^j))^2 \times \left| \frac{\left(\hat{\beta}_n(\bar{w}^j) - b\right)^2}{\hat{\sigma}_n^2(b, \bar{w}^j)k/\hat{f}_z(z_0|\bar{w}^j)} - \frac{(\beta(\bar{w}^j) - b)^2}{\sigma^2(b, \bar{w}^j)k/f_z(z_0|\bar{w}^j)} \right| + \sum_{j=1}^{J^*} R_{j,n}(b), \tag{S.15}$$

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

$$\left(\left| \frac{n_j h_{n_j}}{nh_n} - p_j \right| (\Delta x(\bar{w}^j))^2 + \left| (\widehat{\Delta x}_n(\bar{w}^j))^2 - (\Delta x(\bar{w}^j))^2 \right| p_j \right) \left| \frac{\left(\hat{\beta}_n(\bar{w}^j) - b\right)^2}{\hat{\sigma}_n^2(b, \bar{w}^j)k/\hat{f}_z(z_0|\bar{w}^j)} \right|.$$

Since $n_j h_{n_j}/nh_n \to p_j$, it follows from (S.14) 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 $j = 1, \ldots, J^*$,

$$\sup_{b\in\mathbb{R}} \left| \frac{\left(\hat{\beta}_n(\bar{w}^j) - b\right)^2}{\hat{\sigma}_n^2(b, \bar{w}^j)k/\hat{f}_z(z_0|\bar{w}^j)} - \frac{(\beta(\bar{w}^j) - b)^2}{\sigma^2(b, \bar{w}^j)k/f_z(z_0|\bar{w}^j)} \right| = o_p(1). \tag{S.16}$$

The last result holds since it is assumed that there is strong or semi-strong identification for $j = 1, \ldots, J^*$. It also establishes (S.15), which now implies (S.13). □

3 Monte Carlo results for standard and weak-identification-robust confidence sets for FRD designs

In this section, we discuss the performance of standard and robust confidence sets for FRD in Monte Carlo experiments. The model is as in Section 2.1 of the main paper, and specific

parametrizations that we use for our simulations are described below. Standard confidence intervals are based on the usual t-statistic for FRD ($T_n(\beta_0)$) and standard normal critical values. Thus, standard two-sided symmetric confidence intervals with asymptotic coverage probability of $1-\alpha$ are constructed as estimator$\pm z_{1-\alpha/2} \times$ std.err, where z_τ denotes the τ-th quantile of $N(0,1)$ distribution, and they correspond to testing $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$. Standard one-sided confidence intervals are constructed as $(-\infty, \text{estimator} - z_\alpha \times \text{std.err}] = (-\infty, \text{estimator} + z_{1-\alpha} \times \text{std.err}]$, and they correspond to testing $H_0 : \beta \geq \beta_0$ against $H_0 : \beta < \beta_0$. Robust confidence sets are constructed as discussed in Section 2.4 of the main paper using robust t-statistic $T_n^R(\beta_0)$. Robust two-sided confidence sets consist of all values β_0 such that $|T_n^R(\beta_0)| < z_{1-\alpha/2}$, and are computed analytically by solving (11) in the main paper.

Robust one-sided confidence sets consist of all values β_0 that satisfy the inequality $T_n^R(\beta_0) > \hat{cv}_{n,1-\alpha}(\beta_0, \hat{Q}_n(\beta_0))$, where $\hat{cv}_{n,1-\alpha}(\beta_0, \hat{Q}_n(\beta_0))$ denotes data-dependent critical values discussed in Section 2.4 of the main paper. In practice one-sided robust confidence sets can be computed numerically by checking the above inequality over a grid of values for β_0. However, to evaluate their coverage probabilities in a Monte Carlo experiment, one can simply compute the relative frequency of occurrence of the event $T_n^R(\beta) > \hat{cv}_{n,1-\alpha}(\beta, \hat{Q}_n(\beta))$, where β is the true value used to generate data.

3.1 Data generating process (DGP)

The outcome variable y_i is generated according to the following model with a constant RD effect:

$$\begin{aligned} y_i &= y_{0i} + x_i\beta, \\ y_{0i} &= g_y(u_{yi}, v_i), \\ x_i &= g_x(z_i, u_{xi}, c), \end{aligned}$$

where u_{yi} and u_{xi} are bivariate normal with correlation parameter κ:

$$\begin{pmatrix} u_{yi} \\ u_{xi} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix} \right),$$

the assignment variable x_i follows a normal distribution,

$$z_i \sim N(0, \sigma_z^2),$$

and v_i follows a χ^2-distribution with 3 degrees of freedom,

$$v_i \sim \chi_3^2.$$

Note also that $(u_{yi}, u_{xi})'$, z_i, and v_i are jointly independent. In our setup, y_{0i} captures the outcome in the absence of treatment, and the function $g_y(\cdot, \cdot)$ takes one of the following three forms:

$$g_y(u_{yi}, v_i) \quad = \quad \begin{cases} u_{yi}, & y_{0i} \sim N(0,1), \\ \exp(u_{yi}), & y_{0i} \sim \text{Log-normal}, \\ u_{yi}/v_i, & y_{0i} \sim t_3. \end{cases}$$

Thus, the marginal distribution of the outcome without treatment is either standard normal, log-normal, or a t-distribution with 3 degrees of freedom. Log-normal and t- distributions are used to evaluate the effect of deviations from normality: asymmetries in the first case and heavier tails in the second case.

The function $g_x(\cdot, \cdot, \cdot)$ controls when and what kind of treatment is received, and it can take one of the two following forms. In the first case,

$$g_x(z_i, u_{xi}, c) = u_i \times 1\{z_i \leq 0\} + (u_i + c) \times 1\{z_i > 0\}.$$

This is a DGP with continuous treatment variables (thus, there are treatments of different intensity). In this case, $x^+ - x^- = c$, and the concentration parameter (equations (7) and (8) in the main paper) is given by

$$\frac{nh_n c^2}{2k\sqrt{2\pi\sigma_z^2}},$$

since $z_0 = 0$, $f_z(0) = 1/\sqrt{2\pi\sigma_z^2}$, $Var(x_i|z_i) = 1$, so $\sigma_x^2 = 2$. Thus, weaker designs can be generated by reducing the value of $|c|$ or by increasing the value of σ_z^2. Alternatively, treatment assignment can be generated using

$$g_x(z_i, u_{xi}, c) = 1\{u_i \leq 0\} \times 1\{z_i \leq 0\} + 1\{u_i \leq c\} \times 1\{z_i > 0\}.$$

With this DGP, the treatment variable is binary. Let $\Phi(\cdot)$ denote the standard normal CDF. Then, $x^+ - x^- = \Phi(c) - \Phi(0)$, and $\sigma_x^2 = \Phi(0)(1 - \Phi(0)) + \Phi(c)(1 - \Phi(c))$. Hence, the concentration parameter is given by

$$\frac{nh_n(\Phi(c) - \Phi(0))^2}{2k\sqrt{2\pi\sigma_z^2}\left(\Phi(0)(1 - \Phi(0)) + \Phi(c)(1 - \Phi(c))\right)}.$$

Similarly to the continuous case DGP, the concentration parameter is increasing in c, however, it is now bounded from above for fixed nh_n and σ_z^2, since $\lim_{c\to\infty} \Phi(c) = 1$.

In our DGP, u_{xi} determines whether the treatment is received, and therefore the parameter κ captures the degree of endogeneity of treatment.

Observations are simulated to be independent across i's. The number of Monte Carlo replications is set to 10,000. Our sample size is set to $n = 1,000$. Our base bandwidth value has been chosen as $h_n = n^{-1/4} \approx 0.1778$. We also explore sensitivity of the results to bandwidth choices by also using $h_n = 0.0889$ and 0.8891. We use the uniform kernel function $K(z) = 1/2 \times 1\{-1 \leq z \leq 1\}$, which corresponds to $k = 4$.

We use the following parameter values:

$$\beta = 0,$$

$$\begin{aligned} \sigma_z &= 1,\ 5, \\ \kappa &= 0.5,\ 0.99, \\ c &= 2,\ 0.5,\ 0.1 \end{aligned}$$

where with $c = 2$ identification is considered to be relatively strong, and it becomes weak as c decreases. The values for our endogeneity parameter (κ) are the same as those used in the weak IV literature (Staiger and Stock, 1997). However, since our DGP is non-linear, κ is different from the asymptotic correlation between estimation errors ρ_{xy} (see Section 2.1 of the main paper), where ρ_{xy} is typically smaller in absolute value than κ. Note that ρ_{xy} directly affects asymptotic rejection probabilities.

3.2 Results

First, we consider the effect of weak identification on the distribution of the usual t-statistic $T_n(\beta)$. Figure 1 shows the densities of $T_n(\beta)$ estimated by kernel smoothing for binary x_i, normal y_{0i}, and $\sigma_z = 1$. As a comparison, we also plot the standard normal density. From panels (a) and (b) constructed using $\kappa = 0.50$ or $\kappa = 0.99$ and $c = 2$ (strong identification), it is apparent that the standard normal distribution is a very good approximation to the distribution of $T_n(\beta)$. When $\kappa = 0.99$, the distribution of $T_n(\beta)$ is slightly skewed to the left. However, the normal approximation should still work reasonably well (as we show below), because there are no substantial deviations of the extreme values of the distribution of $T_n(\beta)$ from those of the standard normal distribution.

Figures 1 (c) and (d) show the density of $T_n(\beta)$ under very weak identification ($c = 0.1$). In this case, the distribution of $T_n(\beta)$ is very different from normal. It is strongly skewed to the left, although when $\kappa = 0.50$ it is also more concentrated around zero. A consequence of concentration is that there will be no size distortions when identification is weak but the degree of endogeneity is small. The picture changes drastically when $\kappa = 0.99$. The distribution of $T_n(\beta)$ is strongly skewed to the left and no longer concentrated as much around zero. One can expect substantial size distortions in this case for two-sided tests or

confidence intervals. Even more severe distortions can be expected for one-sided tests of $H_0 : \beta \geq \beta_0$ against $H_1 : \beta < \beta_0$. However, there will be no size distortions for tests of $H_0 : \beta \leq \beta_0$ against $H_1 : \beta > \beta_0$, since the probability mass is shifted to the left. Similarly, one can expect that the coverage probability of one-sided confidence intervals of the form $(-\infty, \text{estimator} + z_{1-\alpha} \times \text{std.err}]$ will be below the nominal coverage of $1-\alpha$. The discrepancy between the actual and nominal coverage for such intervals would be even larger than for two-sided confidence intervals. At the same time, one can expect that the actual coverage of one-sided confidence intervals of the form $[\text{estimator} - z_{1-\alpha} \times \text{std.err}, \infty)$ will exceed the nominal coverage.

Simulated coverage probabilities for different combinations of the model's parameters are reported in Table 1. With moderate degree of endogeneity ($\kappa = 0.5$) and when identification is relatively strong (the concentration parameter is around 35 or 7), the usual confidence intervals, two-sided and one-sided, have coverage probabilities very close to the nominal ones. Their coverage probabilities remain very close to nominal even when the concentration parameter drops to very small values (0.09 and 0.02), as long as endogeneity remains moderate.

When the degree of endogeneity is very high ($\kappa = 0.99$), the coverage probabilities of the standard confidence intervals deviate from the nominal levels. Even with a large value of the concentration parameter of 35, the simulated coverage of one-sided intervals can be below the nominal level by 5% (while two-sided intervals remain quite accurate). For a concentration parameter around 7, distortions can be up to 10% for one-sided intervals and 5% for two-sided intervals. The situation becomes substantially worse when identification is very weak ($c = 0.1$ and the corresponding values of the concentration parameter below 0.1). In this case we observe severe size distortions for one-sided and two-sided interval. For example, the actual coverage probabilities of the 90%, 95% and 99% two-sided confidence intervals are approximately 51%, 55%, and 62%, respectively when the concentration parameter is around 0.09. More substantial size distortions were observed with the concentration parameter equal to 0.009: the actual coverages of the 90%, 95% and 99% two-sided confidence intervals were

32%, 35%, and 41%, respectively.

As expected, the performance of one-sided intervals was even worse due to the skewness of the distribution of $T_n(\beta)$ in the case of weak identification and strong endogeneity (Figure 1(d)). For example, when $c = 0.1$ and $\kappa = 0.99$, the actual coverage of the 90% one-sided confidence intervals does not exceed 46%, and it goes as low as 28% in the case of the concentration parameter equal to 0.009.

The last rows of Table 1 show coverage when identification is still weak, but less so ($c = 0.5$), which results in the concentration parameter values of 2.22, 0.44, or 0.22 depending on σ_z and the bandwidth. Still, distortions remain serious when κ is large. For example, the actual coverage for two-sided intervals, when nominal coverage is set to 90%, ranges from 83% when the concentration parameter is 2.22, to 55% when it is 0.22.

Table 1 makes clear that, as long as the concentration parameter is similar, coverage will be similar even if some of the primitive parameters such as the bandwidth and σ_z differ. For example, compare the seventh row of Table 1 with the tenth row. In the seventh row, the bandwidth is 0.1778, and σ_z is equal to one, which corresponds to the concentration parameter of approximately 0.09. In the tenth row, σ_z is five, but the bandwidth has been increased five times leaving the concentration parameter unchanged. The actual coverage probabilities of the standard confidence intervals are equal in both cases.[1]

Table 1 also clearly demonstrates how the magnitude of the concentration parameter relates to the degree of distortions observed. When the concentration parameter is relatively large as in the first and fifth rows of Table 1, the coverage probabilities are close to the nominal ones. On the other hand, the closer to zero the concentration parameter is, the more severe the distortions. It is important to note, however, that the degree of endogeneity is kept the same. Even with equivalent concentration parameters, if there is a higher degree of endogeneity then distortions will be more severe. We can see this by comparing the third and seventh rows.

We have also computed the simulated coverage probabilities of the weak identification

[1]Note that in each experiment (i.e. for each combination of parameters), we used the same sequence of primitive random variables by controlling the random numbers generator.

robust confidence sets. We find that regardless of the strength of identification and degree of endogeneity, the simulated coverage probabilities of two-sided and one-sided robust confidence sets are uniformly very close to the nominal coverage probabilities. This supports our claim that the inference based on the null-restricted statistic $T_n^R(\beta_0)$ does not suffer from size distortions.

Table 2 repeats the exercises presented in Table 1 when the treatment variable is binary rather than continuous. The distortions observed in this case are less severe under all specifications. The reason is that κ does not map exactly to the asymptotic correlation between the estimation errors ρ_{xy}, which controls directly asymptotic rejection probabilities (see Section 2.2 of the main paper). This is due to the non-linear nature of the DGP. For example, when $\kappa = 0.99$ and $\sigma_z = 5$, the implied correlation between $\widehat{\Delta y}_n$ and $\widehat{\Delta x}_n$ is approximately 0.80. Under our standard choice of bandwidth, the coverage probabilities of the 90%, 95%, and 99% confidence sets are 78.7%, 84.9%, and 92.2%, respectively.

Table 3 presents the results for non-normal DGPs. Again, the distortions appear less severe than those in Table 1 due to the non-linear mapping of κ to the degree of endogeneity. We report the implied value of ρ_{xy} in Table 3. Nevertheless, the table demonstrates that even when the endogeneity is not as severe and the system is non-normal, size distortions are still present.

In the main body of the paper, we have discussed the possibility that the robust confidence sets may cover the entire real line or be the union of two half-lines. Table 4 demonstrates the likelihood of this possibility under different scenarios. When identification is relatively strong ($c = 2$), the form of the robust confidence sets complies to the standard one. By contrast, the shape of weak-identification-robust confidence sets is non-standard when identification becomes weaker. As reported in Table 4, in the case of a relatively strong FRD ($c = 2$), the probabilities that the robust confidence sets are unbounded are very small regardless of the value of κ, and even negligible in the case of continuous treatment. For binary treatment, the probability of seeing unbounded confidence sets when $c = 2$ varies between 0.0007 and 0.11 depending on the nominal coverage and the value of κ. On the other

hand, in the case of a weak FRD ($c = 0.1$ or $c = 0.5$), unbounded robust confidence sets are obtained with very high probabilities. For example, in the case of continuous treatment and when $c = 0.1$ and $\kappa = 0.99$, the entire real line is obtained with probabilities 23%, 35% and 60% and the union of two half lines with probabilities 65%, 60% and 39% for the confidence sets with nominal coverage of 90%, 95% and 99% respectively.

4 Monte Carlo results for the test of constancy of the RD effect

In this section, we present the simulated size and power of the standard and weak-identification-robust constancy tests. As discussed in Section 3 of the main paper, it is assumed that in addition to y_i, x_i, z_i, the econometrician observes the covariate variable w_i that takes values in $\mathcal{W} = \{\bar{w}^1, \ldots, \bar{w}^J\}$. The econometrician is interested in testing the null hypothesis that the RD effect is independent of w_i. We maintain the same basic design as in Section 3 with normally distributed outcomes in the absence of treatment and continuous treatment variables. However, we now consider three population sub-groups ($J = 3$) with $n_j = 1,000$ for all $j = 1, 2, 3$. We use the uniform kernel for all sub-groups, select the bandwidth according to $h_{n_j} = n_j^{-1/4}$, and use $\sigma_z = 1$ and $\kappa = 0.99$ in all simulations for all categories of w_i. Under H_0 of constancy, we generate data with $\beta_j = 0$ for all $j = 1, 2, 3$. Under the alternative of heterogenous treatment effects, we generate data with $\beta_1 = 0$, $\beta_2 = -1$ and $\beta_3 = 1$, or $\beta_2 = -3$ and $\beta_3 = 3$.

The standard constancy test can be constructed along the lines of the ANOVA F-test. See for example, Casella and Berger (2002, Chapter 11). Using the notation of Section 3 of the main paper, let

$$\begin{aligned} CB_n &= \sum_{j=1}^{J} \left(\hat{\beta}_n(\bar{w}^j) - \bar{\beta}_n \right)^2 / \hat{V}_n(\bar{w}^j), \text{ where} \\ \bar{\beta}_n &= \frac{\sum_{j=1}^{J} \hat{\beta}_n(\bar{w}^j)/\hat{V}_n(\bar{w}^j)}{\sum_{j=1}^{J} 1/\hat{V}_n(\bar{w}^j)}, \text{ and} \end{aligned}$$

$$\hat{V}_n(\bar{w}^j) \quad = \quad \frac{1}{n_j h_{n_j}} \frac{k\hat{\sigma}_n^2(\beta_0, \bar{w}^j)}{\hat{f}_{z,n}(z_0|\bar{w}^j)(\widehat{\Delta x}_n(\bar{w}^j))^2}.$$

Under H_0 of constancy of the RD effect across the covariate's values and under strong identification for all J categories, $CB_n \to_d \chi^2_{J-1}$. Thus, the standard test of constancy will reject H_0 when$CB_n > \chi^2_{J-1,1-\alpha}$. Since the standard test relies on the asymptotic normal approximation for the FRD estimator, one can expect that it will be distorted when identification is weak.

Weak-identification-robust constancy test is proposed in Section 3 of the main paper. The robust statistic $G_n(\beta_0)$ is evaluated over a grid of values for β_0 that covers the interval $[-10, 10]$. The null hypothesis of constancy is rejected by the robust test if the smallest value of $G_n(\beta_0)$ obtained on the grid exceeds $\chi^2_{J,1-\alpha}$. Our theoretical results predict that the robust test has accurate size regardless of the strength of identification, and good power if at least two categories with different RD effects have sufficiently strong identification.

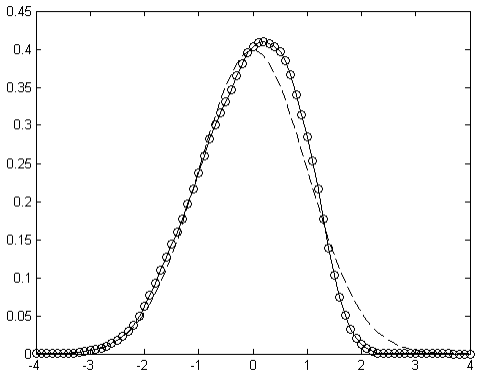
Table 5 reports simulated rejection probabilities for the standard and robust tests when the treatment effect is strongly identified for all sub-groups and when it is weakly identified for some sub-groups. As one can see from the first row of Table 5, which reports the results under H_0, the standard test is under-sized when identification is relatively strong for all three categories, while the robust test has rejection probabilities very close to the desired significance level.

When the treatment effect for one of the groups is only weakly identified as shown in the second row of Table 5, the standard test over-rejects the null hypothesis of equality: the simulated rejection probabilities of the standard test are equal to 43%, 38%, and 31% for the significance levels of 10%, 5%, and 1% respectively. The rejection probabilities for the robust test remain very close to the corresponding significance levels.

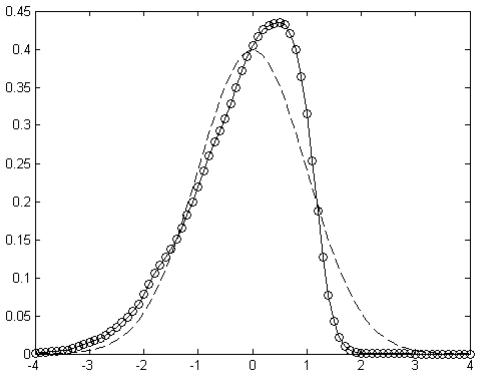
When the treatment effect differs between the groups, rows three and four of Table 5 demonstrate that the robust test has reasonable power to reject the null hypothesis of constancy even when the treatment effect is weakly identified for one or more groups. For example, when the treatment effect is 0, -1, and 1 from groups 1, 2 and 3 respectively, the

robust test rejects the null hypothesis 80%, 70% and 47% of the time for a 10%, 5% and 1% test. When the difference in the treatment effect between the groups is greater, 0, -3 and 3 for example, the robust test rejects the null hypothesis of equality nearly 100% of the time. Rows five and six of Table 5 demonstrate that similar results hold when weak identification is a problem for more than one of the sub-groups. The last row of the table shows that, when identification is strong for all groups, the two tests have comparable power.

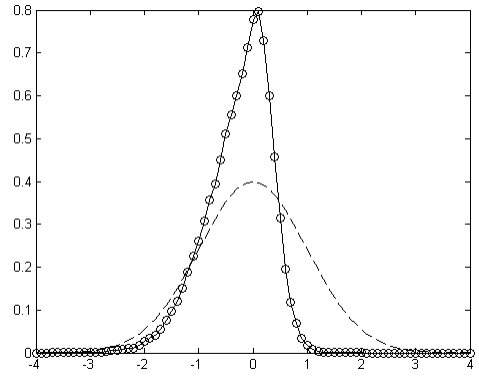
Figure 1: Kernel estimated density of the usual T statistic (solid line) under strong (c = 2) and weak (c = 0.1) identification for different values of the endogeneity parameter against the standard normal PDF (dashed line)



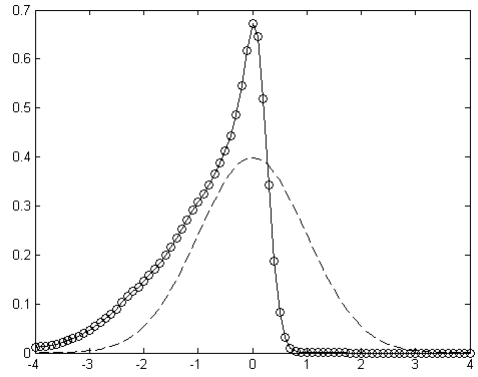
(a) Strong identification, $\kappa = 0.50$



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



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



(d) Weak identification, $\kappa = 0.99$

Table 1: (Continuous x_i & normal y_{0i}) Simulated coverage probabilities of standard and weak-identification-robust confidence sets for different values of the standard deviation of the assignment variable (σ_z), size of discontinuity in treatment assignment (c), degree of endogeneity (κ), and bandwidth (h_n).

						Simulated Coverage			
			Concentration		Nominal	Two-sided		One-sided	
σ_z	c	κ	Parameter	h_n	Coverage	Standard	Robust	Standard	Robust
1	2	0.50	35.48	0.1778	0.90	0.9180	0.9051	0.8784	0.8970
					0.95	0.9583	0.9522	0.9320	0.9505
					0.99	0.9885	0.9931	0.9793	0.9907
5	2	0.50	7.10	0.1778	0.90	0.9262	0.9157	0.8713	0.9025
					0.95	0.9604	0.9693	0.9273	0.9585
					0.99	0.9896	0.9984	0.9820	0.9969
1	0.1	0.50	0.09	0.1778	0.90	0.9600	0.9051	0.9073	0.8970
					0.95	0.9815	0.9522	0.9600	0.9505
					0.99	0.9969	0.9931	0.9938	0.9907
5	0.1	0.50	0.02	0.1778	0.90	0.9403	0.9157	0.8824	0.9025
					0.95	0.9659	0.9693	0.9404	0.9585
					0.99	0.9892	0.9984	0.9831	0.9969
1	2	0.99	35.48	0.1778	0.90	0.9072	0.9051	0.8580	0.8970
					0.95	0.9393	0.9522	0.9072	0.9505
					0.99	0.9722	0.9931	0.9629	0.9907
5	2	0.99	7.10	0.1778	0.90	0.8676	0.9157	0.8189	0.9025
					0.95	0.9011	0.9693	0.8676	0.9585
					0.99	0.9443	0.9984	0.9299	0.9969
1	0.1	0.99	0.09	0.1778	0.90	0.5165	0.9051	0.4625	0.8970
					0.95	0.5573	0.9522	0.5165	0.9505
					0.99	0.6203	0.9931	0.5949	0.9907
5	0.1	0.99	0.02	0.1778	0.90	0.3741	0.9157	0.3264	0.9025
					0.95	0.4113	0.9693	0.3741	0.9585
					0.99	0.4718	0.9984	0.4475	0.9969
5	0.1	0.99	0.009	0.0889	0.90	0.3203	0.9289	0.2829	0.8999
					0.95	0.3521	0.9670	0.3203	0.9569
					0.99	0.4139	0.9847	0.3902	0.9845
5	0.1	0.99	0.09	0.8891	0.90	0.5165	0.9051	0.4625	0.8970
					0.95	0.5573	0.9522	0.5165	0.9505
					0.99	0.6203	0.9931	0.5949	0.9907

Table 1: (Continued)

						Simulated Coverage			
			Concentration		Nominal	Two-sided		One-sided	
σ_z	c	κ	Parameter	h_n	Coverage	Standard	Robust	Standard	Robust
					0.90	0.9367	0.9051	0.8828	0.8970
1	0.5	0.50	2.22	0.1778	0.95	0.9656	0.9522	0.9367	0.9505
					0.99	0.9917	0.9931	0.9845	0.9907
					0.90	0.9362	0.9157	0.8780	0.9025
5	0.5	0.50	0.44	0.1778	0.95	0.9646	0.9693	0.9365	0.9585
					0.99	0.9895	0.9984	0.9823	0.9969
					0.90	0.8287	0.9051	0.7843	0.8970
1	0.5	0.99	2.22	0.1778	0.95	0.8617	0.9522	0.8287	0.9505
					0.99	0.9066	0.9931	0.8908	0.9907
					0.90	0.6783	0.9157	0.6183	0.9025
5	0.5	0.99	0.44	0.1778	0.95	0.7195	0.9693	0.6783	0.9585
					0.99	0.7812	0.9984	0.7585	0.9969
					0.90	0.5535	0.9289	0.4983	0.8999
5	0.5	0.99	0.22	0.0889	0.95	0.5946	0.9670	0.5535	0.9569
					0.99	0.6579	0.9847	0.6347	0.9845
					0.90	0.8287	0.9051	0.7843	0.8970
5	0.5	0.99	2.22	0.8891	0.95	0.8617	0.9522	0.8287	0.9505
					0.99	0.9066	0.9931	0.8908	0.9907

Table 2: (Binary x_i & normal y_{0i}) Simulated coverage probabilities of standard and weak-identification-robust confidence sets for different values of the standard deviation of the assignment variable (σ_z), size of discontinuity in treatment assignment (c), degree of endogeneity (κ), and bandwidth (h_n).

						Simulated Coverage			
			Concentration		Nominal	Two-sided		One-sided	
σ_z	c	κ	Parameter	h_n	Coverage	Standard	Robust	Standard	Robust
1	2	0.50	7.42	0.1778	0.90	0.9376	0.9051	0.8925	0.8970
					0.95	0.9745	0.9522	0.9472	0.9505
					0.99	0.9961	0.9931	0.9912	0.9907
5	2	0.50	1.48	0.1778	0.90	0.9606	0.9157	0.9120	0.9025
					0.95	0.9838	0.9693	0.9625	0.9585
					0.99	0.9966	0.9984	0.9932	0.9969
1	0.1	0.50	0.03	0.1778	0.90	0.9783	0.9051	0.9394	0.8970
					0.95	0.9913	0.9522	0.9783	0.9505
					0.99	0.9987	0.9931	0.9972	0.9907
5	0.1	0.50	0.006	0.1778	0.90	0.9627	0.9157	0.9184	0.9025
					0.95	0.9816	0.9693	0.9628	0.9585
					0.99	0.9942	0.9984	0.9913	0.9969
1	2	0.99	7.42	0.1778	0.90	0.9206	0.9051	0.8704	0.8970
					0.95	0.9546	0.9522	0.9211	0.9505
					0.99	0.9857	0.9931	0.9758	0.9907
5	2	0.99	1.48	0.1778	0.90	0.9141	0.9157	0.8514	0.9025
					0.95	0.9493	0.9693	0.9141	0.9585
					0.99	0.9839	0.9984	0.9750	0.9969
1	0.1	0.99	0.03	0.1778	0.90	0.8201	0.9051	0.7382	0.8970
					0.95	0.8732	0.9522	0.8201	0.9505
					0.99	0.9448	0.9931	0.9204	0.9907
5	0.1	0.99	0.006	0.1778	0.90	0.7872	0.9157	0.7057	0.9025
					0.95	0.8495	0.9693	0.7872	0.9585
					0.99	0.9226	0.9984	0.8976	0.9969
5	0.1	0.99	0.003	0.0889	0.90	0.7325	0.9287	0.6487	0.8997
					0.95	0.7868	0.9668	0.7327	0.9567
					0.99	0.8654	0.9845	0.8385	0.9843
5	0.1	0.99	0.03	0.8891	0.90	0.8201	0.9051	0.7382	0.8970
					0.95	0.8732	0.9522	0.8201	0.9505
					0.99	0.9448	0.9931	0.9204	0.9907

Table 3: (Non-normal y_{0i}) Simulated coverage probabilities of standard and weak-identification-robust confidence sets for different values of the standard deviation of the assignment variable (σ_z), size of discontinuity in treatment assignment (c), degree of endogeneity (κ), correlation between estimation errors for y and x (ρ_{xy}), and different distributions of the outcome without treatment (y_{0i}). Bandwidth is equal to 0.1778.

							Simulated Coverage			
		Distribution			Concentration	Nominal	Two-sided		One-sided	
σ_z	c	of y_{0i}	κ	ρ_{xy}	Parameter	Coverage	Standard	Robust	Standard	Robust
continuous treatment x_i										
5	0.1	Log-normal	0.99	0.85	0.02	0.90	0.7189	0.9207	0.6359	0.9025
						0.95	0.7788	0.9769	0.7189	0.9609
						0.99	0.8643	0.9996	0.8321	0.9989
5	0.1	t_3	0.99	0.64	0.02	0.90	0.8774	0.9267	0.8067	0.9015
						0.95	0.9194	0.9785	0.8774	0.9638
						0.99	0.9663	0.9991	0.9521	0.9987
5	0.5	Log-normal	0.99	0.85	0.44	0.90	0.7701	0.9207	0.6949	0.9025
						0.95	0.8193	0.9769	0.7701	0.9609
						0.99	0.8995	0.9996	0.8725	0.9989
5	0.5	t_3	0.99	0.64	0.44	0.90	0.8837	0.9290	0.8166	0.9057
						0.95	0.9245	0.9814	0.8840	0.9636
						0.99	0.9686	0.9992	0.9556	0.9985
binary treatment x_i										
5	0.1	Log-normal	0.99	-0.6435	0.006	0.90	0.8861	0.9207	0.8106	0.9025
						0.95	0.9255	0.9769	0.8861	0.9609
						0.99	0.9716	0.9996	0.9575	0.9989
5	0.1	t_3	0.99	-0.5327	0.006	0.90	0.9327	0.9320	0.8735	0.9051
						0.95	0.9627	0.9812	0.9327	0.9652
						0.99	0.9877	0.9992	0.9801	0.9978

Table 4: Simulated probabilities for weak-identification-robust confidence sets to be unbounded for different values of the discontinuity parameter (c), different degrees of endogeneity (κ), and different types of treatment variable x_i. The bandwidth is set to 0.1778, assignment variable is standard normal, and the outcome without treatment is normal.

c	κ	nominal coverage	entire real line	two half-lines
continuous treatment x_i				
2	0.50	0.90	0	0.0002
		0.95	0	0.0002
		0.99	0	0.0020
2	0.99	0.90	0	0.0001
		0.95	0	0.0001
		0.99	0	0.0008
0.1	0.50	0.90	0.7282	0.1614
		0.95	0.8494	0.0973
		0.99	0.9681	0.0221
0.1	0.99	0.90	0.2354	0.6544
		0.95	0.3534	0.5909
		0.99	0.5972	0.3908
0.5	0.50	0.90	0.3572	0.2148
		0.95	0.4894	0.1980
		0.99	0.7348	0.1324
0.5	0.99	0.90	0	0.5637
		0.95	0	0.6847
		0.99	0	0.8626
binary treatment x_i				
2	0.50	0.90	0.0061	0.0139
		0.95	0.0183	0.0222
		0.99	0.0630	0.0551
2	0.99	0.90	0.0007	0.0196
		0.95	0.0021	0.0402
		0.99	0.0134	0.1103
0.1	0.50	0.90	0.7348	0.1592
		0.95	0.8532	0.0945
		0.99	0.9689	0.0217
0.1	0.99	0.90	0.7218	0.1745
		0.95	0.8385	0.1077
		0.99	0.9619	0.0282

Table 5: Simulated size and power of the standard and weak-identification-robust tests for constancy of the RD effect across covariates. There are three groups with RD effects β_j and discontinuities in treatment assignments c_j.

Size of Discontinuity			Treatment Effect			Nominal	Rejection probabilities	
c_1	c_2	c_3	β_1	β_2	β_3	Size	Standard	Robust
2	2	2	0	0	0	0.10	0.0309	0.0858
						0.05	0.0070	0.0372
						0.01	0	0.0060
2	2	0.1	0	0	0	0.10	0.4310	0.0745
						0.05	0.3864	0.0335
						0.01	0.3171	0.0047
2	2	0.1	0	-1	1	0.10	0.9739	0.7961
						0.05	0.9615	0.6978
						0.01	0.9238	0.4660
2	2	0.1	0	-3	3	0.10	1	0.9840
						0.05	1	0.9684
						0.01	1	0.8946
2	0.1	0.1	0	0	0	0.10	0.6581	0.0676
						0.05	0.6104	0.0312
						0.01	0.5009	0.0043
2	0.1	0.1	0	-1	1	0.10	0.6068	0.4059
						0.05	0.5794	0.2709
						0.01	0.5188	0.0932
2	2	2	0	-1	1	0.10	0.9531	0.9958
						0.05	0.9239	0.9921
						0.01	0.8251	0.9656

5 Empirical Applications: Additional Tables

Table 6: Angrist and Lavy (1999): Test of equality of RD effect across groups at 5% significance level for different values of the bandwidth

bandwidth	estimated effect		reject H_0 of equality? robust	standard
	religious	secular		
6	−0.0524	−0.1131	no	no
8	−0.0540	−0.0985	no	no
10	−0.0381	−0.0756	no	no
12	−0.0170	−0.0364	no	no
14	−0.0274	−0.0363	no	no
16	−0.0035	−0.0382	no	no
18	0.0052	−0.0505	yes	no
20	0.0107	−0.0523	yes	no
	<= 10% disadvantaged	> 10% disadvantaged		
6	−0.0390	−0.0909	no	no
8	−0.0626	−0.0469	no	no
10	−0.0387	−0.0488	no	no
12	−0.0259	−0.0192	no	no
14	−0.0343	−0.0226	no	no
16	−0.0290	−0.0079	no	no
18	−0.0368	−0.0037	no	no
20	−0.0360	−0.0008	yes	no

Table 7: Urquiola and Verhoogen (2009): Estimated discontinuity in the treatment variable for the first cutoff and F-statistics for testing for potential size distortions for various values of the bandwidth

bandwidth	discontinuity estimates	F-statistic
6	1.388	0.8226
8	−0.387	0.0812
10	−3.107	6.8069
12	−4.779	20.684
14	−6.092	41.037
16	−7.870	84.236
18	−8.934	129.80
20	−9.968	188.43

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.

References

Angrist, J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 313–336.

Card, D., Dobkin, C., and Maestas, N. (2009), "Does Medicare Save Lives?" *Quarterly Journal of Economics*, 124, 597–636.

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, Duxbury Press, Pacific Grove, CA, 2nd ed.

Davidson, J. (1994), *Stochastic Limit Theory*, New York: Oxford University Press.

Feir, D., Lemieux, T., and Marmer, V. (2015), "Weak Identification in Fuzzy Regression Discontinuity Designs," UBC Working paper.

Greenstone, M. and Gallagher, J. (2008), "Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program," *Quarterly Journal of Economics*, 123, 951–1003.

Hahn, J., Todd, P., and Van der Klaauw, W. (1999), "Evaluating the Effect of an Antidiscrimination Law Using a Regresion-Discontinuity Design," NBER Working Paper 7131.

Hoxby, C. M. (2000), "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, 115, 1239–1285.

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

Jacob, B. and Lefgren, L. (2004), "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *Review of Economics and Statistics*, 86, 226–244.

Kane, T. J. (2003), "A Quasi-Experimental Estimate of the Impact of Financial Aid on College-Going," NBER Working Paper 9703.

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

Lehmann, E. L. and Romano, J. P. (2005), *Testing Statistical Hypotheses*, New York: Springer, 3rd ed.

Li, Q. and Racine, J. S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton, New Jersey: Princeton University Press.

Oreopoulos, P. (2006), "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *American Economic Review*, 152–175.

Pitt, M. M. and Khandker, S. R. (1998), "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" *Journal of Political Economy*, 106, 958–996.

Staiger, D. and Stock, J. H. (1997), "Instrumental Variables Regression With Weak Instruments," *Econometrica*, 65, 557–586.

Thistlethwaite, D. L. and Campbell, D. T. (1960), "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 51, 309–317.

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