Regression Discontinuity Design:
A Test for Manipulation of the Running Variable

Research Module in Econometrics and Statistics
Instructor: JProf. Dr. Dominik Liebl
University of Bonn
WS 2019/20

Sofia Badini, Max Schäfer, Caroline Krayer

February 5, 2020

Contents

1 Introduction

Cause and effect have long been studied by researchers from various backgrounds and more recently, experimental and quasi-experimental methods in applied work gained ground in the social sciences, and in economics, especially. To distill causes and effects, questions of the following type are usually asked: *What would have been if a certain state had not occurred?* Unfortunately, answering this question for an individual unit is impossible since the same individual can only be in one state of the world at a particular point in time. This observation is known as the fundamental problem of causal inference. Experimental and quasi-experimental methods have been developed to approach answers to what-if-questions. Their aim is to estimate an effect by comparing outcomes of individual units that exhibit the same set of characteristics and differ only with respect to exposure to treatment.

First introduced by Thistlethwaite and Campbell [18], the Regression Discontinuity Design (RDD) is in specific contexts suited to create two groups of individuals with opposite treatment status and, at least locally, similar characteristics by exploiting a cutoff in a covariate called the running variable. We owe this quasi-random assignment of individuals into the treatment and control group to the defining feature of RDD: The probability of receiving treatment follows an assignment rule where treatment assignment changes discontinuously at the cutoff as a function of the running variable (cf. [8]). Given the running variable being continuously distributed, we can assume that individuals marginally below and above the cutoff are very similar, as long as ending up right above vis-à-vis below the cutoff involves chance. Therefore, comparing outcomes of the treated and untreated near the cutoff informs us about the effect a treatment has.

Treatment effects obtained from appropriate RDDs enjoy high internal validity since individuals are as good as randomly divided into treatment and control groups. However, validity of this causal effect is jeopardized if randomization is not ensured and hence the individuals on either side of the cutoff do not have on average equal characteristics. Eminently, if (some) individuals are able to select into treatment by manipulating their running variable, the treatment effect cannot be recovered. Hence, the researcher must be well-aware of the RDD context, and specifically consider the possibility of manipulation.

Thus, while testing for manipulation can never validate a RDD, it is key to do so in any study exploiting such a treatment assignment rule (cf. [11]), and the literature proposes five distinct tests. Lee [10] uses that in absence of manipulation individuals around the cutoff have similar characteristics that are not influenced by treatment. He tests whether the density functions of observed pre-treatment variables conditional on the running variable are continuous at the cutoff. However, this approach requires data on pre-treatment variables which may not be available.

One of the most popular manipulation tests, proposed by McCrary [13], tests for a discontinuity gap in the running variable's density function at the cutoff with means of a Wald test. The estimate is constructed in two steps by first binning the data and then

smoothing the obtained histogram with local linear regression. An advantage is that it can always be performed as data on the running variable are available and the estimate inherits useful properties of local linear regression. Drawbacks are that we have to choose two smoothing parameters, a binsize in the first step and a bandwidth in the second. Lastly, it is not always suitable in case of a discrete running variable.

In the same spirit, Otsu, Xu and Matsushita [14] test for a discontinuity in the running variable's density but they rely on an empirical likelihood-based test. They construct a localized version of likelihood functions for the density function estimates using boundary-corrected kernels and afterwards apply likelihood maximization. The test itself deploys the fact that under certain assumptions the empirical likelihood function converges in distribution to a chi-squared random variable with one degree of freedom. Thus, no asymptotic variance of the estimator as in McCrary's approach must be estimated.

Also, Cattaneo, Jansson and Ma [2] propose a manipulation test resting on the Wald test. Instead of binning the data like McCrary, the empirical distribution function approaching the cutoff from above and below, respectively, is used as a starting point for local polynomial regression in the second step. Beneficially, this procedure requires a single smoothing parameter (the bandwidth) and all useful properties of local polynomial regression remain valid.

Finally, Frandsen [7] proposes a test which improves over McCrary's for the case of a discrete running variable. Instead of using local linear regression, which relies on some extrapolation, the test only uses support points of the running variable at and immediately adjacent to the cutoff. In case of a sufficiently smooth density function at the cutoff, the probability of being exactly at the cutoff conditional on the running variable taking on the cutoff value – or an adjacent one – lies in a certain interval.

In this paper we focus on the manipulation test proposed by McCrary [13], as it is very popular and lays the foundation for many of the other tests. In Section 2 we introduce a more formal context of Regression Discontinuity Design as well as the problem of manipulation. Section 3 formally derives the manipulation test, discusses its properties and the issue of parameter choice. Subsequently, in a simulation study in Section 4 we asses the validity of the estimator's asymptotic properties in finite samples and the test's performance in different settings. Finally, in Section 5 we challenge the test in an application to real data, building on a paper by Camacho and Conover [1], who study manipulation of social program eligibility in Colombia. Concluding remarks follow.

In all, we shared the work equally and nobody is solely responsible for one section.

2 Regression Discontinuity Design

2.1 Theoretical Framework and Identification

The goal of causal inference is to uncover the effect of a treatment $D_i \in \{0, 1\}$, which we assume to be binary, on an outcome of interest Y_i for each individual i in a sample of size n. Following the potential-outcome framework formalized by Rubin [16], we define the potential outcomes of each individual as $Y_i(1)$ if the individual receives treatment, and $Y_i(0)$ if she does not. The fundamental problem of causal inference implies that we cannot observe both at the same time, but only Y_i defined as $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$.

In Regression Discontinuity Design (cf. [9] and [11]), the assignment to treatment of each individual is determined by the value of a pre-treatment variable R_i, the running variable, being on either side of a fixed cutoff c. Throughout our paper, we stick to the sharp RDD where participation is mandatory and the running variable perfectly predicts treatment assignment. Thus, treatment status is a deterministic function of the running variable and treatment is granted to those whose value of the running variable passes the cutoff: $D_i = \mathbb{1}_{R_i \geq c}$ or $D_i = \mathbb{1}_{R_i \leq c}$, depending on the treatment-assignment rule.

Moreover, we assume individual outcome is a linear function of treatment:

$$Y_i = \alpha_i + \beta_i D_i + \gamma_i X_i = \bar{\alpha} + \bar{\beta} D_i + \bar{\gamma} X_i + \varepsilon_i \tag{2.1}$$

where α_i, β_i and γ_i are random variables with means $\bar{\alpha}$, $\bar{\beta}$ and $\bar{\gamma}$, respectively. X_i is a vector of pre-treatment variables which determine the characteristics of the individual, including R_i. Such variables can be observed or unobserved. Finally, ε_i is assumed to be a purely random error.

Our quantity of interest is the individual-level treatment effect, $Y_i(1) - Y_i(0)$, which cannot be recovered as stated above. However, we can rely on a comparison between groups of individuals having similar characteristics but opposite treatment status in order to estimate the corresponding population-level treatment effect. So our analysis must rely on the population-level variables $Y(1)$, $Y(0)$, Y, R and D.

The fundamental intuition behind RDD is that, in absence of treatment, individuals with running variable values close to the cutoff are similar, and therefore individuals marginally below the cutoff identify the true counterfactual of those marginally above. With true counterfactual we mean the potential outcome that would have resulted had the individuals been assigned to the other treatment status.

The comparability of individuals just below and above the cutoff is a consequence of the RDD given certain weak assumptions. If they are satisfied, the discontinuity in the conditional expectation of the outcome given the running variable at c, $\lim_{r \downarrow c} \mathbb{E}[Y|R = r] - \lim_{r \uparrow c} \mathbb{E}[Y|R = r]$, identifies the Average Treatment Effect (ATE). The ATE is in general defined as $E[Y(1) - Y(0)]$, but in the RDD context we have a weighted version where the weight assigned to each individual is determined by her likelihood to obtain a draw of the running variable close to c. An assumption needed for our inference to be valid is

continuity of the conditional regression functions:

$$\mathbb{E}\left[Y_i(0)|R_i = r\right], \mathbb{E}\left[Y_i(1)|R_i = r\right] \text{ and } f_{R_i}(r) \text{ are continuous in } r \tag{2.2}$$

As noted by Imbens and Lemieux [9], this condition is actually stronger than required, since continuity is only needed at the cutoff value c. More problematically, this condition is untestable and its plausibility hard to evaluate in many real world settings. For example, when individuals can manipulate their running variable – our main case of interest in this paper –, continuity is possibly invalidated.

2.2 Manipulation of the Running Variable

If the treatment assignment rule is public knowledge, individuals benefit from a specific treatment status and have the ability to adjust their behavior accordingly, they could try to influence their running variable's value to move just below or just above the cutoff. As an example, consider a school voucher program for kids from poor households that pays 100 US-Dollars to households above some cutoff of a poverty index score. The higher the poverty index score, the poorer the household is. If a household barely below the cutoff finds out about the program, it could try to influence its poverty index score to sort into treatment; an effort it would not have made if the program did not exist.

At this point it is useful to make a distinction between R_i, the observed value of the running variable, and R_{i0}, the unobserved value the individual would obtain were there no treatment program. Manipulation of the running variable occurs if $R_i \neq R_{i0}$ for some individual i. It can lead $\mathbb{E}\left[Y_i(0)|R_i\right]$ and $\mathbb{E}\left[Y_i(1)|R_i\right]$ to be discontinuous at the cutoff, despite continuity of $\mathbb{E}\left[Y_i(0)|R_{i0}\right]$ and $\mathbb{E}\left[Y_i(1)|R_{i0}\right]$, as individuals try to achieve a certain value of the running variable precisely because the treatment is in place.

If individuals benefit from being assigned to treatment and manipulation is possible, they may self-select into treatment based on anticipated gains by achieving a value of R_i that makes the probability of receiving treatment jump to one. Therefore, it becomes likely to observe more people slightly above than slightly below the cutoff, or vice versa. A problem for causal inference may arise when those slightly above the cutoff no longer constitute the true counterfactual for those slightly below. However, we can make a distinction between partial and perfect manipulation.

Under perfect manipulation, the running variable is completely under the agent's control. In this case, identification of the treatment effect is not possible at all because individuals can willingly decide what their treatment status will be. Therefore, even after controlling for the observed pre-treatment variables, we should expect significant differences between the unobserved pre-treatment variables of individuals who self-select into treatment and those who do not. Comparability of individuals at the cutoff does not hold anymore.

Under partial manipulation, although the agent can influence the running variable in some sense, there is still an idiosyncratic element that cannot be influenced. Lee [10]

shows that in this case a valid estimation is still possible. In the case of non-random self-selection into treatment where the running variable has some random chance component, the inability to precisely control the running variable leads to random variation around the cutoff. An important condition for such local randomization to occur is the continuity of the running variable's density function, conditional on the individuals' unobserved pre-treatment variables W, which means $f_{R|W}(r|w)$ is continuous in r. Lee proves that, together with some other mild conditions, this implies that the probability of the value of R_i being just above or just below the cutoff is equal, and that the discontinuity in the conditional expectation of the outcome given the running variable at the cutoff still identifies the treatment effect. Therefore, partial manipulation does not hinder the validity of a RDD, while perfect manipulation does and must be tested for.

For his manipulation test, McCrary exploits the fact that continuity of the running variable's conditional density function given the unobserved characteristics of the individuals W, $f_{R|W}(r|w)$, implies continuity of the running variable's density function, $f_R(r)$. He then tests for the null-hypothesis of continuity of $f_R(r)$ at $r = c$ by means of a Wald test. The intuition is that, if perfect manipulation is present, we expect surprisingly many people barely qualifying for treatment and surprisingly few people barely not qualifying. Therefore, we would observe a discontinuity in the density of the running variable at the cutoff.

An important assumption the test requires is monotonic manipulation of the running variable, so either $R_i \leq R_{i,0}$ for each individual i or $R_i \geq R_{i,0}$ for each individual i. All individuals should be willing to manipulate their running variable in one direction only, so there should be a treatment status preferred by everyone. If some individuals prefer to be treated and some prefer not to when perfect manipulation is possible, the discontinuity at the cutoff disappears when the proportion of the two groups is roughly balanced. McCrary's test does not reject, which will be taken as a hint of the validity of RDD. However, self-selection makes causal inference more cumbersome. If treatment status is correlated with treatment outcome, the weighted ATE estimation is biased.

Similarly, the test results may be misleading if perfect manipulation moves significantly more individuals on a side of the cutoff, but at random. Recall the school voucher program example discussed before, and say, manipulation by households is impossible but the government wants to cut expenditures and therefore assigns 20 percent of eligible households a lower poverty index score at random. Then it is still possible to estimate the treatment effect of the school voucher program.

A key takeaway is that McCrary's test cannot process the intention of agents. It will simply detect whether a discontinuity is present at the cutoff, not the reason why it is or is not. Therefore, this test is neither a necessary nor a sufficient condition for the validity of RDD in general and should complement further investigation of manipulation.

3 The Manipulation Test

3.1 Kernel Density Estimation and Boundary Effects

To test for manipulation, we want to estimate the discontinuity of the running variable's density function at the cutoff c. Using all data the estimate depends starkly on the imposed functional form as the density estimate also relies on observations far away from c. To relax functional form assumptions and to focus on data closer to c, non-parametric methods are commonly used.

A standard method to estimate a density function at a particular point of the domain is kernel density estimation (KDE), where the estimate is a weighted average of adjacent data. The kernel is a function that assigns more weight to observations in the proximity of the point of interest. Which data are considered is defined by the bandwidth – a fixed or variable parameter stating the range of values around the point of interest the kernel uses. Applying this procedure for all values delivers an estimate of the density function. However, discontinuities are not fully acknowledged as they are being smoothed over by traditional KDE. To avoid this feature in general, the density function is estimated separately on both sides of the cutoff.

In RDD, when the density function is estimated separately for data on either side of the cutoff, c constitutes a boundary point. As is known (cf. [12]), traditional KDE produces severely biased estimates at these boundary points as it places positive probability mass behind the boundary and therefore tends to underestimate the density at boundaries. As a result, traditional KDE suffers from a slower rate of convergence at boundary points compared to points in the interior. To circumvent boundary effects, robust techniques have been developed within the class of KDE. Schuster [17] first formalized the idea of reflection methods where probability mass that has been falsely placed behind the boundary is returned to the other side. One can also make adjustments to the kernel estimate directly to cope with the bias, and Rice [15] proposes to use a linear combination of two (or more) kernels.

A simple approach featuring automatic boundary corrections is local polynomial regression. Fan and Gijbels [6] show that local polynomial estimators can be transformed into standard KDE where the kernel depends not only on the point of interest but further on its location in the density function's support. Hence, the kernel is automatically transformed such that it assigns probability mass only to points inside the support. Consequently, the rate of convergence is not influenced by estimation at a boundary point and additional boundary modifications become unnecessary. Therefore, we focus on local polynomial regression.

3.2 Local Polynomial Regression Framework

We present the general framework of local polynomial regression which too uses kernels to weigh data according to their distance to the point of interest. For a rigorous treatment

of local polynomial regression see [6], [5] and [3] on which this section builds upon.

Assume we have n independent and identically distributed observation pairs (X_i, Y_i) randomly sampled from a population (X, Y) where the dependent variable Y is related to the regressor X through the (unknown) regression function m: $Y_i = m(X_i) + \varepsilon_i$. We are interested in estimating the conditional expectation function $m(x_0) = E[Y|X = x_0]$ and its first p derivatives non-parametrically at some point x_0, and assume $E[\varepsilon] = 0$, fixed and finite variance and orthogonality of X and ε. Local polynomial regression fits a polynomial of order p to a localized subset of the data to estimate m at some point x_0. Thereby it assigns more weight to observations whose value of X is closer to x_0 where weights are determined by a kernel function K. Let h be the bandwidth.

To derive the local polynomial estimator, suppose the function m is $(p+1)$-times continuously differentiable such that it can be locally approximated using a Taylor approximation of order p for some z in a neighborhood of x_0:

$$m(z) = \sum_{k=0}^{p} \frac{m^{(k)}(x_0)}{k!}(z - x_0)^k \equiv \sum_{k=0}^{p} \phi_k(x_0)(z - x_0)^k. \tag{3.1}$$

From here it follows that $k!\ \widehat{\phi}_k(x_0)$ is an estimator for the k-th derivative of the regression function $m^{(k)}(x_0)$ where $k \in \{0, \ldots, p\}$. In particular, the estimator for the intercept $\widehat{\phi}_0(x_0)$ is an estimator for $m(x_0)$, the regression function itself. The estimates $\widehat{\phi}_k(x_0)$ result from the following minimization problem:

$$\min_{\phi_0,\ldots,\phi_p} \sum_{i=1}^{n} \left(Y_i - \sum_{k=0}^{p} \phi_k \left(X_i - x_0\right)^k \right)^2 K\left(\frac{X_i - x_0}{h}\right). \tag{3.2}$$

In line with McCrary, we focus on local linear regression as any function can be approximated by a linear function in a neighborhood of a point of interest. We derive the solution of the minimization problem (3.2) for $p = 1$. Define

$$X = \begin{pmatrix} 1 & X_1 - x_0 \\ \vdots & \vdots \\ 1 & X_n - x_0 \end{pmatrix},\ Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},\ \widehat{\phi}(x_0) = \begin{pmatrix} \widehat{\phi}_0(x_0) \\ \widehat{\phi}_1(x_0) \end{pmatrix} \tag{3.3}$$

and write the weighing regime determined by the kernel as a weight matrix W:

$$W = diag\left(K\left(\frac{X - x_0}{h}\right)\right). \tag{3.4}$$

We then arrive at a standard weighted least squares problem and are able to deploy general results from least squares regression theory to obtain the solution:

$$\widehat{\phi}(x_0) = \left(X'WX\right)^{-1} X'WY = \begin{pmatrix} S_{n,0}(x_0) & S_{n,1}(x_0) \\ S_{n,1}(x_0) & S_{n,2}(x_0) \end{pmatrix}^{-1} \begin{pmatrix} T_{n,0}(x_0) \\ T_{n,1}(x_0) \end{pmatrix} \tag{3.5}$$

with

$$S_{n,k}(x_0) = \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)^k \text{ for } k \in \{0,1,2\}$$

$$T_{n,k}(x_0) = \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)^k\, Y_i \text{ for } k \in \{0,1\}$$

In particular, the estimator of the regression function itself at a certain point x_0 writes

$$\widehat{\phi}_0(x_0) = \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)\frac{S_{n,2}(x_0) - S_{n,1}(x_0)(X_i - x_0)}{S_{n,2}(x_0)S_{n,0}(x_0) - S_{n,1}(x_0)^2}\, Y_i. \tag{3.6}$$

3.3 Derivation of the Test Statistic

After deriving the local linear estimator, we build on this result to estimate the density of a random variable. To apply regression methods to density estimation we require a regressor and a regressand. Binning the data, counting data points within each bin and centering them produces observation pairs (bin center, bin count) – regressor and regressand – and therefore creates a suitable context for regression analysis.

Hence, following McCrary, estimation of (potential) discontinuities in the density function f proceeds in two steps. First, we bin data on the running variable in equally sized bins such that no bin contains values from both sides of c. Then, we construct a histogram with height equal to the normalized number of observations contained in each bin. The second step involves fitting the obtained histogram with local linear regression for bin center and bin height pairs both right and left of the cutoff. The difference of the estimated densities at points arbitrarily close to the cutoff from both sides depicts the estimator of the discontinuity at c.

Step 1: Data Binning and Histogram Construction

To construct the histogram we bin data of the running variable R_i. That is, we divide the support of the density function into disjoint intervals on the left and right side of the cutoff c. Then we move observations in each bin to the bin center which gives a discretized running variable. Formally, for some uniform and positive binsize b and number of bins J_0 and J_1 to the left and right of c, respectively, we construct bins $(b_j, b_{j+1}]$ with $j = 1, \ldots, J = J_0 + J_1$ and $b_j = c - (J_0 - j + 1)\, b$ (cf. [11]).

We choose values for b, J_0 and J_1 such that the support of f is covered, and fix b_1 to ensure no bin contains values from both left and right of the cutoff. The number of observations N_j in each bin results from $N_j = \sum_{i=1}^{n} \mathbb{1}_{(b_j, b_{j+1}]}(R_i)$ and normalizing them yields $\frac{1}{nb}N_j$. Finally, we construct the histogram by placing the bin centers X_j on the x-axis and the normalized number of observations $\frac{1}{nb}N_j$ centered at X_j on the y-axis. This gives a scatterplot $(X_j, \frac{1}{nb}N_j)$ which serves as an approximation of the underlying density function at bin centers, $f(X_j)$, as shown formally in Appendix A.

Step 2: Local Linear Smoothing of the Histogram

In the second step, we smooth the histogram by means of local linear regression using bin centers X_j as regressors and bin heights $\frac{1}{nb}N_j$ as the dependent variable. We draw from the local polynomial regression framework introduced in Section 3.2 and consider the special case of fitting a linear model locally. Following McCrary, we adjust the weighted least squares problem to the RDD context. We only consider observations within the bandwidth on the side of the cutoff on which the bin center of interest lies. The reason is that the density function to the left and right of the cutoff may be different. In this case, relying on observations from the other side biases the estimate of the density function. To cope with this, we introduce indicator functions in the minimization problem to estimate the value of the density function at some point r:

$$\min_{\phi_0,\phi_1} \sum_{j=1}^{J} K\left(\frac{X_j - r}{h}\right)\left(\frac{1}{nb}N_j - \phi_0 - \phi_1(X_j - r)\right)^2 \left(\mathbb{1}_{X_j>c}\mathbb{1}_{r\geq c} + \mathbb{1}_{X_j<c}\mathbb{1}_{r<c}\right). \tag{3.7}$$

Then, the estimators for the density function when approaching the cutoff c from above and below, denoted by $\widehat{f}^+$ and $\widehat{f}^-$, are

$$\begin{aligned}
\widehat{f}^+ \equiv \widehat{\phi}_0^+(c) &= \sum_{X_j>c} K\left(\frac{X_j - c}{h}\right) \frac{S_{n,2}^+(c) - S_{n,1}^+(c)(X_j - c)}{S_{n,2}^+(c)S_{n,0}^+(c) - S_{n,1}^+(c)^2}\, Y_j \\
\widehat{f}^- \equiv \widehat{\phi}_0^-(c) &= \sum_{X_j<c} K\left(\frac{X_j - c}{h}\right) \frac{S_{n,2}^-(c) - S_{n,1}^-(c)(X_j - c)}{S_{n,2}^-(c)S_{n,0}^-(c) - S_{n,1}^-(c)^2}\, Y_j
\end{aligned} \tag{3.8}$$

where

$$\begin{aligned}
S_{n,k}^+(c) &= \sum_{X_j>c} K\left(\frac{X_j - c}{h}\right)(X_j - c)^k \text{ for } k \in \{0,1,2\} \\
S_{n,k}^-(c) &= \sum_{X_j<c} K\left(\frac{X_j - c}{h}\right)(X_j - c)^k \text{ for } k \in \{0,1,2\}\ .
\end{aligned}$$

To add intuition, note that $\widehat{f}^+$ and $\widehat{f}^-$ are the intercepts from the weighted least squares problem $\widehat{\phi}_0^+(c)$ and $\widehat{\phi}_0^-(c)$, respectively. Since we want to estimate the density function on both sides of the cutoff c, the intercept informs us about the (smoothed) height of the underlying bin at c – the value of the estimated density function at the cutoff.

As we aim to estimate the discontinuity gap at c we deploy the difference in the density function estimates at points arbitrarily close to c on either side. Taking the natural logarithm first, we get

$$\widehat{\theta} \equiv \ln(\lim_{r\downarrow c} \widehat{f}(r)) - \ln(\lim_{r\uparrow c} \widehat{f}(r)) \equiv \ln(\widehat{f}^+) - \ln(\widehat{f}^-). \tag{3.9}$$

We must make sure $\widehat{f}^-$ and $\widehat{f}^+$ are strictly greater than zero for $\widehat{\theta}$ to be well-defined.

The natural logarithm is applied to both density estimates to make gap sizes comparable across locations in the support. For example, the same estimate for the gap $\widehat{f}^{+} - \widehat{f}^{-}$ needs to be differently interpreted when the cutoff is located at the tails vis-à-vis the mean of a normal distribution. Indeed, to produce the same estimate a higher share of perfect manipulators is needed at the tails compared to the mean. The logarithmic transformation ensures that comparable shares of manipulators result in comparable gap estimates. It further allows the interpretation of small differences as percentage changes.

Given the estimate (3.9) we need a formal test to evaluate the size of a potential gap. McCrary proposes a Wald test for the null-hypothesis of no discontinuity in the density function at the cutoff. As we will discuss in Section 3.4 the Wald test statistic for H0 : $\theta = 0$ and its asymptotic distribution are given by

$$\frac{\widehat{\theta}}{\widehat{\sigma}_{\theta}} \xrightarrow{d} \mathcal{N}(0,1). \tag{3.10}$$

3.4 Properties

The estimate $\widehat{\theta}$ inherits useful properties of general local linear regression. First, local linear regression is attractive due to its simplicity as it can be transformed in a standard weighted least squares problem which is then easily accessible by conventional econometric methods as already shown in Section 3.2. Second, the most important advantage of local linear regression over traditional KDE is the fact that it automatically accounts for the bias at boundary points. Thus, as explained in Section 3.1, the rate of convergence is not affected when estimating the density function at the cutoff. Third, as proved by Cheng, Fan and Marron [4], local linear regression is asymptotically minimax efficient among the class of linear estimators with locally bounded second derivative. This means that local linear regression has the smallest maximum risk in terms of asymptotic mean squared error among all linear estimators in the class considered. So local linear regression results in the asymptotically most accurate linear estimator with respect to asymptotic mean squared error performance. Hence, it is weakly preferred to other methods.

To perform a Wald test for the null-hypothesis of no discontinuity at the cutoff, we need to further study the limiting distribution of the test statistic (3.10). To do so we make use of the following proposition stated and proved by McCrary:

Proposition 1. *Let f be a density function which, everywhere except at c, has three continuous and bounded derivatives. Let $K(t) = \max\{0, 1 - |t|\}$ be the triangle kernel and suppose that $h \to 0, nh \to \infty, b/h \to 0$, and $h^2\sqrt{nh} \to H \in [0, \infty)$ as $n \to \infty$. Then, if $R_1, R_2, \ldots R_n$ is a random sample with density $f(r)$,*

$$\sqrt{nh}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}\left(B, \frac{24}{5}\left(\frac{1}{f^{+}} + \frac{1}{f^{-}}\right)\right), \text{ where } B = \frac{H}{20}\left(-\frac{f^{+\prime\prime}}{f^{+}} + \frac{f^{-\prime\prime}}{f^{-}}\right). \tag{3.11}$$

The proposition reveals that the estimator is asymptotically normally distributed and with the correct choice of the bandwidth h also consistent. In particular, the standard

error used in the Wald test is

$$\widehat{\sigma}_\theta = \sqrt{\frac{1}{nh}\frac{24}{5}\left(\frac{1}{\widehat{f}^+} + \frac{1}{\widehat{f}^-}\right)}. \tag{3.12}$$

Due to consistency and asymptotic normality, we can use quantiles of the standard normal distribution to derive the test decision. Unfortunately, consistency only obtains with a bandwidth choice that results in $H = \lim_{n\to\infty} h^2\sqrt{nh} = 0$. If h is not chosen appropriately, there is a bias in the standard normal limit which depends on H and on the value and curvature of the density function estimate approaching c from above and below, respectively. In Appendix B we give a formal derivation of this bias.

As we are in a non-parametric setting, the rate of convergence is slower than for parametric estimators. In our case, the estimator's rate of convergence is given by $(nh)^{-1/2}$. Hence, we need a large sample size n to obtain an accurate estimate of the discontinuity gap and convincingly argue with asymptotic properties. We investigate the test's performance with respect to different sample sizes in a simulation study in Section 4.

3.5 The Issue of Parameter Selection

As introduced in earlier sections, estimating the discontinuity involves three parameter choices – the binsize b used to construct the first-step histogram, and the kernel function K and bandwidth h used for local linear regression in the second step.

The binsize b determines the appearance of the histogram and choosing it involves a trade-off between bias and variance. If b is too small, only few observations fall into a single bin and the estimated bin heights will suffer from large variance. On the other hand, if b is too large we might fail to account for important characteristics of the histogram and the few amount of bins biases local linear regression in the second step.

As the sample size n and the sample standard deviation of the running variable s influence the number of observations falling into a specified bin, McCrary starts with computing the binsize as $\widehat{b} = 2sn^{-1/2}$. Then he subjectively undersmoothes this pilot binsize after inspecting the appearance of the resulting histogram estimate. The reason why he relies on this procedure is not clear. It is intuitive that n and s are crucial ingredients but McCrary does not explain their composition in the formula for $\widehat{b}$. However, he claims that given a fixed bandwidth h the estimator's performance in the second step is robust to different choices of binsize provided that h is sufficiently larger than b. The reason is that the density estimate $\widehat{f}^+$ can be written as

$$\sqrt{nh}\left(\widehat{f}^+ - f^+\right) = A_n + \mathbb{E}\left[\widehat{f}^+ - f^+\right] \tag{3.13}$$

where A_n tends towards a normal distribution whose quality is independent of the binsize, as shown by McCrary. If h is sufficiently larger than b, the finite-sample bias in the second term is dominated by the bandwidth chosen, rendering the binsize less important. The same argument holds for $\widehat{f}^-$. We give a formal derivation of the finite sample bias that

shows the influence of b and h in the proof of Proposition 1 in Appendix B.

The estimator's robustness under the above condition emphasizes the need for undersmoothing but also shows that an appropriate bandwidth choice is more important. Nevertheless, it is not clear whether the automatic procedure for the binsize and the one for the bandwidth explained below guarantee this relation between h and b, which may be further invalidated by additional rescaling of the bandwidth. So after computing the two parameters, we have to inspect their scale to ensure asymptotic negligibility of b. This is an important point McCrary misses to emphasize.

Moreover, McCrary implicitly assumes that the running variable is continuous as the first step serves to construct a discretized version of it. However, if the running variable is already discrete we might get problems when using his binning technique. If the automatic procedure results in a number of bins larger than the amount of categories of the running variable, we end up with bins not covering any observation. Consequently, the second step underestimates the density function at this particular point. Hence, we suggest to compute the binsize with the procedure above and compare the resulting amount of bins to the range of the running variable. That is, we have to ensure that we have at most as many bins as categories of the running variable. We study an application of the test in presence of a discrete running variable in the data application in Section 5.

To perform local linear regression in the second step we need to specify the local neighborhood taken into account as well as the weighing scheme. The weighing scheme is specified by the kernel K, a non-negative function with bounded support typically assumed to be symmetric. In density estimation settings it assigns a higher weight to observations with regressor values close to the point of interest. There are many kernels used in practice, but Cheng, Fan and Marron [4] show that the triangle kernel $K(t) = \max\{0, 1 - |t|\}$ is boundary optimal in the sense of minimizing the asymptotic mean squared error at boundary points. Therefore, as we are interested in estimation at a boundary point, we employ the triangle kernel throughout our work.

The size of the local neighborhood taken into account by the kernel is determined by a non-negative smoothing parameter, the bandwidth h, controlling the complexity of the model. Similar to binsize selection there is a bias-variance trade-off in choosing the bandwidth. If h is too small we speak of an undersmoothed bandwidth which leads to only few observations falling into the local neighborhood. This results in small bias but large variance and an overfitted density function, highlighting spuriously fine data structures. On the other hand, if h is too large, the large amount of observations falling into the local neighborhood leads to small variance but potentially large bias and the density curve might miss important features of the data. As noted before, because of its influence on the bias an appropriate choice of h is a crucial ingredient. For the sake of simplicity we stick to a constant bandwidth that is the same on both sides of the cutoff.

Fan and Gijbels [6] show that there are in general two different approaches for selecting the bandwidth: subjective choice and automatic procedures. Analogously to binsize selection, subjective choice of the bandwidth is mainly based on visual inspection of the

data and the density function estimated with some pilot bandwidth. The plot is investigated with respect to the model complexity needed and the pilot bandwidth is adjusted accordingly. Automatic selection procedures involve minimizing some criterion function with respect to the bandwidth and then plugging in estimates of unknown quantities in the obtained expression for h.

One commonly used plug-in method resulting in a theoretically optimal constant bandwidth involves minimizing the mean integrated squared error and then plugging in asymptotic expressions for unknown quantities. Following Fan and Gijbels [6], this asymptotically optimal bandwidth is proportional to $n^{-1/5}$ in the linear setting. But using a bandwidth proportional to $n^{-1/5}$ results in $H \neq 0$ in (3.11) and thus an asymptotic bias. We give a formal explanation for the bias' presence in Appendix C.

In the end, we have to choose another bandwidth that is good for testing purposes in the sense of minimizing the asymptotic bias. Following Wasserman [19], typical solutions are either estimation of the bias or undersmoothing the bandwidth such that the bias of $\widehat{\theta}$ decreases asymptotically relative to its variance and the bias in the standard normal limit vanishes. Bias estimation is complicated as it involves estimating the density function's second derivative. For this, we either need to impose functional form assumptions or choose additional smoothing parameters, possibly leading to further bias. Hence, we focus on undersmoothing.

In general, our bandwidth selection procedure follows McCrary's. He uses an automatic procedure to get some pilot bandwidth and then applies undersmoothing. The automatic procedure is based on the rule-of-thumb bandwidth selector by Fan and Gijbels [6] that uses the expression for the asymptotically optimal bandwidth and plugs in estimates of unknown quantities. More formally, using the histogram, McCrary computes the bandwidth as the average of the following quantity for both sides of the cutoff:

$$\kappa \left(\frac{\tilde{\sigma}^2 (b-a)}{\sum_{j=1}^{J} \tilde{f}''(X_j)^2} \right)^{1/5} \tag{3.14}$$

Here, $\tilde{f}$ is a global polynomial of order four fitted to the histogram on the corresponding side of the cutoff, and $\tilde{\sigma}^2$ denotes the mean squared error of this polynomial fitting. The constant $\kappa = 3.348$ is specific for the triangle kernel and the boundary case, and we choose $b - a = X_J - c$ for the right side and $b - a = c - X_1$ for the left.

As the resulting bandwidth is proportional to $n^{-1/5}$, we cannot use it without prior undersmoothing. McCrary recommends a rescaling of 50 percent but does not further discuss the issue. The problem is that the optimal amount of undersmoothing is unknown and subjective criteria are necessary. Also, as mentioned above, we have to pay attention that undersmoothing the bandwidth does not invalidate the binsize's asymptotic negligibility. From a practical point of view this is rather unfavorable and time-consuming. We investigate the influence of undersmoothing in our simulation study in Section 4.

4 Simulation Study

In this section we show by means of a simulation study how McCrary's test performs in different settings. First, we investigate the power of the test by varying the discontinuity size, the number of observations and the distributions the initial values of the running variable are drawn from. Second, we inspect the validity of the test statistic's asymptotic normality and consistency, considering different sample sizes and bandwidths.

4.1 Data Generating Process

To assess and challenge the estimator at hand we require a data generating process (DGP) that creates a discontinuity in the running variable's density function. For that, we randomly and independently draw values from a fixed distribution and apply a selection rule to the data. Observations in a defined neighborhood below the cutoff face an unfair coin flip and may earn an additional value which moves them across the cutoff if the coin evaluates to, say heads. Then, the new assigned running variable value is the initial one plus the neighborhood's length. Manipulation always occurs from left to right. In case of no success in the coin flip, the initial value is untouched. Hence, we ensure that there is no bunching of values at one particular number but that the distribution of values behind the cutoff is smooth. What changes is the probability mass of values within the neighborhood below and above the cutoff and the density function witnesses a jump. We influence the gap size by increasing the probability of success in the coin flip, leading to a larger share of perfect manipulators. In the following, when referring to the share of perfect manipulators we mean the share within the specified neighborhood below the cutoff. To provide economic intuition to the discontinuity gap, we choose the DGP to be rooted in a potential real world setting. Interpreting the density estimates becomes more straight-forward but we lose the ability to precisely control the gap size. See Figure D.1 in Appendix D for a visualization of data manipulated in accordance with the DGP, and the respective density estimation described in Section 3.3.

In the RDD context, our selection rule mirrors a situation where some units have perfect control over treatment assignment, have knowledge of the cutoff value and expect benefits from treatment. Consider again the school voucher program example from Section 2.2. Say, all households want to send their kids to school, they can illicitly increase their poverty index score by a fixed amount and the cutoff is known to a fraction of them. Assume households with poverty index scores far below the cutoff value have little incentives or face moral objections to manipulate their running variable. Even if they know the cutoff, these households decide against selecting into treatment. Further, some households barely below the cutoff may not select into treatment as they are not aware of the program, even if they have an incentive to do so. Finally, only households that have a poverty index score within a certain distance to the cutoff and are aware of the program will manipulate their score and move into treatment. In our simulation study, we can directly control the share of aware households and the lower bound of the poverty

index score that is required to provide enough incentives for manipulation.

4.2 Power of the Test

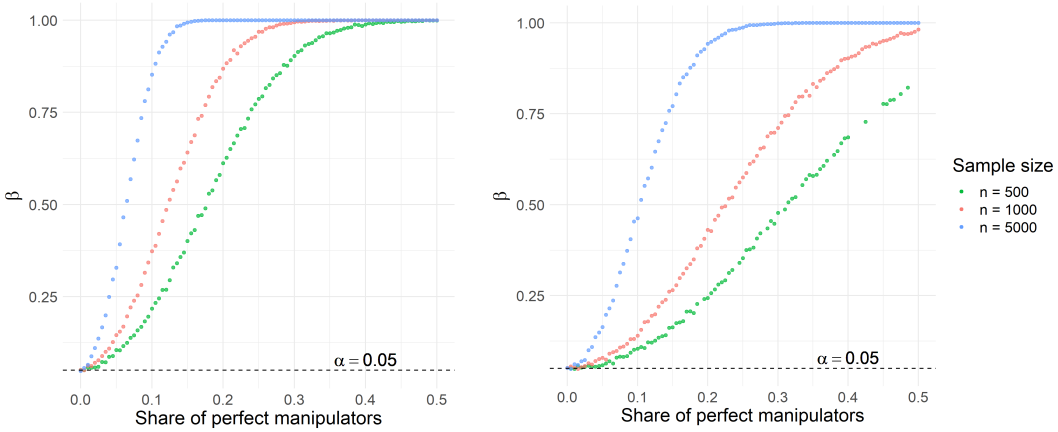
In the first part of our simulation study we investigate the power of McCrary's test as the gap size becomes larger. That is, we inspect the probability that the test rejects the null-hypothesis of continuity when the alternative hypothesis is true when increasing the share of manipulators.

We carry out a Monte Carlo simulation with 5000 draws to investigate how the power of the test evolves as the share of perfect manipulators increases up to 50 percent. We choose a significance level of $\alpha = 0.05$ and compare the results for different sample sizes and different distributions from which the running variable is drawn before being manipulated. We use the standard normal distribution as a benchmark. Then, we deploy a uniform distribution to inspect the test's performance in presence of a flat density function. Finally, we compare the results to those of a F-distribution, which may better mimic income or wealth distributions as they are usually skewed to the right. To make the results for different distribution functions comparable, we fix the neighborhood below the cutoff facing the unfair coin flip in such a way that we arrive at the same absolute number of manipulators. When implementing the test, the bandwidth choice follows the automatic procedure described in Section 3.5 without undersmoothing. We further investigate the issue of undersmoothing in the second part of our simulation study in Section 4.3.

As the estimation error decreases with the sample size n, for each distribution we expect the power function to reach one faster when increasing n, that is we expect to require a lower share of perfect manipulators to always reject the null-hypothesis.

The results for the normal and uniform distribution are shown in Figure 4.1, while those for the other distributions we investigate are in Figure D.2 in Appendix D. As expected, the power approaches one faster as the sample size n increases. In Subfigure (a) of Figure 4.1 we display the results for a standard normal distribution. We choose the cutoff to be -0.25 and fix the size of the neighborhood in which individuals face the coin flip at 0.5. For a sample size of 5000 observations the power exceeds 0.99 for a share of perfect manipulators of 15 percent. That is, if 2.6 percent of all individuals (131 in total) manipulate the test rejects for 99 percent of all Monte Carlo draws. In comparison, with a sample size of 500 we require a share of perfect manipulators of 40 percent, that is around 7 percent of the full sample, for the same rejection rate.

The powerfunction of the uniform distribution in Subfigure (b) depicts the same pattern but the power converges to one slower compared to the normal distribution for every sample size. For instance, we require a 25 percent share of perfect manipulators for 5000 observations to be sufficiently close to one. A potential reason is that for a uniform random variable defined on the interval $[0, 1]$ the density function's value is always higher than for a standard normal random variable. For the same absolute difference in density



(a) Normal distribution

(b) Uniform distribution

Figure 4.1: Power of the Test for Different Distributions and Sample Sizes.

Notes: Monte Carlo simulation with 5000 draws. Before being manipulated, the running variable is drawn from a standard normal distribution (left), and a uniform distribution with bounds 0 and 1 (right). The cutoff is -0.25 for the standard normal distribution and 0.5 for the uniform distribution.

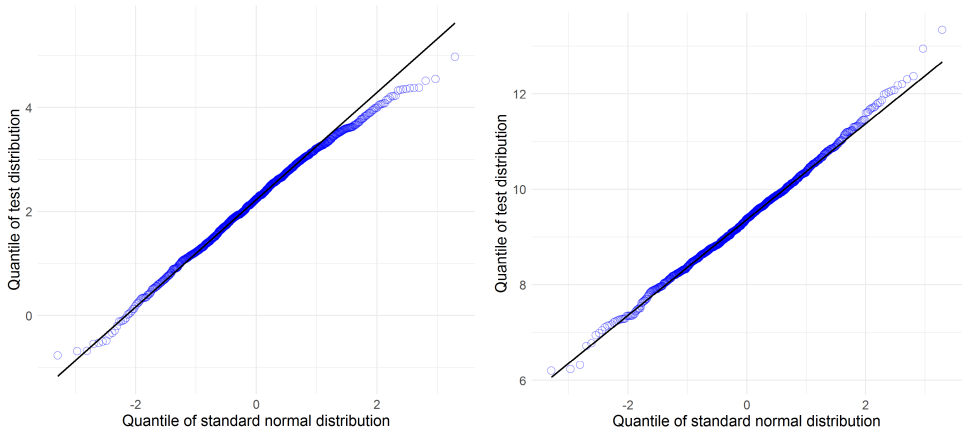
estimates at the cutoff, applying the logarithmic transformation to larger values of density estimates results in a smaller $\widehat{\theta}$ that makes the gap harder to detect.

Additionally, for a sample size of 500, the powerfunction of the uniform distribution looks more sparse. A general problem we encounter is that, if the sample size is small and/or the share of perfect manipulators is large, too few observations may remain just below one side of the cutoff after the running variable is manipulated. In our case, this results in $\widehat{f}^{-} = 0$ and therefore $\widehat{\theta}$ not being defined. Manipulation of the running variable decreases the first derivative of the density function in a neighborhood left of the cutoff. If the density function's derivative in that region is already non-positive, the density estimate approaching the cutoff from below may be zero. The issue emerges in the case of a uniform random variable and for the F-distribution, where we fix the cutoff in the decreasing region of the density function.

4.3 Asymptotic Normality and Consistency

The correctness of the test decision and thus also the test's power depend on the assumptions of asymptotic normality and consistency of the test statistic (3.10). If these are not fulfilled, the use of quantiles of the standard normal distribution might lead to misleading test decisions. Hence, to complement the study of the test's power, we investigate the validity of the standard normal approximation in a second part.

The test statistic's asymptotic standard normality builds on asymptotic normality of (3.11) and consistency of $\widehat{\theta}$. As discussed in Section 3.4, for these to hold we need a sufficiently large sample size as well as the right bandwidth choice. If the bandwidth is not chosen appropriately, the asymptotic bias also depends on the curvature and value of the density function approaching the cutoff from above and below, respectively. However,



(a) 500 sample observations (b) 20,000 sample observations

Figure 4.2: Quantile-Quantile Plots for Different Sample Sizes.

Notes: Monte Carlo simulation with 1000 draws of the test statistic. Before being manipulated the running variable is drawn from a standard normal distribution. We select a share of manipulators of 20 percent.

our DGP makes it hard to control the curvature systematically and in general it is difficult to disentangle the effect of the two. Hence, we focus on the influence of the sample size and bandwidth choice.

We first perform McCrary's test for different sample sizes and afterwards, in line with Section 3.5, consider the influence of undersmoothing the pilot bandwidth. To assess the test statistic's underlying distribution, we use quantile-quantile plots that plot the sample quantiles of the test distribution against the theoretical quantiles of the standard normal distribution. We compute the quantiles of the test distribution by generating 1000 draws of the test statistic. Additionally, we plot a straight line that would correspond to a linear mapping between the sample quantiles and the theoretical ones. If the approximation of the test statistic by a normal random variable holds, we expect the quantiles to lie close to the line. If it is also correctly centered at zero, we expect the line to coincide with the 45°-line. In the following, we only display the results for the case of an initial standard normal distribution as the pattern for the other distributions is similar.

To examine the influence of different sample sizes, we perform the manipulation test for 500, 1000, 5000, 10,000 and 20,000 sample observations. As expected, with an increasing sample size the plotted quantiles approach the straight line evermore, rendering the normal approximation more valid. Figure 4.2 shows the plots for the cases of 500 and 20,000 observations. With only 500 observations the quantiles are still apart, especially at the tails. But with 20,000 observations, a linear mapping between the sample and theoretical quantiles seems to become more valid. The plots for the intermediate sample sizes in Figure D.3 in Appendix D underline this finding. However, the straight line does not correspond to the 45°-line in all cases, implying that the normal approximation in question is not centered at zero.

In line with the discussion on the asymptotic bias in Section 3.5 we expect to achieve

a better centering of the normal approximation when undersmoothing the bandwidth. As the correct amount of undersmoothing is unknown, we compute the pilot bandwidth and then take different percentages of it to perform the test. We pay attention to perform not too much undersmoothing to ensure asymptotic negligibility of the binsize. This is especially important in small samples. We observe an improvement of the centering when shrinking the bandwidth, but we never achieve the sample median coinciding with zero. This is highlighted in Figure D.5 in Appendix D, where we plot the limiting distribution and the test statistic's distribution using the pilot bandwidth and 50 percent undersmoothing thereof.

In Figure D.4 in Appendix D we additionally display the quantile-quantile plots for a sample size of 20,000 and selected degrees of undersmoothing. We observe an improvement in the centering and a slight tendency to come closer to the line, meaning that also the normal approximation in general slightly improves with undersmoothing. However, the improvement is not definite and even more inconclusive in smaller samples. Comparing the different degrees of undersmoothing considered, we realize that McCrary's recommendation of 50 percent is quite arbitrary. Choosing 50 percent undersmoothing (marginally) improves over no undersmoothing but other degrees provide comparable results. So his suggestion should only be taken as a benchmark and further investigation of the appropriate amount of undersmoothing is necessary in either case.

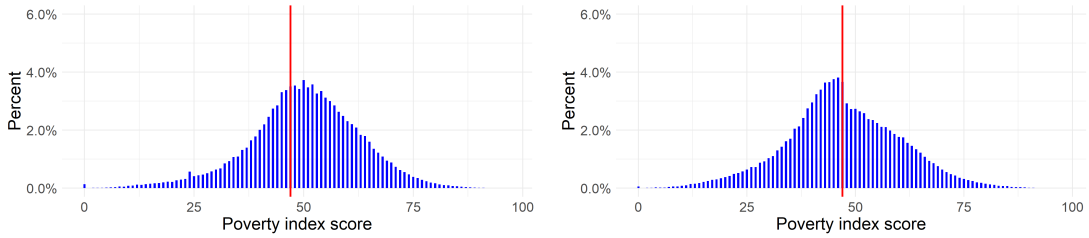
Finally, to see whether the test's size is affected by undersmoothing, we compute the probability of rejecting H0 if no manipulation is present, a Type I error. We choose a significance level of $\alpha = 0.05$ and observe that for 20,000 sample observations and any bandwidth considered the test's size is roughly equal to α. Table D.1 in Appendix D shows detailed results. In general, this finding reassures us that undersmoothing does not lead to an increase in incorrect rejections.

5 Application to Real Data

5.1 Manipulation of Social Program Eligibility in Colombia: The Case of SISBEN

The McCrary test is helpful to assess the validity of a RDD as perfect manipulation is problematic for causal inference. Sometimes the density function of the running variable is of interest per se. Camacho and Conover [1] study an example where a discontinuity in the density function of the running variable helps uncover manipulation of the Census of the Poor, a score-based targeting system for social welfare programs in Colombia.

The Census of the Poor, known as SISBEN in Colombia, is a national survey mandated by the Colombian government and managed by municipalities. It collects comprehensive information on, among many other characteristics, individuals' income, education, housing and employment status. It was designed in the early 1990s to measure long-term quality of life and uses such information to assign a poverty index score, rang-



(a) 1997 (b) 1998

Figure 5.1: Poverty Index Score Distribution.

Notes: The sample for each year is restricted to urban families and to the three poorest categories of neighborhood. In 1997 the algorithm was revealed to local officers. In 1998 a gap emerges at the cutoff of 47, indicated by the red line.

ing from 0 (poorest) to 100 (least poor), to each family. The score is used to decide on individuals' eligibility for a number of social welfare programs targeted at improving living conditions for those in need of assistance. Between 1993 and 2004, when the first SISBEN was implemented, the most common cutoff for urban families to qualify was a score of 47 or less. The algorithm determining the score was revealed to municipal officers in July 1997.

Camacho and Conover analyze the first SISBEN dataset, which covers around 18 million individual observations in urban areas and includes both the answers to all the survey questions and the poverty index score assigned to each family. The dataset is restricted to the poorest three out of six categories of neighborhood, as Colombian neighborhoods are geographically stratified according to their level of wealth.

The authors detect the emergence of a sharp discontinuity in the score density, precisely at 47, after the algorithm was revealed. Figure 5.1 shows the poverty index score distribution for 1997 and 1998, where a discontinuity gap emerges. The evolution of the poverty index score distribution from 1994 until 2003 is shown in Appendix E.

5.2 Estimation of the Gap Size

After a first visual inspection we want to find out whether McCrary's test gives additional evidence for the presence of manipulation. As the running variable is already discrete, we bypass the automatic binning procedure and take as many bins as possible score values (100). Binning the data as suggested by McCrary would result in bins containing zero observations, and therefore biased estimates. We compare the test decisions for the pilot bandwidth and for a 50 percent undersmoothing of the same pilot bandwidth, the amount recommended by McCrary.

Camacho and Conover also inspect the size of the gap, but their procedure differs from McCrary's. They estimate the gap directly through local linear regression. For each year between 1994 and 2003, they regress the normalized height of the bins on the score values and on a dummy variable, which takes the value one if the score is less or equal

	Camacho and Conover		McCrary's test			
			Pilot bandwidth		50% undersmoothing	
Year	Estimate	SE	Estimate	SE	Estimate	SE
1994	0.037	[0.086]	-0.010	[0.007]	0.014	[0.010]
1995	0.080	[0.083]	-0.026***	[0.004]	0.004	[0.006]
1996	0.008	[0.121]	-0.002	[0.006]	-0.001	[0.009]
1997	0.021	[0.097]	-0.006	[0.005]	-0.016*	[0.007]
1998	0.868***	[0.019]	-0.256***	[0.005]	-0.219***	[0.007]
1999	1.208***	[0.145]	-0.314***	[0.004]	-0.271***	[0.006]
2000	1.424***	[0.154]	-0.374***	[0.003]	-0.313***	[0.004]
2001	1.682***	[0.150]	-0.510***	[0.003]	-0.450***	[0.005]
2002	1.571***	[0.132]	-0.571***	[0.004]	-0.486***	[0.006]
2003	1.553***	[0.132]	-0.542***	[0.008]	-0.459***	[0.012]

Table 5.1: Size of the Discontinuity at the Cutoff

Notes: Camacho and Conover employ local linear regression to estimate the gap directly.
The test decisions of McCrary's test are compared for two different choices of bandwidth: the pilot bandwidth h proposed by McCrary and the rescaled bandwidth $0.5h$.
We adjust the significance levels for multiple testing using a Bonferroni correction.
*** Significant at the 1 percent level * Significant at the 10 percent level

than 47 and zero otherwise. The dummy's regression coefficient indicates the size of the gap after controlling for the demeaned score. The idea is to detect a spike in the value of the density function at the cutoff, when the dummy variable jumps from zero to one. The authors also use the bandwidth selection procedure suggested by McCrary, without undersmoothing. We find that their estimated coefficients are numerically equivalent to the difference in height at the cutoff $\widehat{f}^- - \widehat{f}^+$ when using the pilot bandwidth without undersmoothing.

Table 5.1 shows the gap estimates and their standard errors by Camacho and Conover from 1994 to 2003. Additionally, we provide the estimator $\widehat{\theta}$ and its standard error $\widehat{\sigma}_\theta$ for all years using the pilot bandwidth and 50 percent undersmoothing thereof. The null-hypothesis of continuity is rejected for all years after 1997, irrespective of the method used. However, the test decisions for years between 1994 and 1997 appear to be more sensitive to the choice of bandwidth, even after applying a Bonferroni correction to account for the problem of multiple testing. Indeed, Frandsen [7] mentions that in presence of a discrete running variable the test may reject the null-hypothesis of continuity too often. McCrary's test relies on an increasing number of bins for local linear regression near the cutoff when increasing the sample size, which is not the case with bins defined as the fixed number of categories of the discrete running variable.

5.3 Further Evidence of Manipulation

A test rejection is not a proof of manipulation. Additional evidence and different explanations that could generate the observed discontinuity in the data need to be considered.

Manipulation could have occurred in different ways and at different stages of the surveying process: the answers or the score could have been changed by local officers, possibly instructed by a politician, or respondents could have lied. The authors mention anecdotal evidence of people hiding their assets or borrowing children to decrease their score. However, the complexity of the score algorithm made perfect manipulation at the individual level very difficult. The score distribution reconstructed by the authors using the algorithm and the respondents' answers also changes discontinuously at the cutoff, which suggests issues at the data entry stage rather than true score overwriting.

Camacho and Conover note that a way to game the algorithm could have been learning a combination of answers resulting in a poverty index score below the cutoff and repeatedly using such combination for different households. Indeed, they identify suspicious repeated answers appearing in surveys conducted after 1998. As an example, they report the case of a municipality in which 45,000 individuals were interviewed on the same day and all scored 31. These individuals had the same answers for questions related to education, earnings, housing and possessions, as well as the same survey supervisor and data-entry person, suggesting a form of centralized manipulation. Depending on different sample restrictions, the results identify between 50,000 and 819,000 households (corresponding to between 178,000 and 2.8 million people) with suspicious similarities in their answers, of which between 77 percent and 95 percent fall below the cutoff.

Camacho and Conover underline these empirical findings with a theoretical framework in which local mayors attempt to influence their chances of reelection by changing the SISBEN scores. They show that the discontinuity at the cutoff is larger in municipalities with the most political competition, where the benefit from an additional vote is higher.

The authors further investigate other potential reasons justifying the data pattern. First, they exclude that the score algorithm itself may mechanically generate a larger number of observations below the eligibility cutoff. Second, they rule out that the distribution of the poverty index score may have resulted from changes in macroeconomic and labor market conditions by examining an alternative dataset designed to measure living conditions where no incentives to manipulation were present. Third, they disprove self-selection in the interviewing process, with rich municipalities conducting interviews earlier and poor municipalities conducting interviews later, thus shifting the data distribution to the left in later years.

In all, there is substantial evidence of manipulation occurring during the first Census of the Poor in Colombia. The analysis suggests large-scale corruption at the local government level, with strong hints of mass manipulation of the data at the entry stage and estimated costs of corruption roughly equal to 7 percent of the National Health and Social Security budget, as an estimated three million people (8 percent of the Colombian population at the time) had their score lowered. The Colombian government has taken measure to avoid the same issue in the implementation of the second Census of the Poor, which started in 2003. Among them, a new questionnaire was provided and the score algorithm was kept secret.

6 Conclusion

In this paper we deal with the manipulation test by McCrary [13] which is commonly used in RDD applications to detect manipulation of the running variable. Intuitively, in case of perfect manipulation we expect many individuals barely qualifying for treatment and few barely not qualifying. This would generate a discontinuity at the cutoff in the running variable's density function which we would not expect in absence of manipulation. Therefore, we test for the null-hypothesis of continuity by means of a Wald test.

To investigate the test's ability to reject the null-hypothesis and highlight the test's properties, we design a simulation study in Section 4 where we draw a running variable from various distributions and apply a selection rule that creates such a discontinuity. In line with asymptotic properties derived in Section 3.4, the test's power increases both with the number of observations and the gap size. As the correctness of test decisions depends upon the test statistic's asymptotic normality and consistency, which are in turn influenced by parameter choice, we investigate the quality of the standard normal limiting distribution for different sample sizes and different degrees of undersmoothing. We show that normality improves with an increasing sample size but the influence of undersmoothing is inconclusive. However, undersmoothing the bandwidth improves the centering of the test statistic's distribution in general, but does not necessarily result in the correct centering at zero. Further, our results suggest that the test's size is robust to bandwidth choice.

Moreover, to study manipulation of social program eligibility in Colombia, in Section 5 we apply the test to a large administrative dataset where the running variable is already discrete. We find that the test rejects for all years in which we strongly believe that manipulation is present, while for years with no further evidence of manipulation the test seems to reject the null-hypothesis too often.

In the end, both the simulation study and the application to real data highlight that parameter choice is non-trivial and the researcher should not rely on rules-of-thumb alone. The test's performance depends on the data structure at hand and further investigation of manipulation is always required. Despite these practical limitations, our simulation study and application to real data showcase the relevance for this test as an important source of information when evaluating the validity of a RDD.

References

[1] A. Camacho and E. Conover. Manipulation of social program eligibility. American Economic Journal: Economic Policy, 3(2):41–65, 2011.

[2] M. D. Cattaneo, M. Jansson, and X. Ma. Simple local polynomial density estimators. Journal of the American Statistical Association, 0(0):1–7, 2019.

[3] M.-Y. Cheng. Boundary aware estimators of integrated density derivative products. Journal of the Royal Statistical Society, Series B, 59(1):191–203, 1997.

[4] M.-Y. Cheng, J. Fan, and J. S. Marron. On automatic boundary corrections. The Annals of Statistics, 25(4):1691–1708, 1997.

[5] J. Fan, T. Gasser, I. Gijbels, M. Brockmann, and J. Engel. Local polynomial fitting: A standard for nonparametric regression. North Carolina State University. Dept. of Statistics, Institute of Statistics mimeo series, 1993.

[6] J. Fan and I. Gijbels. Local Polynomial Modelling and its Applications. Chapman and Hall, New York, 1996.

[7] B. Frandsen. Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete. In M. D. Cattaneo and J. C. Escanciano, editors, Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics), volume 38, pages 281–315. Emerald Group Publishing, 2017.

[8] J. Hahn, P. Todd, and W. van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica, 69(1):201–209, 2001.

[9] G. W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142(2):615–635, 2008.

[10] D. S. Lee. Randomized experiments from non-random selection in U.S. house elections. Journal of Econometrics, 142(2):675–697, 2008.

[11] D. S. Lee and T. Lemieux. Regression discontinuity designs in economics. Journal of Economic Literature, 48(2):281–355, 2010.

[12] J. Marron and D. Ruppert. Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society, Series B, 56(4):653–671, 1994.

[13] J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. Journal of Econometrics, 142(2):698–714, 2008.

[14] T. Otsu, K.-L. Xu, and Y. Matsushita. Estimation and inference of discontinuity in density. Journal of Business and Economic Statistics, 31(4):507–524, 2014.

[15] J. Rice. Boundary modification for kernel regression. Communications in Statistics - Theory and Methods, 7(13):893–900, 1984.

[16] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66:688–701, 1974.

[17] E. F. Schuster. Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics - Theory and Methods, 14(5):1123–1136, 1985.

[18] D. L. Thistlethwaite and D. T. Campbell. Regression-discontinuity analysis: An alternative to the ex-post facto experiment. Journal of Educational Psychology, 51(6):309–317, 1960.

[19] L. Wasserman. All of Nonparametric Statistics. Springer, New York, 2006.

Appendices

A Convergence of the First-Step Histogram

Consider the notation of Section 3 and denote the distribution function of the running variable as F. Then, as formally shown by Cheng [3], the first-step histogram serves as an approximation of the density function in the following way:

$$\begin{aligned}\frac{1}{nb}N_j &= \frac{1}{nb}\sum_{i=1}^{n}\mathbb{1}_{(X_j-b/2,X_j+b/2]}(R_i)\\ &\xrightarrow{p}\frac{1}{b}\int_{X_j-b/2}^{X_j+b/2}f(u)du\\ &=\frac{1}{b}\left(F\left(X_j+\frac{b}{2}\right)-F\left(X_j-\frac{b}{2}\right)\right)\\ &\approx f(X_j)\end{aligned}$$

Convergence in probability follows from the weak law of large numbers and the approximate equality in the last line holds through the approximation with a symmetric difference quotient.

B Proof of Proposition 1

We derive the expression for the estimator's asymptotic bias and refer to the original proof by McCrary [13] for the derivation of the variance and the normal limit.

Proof. We start with deriving the asymptotic bias for the density function estimate when approaching the cutoff from the right side, $\widehat{f}^+$.
Consider the notation of Section 3 and the assumptions of Proposition 1. Define $Y_j = \frac{1}{nb}\sum_{i=1}^{n}\mathbb{1}_{(b_j,b_{j+1}]}(R_i)$ and $t_j = \frac{X_j-c}{h}$. Then, with the results in Section 3.3 we can write the density function estimate approaching c from above as

$$\begin{aligned}\widehat{f}^+ &= \sum_{X_j>c}K\left(\frac{X_j-c}{h}\right)\frac{S^+_{n,2}(c)-S^+_{n,1}(c)(X_j-c)}{S^+_{n,2}(c)S^+_{n,0}(c)-S^+_{n,1}(c)^2}\;Y_j\\ &=\sum_{j=1}^{J}K(t_j)\;\mathbb{1}_{t_j>0}\;\frac{S^+_{n,2}(c)-S^+_{n,1}(c)(X_j-c)}{S^+_{n,2}(c)S^+_{n,0}(c)-S^+_{n,1}(c)^2}\left(\frac{1}{nb}\sum_{i=1}^{n}\mathbb{1}_{(b_j,b_{j+1}]}(R_i)\right)\\ &=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}K(t_j)\;\mathbb{1}_{t_j>0}\;\frac{S^+_{n,2}(c)-S^+_{n,1}(c)(X_j-c)}{S^+_{n,2}(c)S^+_{n,0}(c)-S^+_{n,1}(c)^2}\;\frac{1}{b}\;\mathbb{1}_{(b_j,b_{j+1}]}(R_i)\\ &\equiv\frac{1}{n}\sum_{i=1}^{n}Z_{in} \qquad (1)\end{aligned}$$

As the kernel K is a Riemann-integrable function, we can use a Riemann approximation to transform $S^+_{n,k}(c) = \sum_{j=1}^{J}K(t_j)\,\mathbb{1}_{t_j>0}(X_j-c)^k$ in a finite integral. As we stick to

the triangle kernel $K(t_j) = \max\{0, 1 - |t_j|\}$ by assumption, the support of K is equal to the interval $[0, 1]$, which we partition into smaller intervals of length $\frac{b}{h}$ for the Riemann approximation. The definition of the bin centers X_j leads to t_j being the midpoint of the smaller interval $\left[\frac{b}{h}(j-1), \frac{b}{h}j\right]$ for $j \in \{1, \ldots, \frac{h}{b}\}$. So we can perform a Riemann approximation using the midpoint rule to get

$$\begin{aligned} S^+_{n,k}(c) &= \sum_{j=1}^{J} K(t_j)\ (X_j - c)^k\ \mathbb{1}_{t_j>0} \\ &= h^k \sum_{j=1}^{J} \max\{0, 1 - |t_j|\}\ t_j^k\ \mathbb{1}_{t_j>0} \\ &= h^k \frac{h}{b} \sum_{j=1}^{J} \frac{b}{h}(1 - t_j)\ t_j^k\ \mathbb{1}_{t_j \in [0,1]} \\ &= \frac{h^{k+1}}{b}\left(\int_0^1 t^k (1-t)\, dt + \mathcal{O}\left(\frac{b^2}{h^2}\right)\right) \\ &= \frac{h^{k+1}}{b}\left(\frac{1}{k+1} - \frac{1}{k+2}\right) + \mathcal{O}\left(h^{k-1}b\right). \end{aligned} \tag{2}$$

Hereby, the error term in the fourth line follows from applying the midpoint rule in the Riemann approximation. In general, when approximating a function at the midpoint of partitioned intervals, $\frac{M(b-a)^3}{24m^2}$ is an upper bound to the error, where M is the maximum of the absolute value of the function's second derivative on the interval $[a, b]$ and m is the partitioned intervals' length. In our case, we approximate the function $g(t) = t^k(1-t)$ on the interval $[0, 1]$, so M is of order $\mathcal{O}(1)$ and the size of the partitioned intervals equals $m = \frac{b}{h}$, which leads to an error of order $\mathcal{O}\left(\frac{b^2}{h^2}\right)$.
For $k \in \{0, 1, 2\}$ we obtain

$$S^+_{n,0}(c) = \frac{h}{2b} + \mathcal{O}\left(\frac{b}{h}\right),\ S^+_{n,1}(c) = \frac{h^2}{6b} + \mathcal{O}(b) \text{ and } S^+_{n,2}(c) = \frac{h^3}{12b} + \mathcal{O}(hb). \tag{3}$$

Using this, McCrary shows that

$$\frac{S^+_{n,2}(c) - S^+_{n,1}(c)(X_j - c)}{S^+_{n,2}(c)S^+_{n,0}(c) - S^+_{n,1}(c)^2} = \frac{b}{h}\ 6(1 - 2t_j) + \mathcal{O}\left(\frac{b^2}{h^2}\right). \tag{4}$$

Now with this result, we can take the expectation of (1) and use that the Z_i's are identically distributed, which is the case as the R_i's follow the same distribution by assumption. Pulling out deterministic terms gives

$$\mathbb{E}\left[\widehat{f}^+\right] = \mathbb{E}[Z_{in}] = \sum_{j=1}^{J} K(t_j)\ \mathbb{1}_{t_j>0}\ \frac{b}{h}\ 6(1 - 2t_j)\ \mathbb{E}\left[\frac{1}{b}\mathbb{1}_{(b_j, b_{j+1}]}(R_i)\right] + \mathcal{O}\left(\frac{b}{h}\right) \tag{5}$$

Using the fact that the R_i's are identically distributed with density function f and corresponding distribution function F, and applying an approximation with a symmetric

difference quotient we can transform the expectation into

$$\begin{aligned}\mathbb{E}\left[\frac{1}{b}\mathbb{1}_{(b_j,b_{j+1}]}(R_i)\right] &= \frac{1}{b}\left(F\left(b_{j+1}\right) - F\left(b_j\right)\right)\\ &= f(X_j) + \mathcal{O}\left(b^2\right)\\ &= f(c+ht_j) + \mathcal{O}\left(b^2\right) \qquad (6)\end{aligned}$$

The error term results from the approximation using a symmetric difference quotient. Plugging this into (5) we get

$$\begin{aligned}\mathbb{E}\left[\widehat{f}^{+}\right] &= \sum_{j=1}^{J} \frac{b}{h}\; K(t_j)\; \mathbb{1}_{t_j>0}\; 6(1-2t_j)\; f(c+ht_j) + \mathcal{O}\left(b^2\right) + \mathcal{O}\left(\frac{b}{h}\right)\\ &= \sum_{j=1}^{J} \frac{b}{h}\; (1-t_j)\; 6(1-2t_j)\; f(c+ht_j)\; \mathbb{1}_{t_j\in[0,1]} + \mathcal{O}\left(b^2\right) + \mathcal{O}\left(\frac{b}{h}\right). \qquad (7)\end{aligned}$$

Using the same interval partitioning as above, we can again exploit a Riemann approximation with the midpoint rule to turn (7) into a finite integral. The function of interest is now $g(t) = 6(1-t)(1-2t)\; f(c+ht)$, so that the upper bound of the absolute value of its second derivative is of order $\mathcal{O}\left(h^2\right)$ and the resulting overall error is of order $\mathcal{O}\left(b^2\right)$. Thus, the Riemann approximation gives

$$\sum_{j=1}^{J} \frac{b}{h}\; (1-t_j)\; 6(1-2t_j)\; f(c+ht_j)\; \mathbb{1}_{t_j\in[0,1]} = \int_0^1 6(1-t)(1-2t)\; f(c+ht)\; dt + \mathcal{O}\left(b^2\right) \qquad (8)$$

Additionally, as f is three times continuously differentiable by assumption, we can use a second order Taylor approximation at a point strictly larger than c and then let this point converge to c. Applying the fact that f^+ is the density function's value approaching c from above, the Taylor approximation gives

$$f(c+ht) = f^+ + f^{+\prime} ht + \frac{1}{2}\; f^{+\prime\prime} h^2 t^2 + \mathcal{O}\left(h^3\right). \qquad (9)$$

Thus, equation (7) results in

$$\begin{aligned}\mathbb{E}\left[\widehat{f}^{+}\right] &= f^+ \int_0^1 6(1-t)(1-2t)\; dt + h f^{+\prime} \int_0^1 t\; 6(1-t)(1-2t)\; dt\\ &\quad + \frac{1}{2}\; h^2 f^{+\prime\prime} \int_0^1 t^2\; 6(1-t)(1-2t)\; dt + \mathcal{O}\left(h^3\right) + \mathcal{O}\left(\frac{b}{h}\right) + \mathcal{O}\left(b^2\right)\\ &= f^+ - \frac{1}{2}\frac{1}{10} h^2 f^{+\prime\prime} + \mathcal{O}\left(h^3\right) + \mathcal{O}\left(\frac{b}{h}\right) + \mathcal{O}\left(b^2\right). \qquad (10)\end{aligned}$$

Rearranging and multiplying by $\sqrt{nh}$ gives

(11)
$$\sqrt{nh}\,\mathbb{E}\left[\widehat{f}^{+}-f^{+}\right]=\frac{h^2\sqrt{nh}}{20}\left(-f^{+''}\right)+\mathcal{O}\left(h^3\right)+\mathcal{O}\left(\frac{b}{h}\right)+\mathcal{O}\left(b^2\right)\xrightarrow{n\to\infty}\frac{H}{20}\left(-f^{+''}\right).$$

As $\frac{\partial}{\partial f^{+}}\ln\left(f^{+}\right)\neq 0$ exists, we can finally use a first order Taylor approximation of the natural logarithm at f^{+} to obtain

$$\ln\left(\widehat{f}^{+}\right)=\ln\left(f^{+}\right)+\frac{\partial}{\partial f^{+}}\ln\left(f^{+}\right)\left(\widehat{f}^{+}-f^{+}\right)=\ln\left(f^{+}\right)+\frac{1}{f^{+}}\left(\widehat{f}^{+}-f^{+}\right). \tag{12}$$

Together with (11) this leads to

$$\sqrt{nh}\,\mathbb{E}\left[\ln\left(\widehat{f}^{+}\right)-\ln\left(f^{+}\right)\right]\xrightarrow{n\to\infty}\frac{H}{20}\left(-\frac{f^{+''}}{f^{+}}\right). \tag{13}$$

Analogous derivations for the density function estimate approaching the cutoff from below, $\widehat{f}^{-}$, result in

$$\sqrt{nh}\,\mathbb{E}\left[\ln\left(\widehat{f}^{-}\right)-\ln\left(f^{-}\right)\right]\xrightarrow{n\to\infty}\frac{H}{20}\left(-\frac{f^{-''}}{f^{-}}\right). \tag{14}$$

Combining the results in equations (13) and (14) we get

$$\begin{aligned}\sqrt{nh}\,\mathbb{E}\left[\widehat{\theta}-\theta\right]&=\sqrt{nh}\,\mathbb{E}\left[\left(\ln\left(\widehat{f}^{+}\right)-\ln\left(\widehat{f}^{-}\right)\right)-\left(\ln\left(f^{+}\right)-\ln\left(f^{-}\right)\right)\right]\\&\xrightarrow{n\to\infty}\frac{H}{20}\left(-\frac{f^{+''}}{f^{+}}+\frac{f^{-''}}{f^{-}}\right)\end{aligned} \tag{15}$$

which completes the proof for the estimator's asymptotic bias. □

C Asymptotic Bias in the Standard Normal Limit

Suppose we want to test the null-hypothesis that the transformed discontinuity gap at the cutoff equals θ. Let $\widehat{\theta}$ be an estimator for this parameter and $\widehat{\sigma}_\theta$ an estimate for the estimator's standard deviation. Then, following Wasserman [19] the Wald test statistic can be written as

$$\frac{\widehat{\theta}-\theta}{\widehat{\sigma}_\theta}=\frac{\widehat{\theta}-\mathbb{E}[\widehat{\theta}]}{\widehat{\sigma}_\theta}+\frac{\mathbb{E}[\widehat{\theta}]-\theta}{\widehat{\sigma}_\theta}=\frac{\widehat{\theta}-\mathbb{E}[\widehat{\theta}]}{\widehat{\sigma}_\theta}+\frac{Bias(\widehat{\theta})}{\widehat{\sigma}_\theta}\equiv Z_n+\frac{Bias(\widehat{\theta})}{\widehat{\sigma}_\theta}.$$

With an increasing sample size n, Z_n typically converges to a standard normal random variable. But the second term does not vanish even with large sample sizes if we use an asymptotically optimal bandwidth proportional to $n^{-1/5}$. The reason is that in non-parametric settings using a mean (integrated) squared error optimal bandwidth balances bias and variance of the estimator, that is accepts an asymptotic bias in order to reduce

asymptotic variance. As a result, the second term does not decrease with an increasing sample size n and thus introduces a bias in the standard normal limit of Z_n.

D Additional Figures for the Simulation Study

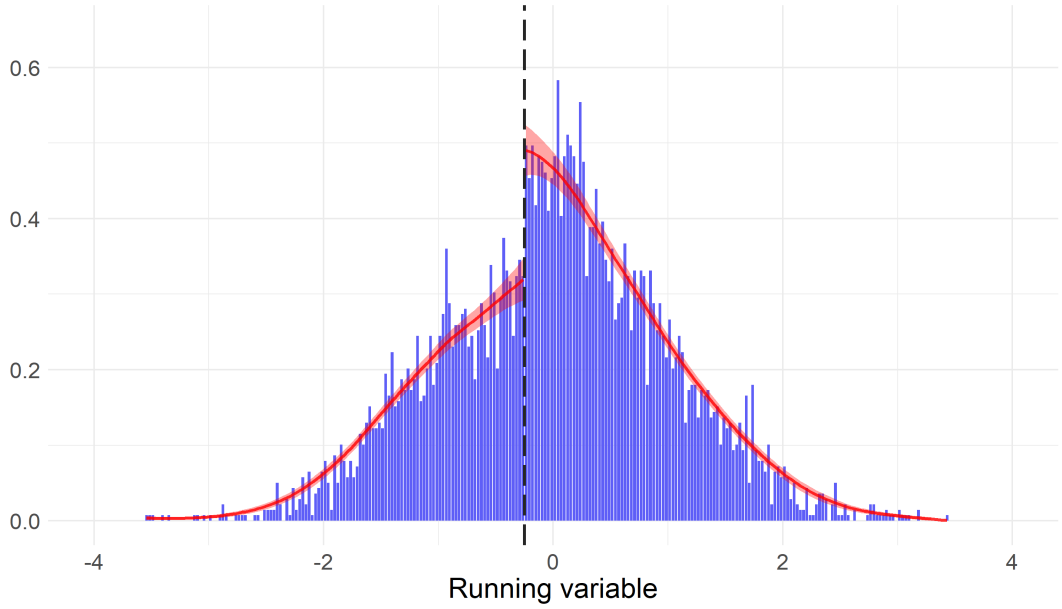
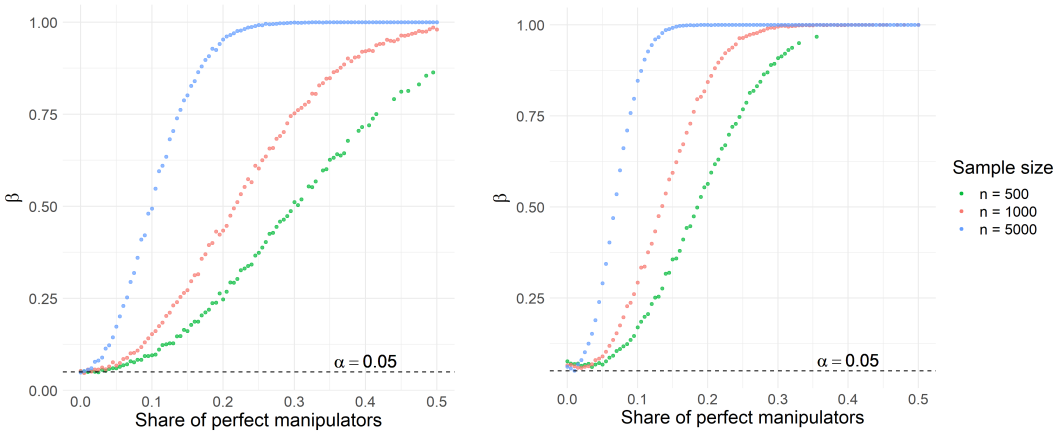


Figure D.1: Density Function estimated with Local Linear Regression.

Notes: Density function of a hypothetical running variable estimated on both sides of the cutoff (-0.25) separately with local linear regression. We draw 5000 initial values from a standard normal distribution and apply the selection rule with a share of perfect manipulators of 20 percent in accordance with the data generating process. The shaded red area is the 95 percent confidence band. We follow the rule-of-thumb parameter selection and construct 260 bins while using a bandwidth of 0.767.

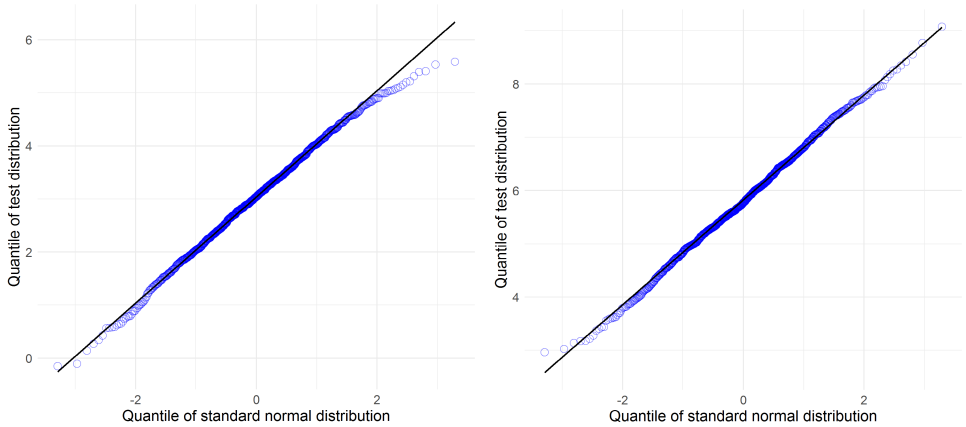


(a) Uniform distribution

(b) F-distribution

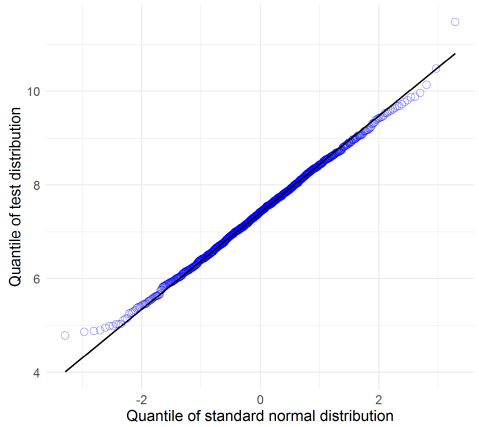
Figure D.2: Power of the Test for Different Distributions and Sample Sizes.

Notes: Monte Carlo simulation with 5000 draws. Before being manipulated, the running variable is drawn from a uniform distribution with bounds 0 and 3.5 (left), and a F-distribution with degrees of freedom 10 and 20 (right). The cutoff is 2 for the uniform distribution and 1.3 for the F-distribution.



(a) 1000 sample observations

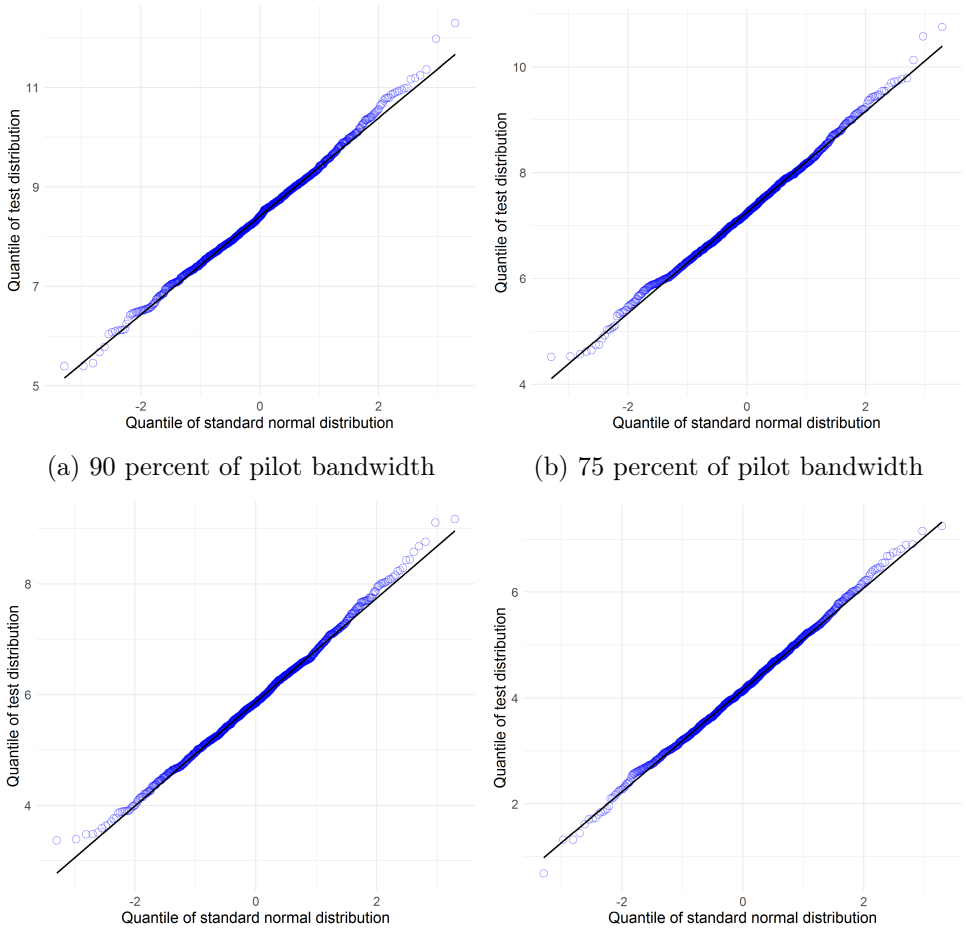
(b) 5000 sample observations



(c) 10,000 sample observations

Figure D.3: Quantile-Quantile Plots for Different Sample Sizes.

Notes: Monte Carlo simulation with 1000 draws of the test statistic. Before being manipulated the running variable is drawn from a standard normal distribution. We select a share of manipulators of 20 percent.



(a) 90 percent of pilot bandwidth

(b) 75 percent of pilot bandwidth

(c) 50 percent of pilot bandwidth

(d) 25 percent of pilot bandwidth

Figure D.4: Quantile-Quantile Plots for Different Degrees of Undersmoothing.

Notes: Monte Carlo simulation with 1000 draws of the test statistic, each with 20,000 observations. Before being manipulated the running variable is drawn from a standard normal distribution. We select a share of manipulators of 20 percent. Undersmoothing is performed by taking a percentage of the pilot bandwidth proposed by McCrary.

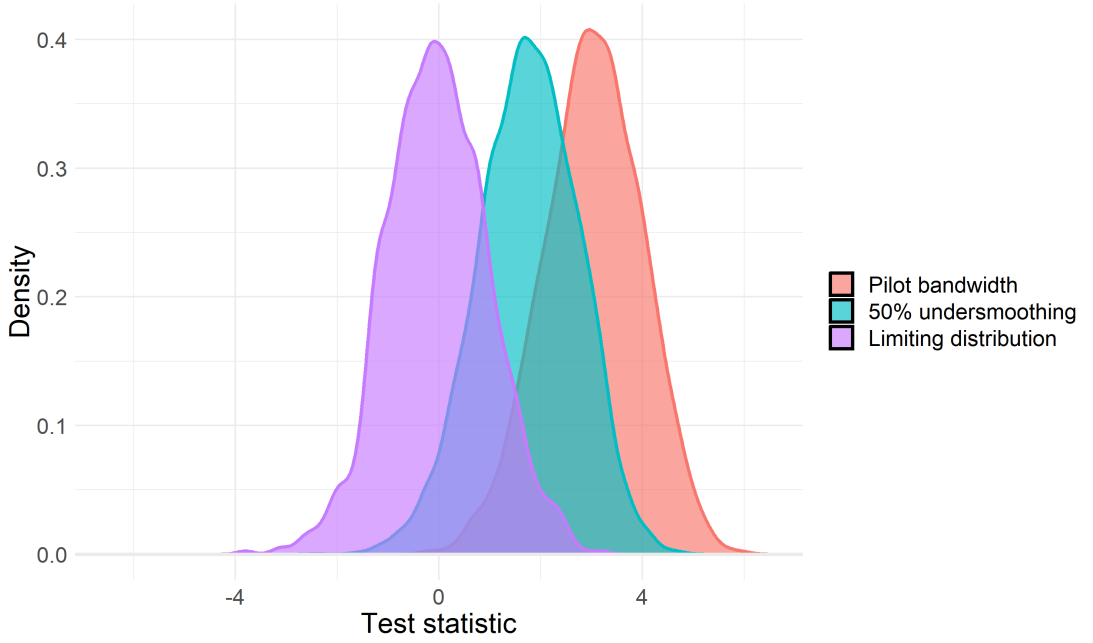


Figure D.5: Consistency and Limiting Distribution.

Notes: Monte Carlo simulation with 5000 draws, each with 1000 observations. We show the test statistic's distributions for the pilot bandwidth and for a 50 percent undersmoothing. For comparison, we depict the correctly centered limiting distribution as well.

Share of pilot bandwidth	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
Normal distribution	0.043	0.052	0.051	0.049	0.056	0.047	0.049	0.058	0.041	0.037	0.040	0.050	0.042	0.046	0.052	0.049
Uniform distribution	0.051	0.035	0.046	0.043	0.053	0.043	0.042	0.030	0.045	0.057	0.047	0.054	0.055	0.056	0.042	0.051
F-distribution	0.046	0.043	0.058	0.054	0.052	0.057	0.049	0.068	0.052	0.044	0.059	0.053	0.054	0.062	0.054	0.075

Table D.1: Size of the Test.

Notes: Monte Carlo simulation with 1000 draws, each with 20,000 observations. Before being manipulated the running variable is drawn from a standard normal distribution, a uniform distribution with bounds 0 and 1 and a F-distribution with degrees of freedom 10 and 20. The significance level of the test is $\alpha = 0.05$.

E Additional Figures for the Application to Real Data

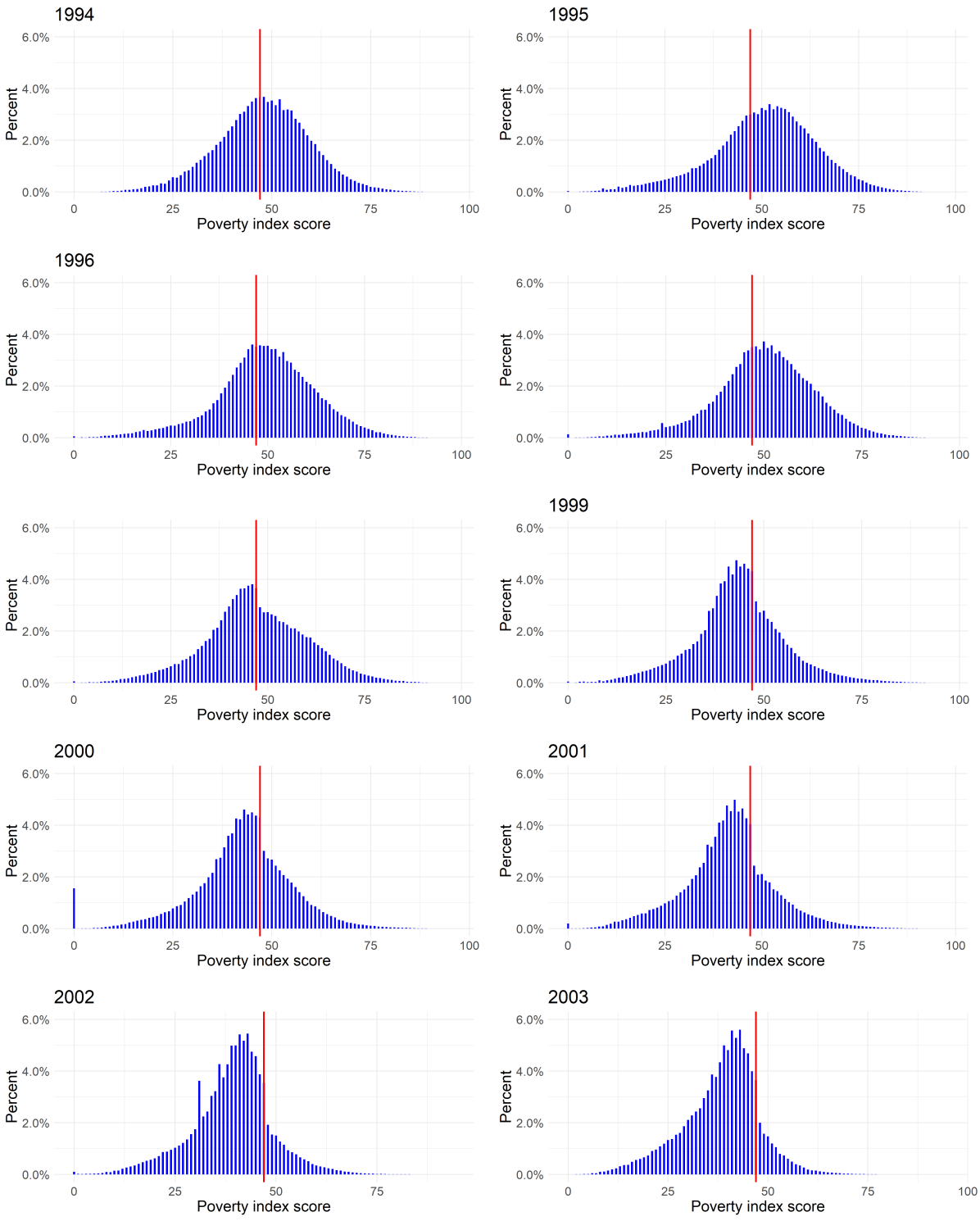


Figure E.1: Poverty Index Score Distribution 1994-2003.

Notes: The algorithm was revealed to local officers in 1997. The sample for each year is restricted to urban families and to the three poorest categories of neighborhood. The red line represents the cutoff of 47.