Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 142 (2008) 698–714
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

Manipulation of the running variable in the regression discontinuity design: A density test

Justin McCrary*

University of Michigan, 735 S. State St. # 5220, Ann Arbor, MI 48109, USA

Available online 29 May 2007

Abstract

Standard sufficient conditions for identification in the regression discontinuity design are continuity of the conditional expectation of counterfactual outcomes in the running variable. These continuity assumptions may not be plausible if agents are able to manipulate the running variable. This paper develops a test of manipulation related to continuity of the running variable density function. The methodology is applied to popular elections to the House of Representatives, where sorting is neither expected nor found, and to roll call voting in the House, where sorting is both expected and found.

JEL classification: C14

Keywords: Regression discontinuity design; Local linear density estimator

1. Introduction

One reason for the increasing popularity in economics of regression discontinuity applications is the perception that the identifying assumptions are quite weak. However, while some applications of the design can be highly persuasive, many are subject to the criticism that public knowledge of the treatment assignment rule may invalidate the continuity assumptions at the heart of identification.

Consider a hypothetical example. A doctor plans to randomly assign heart patients to a statin and a placebo to study the effect of the statin on heart attack within 10 years. The doctor randomly assigns patients to two different waiting rooms, A and B, and plans to give those in A the statin and those in B the placebo. If some of the patients learn of the planned treatment assignment mechanism, we would expect them to proceed to waiting room A. If the doctor fails to divine the patients' contrivance and follows the original protocol, random assignment of patients to separate waiting rooms may be undone by patient sorting after random assignment. In the regression discontinuity context, an analogous evaluation problem may occur in the common case where the treatment assignment rule is public knowledge (cf., Lee, 2007).

In this paper, I propose a formal test for sorting of this type. The test is based on the intuition that, in the example above, we would expect for waiting room A to become crowded. In the regression discontinuity

*Tel.: +1 734 615 7549; fax: +1 734 763 9181.
E-mail address: jmccrary@umich.edu

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2007.05.005

context, this is analogous to expecting the running variable to be discontinuous at the cutoff, with surprisingly many individuals just barely qualifying for a desirable treatment assignment and surprisingly few failing to quality. This test will be informative when manipulation of the running variable is monotonic, in a sense to be made specific below.

The proposed test is based on an estimator for the discontinuity at the cutoff in the density function of the running variable. The test is implemented as a Wald test of the null hypothesis that the discontinuity is zero. The estimator, which is a simple extension of the local linear density estimator (Cheng et al., 1997), proceeds in two steps. In the first step, one obtains a finely gridded histogram. In the second step, one smooths the histogram using local linear regression, separately on either side of the cutoff. To efficiently convey sensitivity of the discontinuity estimate to smoothing assumptions, one may augment a graphical presentation of the second-step smoother with the first-step histogram, analogous to presenting local averages along with an estimated conditional expectation.

This test complements existing specification checks in regression discontinuity applications. Authors routinely report on the smoothness of pre-determined characteristics around the cutoff (e.g., DiNardo and Lee, 2004). If the particular pre-determined characteristics the researcher has at disposal are relevant to the problem, this method should be informative about any sorting around the discontinuity. However, in some applications pre-determined characteristics are either not available, or those which are available are not relevant to the outcome under study. By way of contrast, the density test may always be conducted, since data on the running variable is required for any analysis. The method is also useful in applications where a discontinuous density function is itself the object of interest. For example, Saez (1999, 2002) measures tax avoidance using the discontinuity in the density of income reported to the Internal Revenue Service.

To show how the estimator works in practice, I apply the methodology to two distinct settings. The first setting is popular elections to the United States House of Representatives, considered in Lee's (2001, 2007) incumbency study. In this context, it is natural to assume that the density function of the democratic vote share is continuous at 50%. The data do not reject this prediction.[1] The second setting is roll call votes in the House. In this context, the vote tally for a given bill is expected to be subject to manipulation. Although the number of representatives would seem to make coordination between members difficult, these problems are overcome by a combination of the repeated game aspect of roll call votes and the fact that a representative's actual vote becomes public knowledge, enabling credible commitments and vote contracting. In this setting, the density test provides strong evidence of manipulation.

The remainder of the paper is organized as follows. Section 2 defines manipulation and distinguishes between partial and complete manipulation. Section 3 describes the estimator and discusses smoothing parameter methods and inference procedures. Section 4 motivates the manipulation problem with a hypothetical job training program. Section 5 presents the results of a small simulation study. Section 6 presents the empirical analysis, and Section 7 concludes. Appendix A gives a proof of the proposition of Section 3, and Appendix B describes the data.

2. Identification under partial and complete manipulation

Let Y_i denote an outcome and D_i a binary treatment. The outcome depends on treatment according to

$$Y_i = \alpha_i + \beta_i D_i = \overline{\alpha} + \overline{\beta} D_i + \varepsilon_i, \tag{1}$$

where α_i and β_i are random variables with means $\overline{\alpha}$ and $\overline{\beta}$, respectively, and $\varepsilon_i = \alpha_i - \overline{\alpha} + (\beta_i - \overline{\beta})D_i$ (cf., appendices of Card, 1999). In counterfactual notation, $\alpha_i = Y_{i0}$ and $\beta_i = Y_{i1} - Y_{i0}$, where Y_{i0} is the outcome that would obtain, were $D_i = 0$, and Y_{i1} is the outcome that would obtain, were $D_i = 1$. Eq. (1) is viewed as a structural equation, in the sense that the manner in which i is induced into participation in the program does not affect (Y_{i0}, Y_{i1}) under exogeneity.[2] As noted by Hahn et al. (2001, hereinafter HTV), and following Imbens and Angrist (1994), the average β_i for a specific subpopulation is identifiable under continuity of the

[1] However, see Snyder (2005).

[2] This is Heckman's (2005, p. 11) assumption (A-2). In the statistics literature, this is subsumed under the stable unit treatment value assumption (SUTVA). See Rubin (1980, 1986). I also abstract from general equilibrium effects.

conditional expectations of α_i and β_i, given an underlying index. This index is here termed the "running variable" and denoted R_i.

Underlying R_i is an unobservable index R_{i0} that is the running variable that would obtain, were there no program. This may be different from R_i. The running variable is *manipulated* when $R_i \neq R_{i0}$. Although R_{i0} is not observed, it is well-defined conceptually. For example, van der Klaauw (2002) studies the impact of scholarships on students' enrollment decisions, where scholarships are assigned discontinuously on the basis of a linear combination of SAT and high school grade point average (GPA). It is straightforward to conceptualize of the linear combination of the ith student's SAT and GPA that would obtain, if the university in question did not run such a scholarship program.

I interpret the identification results of HTV as holding under continuity assumptions pertaining to the unobservable index R_{i0}. Formally, throughout the paper I assume that

$$\mathrm{E}[\alpha_i|R_{i0}=r], \quad \mathrm{E}[\beta_i|R_{i0}=r], \quad \text{and} \quad f_{R_{i0}}(r) \text{ are continuous in } r, \tag{A0}$$

where $f_{R_{i0}}(r)$ is the density of R_{i0}. Although this assumption is very weak, it is sufficient for a regression discontinuity estimator based on the index R_{i0} to identify a local average treatment effect.[3] The conditional expectation restrictions in (A0) are HTV's identification assumptions, but (A0) is stronger than their assumptions because of the additional restriction that the density function be continuous. For most settings in which continuity of the conditional expectations is plausible, continuity of the density will be plausible.

If there is no manipulation, then (A0) holds with R_i replacing R_{i0}, and identification of meaningful parameters can be obtained. Sufficient conditions for lack of manipulation include timing, such as when the program is announced simultaneously with implementation, and lack of agent interest in obtaining any particular training assignment, for example. However, when individuals know of the selection rule for treatment, are interested in being treated, and have time to adjust their behavior accordingly, manipulation can be important. In Section 4, below, I give an example of a job training program where manipulation is expected and show how manipulation leads to erroneous inferences. The density test detects manipulation easily in this setting. In the example, the identification problem arises because the incentives of the program lead to sorting on the running variable. Generally, manipulation can lead $\mathrm{E}[\alpha_i|R_i=r]$ and $\mathrm{E}[\beta_i|R_i=r]$ to be discontinuous at the cutoff, despite continuity of $\mathrm{E}[\alpha_i|R_{i0}=r]$ and $\mathrm{E}[\beta_i|R_{i0}=r]$.

Only some varieties of manipulation lead to identification problems. I draw a distinction between *partial* and *complete* manipulation. Partial manipulation occurs when the running variable is under the agent's control, but also has an idiosyncratic element. Typically, partial manipulation of the running variable does not lead to identification problems. Examples of regression discontinuity settings where partial manipulation is arguably plausible include van der Klaauw (2002) and DiNardo and Lee (2004), for example.[4] Complete manipulation occurs when the running variable is entirely under the agent's control. Typically, complete manipulation of the running variable does lead to identification problems. Examples of regression discontinuity settings in which complete manipulation is a potential threat to validity include Hahn et al. (1999) and Jacob and Lefgren (2004), for example.[5]

[3] For discussion of the local average treatment effect parameter, see Angrist et al. (1996) and Heckman et al. (2006), for example.

[4] van der Klaauw (2002) studies the effect of scholarships on enrollment for a college that assigns scholarships discontinuously using an index that is a linear combination of SAT score and high school grade point average (p. 1255). van der Klaauw does not state whether students could have had prior knowledge of the formula used, but it seems plausible that even if they had, it would be difficult to control precisely the value of such an index. Similarly, it might be difficult to control one's grade point average perfectly. DiNardo and Lee (2004) study the impact of unionization on establishment outcomes. Firms become unionized based on a majority vote of the employees. While firms and unions certainly attempt to manipulate the vote tally, it would be difficult for either to do so perfectly, particularly since union certification elections are secret ballot.

[5] Hahn et al. (1999) study the impact of equal employment opportunity laws on employment of racial minorities, taking advantage of the fact that the 1964 Civil Rights Act, as amended, covers only those firms with 15 or more employees. Employers presumably maintain perfect control over labor inputs. This raises the possibility that a firm owner with a taste for discrimination, who would otherwise find it profit-maximizing to employ 15, 16, or 17 employees, for example, would elect to employ 14 employees to preclude the possibility of litigation alleging violations of the Civil Rights Act (cf., Becker, 1957). Jacob and Lefgren (2004) study the impact of summer school and grade retention on test scores, where the treatments depend discontinuously on separate pre-tests. In that context, because the treatment assignment rule is public knowledge, it is possible that those grading the pre-test would be motivated to influence a student's treatment assignment by strategically mismeasuring the student's actual score (see authors' discussion, p. 231).

Propositions 2 and 3 of Lee (2007) establish that, under mild regularity conditions, identification of meaningful parameters can be obtained under partial manipulation. As Lee notes, the critical assumption underlying both propositions is that the conditional density function $f_{R|W}(r|w)$ be continuous in r, where W represents potential confounders (''types''). This is an intuitive identifying assumption: if the running variable has a continuous density conditional on type, then for every type of person the chance of a running variable draw just above the cutoff is equal to the chance of a running variable draw just below the cutoff. The assumption is not directly testable, since types are unobserved. However, Lee stresses the important idea that this assumption implies continuity of the conditional expectation of any baseline characteristic in the running variable. It is thus easy to test the identifying assumption using standard estimators for conditional expectations, such as local linear regression or global polynomial regression. Such tests are already commonly reported in applications.

This paper develops a complementary testing procedure. The idea of the test is that continuity in r of the conditional density $f_{R|W}(r|w)$ implies continuity of $f_R(r)$, the density of the running variable. Thus, a natural specification test in applications is a test of the continuity of the running variable density function.

The density test may not be informative unless the existence of the program induces agents to adjust the running variable in one direction only. Manipulation is *monotonic* if either $R_i \geqslant R_{i0}$ for all i or $R_i \leqslant R_{i0}$ for all i. Consider a hypothetical example based on the Jacob and Lefgren (2004) study, in which the probability of summer school is a discontinuous function of test scores, and teachers are in charge of grading examinations for summer school. Assume students attend summer school if and only if assigned to attend, so that in the absence of manipulation, the local average treatment effect equals the average treatment effect (ATE). Let the ATE be zero, but assume students have heterogeneous treatment effects of summer school; summer school helps half and harms half. Teachers discern these treatment effects, and manipulate the scores of those who would be helped and who just barely passed, so that they fail and have to go to summer school. Similarly, teachers manipulate the scores of those who would be harmed and who just barely failed, so that they pass and avoid going to summer school. Estimated treatment effects of the program would be positive, because of teacher manipulation of scores. However, because the manipulation is non-monotonic, and because those whose scores are adjusted up are equally numerous as those whose scores are adjusted down, the density test will fail to detect manipulation.

The density test could also fail, even when there is no failure of identification. Assume teachers give bonus points to some of those who just barely fail the exam (perhaps to reduce the size of summer school classes), and subtract points from no student. Then the density test would suggest a failure of identification. However, if teachers select at random which students receive bonus points, then an ATE would be identified. These examples clarify that a running variable with a continuous density is neither necessary nor sufficient for identification except under auxiliary assumptions.[6]

3. Estimation and inference procedures

To estimate potentially discontinuous density functions, economists have used either traditional histogram techniques (DiNardo and Lee, 2004; Saez, 2002), or kernel density estimates which smooth over the point of potential discontinuity (DiNardo et al., 1996; Saez, 1999; Jacob and Lefgren, 2004). Neither procedure allows for point estimation or inference. One could estimate a kernel density function separately for points to the left and right of the point of discontinuity, but at boundaries a kernel density estimator is badly biased, as is well-known (e.g., Marron and Ruppert, 1994).

One method that corrects for boundary bias is the local linear density estimator developed by Cheng et al. (1993) and Cheng (1994).[7,8] The grounds for focusing on the local linear density estimator are theoretical and

[6] I thank the editors for their emphasis of this important point.

[7] Published papers describing the local linear density approach include Fan and Gijbels (1996), Cheng (1997a, b), and Cheng et al. (1997). The general idea of ''pre-binning'' the data before density estimation, and the conclusion that estimators based on pre-binned data do not suffer in terms of practical performance despite theoretical loss of information, are both much older than the idea of local linear density estimation; see, for example, Jones (1989) and references therein.

[8] Competing estimators for estimating a density function at a boundary are also available. Estimators from the statistics literature include modified kernel methods (see, e.g., Chu and Cheng, 1996; Cline and Hart, 1991) and wavelet methods (for references, see Hall et

practical. Theoretically, the estimator weakly dominates other proposed methods. Cheng et al. (1997) show that for a boundary point the local linear method is 100% efficient among linear estimators in a minimax sense.[9] Practically, the first-step histogram is of interest in its own right, because it provides an analogue to the local averages typically accompanying conditional expectation estimates in regression discontinuity applications. Moreover, among nonparametric methods showing good performance at boundaries, local linear density estimation is simplest.

3.1. Estimation

Implementing the local linear density estimator involves two steps. The first step is a very undersmoothed histogram. The bins for the histogram are defined carefully enough that no one histogram bin includes points both to the left and right of the point of discontinuity. The second step is local linear smoothing of the histogram. The midpoints of the histogram bins are treated as a regressor, and the normalized counts of the number of observations falling into the bins are treated as an outcome variable. To accommodate the potential discontinuity in the density, local linear smoothing is conducted separately for the bins to the right and left of the point of potential discontinuity, here denoted c.

The first-step histogram is based on the frequency table of a discretized version of the running variable,

$$g(R_i) = \left\lfloor \frac{R_i - c}{b} \right\rfloor b + \frac{b}{2} + c \in \left\{ \ldots, c - 5\frac{b}{2}, c - 3\frac{b}{2}, c - \frac{b}{2}, c + \frac{b}{2}, c + 3\frac{b}{2}, c + 5\frac{b}{2}, \ldots \right\}, \tag{2}$$

where $\lfloor a \rfloor$ is the greatest integer in a.[10,11] Define an equi-spaced grid $X_1, X_2, \ldots, X_J$ of width b covering the support of $g(R_i)$ and define the (normalized) cellsize for the jth bin, $Y_j = (1/nb)\sum_{i=1}^{n} \mathbf{1}(g(R_i) = X_j)$.[12,13] The first-step histogram is the scatterplot (X_j, Y_j). The second step smooths the histogram using local linear regression. Formally, the density estimate at r is given by $\widehat{f}(r) = \widehat{\phi}_1$, where $(\widehat{\phi}_1, \widehat{\phi}_2)$ minimize $L(\phi_1, \phi_2, r) = \sum_{j=1}^{J} \{Y_j - \phi_1 - \phi_2(X_j - r)\}^2 K((X_j - r)/h)\{\mathbf{1}(X_j > c)\mathbf{1}(r \geqslant c) + \mathbf{1}(X_j < c)\mathbf{1}(r < c)\}$, $K(\cdot)$ is a kernel function, here chosen as the triangle kernel $K(t) = \max\{0, 1 - |t|\}$, and h is the bandwidth, or the window width defining which observations are included in the regression.[14] In words, the second step smooths the histogram by estimating a weighted regression using the bin midpoints to explain the height of the bins, giving most weight

(*footnote continued*)

al., 1996). Among the better-known methods, one with good properties is Rice (1984). Boundary folding methods are also used (see, for example, Schuster, 1985), but their properties are not favorable. Marron and Ruppert (1994) give a three-step transformation method. An older method with favorable properties is the smoothed histogram approach developed by Gawronski and Stadtmüller (1980, 1981) and recently explored by Bouezmarni and Scaillet (2005). These last authors also discuss the use of asymmetric kernels for circumventing the boundary bias of kernel estimators. Bouezmarni and Scaillet appear to be the first authors in economics to estimate a density function at a boundary using a nonparametric method, but they do not discuss local linear density estimation. Parametric models involving discontinuous density functions have been studied extensively in economics; see Aigner et al. (1976) for an early paper and Chernozhukov and Hong (2004) for references.

[9] Fan and Gijbels (1996) give a good discussion of this result and discuss further results regarding deeper senses of efficiency.

[10] The greatest integer in a is the unique integer k such that $k \leqslant a < k + 1$ ("round to the left"). In software, this is typically known as the `floor` function, which is not the same as the `int` function, because negatives are handled differently.

[11] Eq. (2) will result in observations with $R_i = c$ being assigned to the bin $c + b/2$, which is valid if ties are assigned to treatment. If ties are assigned to control, re-define $g(R_i) = \lceil (R_i - c)/b \rceil b - b/2 + c$, where $\lceil a \rceil$ is 1 plus the greatest integer in a.

[12] Defining a grid covering the support of $g(R_i)$ is necessary to account for "zero count" bins.

[13] Note that these values of X_j are deterministic in the sense that c and b are treated as constants. The endpoint X_1 (X_J) may always be chosen arbitrarily small (large) so that it is well beyond the support of $g(R_i)$ with no consequences for estimation of the overall density anywhere within $[\underline{R} + h, \overline{R} - h]$, where $[\underline{R}, \overline{R}]$ is the support of the original R_i. Thus, without loss of generality, we may also define $X_j = l + (j - 1)b$, where $l = \lfloor (R^{\min} - c)/b \rfloor b + (b/2) + c$, and $J = \lfloor (R^{\max} - R^{\min})/b \rfloor + 2$. However, if global polynomial fitting is used, as in the automatic bandwidth selector discussed in Section 3.2, below, then the grid should fall strictly in the range $[\underline{R}, \overline{R}]$. This is not necessary if modeling a density with unbounded support.

[14] Given its generally minimal role in performance, the kernel function may be chosen on the basis of convenience. However, the triangle kernel is boundary optimal (Cheng et al., 1997). At interior points, where the Epanechnikov kernel $K(t) = \max\{0, 0.75(1 - t^2)\}$ is optimal, the local linear density estimator is primarily used for graphical purposes and informal inference. Hence there is little cost to using the triangle kernel everywhere, and this is the convention I adopt for Sections 4–6, below.

to the bins nearest where one is trying to estimate the density. It is straightforward to estimate the entire density function, $f(r)$, by looping over evaluation points r.

Define the parameter of interest to be the log difference in height, or

$$\theta = \ln \lim_{r \downarrow c} f(r) - \ln \lim_{r \uparrow c} f(r) \equiv \ln f^{+} - \ln f^{-}. \tag{3}$$

While one can estimate f^{+} and f^{-} using $\widehat{f}(r)$ for r just above and below c, respectively, it is easier and more accurate to estimate two separate local linear regressions, one on either side of c, with $X_j - c$ as regressor. The log difference of the coefficients on the intercepts then estimates θ. Formally,

$$\begin{aligned}\widehat{\theta} &\equiv \ln \widehat{f^{+}} - \ln \widehat{f^{-}} \\ &= \ln \left\{ \sum_{X_j > c} K\left(\frac{X_j - c}{h}\right) \frac{S^{+}_{n,2} - S^{+}_{n,1}(X_j - c)}{S^{+}_{n,2} S^{+}_{n,0} - \left(S^{+}_{n,1}\right)^2} Y_j \right\} \\ &\quad - \ln \left\{ \sum_{X_j < c} K\left(\frac{X_j - c}{h}\right) \frac{S^{-}_{n,2} - S^{-}_{n,1}(X_j - c)}{S^{-}_{n,2} S^{-}_{n,0} - \left(S^{-}_{n,1}\right)^2} Y_j \right\},\end{aligned} \tag{4}$$

where $S^{+}_{n,k} = \sum_{X_j > c} K((X_j - c)/h)(X_j - c)^k$ and $S^{-}_{n,k} = \sum_{X_j < c} K((X_j - c)/h)(X_j - c)^k$. Under standard nonparametric regularity conditions, $\widehat{\theta}$ is consistent and asymptotically normal.

Proposition. *Let $f(\cdot)$ 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 $h^2\sqrt{nh} \longrightarrow H \in [0, \infty)$. Then if $R_1, R_2, \ldots, R_n$ is a random sample with density $f(r)$,*

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

The proof, given in Appendix A, builds on an unpublished proof of Cheng (1994).

The proposition implies an approximate standard error for $\widehat{\theta}$ of

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

As shown in the simulation study in Section 5, below, t-tests constructed using this standard error are very nearly normally distributed under the null hypothesis.

However, the normal distribution in question is not quite centered at zero if the bandwidth is of order $n^{-1/5}$, the rate which minimizes the asymptotic mean-squared error. This is typical of a nonparametric setting; a tuning parameter that is good for estimation purposes is not necessarily good for testing purposes (Pagan and Ullah, 1999). Practically, this means that a confidence region for $\widehat{\theta}$ constructed using the standard error above will give good coverage accuracy for the probability limit of $\widehat{\theta}$, as opposed to good coverage accuracy for θ. Two approaches are taken in the literature to circumvent this problem. First, relative to a bandwidth which is believed to minimize the mean-squared error, one can choose a bandwidth smaller than that. The hope is that the bias is thereby sufficiently reduced that it may be ignored. Second, one can estimate the bias.[15] This bells the cat in that it requires choosing another bandwidth. Following Horowitz (2001) and Hall (1992), I focus on undersmoothing. A simple undersmoothing method is to take a reference bandwidth and to divide it by 2

[15] In their survey, Härdle and Linton (1994) discuss only undersmoothing. Pagan and Ullah (1999) discuss a variety of procedures, but do not provide recommendations. In the related context of local linear regression, Fan and Gijbels (1996) recommend estimating the bias using a two-step procedure; a pilot bandwidth is required for this procedure. Davison and Hinkley (1997) suggest the use of the bootstrap to estimate the bias of the kernel density estimator, but Hall (1992) shows that this method performs badly.

(Hall, 1992). Section 5 presents simulation evidence on the success of this strategy in connection with the reference bandwidth described in the next subsection.

3.2. Binsize and bandwidth selection

For a fixed bandwidth, the estimator described above is robust to different choices of binsize provided that $h/b > 10$, say. To understand this robustness, decompose $\widehat{f}^+$ as

$$\begin{aligned}\sqrt{nh}(\widehat{f}^{\,+} - f^+) &= \frac{1}{\sqrt{h/b}} \sum_{X_j > c} K\left(\frac{X_j - c}{h}\right) \frac{\chi_2 - \chi_1(X_j - c)}{\chi_2\chi_0 - (\chi_1)^2} \sqrt{nb}\left(Y_j - \frac{1}{b}p_j\right) \\ &\quad + \frac{1}{h/b} \sum_{X_j > c} K\left(\frac{X_j - c}{h}\right) \frac{\chi_2 - \chi_1(X_j - c)}{\chi_2\chi_0 - (\chi_1)^2} \sqrt{nh}\left(\frac{1}{b}p_j - f^+\right) \\ &\equiv A_n + \mathrm{E}[\widehat{f}^{\,+} - f^+], \end{aligned} \tag{6}$$

where $\chi_k = (1/(h/b))\sum_{X_j > c} K((X_j - c)/h)(X_j - c)^k$, $k = 0, 1, 2$ and $(1/b)p_j = \int_{-1/2}^{1/2} f\,(X_j + bu)\,\mathrm{d}u$.[16] As shown formally in Appendix A, A_n tends towards a normal distribution. The quality of the normal approximation does not turn on the magnitude of b. Intuitively, the second-step smoother averages over the Y_j, which are themselves averages. If b is small, then the Y_j are not particularly normal, but the second-step smoothing compensates. If b is large, then the Y_j are very nearly normal, and not much averaging needs to happen in the second step. The second sum in this decomposition gives the finite-sample bias of the estimator. Two Taylor approximations and the algebra of regressions show that

$$\mathrm{E}[\widehat{f}^{\,+} - f^+] = \sum_{X_j > c} \frac{b}{h} K(t_j) \frac{\chi_2 - \chi_1 h t_j}{\chi_2\chi_0 - (\chi_1)^2} \sqrt{nh}\left\{h^2 t_j^2 f''^+ + \mathrm{O}(h^3) + \mathrm{O}(b^2)\right\}, \tag{7}$$

where $t_j = (X_j - c)/h$. Since the t_j sequence is b/h apart, this is a Riemann approximation to the area under a curve. The height of the curve in question is dominated by the h^2 term since $h > b$. The analysis for $\sqrt{nh}(\widehat{f}^{\,-} - f^-)$ is symmetric. Thus, good performance of $\widehat{\theta}$ does not appear to require a careful choice of binsize. This point is substantiated in the simulation study in Section 5, below.

Good performance of $\widehat{\theta}$ does require a good choice of bandwidth, however. Probably the best method of bandwidth selection is visual inspection of the first-step histogram and the second-step local linear density function estimate, under a variety of choices for b and h. With software, it is easy to inspect both functions within a few seconds.[17] One of the practical advantages of the two-step estimation method described here is visual. Suppose that as part of a pilot investigation, one has estimated the first-step histogram using binsize b and the second-step local linear smoother using bandwidth h. Graphically, superimposing the local linear smoother on the scatterplot (X_j, Y_j) reveals rapidly the likely consequences of choosing a different bandwidth. The effectiveness of subjective bandwidth choice has been noted in related contexts by Pagan and Ullah (1999) and Deaton (1997), for example.

Less subjective methods include cross-validation (Stone, 1974, 1977) and plug-in estimators. Cheng (1997a) proposes a plug-in bandwidth selector tailored to local linear density estimation, analogous to the Sheather and Jones (1991) selector that is popular in standard density estimation settings. Her method requires estimating the integral of the squared second derivative, $\int (f^{(2)}(r))^2\,\mathrm{d}r$. As is standard in the literature, she uses a bandwidth other than h to estimate $\int (f^{(2)}(r))^2\,\mathrm{d}r$; to find the optimal bandwidth for this ancillary task requires approximating $\int f^{(2)}(r) f^{(4)}(r)\,\mathrm{d}r$, and we are back where we started. Cheng (1994, Section 4.5.2) notes that the method fares poorly in the boundary setting, where the integrals are (particularly) hard to estimate with any accuracy, and suggests further modifications.

[16] An analogous decomposition can be used to motivate an estimator that replaces takes the log of the histogram counts before smoothing. Due to the covariance structure of the Y_j and the nonlinearity of $\ln(\cdot)$, a rigorous demonstration of asymptotic normality does not appear straightforward unless one fixes b and redefines the parameter of interest. Nonetheless, such an estimator is consistent whenever $\widehat{\theta}$ is, and has the same asymptotic variance as $\widehat{\theta}$, provided $nb \to \infty$.

[17] Software (STATA version 9) is available from the author for a period of 3 years from the date of publication.

To be practical, bandwidth selection rules need to be easy to implement. My own view is that the best method is subjective choice, guided by an automatic procedure, particularly if the researcher agrees to report how much the chosen bandwidth deviates from the recommendations of the automatic selector. Here is a simple automatic bandwidth selection procedure that may be used as a guide:

1. Compute the first-step histogram using the binsize $\widehat{b} = 2\widehat{\sigma}n^{-1/2}$, where $\widehat{\sigma}$ is the sample standard deviation of the running variable.
2. Using the first-step histogram, estimate a global 4th order polynomial separately on either side of the cutoff. For each side, compute $\kappa[\check{\sigma}^2(b-a)/\sum \check{f}''(X_j)^2]^{1/5}$, and set $\widehat{h}$ equal to the average of these two quantities, where $\kappa \doteq 3.348$, $\check{\sigma}^2$ is the mean-squared error of the regression, $b-a$ equals $X_J - c$ for the right-hand regression and $c - X_1$ for the left-hand regression, and $\check{f}''(X_j)$ is the estimated second derivative implied by the global polynomial model.[18]

The second step of this algorithm is based on the rule-of-thumb bandwidth selector of Fan and Gijbels (1996, Section 4.2). After implementing this selector, displaying the first-step histogram based on $\widehat{b}$ and the curve $\widehat{f}(r)$ based on $\widehat{h}$ provides a very detailed sense of the distribution of the running variable, upon which subjective methods can be based. The selection method outlined in the above algorithm is used in the simulation study in Section 5, below, where an automatic method is needed. In the empirical work in Section 6, where subjective methods are feasible, this selection method is used as a guide.

4. Theoretical example

To motivate the potential for identification problems caused by manipulation, consider a simple labor supply model. Agents strive to maximize the present discounted value of utility from income over two periods. Each agent chooses to work full- or part-time in each period. Part-time work requires supplying a fraction f_i of full-time labor supply and receiving a fraction f_i of full-time income. Each worker has a different fraction f_i, which is determined unilaterally by the employer prior to period 1 on the basis of production technology. Earnings in period 1 are given by $R_i = \alpha_i H_i$, where $H_i = 1$ if the individual works full-time and $H_i = f_i$ if the individual works part-time. Between periods 1 and 2, a job training program takes place. Agents are eligible for participation if they pass a means test based on period 1 income: program participation is indicated by $D_i = \mathbf{1}(R_i \leqslant c)$, where c is the earnings threshold. Earnings in period 2 are given by $Y_i = \alpha_i + \beta_i D_i$, as in Eq. (1).

If the program did not exist, agents would supply full labor in both periods. In the notation of Section 2, above, this means that $R_{i0} = \alpha_i$. However, the existence of the program raises the possibility that agents will manipulate the running variable, withholding labor supply to meet the means test and gain access to job training. Schematically, the decision problem can be represented as in Fig. 1, where δ is the discount factor. For well-paid agents with $\alpha_i > c/f_i$, the model predicts $H_i = 1$; for such an agent, reducing labor supply is never worth it, because even under part-time work, the agent will not satisfy the means test. For poorly paid agents with $\alpha_i \leqslant c$, the model similarly predicts $H_i = 1$, but for a different reason: such an agent satisfies the means test for the program, even if working full-time. The remaining agents, those with latent wages satisfying $c < \alpha_i \leqslant c/f_i$, may find it worthwhile to reduce labor supply, because otherwise they will fail the means test. These agents reduce labor supply in response to the program if and only if $u(f_i\alpha_i) + \delta u(\alpha_i + \beta_i) > u(\alpha_i) + \delta u(\alpha_i)$. There will always exist a value β_i large enough to induce an agent to select $H_i = f_i$. If β_i and α_i are correlated, as would be expected in the general case, then this leads the conditional expectation of counterfactual outcomes in R_i to be discontinuous. A necessary condition for the utility inequality above to hold is $\beta_i > 0$. Under concave utility a sufficient condition is $\beta_i > (u(\alpha_i) - u(f_i\alpha_i))/\delta u'(\alpha_i)$.Under linear utility, this condition is

[18]The constant κ is based on various integrals of the kernel used in the second step. The standard formula (see Fan and Gijbels, 1996, Eq. (4.3)) does not apply to the boundary case (see Fan and Gijbels, Eqs. (3.20) and (3.22)). The constant cited is specific to the triangle kernel in the boundary case.

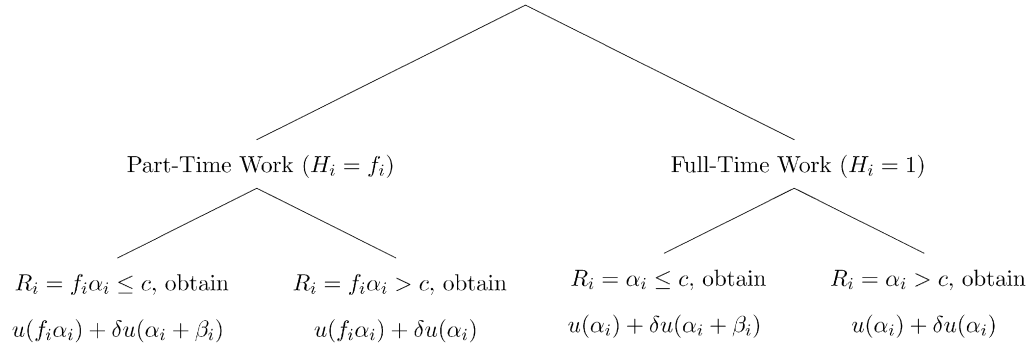

Fig. 1. The agent's problem.

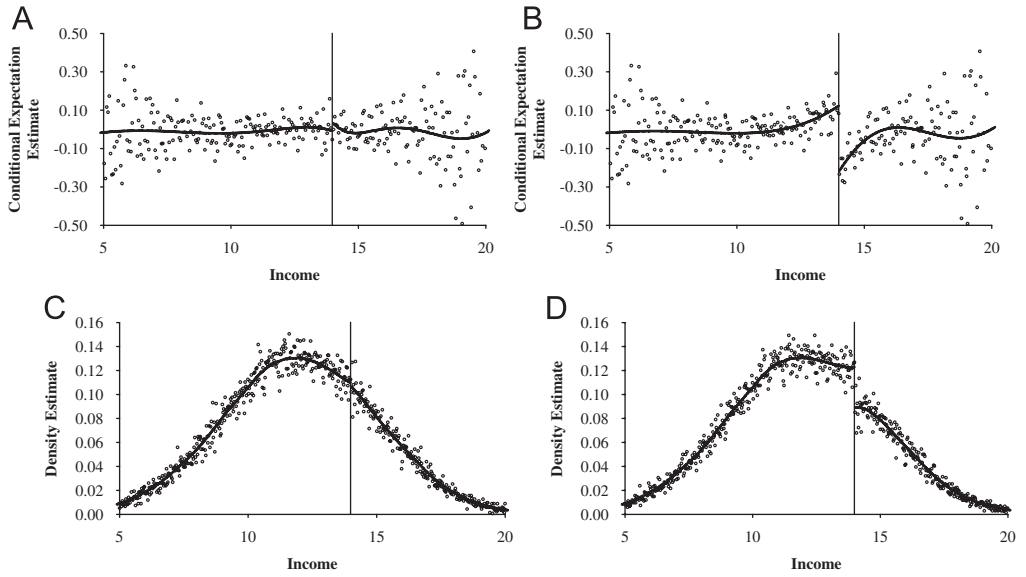

Fig. 2. Hypothetical example: gaming the system with an income-tested job training program: (A) conditional expectation of returns to treatment with no pre-announcement and no manipulation; (B) conditional expectation of returns to treatment with pre-announcement and manipulation; (C) density of income with no pre-announcement and no manipulation; (D) density of income with pre-announcement and manipulation.

also necessary, and we may characterize those who reduce their labor supply as those with $c < \alpha_i \leqslant c/f_i$ and $\beta_i > \alpha_i(1 - f_i)/\delta$.

Fig. 2 shows the implications of these behavioral effects using a simulated data set on 50,000 agents with linear utility. The simulation takes (α_i, β_i) to be distributed as independent normals, with $\mathrm{E}[\alpha_i] = 12$, $V[\alpha_i] = 9$, $\mathrm{E}[\beta_i] = 0$, and $V[\beta_i] = 1$, and the f_i distribution to be uniform on $[0, 1]$ and independent of (α_i, β_i). The earnings threshold is set at $c = 14$.

This data generating process is consistent with (A0). If the program did not exist, then period 1 earnings would be $R_{i0} = \alpha_i$. The conditional expectation of α_i given R_{i0} is thus just the 45° line, which is continuous; the conditional expectation of β_i given R_{i0} is flat, which is likewise continuous; and the density of R_{i0} is the normal density, hence continuous. Panel A of Fig. 2 is a local linear regression estimate of the conditional expectation of β_i given R_{i0}. The smoothness of the conditional expectation indicates the validity of (A0).

However, even though (A0) is satisfied, agents' endogenous labor supply creates an identification problem. The actual running variable is not R_{i0}, but R_i, which is manipulated by those agents who find it worthwhile to do so. Panel B gives a local linear regression estimate of the conditional expectation of β_i given R_i. This panel

highlights the identification problem. The estimated curve is strongly discontinuous near the earnings threshold—those agents who stand to gain from the program self-select to supply less labor and hence are displaced from just to the right of the earnings threshold to just to the left, leading to sample selection effects which operate discontinuously at the earnings threshold.

In empirical work, it is not possible to estimate conditional expectations such as those in panels A and B, because $\beta_i = Y_{i1} - Y_{i0}$ is unobservable. However, it is possible to carry out a density test. Panel C presents an estimate of the density function of R_{i0}, estimated using the local linear density estimation technique described in Section 3, above. The density function is estimated and plotted for evaluation points $r = X_1, X_2, \ldots, X_J$. The bandwidth and binsize were chosen subjectively following inspection of the automatic choices delivered by the algorithm outlined in Section 3.2, above.[19] The density estimate is consistent with continuity at the earnings threshold, as expected.

Panel D instead gives the density function of R_i.[20] In contrast with panel C, the estimated curve is strongly discontinuous at the earnings threshold. The graph furnishes evidence of the economic behavior described by the model above: agents self-select into the job training program by manipulating the value of the running variable that will determine treatment assignment. This leads there to be slightly too few agents just above the means test threshold, and slightly too many agents just below.

5. Simulation evidence

Table 1 presents the results of a small simulation study on the performance of $\widehat{\theta}$ as an estimator and as part of a testing procedure. In the table, "Design I" corresponds to the data generating process underlying panel C from Fig. 2—50,000 independent draws from the N(12, 3) distribution. There are 1000 replication data sets used. For each data set, I calculate $\widehat{\theta}$ using the binsize and bandwidth produced by the algorithm specified in Section 3.2 ("A. Basic, Basic"). In addition to the "basic" implementation of the algorithm, I consider a modified rule that undersmooths the bandwidth, setting it equal to half the size of the basic bandwidth ("B. Basic, Half"). This allows assessment of the bias reduction that comes with undersmoothing. Finally, I consider two non-basic binsizes, corresponding to half the basic binsize width ("C. Half, Basic") and twice the basic binsize width ("D. Twice, Basic"). This is to assess the robustness of the estimator to binsize choices.

The simulation corroborates the good performance suggested by the theoretical work of Section 3. The estimator has generally small bias which declines as the bandwidth shrinks. As well, the standard error suggested by the proposition represents well the approximate underlying standard deviation of the estimator. Importantly, t-tests using the proposition have size of roughly 6%.

Approximating the distribution of $\sqrt{nh}(\widehat{\theta} - \theta)$ with a normal distribution is highly accurate. Fig. 3 presents the normal Q–Q plot for the t-test of the (true) null hypothesis of continuity, where the t-tests stem from the 1000 replications reported in rows A and B of Table 1. Panel A (B) corresponds to row A (B). It is clear from the figure that the quality of the fit is quite good, even far out into the tails where it is most relevant for testing. Comparing panels A and B in the figure, we see that undersmoothing nearly eliminates the estimator's bias.

Perhaps surprisingly, these happy results carry over to much smaller samples. Design II reports results for 1000 replications of data sets with only 1000 observations from the same data generating process as Design I. The bias of the estimator remains manageable, and the accuracy of the variance estimates is striking. The size of tests using the estimation scheme proposed, even with such small sample sizes, is roughly 4–7%. Space precludes the presentation of any further normal Q–Q plots, but these are similar to those shown in Fig. 3, in that neither skewness nor fat tails is indicated.

Finally, these results also carry over to much more challenging density functions with multiple modes. Design III reports results for 1000 replications of data sets with 10,000 observations from a 75–25 mixture of normals with mean 0 and variance 1 and mean 4 and variance 1. The cutoff point was taken to be at 2. This is a challenging point for local linear density estimation in this setting, because it is just to the left of a local minimum of the true density, where the density function is strongly quadratic. However, the estimator

[19]The recommended binsize and bandwidth were $b = 0.03$ and $h = 1.5$, and I chose $b = 0.05$ and $h = 0.9$.

[20]In the interest of maintaining comparability, the binsize and bandwidth are kept the same in panels C and D.

Table 1
Simulation results

Design	Rule for binsize, bandwidth	Range of binsizes (b)	Range of bandwidths (h)	Estimator		Proposition standard error	
				Bias	Std. Dev.	Mean	Size, t-test
I	A. Basic, Basic	[0.027, 0.027]	[1.45, 1.56]	−0.0064	0.0353	0.0345	0.0630
	B. Basic, Half	[0.027, 0.027]	[0.73, 0.78]	−0.0018	0.0513	0.0489	0.0600
	C. Half, Basic	[0.013, 0.013]	[1.45, 1.54]	−0.0063	0.0354	0.0346	0.0640
	D. Twice, Basic	[0.053, 0.054]	[1.46, 1.61]	−0.0066	0.0351	0.0343	0.0600
II	A. Basic, Basic	[0.182, 0.196]	[2.44, 3.45]	−0.0420	0.1800	0.1763	0.0580
	B. Basic, Half	[0.183, 0.196]	[1.22, 1.72]	−0.0059	0.2564	0.2532	0.0430
	C. Half, Basic	[0.091, 0.098]	[2.46, 3.44]	−0.0424	0.1793	0.1757	0.0670
	D. Twice, Basic	[0.366, 0.393]	[2.35, 3.46]	−0.0423	0.1809	0.1775	0.0670
III	A. Basic, Basic	[0.040, 0.040]	[0.851, 1.01]	0.0252	0.1598	0.1484	0.0650
	B. Basic, Half	[0.040, 0.040]	[0.426, 0.506]	0.0011	0.2079	0.2010	0.0560
	C. Half, Basic	[0.020, 0.020]	[0.812, 0.950]	0.0222	0.1608	0.1516	0.0610
	D. Twice, Basic	[0.080, 0.081]	[0.912, 1.11]	0.0307	0.1575	0.1440	0.0690

Notes: Simulation results for three different data generating processes and four different binsize and bandwidth selection rules. See text for details.

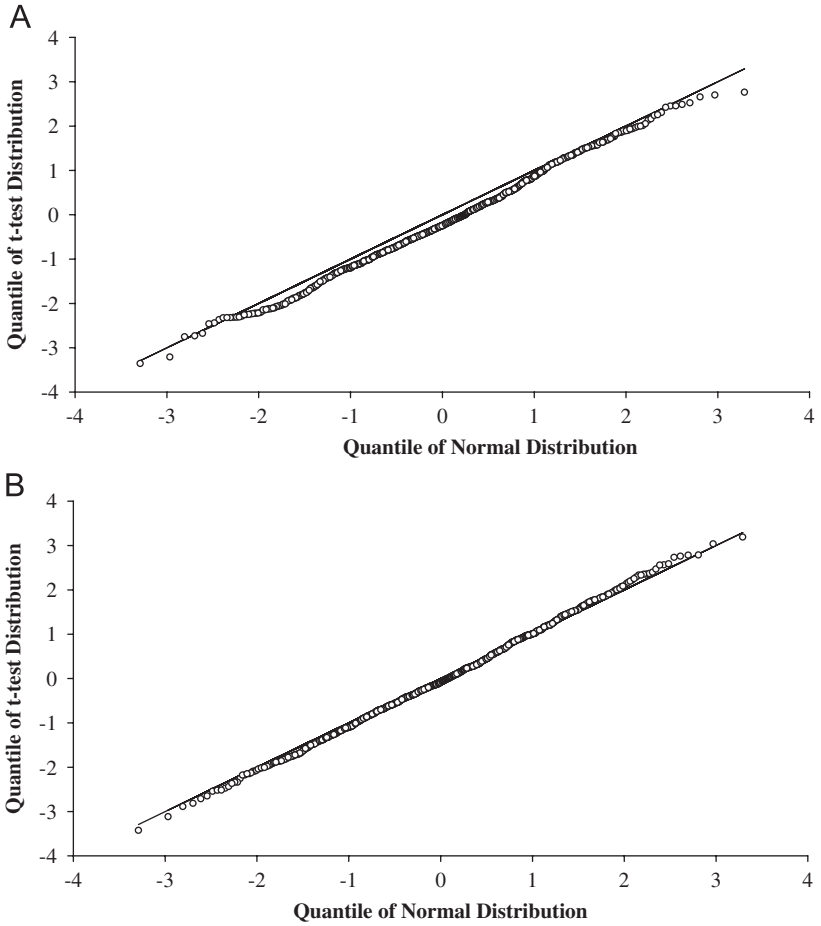

Fig. 3. Quality of normal approximation: (A) t-test based on proposition standard error, no undersmoothing; (B) t-test based on proposition standard error, undersmoothing.

continues to enjoy bias of small magnitude, and t-tests using the estimator and its estimated standard error lead to size of 5–7%.

6. Empirical example

One of the better examples of the regression discontinuity design is the incumbency study of Lee (2001). Political scientists have postulated that there is an incumbency advantage for both parties and individual candidates, whereby having won the election once makes it easier to win the election subsequently. Credibly establishing the magnitude of any incumbency advantage is challenging because of strong selection effects. Lee notes that in a two-party system with majority rule, incumbency is assigned discontinuously at 50% on the basis of the popular vote and uses the regression discontinuity design to assess the party incumbency effect for popular elections to the United States House of Representatives.

The complete manipulation phenomena described in Section 2 seems unlikely to occur in this instance, because voters are unlikely to be able to coordinate to manipulate the vote tally, and because democratic safeguards are presumably sufficient to prevent vote fraud.[21] Thus, a natural expectation is for the density function of the vote share to be smooth. I test this notion formally using the techniques outlined above.

Specifically, using data on the votes cast for each candidate in contested elections to the US House of Representatives involving a democratic candidate, 1900–1990, I estimate the density function of the "democratic margin", defined as the fraction of all votes (vote share) received by the democratic candidate, less the largest vote share received by any other candidate in the election.[22] Defined in this way, the democratic candidate wins the election if and only if the democratic margin is positive.[23]

Fig. 4 gives an estimate of the density function of the democratic margin. The curve was estimated using the estimator outlined in Section 3, with evaluation points $r = X_1, X_2, \ldots, X_J$. The binsize and bandwidth were chosen subjectively after using the automatic procedure outlined in Section 3.2 as a pilot estimate. The automatic procedure in this case seems to oversmooth at the mode in this setting.[24] The estimated curve gives little indication of strong discontinuity near zero. Indeed, the density appears generally quite smooth. Importantly, the first-step histogram reveals that this is not the result of oversmoothing. The estimated parameter $\widehat{\theta}$ is presented in Table 2, along with the proposition standard error. As expected, a t-test of the null hypothesis of continuity fails to reject.

The complete manipulation problem described in Section 2 is unlikely to occur in a fair popular election, because coordination of voters is difficult and there is little discretion in measuring the vote tally. However, in other election contexts, coordination is feasible and complete manipulation may be a concern.

A leading example of this type of coordination is roll call voting in the House of Representatives. Coordination is expected in this context. First, the volume of bills before the House and the long tenure of most representatives conspire to create a repeated game. Second, a representative's vote is public knowledge, allowing for credible commitments to contracts over voting. In such a context, side payments for a representative's vote do not have to involve (illegal) monetary compensation, but may pertain simply to votes on future bills. Riker's (1962) size principle then implies that the most likely bills to be put to vote on the House floor are those expected to narrowly pass.

Fig. 5 presents an estimated density function for the percent voting "yeay" on all roll call votes in the House from 1857–2004.[25,26] The curve was estimated using the estimator outlined in Section 3, with evaluation points

[21] Democratic safeguards may not always be sufficient. Greenberg (2000) discusses the famously contested 1960 presidential election between Richard Nixon and John F. Kennedy. See also Snyder (2005), who uses the estimator described here to analyze close elections to the United States House involving an incumbent.

[22] 1591 elections during this period involve a single candidate. Of the contested elections, 95.3% involve a Democratic candidate, and 92.5% involve a Republican candidate.

[23] This definition of the running variable is slightly different from that in Lee (2007), but differs little as a practical matter, particularly for the post-1948 period pertaining to Lee's study.

[24] I use a binsize of $b = 0.004$ and a bandwidth of $h = 0.02$. The automatic procedure would select $b = 0.004$ and $h = 0.13$.

[25] Stratifying the votes into before and after 1900 subperiods results in highly similar estimates with less precision.

[26] The density estimator is allowed to be discontinuous at 50% but nowhere else, despite the existence of bills which require a supermajority vote for passage (e.g., two-thirds approval for constitutional amendments and veto overrides), so 50% is not the cutoff for passage for all bills. However, bills requiring a supermajority for passage are rare, and the data do not allow me to determine the cutoff for the given bill. Consequently, I focus on the potential discontinuity at 50%, viewing this as being slightly attenuated due to the unobserved supermajority bills.

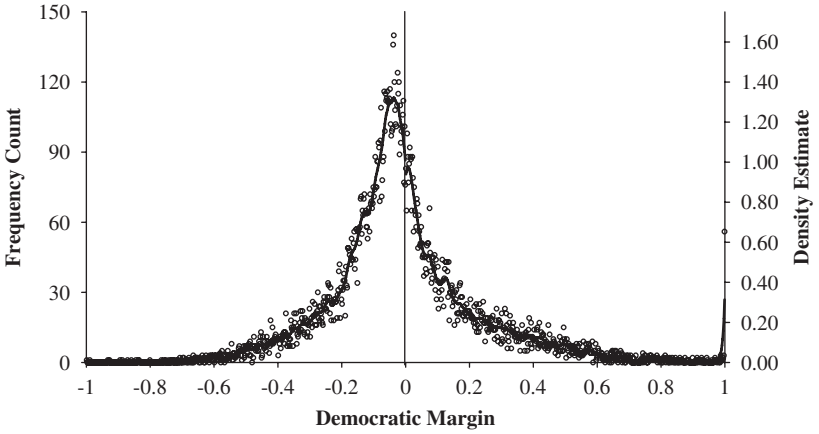

Fig. 4. Democratic vote share relative to cutoff: popular elections to the House of Representatives, 1900–1990.

Table 2
Log discontinuity estimates

	Popular elections	Roll call votes
	−0.060	0.521
	(0.108)	(0.079)
N	16,917	35,052

Note: Standard errors in parentheses. See text for details.

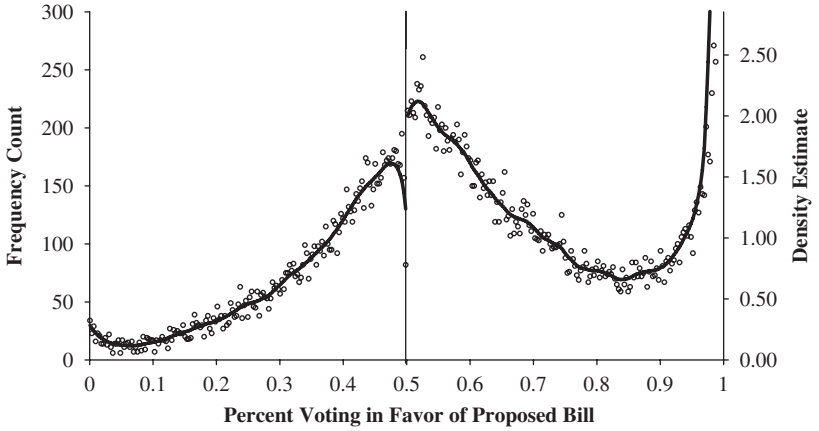

Fig. 5. Percent voting yeay: roll call votes, U.S. House of Representatives, 1857–2004.

$r = X_1, X_2, \ldots, X_J$. The binsize and bandwidth were again chosen subjectively after using the automatic procedure. Much more so than the vote share density, the roll call density exhibits very specific features near the cutoff point that are hard for any automatic procedure to identify.[27]

The figure strongly suggests that the underlying density function is discontinuous at 50%. Outcomes within a handful of votes of the cutoff are much more likely to be won than lost; the first-step histogram indicates that the passage of a roll call vote by 1–2 votes is 2.6 times more likely than the failure of a roll call vote by 1–2

[27]I use a binsize of $b = 0.003$ and a bandwidth of $h = 0.03$. The automatic procedure would select $b = 0.0025$ and $h = 0.114$.

votes. Although the magnitude of the effect is not as extreme, the second-step smoother corroborates the suggestion of the first-step histogram. Table 2 presents the estimated log discontinuity in the density, which is a large 52%. The effect is precisely estimated, with a t-ratio of 6.6.

These empirical results are consistent with a manipulation hypothesis. In particular, the results suggest that it would be a mistake to view the majority vote election procedure in the US House of Representatives as generating quasi-random assignment of policy decisions emerging from the House.

7. Conclusion

This paper describes identification problems encountered in the regression discontinuity design pertaining to manipulation of the running variable and describes a simple test for manipulation. The test involves estimation of the discontinuity in the density function of the running variable at the cutoff. Consistency and asymptotic normality of the log discontinuity in the density at the cutoff was demonstrated theoretically, and inference procedures discussed. The methodology was applied to two distinct settings, one in which manipulation is unexpected and is not detected, and another in which manipulation is expected and demonstrated.

The context of most regression discontinuity applications is such that the treatment assignment rule is public knowledge. I have argued that this will often make it plausible that the agents under study engage in manipulation of the running variable in order to obtain desirable treatment assignments, and I have emphasized that manipulation will often lead to violations of the assumptions necessary for identification.

The standard specification test used currently in regression discontinuity applications is a test for continuity of the conditional expectation of pre-determined characteristics in the running variable at the cutoff. Such tests are a natural and powerful way to assess the plausibility of the identifying assumptions. The density test proposed here complements these methods and is expected to be powerful when manipulation is monotonic, as discussed above. The density test may be particularly important for applications where pre-determined characteristics are not available, or are not relevant to the substantive topic studied.

Acknowledgments

I thank two anonymous referees for comments, the editors for multiple suggestions that substantially improved the paper, Jack Porter, John DiNardo, and Serena Ng for discussion, Jonah Gelbach for computing improvements, and Ming-Yen Cheng for manuscripts. Any errors are my own.

Appendix A. Proof of proposition

Because of the linearity of $Y_j = (1/nb)\sum_{i=1}^{n} \mathbf{1}(g(R_i) = X_j)$, we have

$$\begin{aligned}
\widehat{f}^{+} &= \frac{S^{+}_{n,2}T^{+}_{n,0} - S^{+}_{n,1}T^{+}_{n,1}}{S^{+}_{n,2}S^{+}_{n,0} - S^{+}_{n,1}S^{+}_{n,1}} \\
&= \sum_{j=1}^{J} K(t_j)\mathbf{1}(t_j > 0)\frac{S^{+}_{n,2} - S^{+}_{n,1}\, h\, t_j}{S^{+}_{n,2}S^{+}_{n,0} - S^{+}_{n,1}S^{+}_{n,1}}\, Y_j \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} K(t_j)\mathbf{1}(t_j > 0)\frac{S^{+}_{n,2} - S^{+}_{n,1} h t_j}{S^{+}_{n,2}S^{+}_{n,0} - S^{+}_{n,1}S^{+}_{n,1}}\frac{1}{b}\mathbf{1}(g(R_i) = X_j) \\
&\equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} Z_{ijn} \\
&\equiv \frac{1}{n}\sum_{i=1}^{n} Z_{in}, \qquad \text{(A.1)}
\end{aligned}$$

where $t_j = (X_j - c)/h$, $S^+_{n,k} \equiv h^k \sum_{j=1}^J K(t_j)\mathbf{1}(t_j > 0)t_j^k$, and $T^+_{n,k} \equiv h^k \sum_{j=1}^J K(t_j)\mathbf{1}(t_j > 0)t_j^k Y_j$, and analogously for $\widehat{f}^-$. The proof proceeds by calculating $\mathrm{E}[\widehat{f}^+]$ and $V[\widehat{f}^+]$ and verifying the skewness condition of the Lyapunov central limit theorem (Rao, 1965, p. 107), which applies since Z_{in} is independent of $Z_{i'n}$ for $i' \neq i$. Independence follows since Z_{in} is just a transformation of R_i and since $X_1, X_2, \ldots, X_J$ are constants (see footnote 13). By Riemann approximation (see Cheng, 1994, Lemma 4, for example), we have $S^+_{n,k} = (h^{k+1}/b)S^+_k + \mathrm{O}(h^{k-1}b)$, where $S^+_k = \int_0^\infty t^k K(t)\,\mathrm{d}t$, $k = 0, 1, 2, \ldots$ For the triangle kernel with $k = 0, 1, 2$, S^+_k is equal to $\frac{1}{2}$, $\frac{1}{6}$, and $\frac{1}{12}$, respectively. We have

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

Using Taylor and Riemann approximation we have

$$\begin{aligned}
\mathrm{E}[\widehat{f}^+] = \mathrm{E}[Z_{in}] &= \sum_{j=1}^J \frac{b}{h} K(t_j)\mathbf{1}(t_j > 0)6(1 - 2t_j)f(c + ht_j) + \mathrm{O}\left(\frac{b}{h}\right) + \mathrm{O}(b^2) \\
&= \int_0^1 (1 - t)6(1 - 2t)f(c + ht)\,\mathrm{d}t + \mathrm{O}\left(\frac{b}{h}\right) + \mathrm{O}(b^2) \\
&= f^+ - h^2 \frac{1}{2}\frac{1}{10} f^{+\prime\prime} + \mathrm{O}(h^3) + \mathrm{O}\left(\frac{b}{h}\right) + \mathrm{O}(b^2),
\end{aligned} \tag{A.3}$$

$$\begin{aligned}
V[\widehat{f}^+] = \frac{1}{n} V[Z_{in}] &= \frac{1}{n}(\mathrm{E}[Z_{in}^2] - E^2[Z_{in}]) = \frac{1}{n}\left(\sum_{j=1}^J \sum_{k=1}^J \mathrm{E}[Z_{ijn} Z_{ikn}] - \mathrm{E}^2[Z_{in}]\right) \\
&= \frac{1}{n}\left(\sum_{j=1}^J \frac{b^2}{h^2} K^2(t_j)\mathbf{1}(t_j > 0)36(1 - 2t_j)^2 \frac{1}{b^2} p_j - \mathrm{E}^2[Z_{in}]\right) + \mathrm{O}\left(\frac{b^2}{nh^2}\right) \\
&= \frac{1}{nh}\left(\int_0^1 (1 - t)^2 36(1 - 2t)^2 f(c + ht)\,\mathrm{d}t - h\mathrm{E}^2[Z_{in}]\right) + \mathrm{O}\left(\frac{b^2}{nh^2}\right) + \mathrm{O}\left(\frac{b}{n}\right) \\
&= \frac{1}{nh}\frac{24}{5} f^+ + \mathrm{O}\left(\frac{1}{n}\right)
\end{aligned} \tag{A.4}$$

since $\mathrm{E}[Y_j] = f(X_j) + \mathrm{O}(b^2)$, where the only terms from the double summation which matter are those for which $j = k$ since the histogram bins are mutually exclusive. For the Lyapunov condition, calculate

$$\begin{aligned}
\mathrm{E}[|Z_{in} - \mathrm{E}[Z_{in}]|^3] \leqslant 8\mathrm{E}[|Z_{in}|^3] &\leqslant 8\mathrm{E}\left[\sum_{j=1}^J \sum_{k=1}^J \sum_{l=1}^J |Z_{ijn}| \cdot |Z_{ikn}| \cdot |Z_{iln}|\right] \\
&= 8 \sum_{j=1}^J \mathrm{E}[|Z_{ijn}|^3] \\
&= 8 \frac{1}{h^2} \sum_{j=1}^J \frac{b}{h} K^3(t_j)\mathbf{1}(t_j > 0)6^3 |1 - 2t_j|^3 f(c + ht_j) + \mathrm{O}\left(\frac{b}{h}\right) \\
&= 8 \frac{1}{h^2} \int_0^1 (1 - t)^3 6^3 |1 - 2t|^3 f(c + ht)\,\mathrm{d}t + \mathrm{O}\left(\frac{b}{h}\right) = \mathrm{O}\left(\frac{1}{h^2}\right).
\end{aligned} \tag{A.5}$$

Combining the expression for the variance with the skewness bound, we have

$$\frac{(\sum_{i=1}^n \mathrm{E}[|Z_{in} - \mathrm{E}[Z_{in}]|^3])^{1/3}}{(\sum_{i=1}^n V[Z_{in}])^{1/2}} \leqslant \frac{(\mathrm{O}(n/h^2))^{1/3}}{(\mathrm{O}(n/h))^{1/2}} = \mathrm{O}((nh)^{-1/6}) \tag{A.6}$$

so that the Lyapunov condition is satisfied since $nh \to \infty$. Thus, and by symmetry,

$$\sqrt{nh}(\widehat{f}^{+} - f^{+}) \xrightarrow{d} \mathrm{N}\left(B^{+}, \frac{24}{5} f^{+}\right) \quad \text{and} \quad \sqrt{nh}(\widehat{f}^{-} - f^{-}) \xrightarrow{d} \mathrm{N}\left(B^{-}, \frac{24}{5} f^{-}\right), \tag{A.7}$$

where $B^{+} = -H\frac{1}{20} f^{+\prime\prime}$ and $B^{-} = -H\frac{1}{20} f^{-\prime\prime}$. To strengthen this result to joint asymptotic normality, define $U_{in} = \lambda^{+} Z_{in}^{+} + \lambda^{-} Z_{in}^{-}$, where the Z_{in} from above is redefined to be Z_{in}^{+} and Z_{in}^{-} denotes the analogous quantity to the left of c. Observe that U_{in} is independent of $U_{i'n}$ for all $i' \neq i$. Then we have $\mathrm{E}[U_{in}] = \lambda^{+} f^{+} + \lambda^{-} f^{-} + \mathrm{O}(h^2)$ and $V[U_{in}] = (\frac{24}{5})(\lambda^{+})^2 (f^{+}/h) + (\frac{24}{5})(\lambda^{-})^2 (f^{-}/h) + \mathrm{o}(1/h)$, where the latter follows since $C[Z_{in}^{+}, Z_{in}^{-}] = -\mathrm{E}[Z_{in}^{+}]\mathrm{E}[Z_{in}^{-}] = -f^{+} f^{-} + \mathrm{O}(h^2)$. Using the results from above, we have $\mathrm{E}[|U_{in} - \mathrm{E}[U_{in}]|^3] \leqslant 8\mathrm{E}[|U_{in}|^3] \leqslant 8|\lambda^{+}|^3 \mathrm{E}[|Z_{in}^{+}|^3] + 8|\lambda^{-}|^3 \mathrm{E}[|Z_{in}^{-}|^3] = \mathrm{O}(1/h^2)$ and it is then straightforward to verify the Lyapunov condition as above. Since this holds for every vector $(\lambda^{+}, \lambda^{-})$, the Cramér–Wold device (White, 2001, p. 114) implies joint asymptotic normality with a diagonal asymptotic variance matrix. Define $\tau(f^{+}, f^{-}) = \ln f^{+} - \ln f^{-} = \theta$, note that $\nabla\tau = (1/f^{+}, -1/f^{-})'$, and apply the delta method to conclude

$$\sqrt{nh}(\widehat{\theta} - \theta) \xrightarrow{d} \mathrm{N}\left(B, \frac{24}{5}\left(\frac{1}{f^{+}} + \frac{1}{f^{-}}\right)\right), \tag{A.8}$$

where $B = B^{+}/f^{+} - B^{-}/f^{-}$.

Appendix B. Data

Data on popular elections to the US House of Representatives are taken from ICPSR Study # 7757. These are the same data used by Lee (2001, 2007), but I have engaged in neither the data augmentation nor the cleaning procedures he conducted. Data on roll call votes are taken from http://www.voteview.com/partycount.htm, a website maintained by Keith T. Poole of the University of California, San Diego. This same website is the basis for the data on DW-Nominate scores. Finally, data on party control of the House are taken from http://arts.bev.net/roperldavid/politics/congress.htm, a website maintained by L. David Roper of the Virginia Polytechnic Institute and State University.

All data and programs are available from the author for a period of 3 years from the date of publication.

References

Aigner, D.J., Amemiya, T., Poirier, D.J., 1976. On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function. International Economic Review 17 (2), 377–396.

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91 (434), 444–455.

Becker, G.S., 1957. The Economics of Discrimination. University of Chicago Press, Chicago.

Bouezmarni, T., Scaillet, O., 2005. Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data. Econometric Theory 21 (2), 390–412.

Card, D.E., 1999. The causal effect of education on earnings. In: Ashenfelter, O., Card, D.E. (Eds.), The Handbook of Labor Economics, vol. 3A. Elsevier, Amsterdam.

Cheng, M.-Y., 1994. On boundary effects of smooth curve estimators (dissertation). Unpublished manuscript Series # 2319, Institute for Statistics, University of North Carolina.

Cheng, M.-Y., 1997a. A bandwidth selector for local linear density estimators. Annals of Statistics 25 (3), 1001–1013.

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

Cheng, M.-Y., Fan, J., Marron, J.S., 1993. Minimax efficiency of local polynomial fit estimators at boundaries. Unpublished manuscript Series # 2098, Institute for Statistics, University of North Carolina.

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

Chernozhukov, V., Hong, H., 2004. Likelihood estimation and inference in a class of nonregular econometric models. Econometrica 72 (5), 1445–1480.

Chu, C., Cheng, P., 1996. Estimation of jump points and jump values of a density function. Statistica Sinica 6 (1), 79–96.

Cline, D.B., Hart, J.D., 1991. Kernel estimation of densities with discontinuities or discontinuous derivatives. Statistics 22 (1), 69–84.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press, New York.
Deaton, A., 1997. The Analysis of Household Surveys: A Microeconomic Approach to Development Policy. World Bank, Washington, DC.
DiNardo, J., Fortin, N., Lemieux, T., 1996. Labor market institutions and the distribution of wages, 1973–1992: a semi-parametric approach. Econometrica 64 (5), 1001–1044.
DiNardo, J.E., Lee, D.S., 2004. Economic impacts of new unionization on private sector employers: 1984–2001. Quarterly Journal of Economics 119 (4), 1383–1441.
Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, New York.
Gawronski, W., Stadtmüller, U., 1980. On density estimation by means of Poisson's distribution. Scandinavian Journal of Statistics 7 (2), 90–94.
Gawronski, W., Stadtmüller, U., 1981. Smoothing histograms by means of lattice- and continuous distributions. Metrika 28 (3), 155–164.
Greenberg, D., 2000. Was nixon robbed? The Legend of the Stolen 1960 Presidential Election. Slate.
Hahn, J., Todd, P., van der Klaauw, W., 1999. Identification and estimation of treatment effects with a regression discontinuity design. NBER Working Paper # 7131.
Hahn, J., Todd, P., van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression discontinuity design. Econometrica 69 (1), 201–209.
Hall, P., 1992. Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. The Annals of Statistics 20 (2), 675–694.
Hall, P., McKay, I., Turlach, B.A., 1996. Performance of wavelet methods for functions with many discontinuities. Annals of Statistics 24 (6), 2462–2476.
Härdle, W., Linton, O., 1994. Applied nonparametric methods. In: Engle, R.F., McFadden, D.L. (Eds.), The Handbook of Econometrics, vol. 4. Elsevier, New York, pp. 2297–2341.
Heckman, J.J., 2005. The scientific model of causality. Sociological Methodology 35 (1), 1–98.
Heckman, J.J., Urzua, S., Vytlacil, E., 2006. Understanding instrumental variables in models with essential heterogeneity. Review of Economics and Statistics 88 (3), 389–432.
Horowitz, J.L., 2001. The bootstrap. In: Heckman, J.J., Leamer, E. (Eds.), The Handbook of Econometrics, vol. 5. Elsevier, New York, pp. 3463–3568.
Imbens, G.W., Angrist, J.D., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475.
Jacob, B.A., Lefgren, L., 2004. Remedial education and student achievement a regression-discontinuity analysis. Review of Economics and Statistics 86 (1), 226–244.
Jones, M.C., 1989. Discretized and interpolated kernel density estimates. Journal of the American Statistical Association 84 (407), 733–741.
Lee, D.S., 2001. The electoral advantage to incumbency and voters' valuation of politicians' experience a regression discontinuity analysis of elections to the U.S. House. NBER Working Paper # 8441.
Lee, D.S., 2007. Randomized experiments from non-random selection in U.S. House Elections. Journal of Econometrics.
Marron, J.S., Ruppert, D., 1994. Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society, Series B 56 (4), 653–671.
Pagan, A., Ullah, A., 1999. Nonparametric Econometrics. Cambridge University Press, New York.
Rao, C.R., 1965. Linear Statistical Inference and its Applications, first ed. Wiley, New York.
Rice, J., 1984. Boundary modification for kernel regression. Communications in Statistics, A 13 (7), 893–900.
Riker, W.H., 1962. The Theory of Political Coalitions. Yale University Press, New Haven.
Rubin, D.B., 1980. Randomization analysis of experimental data: the fisher randomization test: comment. Journal of the American Statistical Association 75 (371), 591–593.
Rubin, D.B., 1986. Statistics and causal inference: which ifs have causal answers. Journal of the American Statistical Association 81 (396), 961–962.
Saez, E., 1999. Do taxpayers bunch at kink points? NBER Working Paper # 7366.
Saez, E., 2002. Do taxpayers bunch at kink points? Unpublished manuscript, University of California, Berkeley.
Schuster, E.F., 1985. Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics, A 14 (5), 1123–1136.
Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Series B 53 (3), 683–690.
Snyder, J., 2005. Detecting manipulation in U.S. House elections. Unpublished manuscript, Haas School of Business, University of California, Berkeley.
Stone, M., 1974. Cross-validation and multinomial prediction. Biometrika 61 (3), 509–515.
Stone, M., 1977. Asymptotics for and against cross-validation. Biometrika 64 (1), 29–35.
van der Klaauw, W., 2002. Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach. International Economic Review 43 (4), 1249–1287.
White, H., 2001. Asymptotic Theory for Econometricians. Academic Press, San Diego.