Manipulation testing based on density discontinuity

Matias D. Cattaneo
University of Michigan
Ann Arbor, MI
cattaneo@umich.edu

Michael Jansson
University of California at Berkeley and CREATES
Berkeley, CA
mjansson@econ.berkeley.edu

Xinwei Ma
University of Michigan
Ann Arbor, MI
xinweima@umich.edu

Abstract. In this article, we introduce two community-contributed commands, `rddensity` and `rdbwdensity`, that implement automatic manipulation tests based on density discontinuity and are constructed using the results for local-polynomial density estimators in Cattaneo, Jansson, and Ma (2017b, Simple local polynomial density estimators, Working paper, University of Michigan). These new tests exhibit better size properties (and more power under additional assumptions) than other conventional approaches currently available in the literature. The first command, `rddensity`, implements manipulation tests based on a novel local-polynomial density estimation technique that avoids prebinning of the data (improving size properties) and allows for restrictions on other features of the model (improving power properties). The second command, `rdbwdensity`, implements several bandwidth selectors specifically tailored for the manipulation tests discussed herein. We also provide a companion R package with the same syntax and capabilities as `rddensity` and `rdbwdensity`.

Keywords: st0522, rddensity, rdbwdensity, falsification test, manipulation test, regression discontinuity

1 Introduction

McCrary (2008) introduced the idea of manipulation testing in the context of regression discontinuity (RD) designs. Consider a setting where each unit in a random sample from a large population is assigned to one of two groups depending on whether one of their observed covariates exceeds a known cutoff. In this context, the two possible groups are generically referred to as control and treatment groups. The observed variable, determining group assignment, is generically referred to as the score, index, or running variable. The key idea behind manipulation testing in this context is that in the absence of systematic manipulation of the unit's index around the cutoff, the density of units should be continuous near this cutoff value. Thus, a manipulation test seeks to formally determine whether there is evidence of a discontinuity in the density of units at the

known cutoff. Presence of such evidence is usually interpreted as empirical evidence of self-selection or nonrandom sorting of units into control and treatment status.

Manipulation testing is useful for falsification of RD designs: see Cattaneo and Escanciano (2017) for an edited volume with a recent overview of the RD literature; see Cattaneo, Titiunik, and Vazquez-Bare (2017c) for a practical introduction to RD designs with a comparison between leading empirical methods; see Calonico, Cattaneo, and Titiunik (2015a) for a discussion of graphical presentation and falsification of RD designs; and see references therein for other related topics. In addition to providing a formal statistical check for RD designs, a manipulation test can be used substantively whenever the empirical goal is to test for self-selection or endogenous sorting of units exposed to a known hard threshold-crossing assignment rule. Thus, flexible data-driven implementations of manipulation tests with good size and power properties are potentially very useful for empirical work in economics and related social sciences.

To implement a manipulation test, the researcher must estimate the density of units near the cutoff to conduct a hypothesis test about whether the density is discontinuous. Three distinct manipulation tests have been proposed in the literature. First, McCrary (2008) introduced a test based on the nonparametric local-polynomial density estimator of Cheng, Jianqing, and Marron (1997), which requires prebinning of the data and hence introduces additional tuning parameters. Second, Otsu, Xu, and Matsushita (2014) proposed an empirical likelihood method using boundary-corrected kernels. Third, Cattaneo, Jansson, and Ma (2017b) developed a set of manipulation tests based on a novel local-polynomial density estimator, which does not require prebinning of the data and is constructed in an intuitive way based on easy-to-interpret kernel functions. The latter procedures are shown to also provide demonstrable improvements in both size and power under appropriate assumptions and relative to the other approaches currently available in the literature. Finally, Frandsen (2017) recently proposed a manipulation test in the context of RD designs with a discrete running variable.

In this article, we discuss data-driven implementations of manipulation tests following the results in Cattaneo, Jansson, and Ma (2017b). We introduce two commands that together give several manipulation test implementations, which depend on i) the restrictions imposed in the underlying data-generating process, ii) the method for bias correction, iii) the bandwidth selection approach, and iv) the method to estimate standard errors (SEs), among many other alternatives. Specifically, our command `rddensity` implements two distinct manipulation tests given a choice of bandwidth and SE estimator: one test is constructed using the basic Wald statistic, which requires undersmoothing, while the other test uses robust bias-correction (Calonico, Cattaneo, and Farrell Forthcoming) to obtain valid statistical inference. This command also allows for both unrestricted and restricted models, where the cumulative distribution function and higher-order derivatives are assumed to be equal for both groups (thereby increasing the power of the test). These methods can also be used under additional assumptions in the context of RD designs with a discrete running variable. We review the main aspects of these different approaches to manipulation testing in section 2 below, but we refer the reader to Cattaneo, Jansson, and Ma (2017b) and its supplemental appendix for most of the technical and theoretical discussion.

The command `rddensity` also offers a plot of the manipulation test. To implement this plot, a density estimate must be constructed not only at the cutoff point but also at nearby evaluation points, which may also be affected by boundary bias. Thus, to construct this plot in a principled way, `rddensity` uses the package `lpdensity`, which implements local-polynomial–based density estimation methods. This density estimation package must be installed to construct the density plot; see Cattaneo, Jansson, and Ma (2017a) for further details. If `lpdensity` is not installed, then `rddensity` issues an error message when trying to construct a manipulation test plot.

To complement the command `rddensity`, and because in empirical applications researchers often want to select the bandwidth entering the manipulation test in a data-driven and automatic way, we introduce the companion command `rdbwdensity`, which provides several bandwidth selection methods based on asymptotic mean squared error (AMSE) minimization. Our implementation accounts for whether the unrestricted or the restricted model is used for inference, and it also allows for both different bandwidths on either side of the cutoff (whenever possible) and a common bandwidth for both sides. By default, `rddensity` uses the companion command `rdbwdensity` to estimate the bandwidth(s) whenever the user does not provide a specific choice, thereby giving fully automatic and data-driven inference procedures implemented by `rddensity`.

The commands `rddensity` and `rdbwdensity` complement the recently introduced Stata community-contributed commands and R functions `rdrobust`, `rdbwselect`, and `rdplot`, which are useful for graphical presentation, estimation, and inference in RD designs and use nonparametric local-polynomial techniques. For an introduction to the latter commands, see Calonico, Cattaneo, and Titiunik (2014a, 2015b) and Calonico et al. (2017). Together, the five commands offer a complete toolkit for empirical work using RD designs. In addition, see Cattaneo, Titiunik, and Vazquez-Bare (2016) for Stata commands and R functions (`rdrandinf`, `rdwinselect`, `rdsensitivity`, `rdrbounds`) implementing randomization-based inference methods for RD designs under a local randomization assumption.

The rest of this article is organized as follows. In section 2, we provide a brief review of the methods implemented in our two commands. In sections 3 and 4, we describe the syntax of `rddensity` and `rdbwdensity`, respectively. In section 5, we illustrate some of the functionalities of our commands using real data from Cattaneo, Frandsen, and Titiunik (2015). In section 6, we report results from a small-scale simulation study, tailored to investigate the performance of our testing procedure. We also provide a companion R package with the same functionalities and syntax.

The latest version of this software, as well as other related software for RD designs, can be found at https://sites.google.com/site/rdpackages/.

2   Methods overview

This section offers a brief overview of the methods implemented in our commands `rddensity` and `rdbwdensity`. We closely follow the results in Cattaneo, Jansson, and Ma (2017b), including those in their supplemental appendix. Regularity conditions and most of the technical details are not discussed here to conserve space and ease the exposition.

2.1   Setup and notation

We assume that $\{X_1, X_2, \ldots, X_n\}$ is a random sample of size n from the random variable X with the cumulative distribution function (c.d.f.) and probability density function given by $F(x)$ and $f(x)$, respectively. The random variable X_i denotes the score, index, or running variable of unit i in the sample. Each unit is assigned to control or treatment depending on whether its observed index exceeds a known cutoff denoted by $\overline{x}$. That is, treatment assignment is given by

$$\begin{array}{ll} \text{unit } i \text{ assigned to control group} & \text{if } X_i < \overline{x} \\ \text{unit } i \text{ assigned to treatment group} & \text{if } X_i \geq \overline{x} \end{array}$$

where the cutoff point $\overline{x}$ is known and, of course, we assume enough observations for each group are available (for example, $f(x) > 0$ near $\overline{x}$ and the sample is large enough). In the specific case of RD designs, manipulation testing can be used for sharp RD designs (where treatment assignment and treatment status coincide) and for fuzzy RD designs (where treatment assignment and treatment status differ). In the latter case, of course, the test applies to the intention-to-treat mechanism because units can select into treatment or control status beyond the hard-thresholding rule for treatment assignment (that is, assigned to control group if $X_i < \overline{x}$ and assigned to treatment group if $X_i \geq \overline{x}$).

A manipulation test in this context is a hypothesis test on the continuity of the density $f(\cdot)$ at the cutoff point $\overline{x}$. Formally, we are interested in the testing problem

$$\mathsf{H}_0 : \lim_{x \uparrow \overline{x}} f(x) = \lim_{x \downarrow \overline{x}} f(x) \qquad \text{versus} \qquad \mathsf{H}_a : \lim_{x \uparrow \overline{x}} f(x) \neq \lim_{x \downarrow \overline{x}} f(x)$$

To construct a test statistic for this hypothesis testing problem, we follow Cattaneo, Jansson, and Ma (2017b) and estimate the density $f(x)$ using a local-polynomial density estimator based on the c.d.f. of the observed sample. This estimator has several interesting properties, including the fact that it does not require prebinning of the data and is quite intuitive in its implementation (for example, simple second-order kernels can be used). Importantly, this estimator also permits incorporating restrictions on the c.d.f. and higher-order derivatives of the density, leading to new manipulation tests with better power properties in applications. For an introduction to conventional local-polynomial techniques, see, for example, Fan and Gijbels (1996).

The manipulation test statistics implemented in `rddensity` take the form

$$T_p(h) = \frac{\widehat{f}_{+,p}(h) - \widehat{f}_{-,p}(h)}{\widehat{V}_p(h)} \qquad \widehat{V}_p^2(h) = \widehat{\mathbb{V}}\left\{ \widehat{f}_{+,p}(h) - \widehat{f}_{-,p}(h) \right\}$$

where $T_p(h) \overset{a}{\sim} \mathcal{N}(0, 1)$ under appropriate assumptions, and the notation $\widehat{\mathbb{V}}\{\cdot\}$ is meant to denote some plug-in consistent estimator of the population quantity $\mathbb{V}\{\cdot\}$. The parameter h is the bandwidth(s) used to localize the estimation and inference procedures near the cutoff point $\overline{x}$. The statistics may be constructed in several different ways, as we discuss in more detail below. In particular, given a choice of bandwidth(s), two main ingredients to construct the test statistic $T_p(h)$ are i) the local-polynomial density estimators $\widehat{f}_{+,p}(h)$ and $\widehat{f}_{-,p}(h)$, and ii) the corresponding SE estimator $\widehat{V}_p(h)$. These estimators also depend on the choice of polynomial order p, the choice of kernel function $K(\cdot)$, and the restrictions imposed in the model, among other possibilities. The SE formulas $\widehat{V}_p(h)$ could be based on either an asymptotic plug-in or a jackknife approach, and its specific form will depend on whether additional restrictions are imposed to the model. A crucial ingredient is, of course, the choice of bandwidth h, which determines which observations near the cutoff $\overline{x}$ are used for estimation and inference. This choice can either be specified by the user or be estimated using the available data. Our commands allow, when possible, for different bandwidth choices on either side of the cutoff $\overline{x}$. A common bandwidth on both sides of the cutoff is always possible.

In the following two subsections, we discuss the alternatives for estimation and inference: i) unrestricted inference, ii) restricted inference, iii) SE estimation in both cases, and iv) bandwidth selection in both cases. In closing this section, we also offer a brief review of the different data-driven inference methods implemented in our Stata (and R) commands.

2.2 Unrestricted testing

In unrestricted testing, the manipulation test becomes a standard two-sample problem where the estimators $\widehat{f}_{+,p}(h)$ and $\widehat{f}_{-,p}(h)$ are unrelated. Thus, the SE formula reduces to $\widehat{V}_p^2(h) = \widehat{\mathbb{V}}\{\widehat{f}_{+,p}(h) - \widehat{f}_{-,p}(h)\} = \widehat{\mathbb{V}}\{\widehat{f}_{+,p}(h)\} + \widehat{\mathbb{V}}\{\widehat{f}_{-,p}(h)\}$. To be more concrete, the density estimators take the form

$$\widehat{f}_{-,p}(h) = \mathbf{e}_1' \widehat{\boldsymbol{\beta}}_{-,p}(h) \qquad \text{and} \qquad \widehat{f}_{+,p}(h) = \mathbf{e}_1' \widehat{\boldsymbol{\beta}}_{+,p}(h)$$

with

$$\widehat{\boldsymbol{\beta}}_{-,p}(h) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \mathbb{1}(X_i < \overline{x}) \left\{ \widehat{F}(X_i) - \mathbf{r}_p(X_i - \overline{x})' \boldsymbol{\beta} \right\}^2 K_h(X_i - \overline{x})$$

$$\widehat{\boldsymbol{\beta}}_{+,p}(h) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \mathbb{1}(X_i \geq \overline{x}) \left\{ \widehat{F}(X_i) - \mathbf{r}_p(X_i - \overline{x})' \boldsymbol{\beta} \right\}^2 K_h(X_i - \overline{x})$$

where $\widehat{F}(X_i)$ denotes the (leave-one-out) c.d.f. estimator for the full sample, $\mathbf{r}_p(x) = (1, x, \ldots, x^p)'$, $\mathbf{e}_1 = (0, 1, 0, 0, \ldots, 0)' \in \mathbb{R}^{p+1}$ is the second unit vector, $K_h(u) = K(u/h)/h$ with $K(\cdot)$ being a kernel function, h is a positive bandwidth sequence, and $\mathbb{1}(\cdot)$ denotes the indicator function. Following the results in Cattaneo, Jansson, and Ma (2017b), it can be shown that $\widehat{\boldsymbol{\beta}}_{-,p}(h)$ and $\widehat{\boldsymbol{\beta}}_{+,p}(h)$ roughly approximate, respectively,

$[F_-, f_-, (1/2!)f_-^{(1)}, \ldots, \{1/(p+1)!\}f_-^{(p)}]$ and $[F_+, f_+, (1/2!)f_+^{(1)}, \ldots, \{1/(p+1)!\}f_+^{(p)}]$, where we use the notation

$$f_-^{(s)} = \lim_{x \uparrow \overline{x}} \frac{\partial^s}{\partial x^s} f(x) \qquad \text{and} \qquad f_+^{(s)} = \lim_{x \downarrow \overline{x}} \frac{\partial^s}{\partial x^s} f(x)$$

$s = 1, 2, \ldots, p$. Thus, the second elements of $\widehat{\boldsymbol{\beta}}_{-,p}(h)$ and $\widehat{\boldsymbol{\beta}}_{+,p}(h)$ are used to construct consistent, boundary-corrected density estimators at the cutoff point $\overline{x}$ entering the numerator of the manipulation test statistic. Notice that this estimation approach avoids prebinning of the data and may be constructed using simple and easy-to-interpret kernels $K(\cdot)$, such as the uniform or triangular kernels.

Consistency, asymptotic normality, and moment approximations for the estimators $\widehat{f}_{-,p}(h)$ and $\widehat{f}_{+,p}(h)$ are derived in Cattaneo, Jansson, and Ma (2017b), where these results are also used to study the asymptotic properties of the unrestricted manipulation tests implemented in our main command, `rddensity`. We discuss the exact implementations at the end of this section.

2.3 Restricted testing

The approach described above treats manipulation testing as a traditional two-sample problem, where the left and right approximations to the density $f(x)$ at the cutoff $\overline{x}$ are done independently. However, in the context of manipulation testing, it may be argued that the c.d.f. $F(x)$ and higher-order derivatives $f^{(s)}(x)$, $s \geq 1$, are equal for both groups at the cutoff even when $f_- \neq f_+$ (that is, H_0 is false). In this case, researchers may wish to incorporate these restrictions to construct a more powerful testing procedure.

A restricted manipulation test is constructed by solving the above weighted (local) least-squares problem with the additional restrictions ensuring that all but the second element in $\boldsymbol{\beta}$ are equal in both groups. It follows that this restricted problem can be represented as a single regression problem,

$$\widehat{\boldsymbol{\beta}}_p^{\mathtt{R}}(h) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+2}} \sum_{i=1}^n \left\{ \widehat{F}(X_i) - \mathbf{r}_p^{\mathtt{R}}(X_i - \overline{x})' \boldsymbol{\beta} \right\}^2 K_h(X_i - \overline{x})$$

where $\mathbf{r}_p^{\mathtt{R}}(x) = \{1, x \cdot \mathbb{1}(x < \overline{x}), x \cdot \mathbb{1}(x \geq \overline{x}), x^2, x^3, \ldots, x^p\}'$. In words, this estimation approach incorporates the restrictions ensuring that the estimated c.d.f. and the estimated higher-order derivatives are equal on both sides of the cutoff point $\overline{x}$. This problem involves estimating only $p + 2$ parameters, rather than $2p + 2$ parameters as is the case for the unrestricted method discussed above. The main advantage of imposing these restrictions is related to power improvements, provided the restrictions are indeed satisfied in the underlying data-generating process.

Therefore, a restricted manipulation test uses the density estimators

$$\widehat{f}_{-,p}^{\mathtt{R}}(h) = \mathbf{e}_1' \widehat{\boldsymbol{\beta}}_p^{\mathtt{R}}(h) \qquad \text{and} \qquad \widehat{f}_{+,p}^{\mathtt{R}}(h) = \mathbf{e}_2' \widehat{\boldsymbol{\beta}}_p^{\mathtt{R}}(h)$$

That is, the density estimators from the left and from the right of the cutoff point $\overline{x}$ are given by the second and third elements in the least-squares vector $\widehat{\boldsymbol{\beta}}_p^{\mathtt{R}}(h)$; $\mathbf{e}_2 =$

$(0, 0, 1, 0, 0, \ldots, 0)' \in \mathbb{R}^{p+1}$ denotes the third unit vector. Furthermore, the SE formula is different because now cross-restrictions are incorporated in the estimation procedure, leading to a different asymptotic variance for $\{\widehat{f}^{\mathtt{R}}_{-,p}(h), \widehat{f}^{\mathtt{R}}_{+,p}(h)\}$ and, consequently, for $\widehat{f}^{\mathtt{R}}_{+,p}(h) - \widehat{f}^{\mathtt{R}}_{-,p}(h)$ as well. This is exactly the source of the power gains of a restricted manipulation test relative to an unrestricted one. In this case, the SE formula $\widehat{V}_p(h) = \widehat{\mathbb{V}}\{\widehat{f}^{\mathtt{R}}_{+,p}(h) - \widehat{f}^{\mathtt{R}}_{-,p}(h)\} \neq \widehat{\mathbb{V}}\{\widehat{f}^{\mathtt{R}}_{+,p}(h)\} + \widehat{\mathbb{V}}\{\widehat{f}^{\mathtt{R}}_{-,p}(h)\}$.

Formal asymptotic properties for restricted local-polynomial density estimation and inference are also discussed in Cattaneo, Jansson, and Ma (2017b), where these results are then used to propose an asymptotically valid restricted manipulation test. We also discuss the exact implementations at the end of this section.

2.4   Standard errors

As mentioned above, the asymptotic variance entering the denominator of the manipulation test statistic $T_p(h)$ will be different depending on whether an unrestricted or a restricted model is used. Our commands rddensity and rdbwdensity allow for both cases. In addition, for each of these cases, the commands provide two distinct consistent SE estimators: i) a plug-in estimator based on the asymptotic variance of the numerator of $T_p(h)$, and ii) a jackknife estimator based on the leading term of an expansion of asymptotic variance of the numerator of $T_p(h)$.

The plug-in estimator is faster because it essentially requires no additional estimation beyond the quantities entering the numerator of $T_p(h)$, but it relies on asymptotic approximations. On the other hand, the jackknife estimator is slower because it requires additional estimation and looping over the data, but according to simulation evidence in Cattaneo, Jansson, and Ma (2017b), it appears to provide a more accurate approximation to the finite-sample variability of the numerator of $T_p(h)$ in both cases (unrestricted model and restricted model). Therefore, our implementations use the jackknife SE estimator by default, but we also offer the plug-in estimator for cases involving relatively large sample sizes.

2.5   Bandwidth selection

rddensity requires the specification of bandwidths for estimation; otherwise, it uses rdbwdensity to construct data-driven bandwidth choices specifically tailored for the manipulation tests discussed in this article. In this subsection, we briefly outline the data-driven implementations provided in rdbwdensity for automatic bandwidth selection.

In the unrestricted model, the user has the option to specify two distinct bandwidths: $h_{\mathtt{l}}$ for left estimation and $h_{\mathtt{r}}$ for right estimation. Of course, one such choice may be equal bandwidths: $h = h_{\mathtt{l}} = h_{\mathtt{r}}$. In the restricted model, however, only a common bandwidth h can be specified because estimation is done jointly by construction. For each of these cases, whenever possible, Cattaneo, Jansson, and Ma (2017b) develop

three distinct approaches to select the bandwidth(s) using the mean squared error (MSE) criterion function, generically denoted by $\text{MSE}(\widehat{\theta}) = \mathbb{E}\{(\widehat{\theta} - \theta)^2\}$, where $\widehat{\theta}$ denotes some estimator and θ denotes its target estimand.

For the specific context considered in this article, Cattaneo, Jansson, and Ma (2017b) develop valid asymptotic expansions of several MSE criterion functions. These results can be briefly summarized as

$$\text{MSE}\left\{\widehat{\theta}(h)\right\} \approx \text{AMSE}\left\{\widehat{\theta}(h)\right\}$$

where

$$\text{AMSE}\left\{\widehat{\theta}(h)\right\} = h^{p+1}B_p^2(\theta) + h^{p+2}B_{p+1}^2(\theta) + \frac{1}{nh}V_p(\theta)$$

with, for the unrestricted model,

$$\widehat{\theta}(h) \text{ representing one of } \left\{\widehat{f}_{-,p}(h);\ \widehat{f}_{+,p}(h);\ \widehat{f}_{+,p}(h) - \widehat{f}_{-,p}(h);\ \widehat{f}_{+,p}(h) + \widehat{f}_{-,p}(h)\right\}$$

and with, for the restricted model,

$$\widehat{\theta}(h) \text{ representing one of } \left\{\widehat{f}_{+,p}^{\mathtt{R}}(h) - \widehat{f}_{-,p}^{\mathtt{R}}(h);\ \widehat{f}_{+,p}^{\mathtt{R}}(h) + \widehat{f}_{-,p}^{\mathtt{R}}(h)\right\}$$

and, of course, with

$$\theta \text{ representing one of } \{f_-;\ f_+;\ f_+ - f_-;\ f_+ + f_-\}$$

as appropriate according to the choice of $\widehat{\theta}(h)$. Crucially, for each combination of estimator $\widehat{\theta}(h)$ and estimand θ, the corresponding bias constants $\{B_p(\theta), B_{p+1}(\theta)\}$ and variance constant $V_p(\theta)$ are different. In all cases, as is usual in nonparametric problems, these constants involve features of both the data-generating process and the nonparametric estimator.

Given a choice of estimator and estimand, and provided that preliminary estimates of the leading asymptotic constants in the associated MSE expansion are available, it is straightforward to construct a plug-in bandwidth selector. In our implementations, we consider the five alternative plug-in rules for bandwidth selection mentioned above, which use simple rules of thumb to approximate the leading constants in the MSE expansions. Technical and methodological details underlying these choices are given in Cattaneo, Jansson, and Ma (2017b) and not reproduced here to conserve space.

Specifically, `rdbwdensity` allows for the following alternative bandwidth selectors. We do not introduce additional notation to reflect the estimation of the leading AMSE constants only to ease the exposition, but our implementations are automatic because they rely on preliminary rule-of-thumb estimators to approximate those constants.

- *Unrestricted model with different bandwidths*: when no restrictions are imposed on the model and $(h_{\mathtt{l}}, h_{\mathtt{r}})$ are allowed to be different, the bandwidths are chosen to minimize the AMSE of the density estimators separately, that is,

$$\widehat{h}_{\mathtt{l},p} = \text{argmin}_{h>0}\text{AMSE}\left\{\widehat{f}_{-,p}(h)\right\} \quad \text{and} \quad \widehat{h}_{\mathtt{r},p} = \text{argmin}_{h>0}\text{AMSE}\left\{\widehat{f}_{+,p}(h)\right\}$$

- *Unrestricted model with equal bandwidth*: when no restrictions are imposed on the model but $(h_{\mathtt{l}}, h_{\mathtt{r}})$ are forced to be equal, the common bandwidth may be chosen in two distinct ways:

  1. Difference of densities:
  $$\widehat{h}_{\mathtt{diff},p} = \mathrm{argmin}_{h>0}\mathrm{AMSE}\left\{\widehat{f}_{+,p}(h) - \widehat{f}_{-,p}(h)\right\}$$

  2. Sum of densities:
  $$\widehat{h}_{\mathtt{sum},p} = \mathrm{argmin}_{h>0}\mathrm{AMSE}\left\{\widehat{f}_{+,p}(h) + \widehat{f}_{-,p}(h)\right\}$$

- *Restricted model*: when the restrictions are imposed on the model, then $h = h_{\mathtt{l}} = h_{\mathtt{r}}$ by construction. Also in this case, the common bandwidth may be chosen in two distinct ways:

  1. Difference of densities:
  $$\widehat{h}^{\mathtt{R}}_{\mathtt{diff},p} = \mathrm{argmin}_{h>0}\mathrm{AMSE}\left\{\widehat{f}^{\mathtt{R}}_{+,p}(h) - \widehat{f}^{\mathtt{R}}_{-,p}(h)\right\}$$

  2. Sum of densities:
  $$\widehat{h}^{\mathtt{R}}_{\mathtt{sum},p} = \mathrm{argmin}_{h>0}\mathrm{AMSE}\left\{\widehat{f}^{\mathtt{R}}_{+,p}(h) + \widehat{f}^{\mathtt{R}}_{-,p}(h)\right\}$$

All the bandwidth selectors above have closed-form solutions, and their specific decay rates (as a function of the sample size) depend on the specific choices. This is an important point because $B_p(\theta)$ or $B_{p+1}(\theta)$ may be 0 depending on the choice of θ and p. For example, if $\theta = f_+ - f_-$ then $B_p(\theta) \propto f_+^{(p)} - f_-^{(p)}$ when $p = 2$, but $B_p(\theta) \propto f_+^{(p)} + f_-^{(p)}$ when $p = 3$, which implies that $B_p(\theta) = 0$ under the plausible assumption that $f_+^{(p)} = f_-^{(p)}$. Following the discussion in Cattaneo, Jansson, and Ma (2017b), we also implement two simple "regularization" approaches that avoid these possible degeneracies:

- *Unrestricted model with different bandwidths*:
  $$\widehat{h}_{\mathtt{l},\mathtt{comb},p} = \mathrm{median}\left(\widehat{h}_{\mathtt{l},p}, \widehat{h}_{\mathtt{diff},p}, \widehat{h}_{\mathtt{sum},p}\right)$$

  $$\widehat{h}_{\mathtt{r},\mathtt{comb},p} = \mathrm{median}\left(\widehat{h}_{\mathtt{r},p}, \widehat{h}_{\mathtt{diff},p}, \widehat{h}_{\mathtt{sum},p}\right)$$

- *Unrestricted model with equal bandwidth*:
  $$\widehat{h}_{\mathtt{comb},p} = \mathrm{min}\left(\widehat{h}_{\mathtt{diff},p}, \widehat{h}_{\mathtt{sum},p}\right)$$

- *Restricted model*:
  $$\widehat{h}^{\mathtt{R}}_{\mathtt{comb},p} = \mathrm{min}\left(\widehat{h}^{\mathtt{R}}_{\mathtt{diff},p}, \widehat{h}^{\mathtt{R}}_{\mathtt{sum},p}\right)$$

2.6 Overview of methods

We have discussed how the density point estimators and corresponding SEs are constructed to form the test statistic $T_p(h)$. As mentioned above, these estimators depend on whether the unrestricted or the restricted model is considered. In addition, the bandwidth $h = (h_\mathtt{l}, h_\mathtt{r})$ may be chosen in different ways, including both cases where the two bandwidths are assumed equal and cases where they are allowed to be different. In this section, we close the discussion of manipulation testing by briefly addressing the issue of critical value (or quantile) choice to form the testing procedure.

As mentioned in passing before, we rely on large-sample approximations to justify a result of the form $T_p(h) \overset{a}{\sim} \mathcal{N}(0,1)$ provided that an appropriate choice of bandwidth h and polynomial order p is used. Specifically, the command `rddensity` implements three distinct methods for inference. Each of these methods takes a different approach to handle the potential presence of a first-order bias in the statistic $T_p(h)$ when a large bandwidth h is used (for example, when any of the MSE-optimal bandwidth choices discussed above are used).

To briefly discuss the three alternative manipulation tests, we let $h_{\mathrm{MSE,p}}$ denote any of the bandwidth choices given above when pth order local-polynomial density estimators are used. The following discussion applies to all cases, whether they are unrestricted or restricted models with equal or unequal bandwidths chosen by any of the methods mentioned before. Let $\alpha \in (0,1)$ and $\chi_1^2(\alpha)$ denote the αth quantile of a chi-squared distribution with 1 degree of freedom. The three methods for inference implemented in `rddensity` are as follows:

- *Robust bias-correction approach.* This approach uses ideas in Calonico, Cattaneo, and Titiunik (2014b) and Calonico, Cattaneo, and Farrell (Forthcoming), which are based on analytic bias-correction coupled with variance adjustments, leading to testing procedures with improved distributional properties in finite samples. In particular, letting $q \geq p + 1$, the manipulation test takes the form

  $$\text{An } \alpha\text{-level test rejects } \mathsf{H}_0 \qquad \text{iff} \qquad T_q^2(h_{\mathrm{MSE,p}}) > \chi_1^2(1 - \alpha)$$

  In words, the construction of $T_q^2(h_{\mathrm{MSE,p}})$ ensures an asymptotically valid distributional approximation because $q \geq p+1$. Thus, in this case, the possible first-order bias of the statistic $T_p^2(h_{\mathrm{MSE,p}})$ (note the change from q to p in the subindex) is removed by using a higher-order polynomial in the estimation of the densities and, of course, adjusting the SE formulas accordingly.

  This approach to manipulation testing is the default option in `rddensity` because it is always theoretically justified and leads to power improvements asymptotically. For example, see the simulations presented in section 6.

- *Small bias approach/undersmoothing approach.* A second approach to inference would be to ignore the impact of a possibly first-order bias in the distributional approximation for the statistic $T_p^2(h_{\mathrm{MSE,p}})$. In some specific cases, such an approach may be justified [for example, if $B_p(\theta)$ happens to be exactly 0 for a choice

of p and θ] and the bandwidth was chosen in an ad-hoc manner, but this cannot be justified in general. Alternatively, this approach is valid if the user uses an undersmoothed bandwidth choice relative to the MSE-optimal bandwidth $h_{\mathrm{MSE,p}}$.

For completeness, `rddensity` also implements a conventional Wald test without bias correction. To describe this procedure, let's suppose h is the bandwidth used. Then the command implements the conventional testing procedure

$$\text{Reject } \mathsf{H}_0 \qquad \text{iff} \qquad T_p^2(h) > \chi_1^2(1-\alpha)$$

which delivers a valid α-level test only when the smoothing bias in approximating $f(x)$ at the cutoff point $\overline{x}$ is indeed small. In practice, this requires choosing an undersmoothed bandwidth; for example, an ad hoc choice is $h = s \cdot h_{\mathrm{MSE,p}}$ for some user-chosen scale value $s \in (0,1)$. This testing procedure is also reported when the option `all` is specified.

3   The rddensity command

This section describes the syntax of the command `rddensity`, which implements the manipulation tests for a choice of bandwidth(s).

3.1   Syntax

`rddensity` *runvar* $\big[$ *if* $\big]$ $\big[$ *in* $\big]$ $\big[$, `c(`*cutoff*`)` `p(`*pvalue*`)` `q(`*qvalue*`)`

   `fitselect(`*fitmethod*`)` `kernel(`*kernelfn*`)` `h(`*hlvalue hrvalue*`)`

   `bwselect(`*bwmethod*`)` `vce(`*vcemethod*`)` `all` `plot` `plot_range(`*xmin xmax*`)`

   `plot_n(`*nl nr*`)` `plot_grid(`*gridmethod*`)` `genvars(`*varname*`)` `level(`*#*`)`

   `graph_options(`*graphopts*`)` $\big]$

   runvar is the running variable (also known as the score or index variable).

3.2   Options

`c(`*cutoff*`)` specifies the RD cutoff. The default is `c(0)`.

`p(`*pvalue*`)` specifies the order of the local polynomial used to construct the density point estimators. The default is `p(2)` (local quadratic approximation).

`q(`*qvalue*`)` specifies the order of the local polynomial used to construct the bias-corrected density point estimators. The default is `q(p(#)+1)` (local cubic approximation for the default `p(2)`).

`fitselect`(*fitmethod*) specifies whether restrictions should be imposed. *fitmethod* may be one of the following:

> `unrestricted` for density estimation without any restrictions (two-sample, unrestricted inference). This is the default option.

> `restricted` for density estimation assuming equal c.d.f. and higher-order derivatives.

`kernel`(*kernelfn*) is the kernel function $K(\cdot)$ used to construct the local-polynomial estimator(s). *kernelfn* may be `triangular`, `uniform`, or `epanechnikov`. The default is `kernel(triangular)`.

`h`(*hlvalue hrvalue*) are the values of the main bandwidths $h_\mathtt{l}$ and $h_\mathtt{r}$, respectively. If only one value is specified, then $h_\mathtt{l} = h_\mathtt{r}$ is used. If not specified, they are computed by the companion command `rdbwdensity`. If two bandwidths are specified, the first bandwidth is used for the data below the cutoff, and the second bandwidth is used for the data above the cutoff.

`bwselect`(*bwmethod*) specifies the bandwidth selection procedure to be used. *bwmethod* may be one of the following:

> `each` specifies bandwidth selection based on the MSE of each density separately, that is, $\widehat{h}_{\mathtt{l},p}$ and $\widehat{h}_{\mathtt{r},p}$. This is available only when the unrestricted model is used.

> `diff` specifies bandwidth selection based on the MSE of the difference of densities, that is, $\widehat{h}_{\mathtt{diff},p}$.

> `sum` specifies bandwidth selection based on the MSE of the sum of densities, that is, $\widehat{h}_{\mathtt{sum},p}$.

> `comb` specifies bandwidth selection as a combination of the alternatives above, that is, either $(\widehat{h}_{\mathtt{l},\mathtt{comb},p}, \widehat{h}_{\mathtt{r},\mathtt{comb},p})$, $\widehat{h}_{\mathtt{comb},p}$, or $\widehat{h}^{\mathtt{R}}_{\mathtt{comb},p}$, depending on the model chosen. This is the default option.

`vce`(*vcemethod*) specifies the procedure used to compute the variance–covariance matrix estimator. *vcemethod* may be one of the following:

> `plugin` for asymptotic plug-in SEs.

> `jackknife` for jackknife SEs. This is the default option.

`all` reports two different manipulation tests (given choices `fitselect`(*fitmethod*) and `bwselect`(*bwmethod*)): the conventional test statistic (not valid when using the MSE-optimal bandwidth choice) and the robust bias-corrected statistic (the default option).

`plot` plots the density around the cutoff (this feature depends on the companion package `lpdensity`). Note that additional estimation (computing time) is needed.

`plot_range`(*xmin xmax*) specifies lower (*xmin*) and upper (*xmax*) endpoints for the manipulation test plot. By default, it is three bandwidths around the cutoff.

plot_n(*nl nr*) specifies the number of evaluation points below (*nl*) and above (*nr*) the
   cutoff to be used for the manipulation test plot. The default is plot_n(10 10).

plot_grid(*gridtype*) specifies the location of evaluation points, whether they are evenly
   spaced (es) or quantile spaced (qs), to be used for the manipulation test plot. The
   default is plot_grid(es).

genvars(*varname*) specifies that new variables should be generated to store estimation
   results for plotting. See the help file for details.

level(*#*) specifies the level of confidence intervals for the manipulation test plot. *#*
   must be between 0 and 100. The default is level(95).

graph_options(*graphopts*) are graph options passed to the plot command.

3.3 Description

rddensity provides an implementation of manipulation tests using local-polynomial
density estimators. The user must specify the running variable. This command permits
fully data-driven inference by using the companion command rdbwdensity, which may
also be used as a standalone command.

4 The rdbwdensity command

This section describes the syntax of the command rdbwdensity. This command imple-
ments the different bandwidth selection procedures for manipulation tests based on the
local-polynomial density estimators discussed above.

4.1 Syntax

rdbwdensity *runvar* $\left[\,if\,\right]$ $\left[\,in\,\right]$ $\left[\,,\,$ c(*cutoff*) p(*pvalue*) fitselect(*fitmethod*)
   kernel(*kernelfn*) vce(*vcemethod*) $\left.\right]$

   runvar is the running variable (also known as the score or index variable).

4.2 Options

c(*cutoff*) specifies the RD cutoff. The default is c(0).

p(*pvalue*) specifies the order of the local polynomial used to construct the point esti-
   mator. The default is p(2) (local quadratic approximation).

`fitselect`(*fitmethod*) specifies the model used for estimation and inference. *fitmethod* may be one of the following:

  `unrestricted` for density estimation without any restrictions (two-sample, unrestricted inference). This is the default option.

  `restricted` for density estimation assuming equal c.d.f. and higher-order derivatives.

`kernel`(*kernelfn*) specifies the kernel function $K(\cdot)$ that is used to construct the local-polynomial estimator(s). *kernelfn* may be `triangular`, `uniform`, or `epanechnikov`. The default is `kernel(triangular)`.

`vce`(*vcemethod*) specifies the procedure used to compute the variance–covariance matrix estimator. *vcemethod* may be

  `plugin` for asymptotic plug-in SEs.

  `jackknife` for jackknife SEs. This is the default option.

4.3   Description

`rdbwdensity` implements several bandwidth selection procedures specifically tailored for manipulation testing based on the local-polynomial density estimators implemented in `rddensity`. The user must specify the running variable.

5   Illustration of methods

We illustrate our community-contributed commands using the same dataset already used in Calonico et al. (2017) and Cattaneo, Titiunik, and Vazquez-Bare (2016), where related RD commands are introduced and discussed. This facilitates comparison across the Stata and R packages available for analysis and interpretation of RD designs; the package introduced herein discusses a falsification approach to RD designs via manipulation testing, while the other packages discuss graphical presentation and falsification, estimation, and inference.

  `rddensity_senate.dta` contains the running variable from a larger dataset constructed and studied in Cattaneo, Frandsen, and Titiunik (2015), focusing on party advantages in U.S. Senate elections for the period 1914–2010. Thus, the unit of observation is a state in the United States. In this section, we focus on the running variable used to analyze the RD effect of the Democratic party winning a U.S. Senate seat on the vote share obtained in the following election for that same seat. This empirical illustration is analogous to the one presented by McCrary (2008) for U.S. House elections. The variable of interest is `margin`, which ranges from -100 to 100 and records the Democratic party's margin of victory in the statewide election for a given U.S. Senate seat, defined as the vote share of the Democratic party minus the vote share of its strongest opponent. This corresponds to the index, score, or running variable. By construction, the cutoff is $\overline{x} = 0$.

First, we load the dataset and present summary statistics:

```
. use rddensity_senate
. summarize margin
    Variable |        Obs        Mean    Std. Dev.        Min         Max
-------------+--------------------------------------------------------------
      margin |      1,390    7.171159    34.32488       -100         100
```

The dataset has a total of 1,390 observations, with an average Democratic party margin of victory of about 7 percentage points.

We now conduct a manipulation test using the command `rddensity` with its default options.

```
. rddensity margin
Computing data-driven bandwidth selectors.

RD Manipulation Test using local polynomial density estimation.

        Cutoff c = 0 |Left of c   Right of c          Number of obs =         1390
---------------------+----------------------          Model         = unrestricted
        Number of obs |     640         750           BW method      =         comb
    Eff. Number of obs |     408         460           Kernel        =   triangular
        Order est. (p) |       2           2           VCE method    =    jackknife
        Order bias (q) |       3           3
           BW est. (h) |  19.841      27.119

Running variable: margin.

---------------------------------------------
            Method |          T        P>|T|
-------------------+-------------------------
            Robust |    -0.8753       0.3814
---------------------------------------------
```

The output contains a variety of useful information. First, the upper-left panel gives basic summary statistics on the data being used, separate for control ($X_i < \bar{x}$) and treatment units ($X_i \geq \bar{x}$). This panel also reports the value of the bandwidth(s) chosen. Second, the upper-right panel includes general information regarding the overall sample size and implementation choices of the manipulation test. Finally, the lower panel reports the results from implementing the manipulation test. In this first execution, the test statistic is constructed using a $q = 3$ polynomial, with different bandwidths chosen for an unrestricted model with polynomial order $p = 2$. Specifically, the bandwidth choices are $(\hat{h}_{\mathtt{l},\mathtt{comb},p}, \hat{h}_{\mathtt{r},\mathtt{comb},p}) = (19.841, 27.119)$, leading to effective sample sizes of $N_- = 408$ and $N_+ = 460$ for control and treatment groups, respectively. The final manipulation test is $T_q(\hat{h}_{\mathtt{l},\mathtt{comb},p}, \hat{h}_{\mathtt{r},\mathtt{comb},p}) = -0.8753$, with a p-value of 0.3814. Therefore, in this application, there is no statistical evidence of systematic manipulation of the running variable.

`rddensity` also offers an automatic plot of the manipulation test. This plot is implemented using the package `lpdensity` for local-polynomial–based density estimation in Stata and R. If the user does not have this package installed, `rddensity` will issue an error and will request the user to install it. (In R, the package `lpdensity` is declared as a dependency, so no error ever occurs.) The following command installs the package `lpdensity` in Stata:

```
. net install lpdensity,
> from(https://sites.google.com/site/nppackages/lpdensity/stata) replace
```

To obtain the default manipulation test plot, you just need to add the `plot` option to `rddensity`, as follows:

```
. rddensity margin, plot
Computing data-driven bandwidth selectors.

RD Manipulation Test using local polynomial density estimation.
      Cutoff c = 0 |Left of c   Right of c           Number of obs =        1390
                    |                                 Model         = unrestricted
      Number of obs |      640         750            BW method     =        comb
Eff. Number of obs  |      408         460            Kernel        =   triangular
     Order est. (p) |        2           2            VCE method    =    jackknife
    Order bias (q)  |        3           3
        BW est. (h) |   19.841      27.119
Running variable: margin.

             Method |        T         P>|T|
             Robust |   -0.8753      0.3814
```

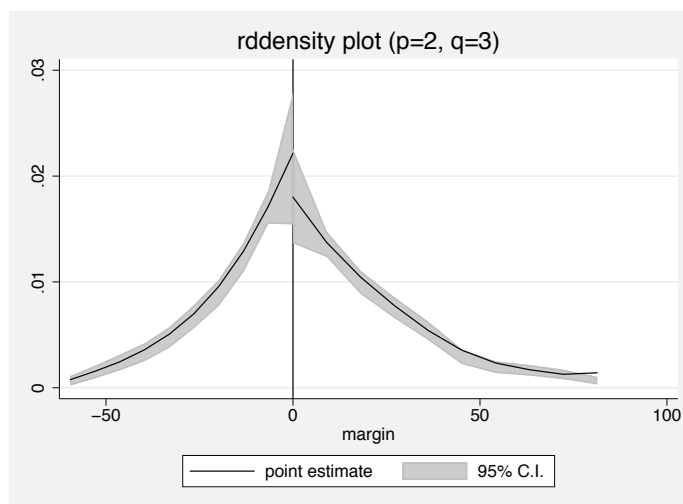The resulting default plot is given in figure 1:

Figure 1. Manipulation test plot (default options)

The basic manipulation test plot can be improved using user-specified options. For example, the following command changes the plot's legends and general appearance. The resulting plot is given in figure 2:

```
. rddensity margin, plot
> graph_options(graphregion(color(white))
> xtitle("Margin of victory") ytitle("Density") legend(off))
Computing data-driven bandwidth selectors.

RD Manipulation Test using local polynomial density estimation.
        Cutoff c = 0 │Left of c   Right of c          Number of obs =         1390
                     │                                 Model         = unrestricted
       Number of obs │     640          750            BW method     =         comb
 Eff. Number of obs  │     408          460            Kernel        =   triangular
      Order est. (p) │       2            2            VCE method    =    jackknife
      Order bias (q) │       3            3
         BW est. (h) │  19.841       27.119
Running variable: margin.

                Method │        T        P>|T|
                       │
                Robust │   -0.8753      0.3814
```


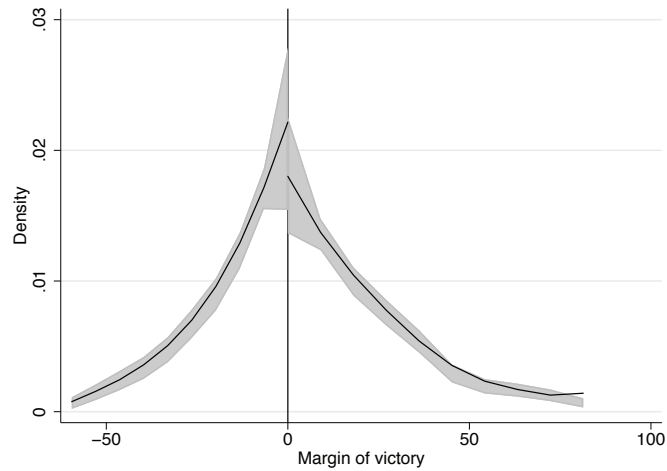
Figure 2. Manipulation test plot (with user options)

To further illustrate some capabilities of rddensity, we consider a few additional runs. We can obtain two distinct statistics, conventional and bias-corrected, by including the option all. This gives the following output:

```
. rddensity margin, all
Computing data-driven bandwidth selectors.

RD Manipulation Test using local polynomial density estimation.
```

Cutoff c = 0	Left of c	Right of c			
Number of obs	640	750			
Eff. Number of obs	408	460			
Order est. (p)	2	2			
Order bias (q)	3	3			
BW est. (h)	19.841	27.119			

```
Number of obs =          1390
Model         = unrestricted
BW method     =          comb
Kernel        =    triangular
VCE method    =     jackknife
```

Running variable: margin.

Method	T	P>|T|
Conventional	−1.6506	0.0988
Robust	−0.8753	0.3814

This second output still uses all the default options, but now it reports two test statistics. The first statistic, labeled `Conventional`, will exhibit asymptotic bias and hence will overreject the null hypothesis of no manipulation when the MSE-optimal or another large bandwidth is used. This is confirmed in the simulation study reported in section 6. The second statistic, labeled `Robust`, implements inference based on robust bias-correction and is the default and recommended option for implementing a manipulation test.

The following output showcases other features of `rddensity`. Here we conduct a manipulation test using the restricted model and plug-in SEs:

```
. rddensity margin, fitselect(restricted) vce(plugin)
Computing data-driven bandwidth selectors.

RD Manipulation Test using local polynomial density estimation.
```

Cutoff c = 0	Left of c	Right of c			
Number of obs	640	750			
Eff. Number of obs	396	362			
Order est. (p)	2	2			
Order bias (q)	3	3			
BW est. (h)	18.753	18.753			

```
Number of obs =          1390
Model         =    restricted
BW method     =          comb
Kernel        =    triangular
VCE method    =        plugin
```

Running variable: margin.

Method	T	P>|T|
Robust	−1.4768	0.1397

In this case, because the restricted model is being used, a common bandwidth for both control and treatment units is selected. This value is $\widehat{h}^{\mathrm{R}}_{\mathrm{comb},p} = 18.753$ with the choice $p = 2$ and is now using the plug-in SE estimator (instead of the jackknife method, as used before). This empirical finding shows that we continue to not reject the null hypothesis of no manipulation (p-value is 0.1397), even when the restricted model is

used, which provides further empirical evidence in favor of the validity of the RD design in this application.

To close this section, we report output from the companion command `rdbwdensity`. This command was used all along implicitly by `rddensity`, but here we use it as a standalone command to illustrate some of its features. The default output for the empirical applications is as follows:

```
. rdbwdensity margin
Bandwidth selection for manipulation testing.
Cutoff c =       0.000 | Left of c  Right of c      Number of obs =        1390
                       |                            Model        = unrestricted
         Number of obs |       640         750      Kernel       =   triangular
      Min Running var. |  -100.000       0.011      VCE method   =    jackknife
      Max Running var. |    -0.079     100.000
   Order loc. poly. (p)|         2           2
Running variable: margin.

               Target | Bandwidth    Variance     Bias^2

         left density |    19.841       0.109      0.000
        right density |    27.569       0.085      0.000
 difference densities |    27.119       0.194      0.000
         sum densities |    19.531       0.194      0.000
```

The output of `rdbwdensity` closely mimics the output from `rddensity`. Notice that the defaults are all the same [for example, $\overline{x} = 0$, $p = 2$, $K(\cdot) =$ triangular, and unrestricted model]. The main results are reported in the lower panel, where now the output includes different estimated bandwidth choices. These choices depend on the MSE criterion function (and the model considered, as discussed previously): the first row (labeled `left density`) reports $\widehat{h}_{\mathrm{l},p}$, the second row (labeled `right density`) reports $\widehat{h}_{\mathrm{r},p}$, the third row (labeled `difference densities`) reports $\widehat{h}_{\mathrm{diff},p}$, and the fourth row (labeled `sum densities`) reports $\widehat{h}_{\mathrm{sum},p}$. Of course, $\widehat{h}_{\mathrm{comb},p}$ may be easily constructed using the above information.

A similar set of results may also be obtained for the restricted model. For this case, the command is `rdbwdensity margin, fitselect(restricted)`, but we do not report these results here to conserve space.

Finally, we briefly illustrate how the two commands can be combined:

```
. quietly rdbwdensity margin
. matrix h = e(h)
. local hr = h[2,1]
. rddensity margin, h(10 `hr´)
RD Manipulation Test using local polynomial density estimation.
```

Cutoff c = 0	Left of c	Right of c			
			Number of obs	=	1390
			Model	=	unrestricted
Number of obs	640	750	BW method	=	manual
Eff. Number of obs	251	464	Kernel	=	triangular
Order est. (p)	2	2	VCE method	=	jackknife
Order bias (q)	3	3			
BW est. (h)	10.000	27.569			

Running variable: margin.

Method	T	P>|T|
Robust	-1.0331	0.3016

First, `rdbwdensity` is used to estimate the bandwidths quietly, but then the left bandwidth is set manually ($h_1 = 10$) while the right bandwidth is estimated ($h_\mathbf{r} = 27.569$) when executing `rddensity`.

The companion replication file (`rddensity_illustration.do`) includes the syntax of all the examples discussed above, as well as additional examples not included here to conserve space. These examples are

1. `rddensity margin, kernel(uniform)`
   Manipulation testing using uniform kernel.

2. `rddensity margin, bwselect(diff)`
   Manipulation testing with bandwidth selection based on MSE of difference of densities.

3. `rddensity margin, h(10 15)`
   Manipulation testing using user-chosen bandwidths $h_1 = 10$ and $h_\mathbf{r} = 15$.

4. `rddensity margin, p(2) q(4)`
   Manipulation testing using $p = 2$ and $q = 4$.

5. `rddensity margin, c(5) all`
   Manipulation testing at cutoff $\overline{x} = 5$ with all statistics.

6. `rdbwdensity margin, p(3) fitselect(restricted)`
   Bandwidth selection for restricted model with $p = 3$.

7. `rdbwdensity margin, kernel(uniform) vce(jackknife)`
   Bandwidth selection with uniform kernel and jackknife SE.

6   Simulation study

This section reports the numerical findings from a simulation study aimed to illustrate the finite sample performance of our new manipulation test.

We consider several implementations of the manipulation test $T_p(h)$, which varies according to the choice of polynomial order ($p = 2$ or $p = 3$) and choice of bandwidth. We analyze the performance of the inference procedure using a grid of bandwidths around the MSE-optimal choice, as well as a data-driven implementation of this optimal bandwidth choice, allowing for both equal and difference bandwidth choices on either side of the threshold. We also investigate the performance of robust bias-correction because we report both the naïve testing procedure without bias correction and its robust bias-corrected version. For implementation, in all cases, we also consider both asymptotic plug-in or jackknife variance estimation.

We consider the data-generating process given by $X_i \sim \sqrt{3/5}\mathcal{T}(5)$, where $\mathcal{T}(k)$ denotes a Student's t distribution with k degrees of freedom and a cutoff point $\overline{x} = 0.5$, which induces an asymmetric distribution. We use a sample size $n = 1000$, and all simulations were based on 2,000 replications with a triangular kernel to implement our density estimators.

6.1   Point estimation and empirical size

In this section, we report simulation results concerning the point-estimation properties of the underlying density estimator at the boundary point $\overline{x}$ as well as the average rejection rate (empirical size) of our proposed manipulation test. All the numerical results are given in table 1.

Table 1. Density estimation and manipulation test (empirical size)

	Bandwidth		Density Estimators						Plug-in SE		Jackknife SE	
	left	right	bias$_-$	bias$_+$	bias	sd	bias/sd	mse	mean	size	mean	size
Grid h_-, h_+												
0.5×	0.538	0.538	0.014	−0.001	−0.015	0.092	0.163	0.816	0.091	0.050	0.091	0.050
0.6×	0.645	0.645	0.021	−0.003	−0.024	0.084	0.292	0.721	0.083	0.056	0.083	0.059
0.7×	0.753	0.753	0.031	−0.005	−0.036	0.078	0.460	0.699	0.078	0.077	0.077	0.078
0.8×	0.861	0.861	0.042	−0.007	−0.049	0.073	0.666	0.735	0.073	0.107	0.072	0.112
0.9×	0.968	0.968	0.054	−0.010	−0.064	0.069	0.921	0.832	0.069	0.143	0.068	0.152
$h_{\text{MSE},p}$	1.076	1.076	0.067	−0.013	−0.080	0.065	1.223	1.000	0.066	0.208	0.064	0.232
1.1×	1.183	1.183	0.081	−0.016	−0.097	0.062	1.567	1.248	0.064	0.312	0.061	0.345
1.2×	1.291	1.291	0.095	−0.019	−0.114	0.059	1.939	1.565	0.061	0.468	0.059	0.497
1.3×	1.399	1.399	0.109	−0.023	−0.132	0.056	2.333	1.940	0.059	0.609	0.056	0.646
1.4×	1.506	1.506	0.121	−0.027	−0.148	0.054	2.737	2.363	0.057	0.747	0.054	0.785
1.5×	1.614	1.614	0.133	−0.031	−0.165	0.052	3.144	2.820	0.056	0.856	0.053	0.876
Est. $\widehat{h}_-$, $\widehat{h}_+$												
$T_p(\widehat{h}_p)$	0.618	0.710	0.023	−0.004	−0.027	0.089	0.303	0.827	0.083	0.084	0.083	0.091
$T_p(\widehat{h}_{p-1})$	0.254	0.218	0.014	−0.001	−0.015	0.142	0.103	1.928	0.141	0.054	0.143	0.050
$T_{p+1}(\widehat{h}_p)$	0.618	0.710	0.007	0.005	−0.002	0.130	0.017	1.593	0.130	0.052	0.131	0.046
Est. $\widehat{h}_- = \widehat{h}_+$												
$T_p(\widehat{h}_p)$	0.596	0.596	0.021	−0.002	−0.023	0.094	0.248	0.880	0.088	0.080	0.087	0.085
$T_p(\widehat{h}_{p-1})$	0.187	0.187	0.009	0.000	−0.008	0.159	0.053	2.395	0.157	0.059	0.160	0.052
$T_{p+1}(\widehat{h}_p)$	0.596	0.596	0.005	0.005	−0.001	0.137	0.005	1.776	0.136	0.056	0.137	0.052

(a) $p = 2$

	Bandwidth		Density Estimators						Plug-in SE		Jackknife SE	
	left	right	bias$_-$	bias$_+$	bias	sd	bias/sd	mse	mean	size	mean	size
Grid h_-, h_+												
0.5×	0.604	0.604	0.003	0.003	0.000	0.133	0.001	1.948	0.134	0.048	0.135	0.046
0.6×	0.725	0.725	0.001	0.004	0.004	0.121	0.030	1.607	0.122	0.045	0.124	0.047
0.7×	0.845	0.845	−0.004	0.005	0.009	0.112	0.082	1.378	0.113	0.052	0.114	0.048
0.8×	0.966	0.966	−0.007	0.007	0.013	0.105	0.128	1.219	0.106	0.048	0.107	0.044
0.9×	1.087	1.087	−0.007	0.008	0.015	0.099	0.153	1.103	0.100	0.050	0.101	0.046
$h_{\text{MSE},p}$	1.208	1.208	−0.006	0.009	0.015	0.094	0.156	1.000	0.095	0.052	0.095	0.049
1.1×	1.328	1.328	−0.003	0.009	0.012	0.090	0.139	0.903	0.090	0.056	0.091	0.054
1.2×	1.449	1.449	0.002	0.010	0.008	0.086	0.094	0.816	0.087	0.052	0.087	0.048
1.3×	1.570	1.570	0.008	0.010	0.002	0.083	0.020	0.746	0.084	0.048	0.083	0.053
1.4×	1.691	1.691	0.016	0.010	−0.007	0.080	0.082	0.697	0.081	0.050	0.080	0.050
1.5×	1.812	1.812	0.026	0.009	−0.016	0.077	0.211	0.676	0.079	0.048	0.077	0.053
Est. $\widehat{h}_-$, $\widehat{h}_+$												
$T_p(\widehat{h}_p)$	1.190	1.237	−0.006	0.008	0.014	0.098	0.140	1.077	0.095	0.051	0.097	0.050
$T_p(\widehat{h}_{p-1})$	0.610	0.705	0.010	0.005	−0.005	0.129	0.040	1.830	0.131	0.048	0.132	0.046
$T_{p+1}(\widehat{h}_p)$	1.190	1.237	0.007	0.006	−0.001	0.129	0.007	1.818	0.134	0.044	0.135	0.040
Est. $\widehat{h}_- = \widehat{h}_+$												
$T_p(\widehat{h}_p)$	1.119	1.119	−0.006	0.008	0.014	0.100	0.136	1.124	0.099	0.048	0.100	0.051
$T_p(\widehat{h}_{p-1})$	0.590	0.590	0.009	0.001	−0.007	0.136	0.055	2.043	0.137	0.050	0.138	0.052
$T_{p+1}(\widehat{h}_p)$	1.119	1.119	0.008	0.006	−0.002	0.134	0.016	1.960	0.140	0.043	0.140	0.038

(b) $p = 3$

We first describe the format of table 1. Starting with the columns, each table reports the following:

 i) the bandwidths used to construct the density estimators on the left and on the right of the cutoff $\overline{x}$ (columns under the label "Bandwidth");

 ii) the average bias of the two density estimators and the difference thereof, simulation variability, standardized bias of the difference of density estimators, and simulation MSE of the difference of density estimators at the cutoff $\overline{x}$ (columns under the label "Density estimators");

 iii) the average of the asymptotic plug-in SE estimator and the empirical size of the associated feasible manipulation test (columns under the label "Plug-in SE"); and

 iv) the average of the jackknife SE estimator and the empirical size of the associated feasible manipulation test (columns under the label "Jackknife SE").

Continuing with the rows, and to better understand the role of the bandwidth choice h_n on the finite sample performance of the density and SE estimators and of the manipulation test, table 1 reports:

 i) a grid of fixed bandwidths constructed around the theoretical infeasible MSE-optimal bandwidth (rows under the label "Grid");

 ii) estimated bandwidths, which are allowed to differ on the two sides of the cutoff, obtained as $\widehat{h}_{\texttt{l,comb},p}$ and $\widehat{h}_{\texttt{r,comb},p}$ (rows under the label "Est. $\widehat{h}_-, \widehat{h}_+$"); and

 iii) estimated bandwidths, which are required to be the same on the two sides of the cutoff, obtained as the smaller of $\widehat{h}_{\texttt{diff},p}$ and $\widehat{h}_{\texttt{sum},p}$ (rows under the label "Est. $\widehat{h}_- = \widehat{h}_+$").

The simulation results allow us to explore the finite sample performance of the different ingredients entering the manipulation test (density, SE, and bandwidth estimators) and of the test itself (and by implication, the quality of the Gaussian distributional approximation). Given the large amount of information, we offer only a summary of the main results we observe from the Monte Carlo experiment.

We first look at nonrandom bandwidths, which help us separate the finite sample performance of the main theoretical results in the article from the impact of bandwidth

estimation. From the grid of bandwidths around the MSE-optimal $h_{\mathrm{MSE,p}}$, we find that i) the simulation variability of the difference of density estimators (the column labeled "sd") is approximated very well by the jackknife SE estimator and reasonably well by the asymptotic plug-in SE estimator (compare with the columns labeled "mean"), and ii) the manipulation test exhibits some empirical size distortion when using the MSE-optimal bandwidth $h_{\mathrm{MSE,p}}$, as expected, but exhibits excellent empirical size when undersmoothing this bandwidth choice. These numerical findings indicate that our main theoretical results concerning bias, variance, and distributional approximations, as well as the consistency of the proposed SE formulas, are borne out in the Monte Carlo experiment.

We now explore the impact of bandwidth selection on estimation and inference in the context of the manipulation test. We focus on the last six rows of table 1. When looking at the different bandwidth estimators, we find that i) our bandwidth estimator $\widehat{h}_p$ tends to deliver smaller values than $h_{\mathrm{MSE,p}}$ on average, a finding that actually helps control the empirical size of the manipulation test, and ii) the robust bias-correction approach [that is, inference based on $T_p(\widehat{h}_{p-1})$] delivers manipulation tests with very good empirical size properties. Because the robust bias-correction approach is theoretically justified and valid for all sample sizes, we recommend it as the best alternative for applications.

To summarize, based on the simulation evidence we obtained, we recommend for empirical work the manipulation test based on the feasible statistic $T_p(\widehat{h}_{p-1})$ constructed using a pth-order local-polynomial density estimator, an MSE-optimal bandwidth choice coupled with robust bias-correction, and the corresponding jackknife SE estimator. This testing procedure exhibited close-to-correct empirical size across all designs we considered and performed as well as (if not better than) all the alternatives we explored. These numerical findings agree with our main theoretical results.

6.2 Empirical power

To complement the numerical results presented above, we also investigate the power of the manipulation test constructed using the local-polynomial density estimator. We continue to use the same data-generating process, but now we scale the data so that the true density satisfies

$$f_+/f_- \in \{0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9,\ 1.0,\ 1.1,\ 1.2,\ 1.3,\ 1.4,\ 1.5\}$$

at the cutoff $\overline{x} = 0.5$. Recall that the sample size is $n = 1000$ and that all simulations are based on 2,000 replications.

The results are given in table 2.

Table 2. Manipulation test (empirical power)

	Bandwidth		Rejection Rate					$\mathsf{H}_0\colon f_+/f_- = 1$ **vs.** H_a: **value below**					
	left	right	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
Plug-in SE													
$T_p(h_{\mathrm{MSE},p})$	1.076	1.076	1.000	0.990	0.922	0.734	0.454	0.212	0.080	0.036	0.060	0.126	0.233
Est. $\widehat{h}_-,\ \widehat{h}_+$													
$T_p(\widehat{h}_p)$	0.615	0.712	0.888	0.715	0.498	0.284	0.154	0.072	0.064	0.078	0.134	0.218	0.330
$T_p(\widehat{h}_{p-1})$	0.254	0.215	0.478	0.316	0.202	0.122	0.074	0.056	0.049	0.060	0.072	0.094	0.130
$T_{p+1}(\widehat{h}_p)$	0.615	0.712	0.474	0.313	0.191	0.112	0.074	0.053	0.050	0.065	0.100	0.143	0.187
Est. $\widehat{h}_- = \widehat{h}_+$													
$T_p(\widehat{h}_p)$	0.594	0.594	0.865	0.686	0.456	0.266	0.138	0.070	0.060	0.074	0.122	0.190	0.285
$T_p(\widehat{h}_{p-1})$	0.185	0.185	0.353	0.239	0.151	0.099	0.066	0.050	0.047	0.060	0.079	0.100	0.131
$T_{p+1}(\widehat{h}_p)$	0.594	0.594	0.458	0.301	0.188	0.114	0.072	0.057	0.055	0.068	0.088	0.117	0.159
Jackknife SE													
$T_p(h_{\mathrm{MSE},p})$	1.076	1.076	1.000	0.990	0.926	0.747	0.480	0.232	0.090	0.045	0.071	0.148	0.265
Est. $\widehat{h}_-,\ \widehat{h}_+$													
$T_p(\widehat{h}_p)$	0.615	0.712	0.866	0.700	0.490	0.288	0.155	0.076	0.067	0.082	0.136	0.231	0.352
$T_p(\widehat{h}_{p-1})$	0.254	0.215	0.459	0.309	0.196	0.114	0.063	0.047	0.044	0.055	0.070	0.098	0.139
$T_{p+1}(\widehat{h}_p)$	0.615	0.712	0.450	0.305	0.196	0.118	0.068	0.051	0.047	0.062	0.104	0.143	0.200
Est. $\widehat{h}_- = \widehat{h}_+$													
$T_p(\widehat{h}_p)$	0.594	0.594	0.844	0.670	0.450	0.266	0.142	0.073	0.063	0.076	0.123	0.202	0.310
$T_p(\widehat{h}_{p-1})$	0.185	0.185	0.345	0.230	0.148	0.097	0.057	0.043	0.044	0.054	0.072	0.100	0.132
$T_{p+1}(\widehat{h}_p)$	0.594	0.594	0.434	0.292	0.188	0.114	0.074	0.052	0.050	0.066	0.088	0.126	0.170

(a) $p = 2$

	Bandwidth		Rejection Rate					$\mathsf{H}_0\colon f_+/f_- = 1$ **vs.** H_a: **value below**					
	left	right	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
Plug-in SE													
$T_p(h_{\mathrm{MSE},p})$	1.208	1.208	0.701	0.456	0.236	0.110	0.056	0.057	0.086	0.143	0.218	0.311	0.412
Est. $\widehat{h}_-,\ \widehat{h}_+$													
$T_p(\widehat{h}_p)$	1.205	1.259	0.680	0.455	0.264	0.138	0.076	0.057	0.090	0.141	0.210	0.306	0.409
$T_p(\widehat{h}_{p-1})$	0.617	0.712	0.472	0.291	0.168	0.098	0.064	0.050	0.055	0.078	0.108	0.152	0.190
$T_{p+1}(\widehat{h}_p)$	1.205	1.259	0.478	0.280	0.149	0.078	0.051	0.050	0.060	0.076	0.106	0.137	0.179
Est. $\widehat{h}_- = \widehat{h}_+$													
$T_p(\widehat{h}_p)$	1.134	1.134	0.658	0.442	0.240	0.124	0.066	0.052	0.084	0.122	0.191	0.272	0.368
$T_p(\widehat{h}_{p-1})$	0.594	0.594	0.447	0.284	0.166	0.102	0.068	0.051	0.056	0.070	0.092	0.122	0.162
$T_{p+1}(\widehat{h}_p)$	1.134	1.134	0.450	0.263	0.148	0.075	0.052	0.052	0.056	0.070	0.098	0.129	0.162
Jackknife SE													
$T_p(h_{\mathrm{MSE},p})$	1.208	1.208	0.646	0.424	0.228	0.107	0.057	0.054	0.083	0.150	0.236	0.336	0.446
Est. $\widehat{h}_-,\ \widehat{h}_+$													
$T_p(\widehat{h}_p)$	1.205	1.259	0.624	0.430	0.256	0.134	0.074	0.053	0.086	0.142	0.232	0.336	0.448
$T_p(\widehat{h}_{p-1})$	0.617	0.712	0.444	0.286	0.174	0.102	0.063	0.041	0.054	0.072	0.112	0.152	0.208
$T_{p+1}(\widehat{h}_p)$	1.205	1.259	0.444	0.274	0.152	0.080	0.047	0.040	0.055	0.080	0.110	0.154	0.204
Est. $\widehat{h}_- = \widehat{h}_+$													
$T_p(\widehat{h}_p)$	1.134	1.134	0.600	0.415	0.236	0.123	0.064	0.049	0.078	0.128	0.210	0.300	0.403
$T_p(\widehat{h}_{p-1})$	0.594	0.594	0.434	0.275	0.170	0.100	0.064	0.045	0.048	0.068	0.092	0.130	0.177
$T_{p+1}(\widehat{h}_p)$	1.134	1.134	0.420	0.267	0.148	0.078	0.047	0.042	0.053	0.072	0.101	0.140	0.184

(b) $p = 3$

Notes: i) Columns "Bandwidth": bandwidths for left and right density estimators under the null hypothesis $f_+/f_- = 1$. For other cases, the bandwidth for estimating f_+ is adjusted proportional to f_+/f_-. ii) $\widehat{h}_p$: estimated MSE-optimal bandwidth for p-order density estimator.

We first describe the format of this table. The two main columns report the following:

i) the bandwidths used to construct the density estimators on the left and on the right of the cutoff $\overline{x}$ (columns under the label "Bandwidth"); and

ii) the density discontinuity and the corresponding rejection rates (columns under the label "Rejection rate"). Note that the column $f_+/f_- = 1$ corresponds to the null hypothesis being true, and the rejection rate under that column is just the empirical size of the test.

Continuing with the rows of table 2, we examine the empirical power with the infeasible MSE-optimal bandwidth as well as the estimated bandwidths, with either the plug-in or the jackknife SE used.

Based on the simulation evidence obtained, we find that using the infeasible MSE-optimal bandwidth will lead to size distortion and the power curve does not attain minimum when the null hypothesis is true (for example, in the $p = 2$ case in table 2, the minimum rejection rate occurs when $f_+/f_- = 1.2$), which again confirms the finding that without bias correction, the manipulation test will not only be inconsistent but also lose power.

For the different bandwidth estimators, we again find that the robust bias-correction approach delivers manipulation tests with very good empirical size properties, because with either method, the power curve achieves minimum when the null hypothesis is true. Also as expected, bias correction will lead to some power loss compared with the $T_p(\widehat{h}_p)$ case.

7 Conclusion

In this article, we discussed the implementation of nonparametric manipulation tests using local-polynomial density estimators, which are useful for falsification of RD designs and for empirical research analyzing whether units are self-selected into a particular group or treatment status. We introduced two commands: `rddensity` and `rdbwdensity`. These commands use ideas from Cattaneo, Jansson, and Ma (2017b). In particular, the first command implements several nonparametric manipulation tests, while the second command provides an array of bandwidth selection methods. Companion R functions are also available from the authors. The latest version of this and related software for RD designs can be found at https://sites.google.com/site/rdpackages/.

8   Acknowledgments

9   References

Calonico, S., M. D. Cattaneo, and M. H. Farrell. Forthcoming. On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik. 2017. rdrobust: Software for regression-discontinuity designs. *Stata Journal* 17: 372–404.

Calonico, S., M. D. Cattaneo, and R. Titiunik. 2014a. Robust data-driven inference in the regression-discontinuity design. *Stata Journal* 14: 909–946.

———. 2014b. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 82: 2295–2326.

———. 2015a. Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association* 110: 1753–1769.

———. 2015b. rdrobust: An R package for robust nonparametric inference in regression-discontinuity designs. *R Journal* 7: 38–51.

Cattaneo, M. D., and J. C. Escanciano, eds. 2017. *Advances in Econometrics: Vol. 38—Regression Discontinuity Designs: Theory and Applications*. Bingley, UK: Emerald.

Cattaneo, M. D., B. R. Frandsen, and R. Titiunik. 2015. Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. Senate. *Journal of Causal Inference* 3: 1–24.

Cattaneo, M. D., M. Jansson, and X. Ma. 2017a. lpdensity: Local polynomial density estimation and inference. Working paper, University of Michigan.

———. 2017b. Simple local polynomial density estimators. Working paper, University of Michigan.

Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare. 2016. Inference in regression discontinuity designs under local randomization. *Stata Journal* 16: 331–367.

———. 2017c. Comparing inference approaches for RD designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management* 36: 643–681.

Cheng, M.-Y., F. Jianqing, and J. S. Marron. 1997. On automatic boundary corrections. *Annals of Statistics* 25: 1691–1708.

Fan, J., and I. Gijbels. 1996. *Local Polynomial Modelling and Its Applications*. New York: Chapman & Hall/CRC.

Frandsen, B. 2017. Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete. In *Advances in Econometrics: Vol. 38—Regression Discontinuity Designs: Theory and Applications*, ed. M. D. Cattaneo and J. C. Escanciano, 29–72. Bingley, UK: Emerald.

McCrary, J. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142: 698–714.

Otsu, T., K.-L. Xu, and Y. Matsushita. 2014. Estimation and inference of discontinuity in density. *Journal of Business and Economic Statistics* 31: 507–524.

About the authors

Matias D. Cattaneo is a professor of economics and a professor of statistics at the University of Michigan.

Michael Jansson is the Edward G. and Nancy S. Jordan Family Professor of Economics at the University of California, Berkeley.

Xinwei Ma is a PhD candidate in economics at the University of Michigan.