ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Selection On Unobservables Designs:
Part 7, Regression Discontinuity Designs[1]

Regression Discontinuity (RD) designs go back in the evaluation literature at least as far as Thistlethwaite and Campbell (1960). Only in the past decade, however, has RD become popular in economics. Most RD designs are basically special cases of IV. Nevertheless, they are probably my favorite selection on unobservables design because the identification (by which I mean the source of variation in the treatment that we are using to identify the treatment effect) is so transparent.

1 Introduction

1.1 Background

Suppose that we want to estimate the effect of some binary treatment D_i on an outcome Y_i. Using the potential outcomes framework, we write $Y_i(0)$ as the potential untreated outcome and $Y_i(1)$ as the potential treated outcome; $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$. Now suppose that the value of D_i – i.e., whether or not an individual gets treated – is completely or partially determined by whether some predictor X_i lies above or below a certain threshold, c. The predictor X_i need not be randomly assigned. In fact, we assume that it is related to the potential outcomes $Y_i(0)$ and $Y_i(1)$, but that this relationship is smooth, i.e., $Y_i(0)$ and $Y_i(1)$ do not jump discontinuously as X_i changes. Any discontinuous change in Y_i as X_i crosses c will thus be interpreted as a causal effect of the treatment D_i. We call X_i the "running variable." or forcing variable

RD designs often arise in administrative situations in which units are assigned a program,

[1]These notes are heavily derived from Imbens and Lemieux (2008).

treatment, or award based upon a numerical index being above or below a certain threshold. For example, a politician may be elected if and only if the differential between the vote share that she receives and the vote share that her opponent receives exceeds 0, a student may be assigned to summer school if and only if his performance on a combination of tests falls below a certain threshold, or a toxic waste site may receive cleanup funds if and only if its hazard rating falls above a certain level. In these cases, individuals or units whose indices X lie directly below the threshold c are considered to be comparable to individuals or units whose indices X lie directly above the threshold c, and we can estimate the treatment effect by taking a difference in mean outcomes for units directly above the threshold and units directly below the threshold.

1.2 The Sharp RD Design

There are two types of RD designs: the sharp design and the fuzzy design. In the sharp RD design (SRD), the probability that $D = 1$ changes from zero to one as running variable crosses c. In other words, no one with $X < c$ gets treated, and everyone with $X \geq c$ gets treated. In the fuzzy RD design, the probability of treatment jumps discontinuously as X crosses c, but it does not jump by 100 percentage points. In other words, either some people with $X < c$ get treated, or some people with $X \geq c$ do not get treated, or (most likely) both. We will focus first on the sharp RD design.

In the sharp RD design, D_i is a deterministic function of X_i: $D_i = \mathbf{1}(X_i \geq c)$.

To estimate the causal effect of D_i on some outcome Y_i, we simply take the difference in mean outcomes on either side of c. Formally, we estimate:

$$\lim_{x \downarrow c} E[Y_i|X_i = x] - \lim_{x \uparrow c} E[Y_i|X_i = x] = \lim_{x \downarrow c} E[Y_i(1)|X_i = x] - \lim_{x \uparrow c} E[Y_i(0)|X_i = x]$$

This represents the average causal effect of D on Y for individuals with $X_i = c$. We will call this effect τ_{SRD}.

$$\tau_{SRD} = E[Y_i(1) - Y_i(0)|X_i = c]$$

To justify this interpretation, we need it to be true that $Y_i(0)$ and $Y_i(1)$ are smooth functions of X_i as X_i crosses c.[2] We make this assumption in the form of a conditional expectation.

Assumption 1

$$E[Y_i(0)|X_i = x] \text{ and } E[Y_i(1)|X_i = x] \text{ are continuous in } x$$

With this assumption we can write

$$\tau_{SRD} = \lim_{x \downarrow c} E[Y_i|X_i = x] - \lim_{x \uparrow c} E[Y_i|X_i = x]$$

and estimate τ_{SRD} as the difference between two regression functions estimated in the neighborhood of c.

Since we never observe $Y_i(0)$ for units with $X_i = c$, we rely upon extrapolating $E[Y_i(0)|X_i = c]$ using units with X_i arbitrarily close to c. The continuity assumption above guarantees that the bias from this extrapolation becomes negligible as we get arbitrarily close to c.

Lee (2008) is an example of the sharp RD design in practice. Lee explores the effects of incumbency using the fact that a politician is elected if and only if he receives more votes than his opponent.[3] Lee uses this fact to examine Congressional districts in which the Democrats won by a few votes in period t to districts in which the Democrats lost by a few votes in period t. He compares party success in these districts in elections held in period $t+1$ to estimate the effect of incumbency on the probability of winning. He finds that party success in period t strongly affects party success in period $t+1$ – the probability of success in $t+1$ rises by approximately 50 percentage points.

[2]Technically, it should suffice to simply have $Y_i(0)$ be a smooth function of X_i as X_i crosses c. In that case we would estimate the average effect of the treatment on the treated at $X_i = c$. But it's hard to imagine an RD scenario in which $Y_i(0)$ is smooth in the running variable at c while $Y_i(1)$ is not smooth in the running variable at c.

[3]This may not truly be a sharp RD design if it includes data from Florida, particularly around 2000.

1.3 The Fuzzy RD Design

The fuzzy RD design (FRD) is similar in concept to the sharp RD design except that D_i is no longer a deterministic function of X_i. Instead, the probability of treatment changes by some nonzero amount as the running variable crosses the threshold c, but this change in probability is less than 100 percentage points. Formally, we write

$$0 < \lim_{x \downarrow c} P(D_i = 1 | X_i = x) - \lim_{x \uparrow c} P(D_i = 1 | X_i = x) < 1$$

This scenario is arguably more common than the sharp RD scenario in that most things in real life are determined by multiple factors, and the influence of the running variable as it crosses the threshold c may be just one of those factors. In the fuzzy RD design there are now two causal effects to be estimated: the effect of crossing the threshold on the probability of treatment and the effect of crossing the threshold on the outcome (in the sharp RD design, the former is known to be 1). Formally, the fuzzy RD estimand is

$$\tau_{FRD} = \frac{\lim_{x \downarrow c} E[Y_i | X_i = x] - \lim_{x \uparrow c} E[Y_i | X_i = x]}{\lim_{x \downarrow c} E[D_i | X_i = x] - \lim_{x \uparrow c} E[D_i | X_i = x]}$$

If this estimator looks somewhat familiar, that's because it should. It's the direct analog of an IV estimator in which the instrument is an indicator for whether X_i lies directly above c. Formally, let $D_i(x^*)$ be the potential treatment status of unit i for a threshold x^* in the neighborhood of c. Note that x^* now represents a potential value for the threshold, not the value of the running variable for unit i. We do this because it is often easier to conceive of manipulating the threshold rather than the running variable.[4] $D_i(x^*)$ is unity if unit i would take the treatment if the threshold were x^*, and zero otherwise. Thus we are considering manipulating the threshold; for example, if individuals are eligible for for free health insurance at age 65, one might imagine changing the threshold for eligibility from 65 to 65.1. When the threshold changes, some people who were previously eligible would now be ineligible. In this context, we need the equivalent of the IV monotonicity assumption:

[4] For example, the running variable may be age.

Assumption 2

$D_i(x^*)$ is non-increasing in x^* at $x^* = c$.

In other words, moving a unit with $X_i = c$ from "intended to treat" to "intended to not treat" by increasing the threshold x^* never results in that unit switching from an untreated status to a treated status (units only "drop out" of treatment or don't change their status as you ratchet up the threshold – they never "drop in" to treatment as you decrease the pool of intended to treat by increasing the threshold). The monotonicity assumption rules out the possibility of defiers, leaving only always-takers, never-takers, and compliers. We define these groups in a manner similar to that of AIR (1996); a complier is a unit such that

$$\lim_{x^* \downarrow X_i} D_i(x^*) = 0 \quad \text{and} \quad \lim_{x^* \uparrow X_i} D_i(x^*) = 1.$$

In other words, a complier is a unit that does not take the treatment when the threshold lies just above X_i and takes the treatment when the threshold lies just below X_i. Analogously, a never-taker is a unit that never takes the treatment when the the threshold is in the neighborhood of X_i (regardless of whether the threshold lies just above or below X_i), and an always-taker is a unit that always takes the treatment when the the threshold is in the neighborhood of X_i (regardless of whether the threshold lies just above or below X_i). Intuitively, the fuzzy RD design measures the average treatment effect for RD compliers:

$$\begin{aligned} \tau_{FRD} &= \frac{\lim_{x \downarrow c} E[Y_i | X_i = x] - \lim_{x \uparrow c} E[Y_i | X_i = x]}{\lim_{x \downarrow c} E[D_i | X_i = x] - \lim_{x \uparrow c} E[D_i | X_i = x]} \\ &= E[Y_i(1) - Y_i(0) | \text{unit } i \text{ is a complier and } X_i = c]. \end{aligned}$$

The logic is the same as in the IV case – the outcomes and treatment statuses for the always-takers and never-takers do not change as the running variable crosses the threshold, so they contribute nothing to either the numerator or the denominator of the fuzzy RD estimator. The only units that have non-zero contribution are the compliers.

DiNardo and Lee (2004) test whether unionization has a direct effect on wages using a fuzzy RD design. They leverage the fact that, in the United States, employees vote on whether they want to unionize (this election is not automatically held – typically organizers must first collect cards from a majority of employees requesting an election from the National Labor Relations Board). If a majority of employees vote in favor of unionizing, then the employer is legally required to recognize the union and bargain in good faith.[5] One might think that crossing the threshold of 50% in the election would generate a sharp RD design, but in fact it does not. First, in a few cases a union is not recognized in spite of winning a majority of the vote because the NLRB invalidates the result. More importantly, in many close losses, union organizers try again and ultimately win a subsequent election. Thus, crossing the threshold of 50% in an initial election changes the probability of union recognition, but it does not change it from zero to one. Instead, the data suggest that the probability of eventual and lasting union recognition rises by approximately 80 percentage points. Interestingly, however, there is no discernible effect on employment, firm survival, or wages as the vote share crosses the 50% threshold. Whether this is due to threat effects (which should be particularly pronounced for compliers), union ineffectiveness, or some other mechanism is not entirely clear.

1.4 The FRD, Matching, and Unconfoundedness

The RD context makes an analysis based on unconfoundedness, i.e. $Y_i(0), Y_i(1) \perp D_i | X_i$, seem attractive. In particular, it is intuitively appealing to match units with X_i close to c and then compare difference in mean outcomes for the treated units in this group and the untreated units in this group – the link to propensity score matching is clear. In the SRD design, this analysis makes sense (in fact, it should reproduce the SRD estimator), but in the FRD design, it does not. That is because, in the FRD design, treated units in the neighborhood of c consist of a mixture of compliers and always-takers, while untreated units in the neighborhood of c consist of a mixture of compliers and never-takers. Thus

[5] In practice, many employers engage in illegal tactics to attempt to forestall elections, including firing union activists.

the treated and untreated units are not, on average, directly comparable. This is the same reason that we do not directly compare treated and untreated units in the context of IV (e.g., in a medical trial) – instead we compare those that we intended to treat to those that we do not intend to treat, and rescale the coefficient by our estimate of the proportion of compliers in the sample.

1.5 External Validity

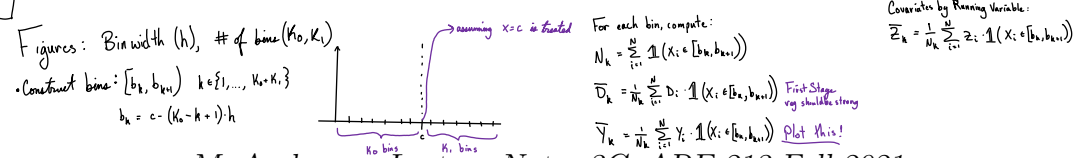
RD estimates are inherently localized. In the SRD design, the effects are estimated for a subpopulation with X_i in the neighborhood of c. In the FRD design, the subpopulation is further restricted – it consists only of compliers in the neighborhood of c. Although the external validity is limited, it is (hopefully) counterbalanced by a relatively high degree of internal validity. Nevertheless, it is important to note and think about the external validity of any RD estimates.

2 Graphical Analysis

2.1 Introduction

A graphical analysis should be *the focus* of any RD paper. The strength of the RD design revolves around the fact that the treatment assignment rule is known (or at least partially known) and that we should be able to see discontinuous changes in the treatment and the outcome (if there is an effect) as the running variable crosses c. Any RD design that fails to exhibit a visually perceptible break in treatment probability at the discontinuity threshold is basically not credible, regardless of the regression results. Conversely, any break that is visually perceptible will almost surely be statistically significant. So with RD papers, the statistical results really take a back seat to the graphical analysis.

There are three types of graphs in RD analyses, though not all analyses will necessarily include all three types. The first type plots outcomes by the running variable, where out-



comes can include both Y_i and D_i. The second type plots covariates by the running variable, and the third type plots the density of the running variable.

2.2 Outcomes by the Running Variable

The first type of graph is basically a histogram-type plot that presents the average value of an outcome at evenly spaced values of the running variable – this is equivalent to running a kernel regression at each of those points using a uniform kernel. Formally, there are two parameters to choose: the binwidth, h, and the number of bins to the left and right of the threshold value, K_0 and K_1.[6] Given these parameters, construct bins $(b_k, b_{k+1}]$ for $k = 1, ..., K = K_0 + K_1$, where

$$b_k = c - (K_0 - k + 1) \cdot h$$

This simply creates K_0 evenly spaced bins of width h below the threshold value (c), and K_1 evenly spaced bins of width h above the threshold value. For each bin, calculate the number of observations lying in that bin:

$$N_k = \sum_{i=1}^{N} \mathbf{1}(b_k < X_i \leq b_{k+1})$$

Then, calculate the average treatment level in the bin (if you have a fuzzy design):

$$\overline{D}_k = \frac{1}{N_k} \sum_{i=1}^{N} D_i \cdot \mathbf{1}(b_k < X_i \leq b_{k+1})$$

Finally, calculate the average outcome in the bin:

$$\overline{Y}_k = \frac{1}{N_k} \sum_{i=1}^{N} Y_i \cdot \mathbf{1}(b_k < X_i \leq b_{k+1})$$

[6] Alternatively, one could choose the binwidth h and the support of the histogram – the two methods are equivalent.

The first plot of interest, particularly in the fuzzy RD design, that of $\overline{D}_k$ against the midpoint of each of the bins $k = 1, ..., K$. The question is whether there is a visual discontinuity in the plot of $\overline{D}_k$ at the threshold c. A visual break implies that crossing the threshold has a significant effect on the probability of treatment – this graph is equivalent to the first stage in an IV analysis. Again, if there is no visual break, then it is unlikely that the statistical analysis will find anything either (and even if it does, it won't be very credible).

The second plot of interest is that of $\overline{Y}_k$ against the midpoint of each of the bins $k = 1, ..., K$. The focus in this plot is on whether there is a visual discontinuity in the outcome as the running variable crosses the threshold c. A visual break implies that crossing the threshold has a significant effect on the outcome, which in turn implies (under our assumptions) that the treatment has a significant effect on the outcome. This graph is equivalent to the reduced form in an IV analysis. In addition to inspecting the threshold for a discontinuity, you should also inspect whether there are any other discontinuities of similar (or greater) magnitude in $\overline{Y}_k$ at other values of the running variable. If there are, and if there is not a clear a priori reason to expect these discontinuities, then the research design is called into question – effectively, we have detected a violation of Assumption 1 (smoothness in expected potential outcomes). Finally, note that it is important not to smooth over the threshold value c – i.e., no bin should cross c. Smoothing over c will tend to minimize any discontinuity at the threshold.

2.3 Covariates by the Running Variable

The second type of graph plots the average values of covariates against the running variable using the same methodology as above. Suppose that we have a covariate Z_i that is related to the outcome but should not be affected by the treatment. Plotting Z_i against the running variable will allow us to determine whether it is balanced across the threshold – this is the equivalent of showing covariate balance after matching on the propensity score

or demonstrating that covariates are balanced across an instrument. Formally, we calculate

$$\overline{Z}_k = \frac{1}{N_k} \sum_{i=1}^{N} Z_i \cdot \mathbf{1}(b_k < X_i \leq b_{k+1})$$

and then plot $\overline{Z}_k$ against the midpoint of each of the bins $k = 1, ..., K$. If the research design is valid, then there should not be any discontinuity in $\overline{Z}_k$ as the running variable crosses the threshold c.

Lee (2008) generates plots of outcomes and covariates by the running variable in the context of Congressional elections. As discussed earlier, he is interested in whether incumbency provides a reelection advantage to the party in power. The first graph, Figure 1, plots $\overline{Y}_k$ against the running variable, with the discontinuity at $c = 0$ (you win if and only if you get more votes than your opponent). This plot demonstrates that incumbency provides an enormous advantage to the party in power – there is a large, discontinuous break in the probability of a Democrat winning the next election if a Democrat just barely won the last election.[7] The second graph, Figure 2, plots $\overline{Z}_k$ against the running variable, with the threshold at $c = 0$. In this case, the covariate Z represents the number of previous election victories of a candidate. If the RD design is valid, then there should be no discontinuity at the threshold (i.e., candidates that just barely won an election should not look any different, in terms of past success, than candidates that just barely lost an election). Lee's figure shows that there is no discontinuity in the covariate, increasing the credibility of the design.

2.4 Density of the Running Variable

The third type of graph arises from a specification test suggested in McCrary (2008). A primary concern in RD designs is that individuals may be able to "game" the assignment rule. That is to say, if individuals understand the assignment mechanism and can manipulate their value of the running variable, then they may be able to place themselves just above (or below)

[7] Because Lee's RD design is a sharp RD design, there is little value in plotting $\overline{D}_k$ against the running variable. We already know what it will look like – it will just change from 0 to 1 as we cross c. If this were a fuzzy RD design, however, the plot of $\overline{D}_k$ would be just as important as the plot of $\overline{Y}_k$.

Figure 1: Candidate's Probability of Winning Election t+1, by Margin of Victory in Election t: local averages and parametric fit. Source: Lee (2008)

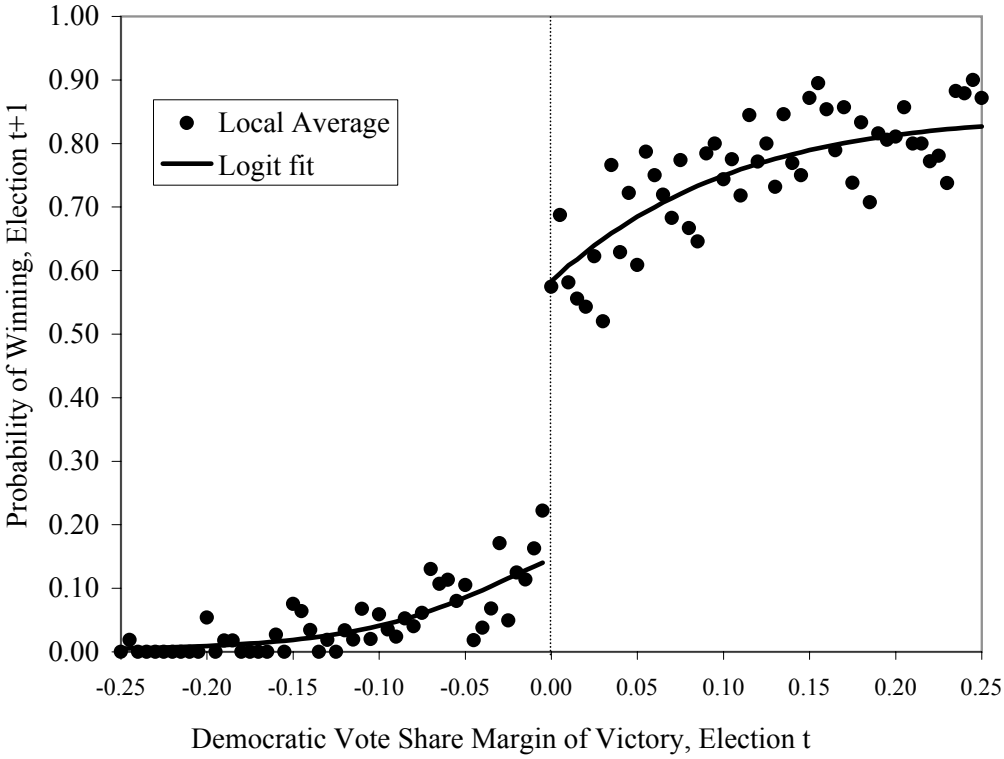
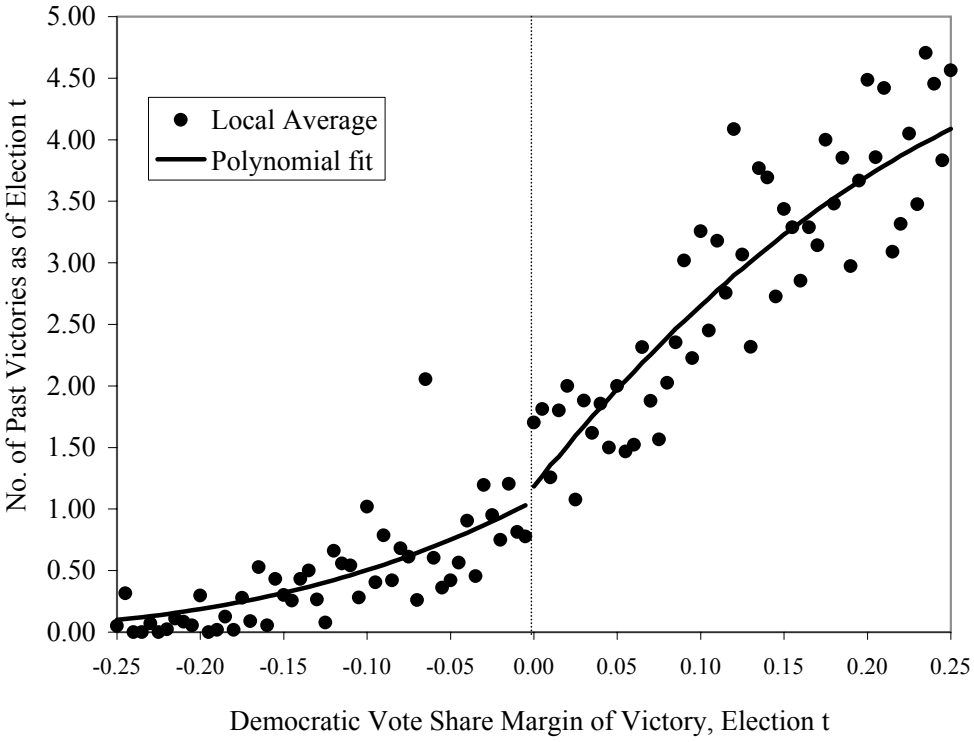


Figure 2: Candidate's Accumulated Number of Past Election Victories, by Margin of Victory in Election t: local averages and parametric fit. Source: Lee (2008)



the threshold c. In that case, the individuals just above the threshold will disproportionately consist of those gaming the rule, and they will not be directly comparable to the individuals lying just below the threshold. For example, consider a welfare program that activates only when income falls below a threshold c. Shrewd families with income in the neighborhood of c will stop working right before income crosses c, so the observations right below c will disproportionately consist of families of this type. This type of family will not be balanced across the discontinuity threshold. Another example would arise if individuals are assigned on the basis of test scores but can re-take the test as many times as necessary. If the researcher uses an individual's maximum test score as the running variable, motivated individuals who re-take the test many times will be more likely to fall right above the discontinuity threshold than right below it.[8]

To address this issue, McCrary suggests a specification test examining the density of the running variable as it crosses c. In practice, this graph can be generated using the same methodology as above, but instead of graphing $\overline{Y}_k$ or $\overline{Z}_k$, just plot the number of observations falling in each bin, i.e. $N_k = \sum_i \mathbf{1}(b_k < X_i \leq b_{k+1})$. If units are manipulating their values of the running variable to fall just above or below c, then we should observe a discontinuity in the distribution of the running variable as it crosses c. If the distribution of the running variable is smooth as it crosses c, then it's unlikely that individuals are gaming the assignment mechanism.

3 Estimation

3.1 Local Linear Regression and the Sharp RD

We now focus on estimating (rather than graphing) the treatment effect in RD designs. Recall from our earlier lectures that kernel regression can be conceptualized as a "local constant estimator" – for a given X^*, it estimates the mean of Y in the neighborhood of

[8] A solution in this case would be to use the individual's first test score as the running variable, assuming these data are available to the researcher.

X^* (hence Y is assumed to be constant in expectation in the neighborhood of X). We also discussed Lowess regression as an example of a "local regression estimator." We use a simpler form of local linear regression to estimate the change in Y_i as the running variable crosses c.

We first choose the bandwidth, h, that will determine the regression sample on either side of the threshold point c. We then fit a linear regression on either side of the threshold point for the samples with $X_i \in (c-h, c)$ and $X_i \in [c, c+h)$. Formally, we compute:

$$\min \alpha_l, \beta_l \sum_{i \ni c-h<X_i<c}^{N} (Y_i - \alpha_l - \beta_l \cdot (X_i - c))^2 \quad \text{and}$$

$$\min \alpha_r, \beta_r \sum_{i \ni c\leq X_i<c+h}^{N} (Y_i - \alpha_r - \beta_r \cdot (X_i - c))^2$$

$\lim_{x \uparrow c} E[Y_i|X_i = x]$ and $\lim_{x \downarrow c} E[Y_i|X_i = x]$ are then estimated as:

$$\hat{\mu}_l(c) = \hat{\alpha}_l + \hat{\beta}_l \cdot (c - c) = \hat{\alpha}_l \quad \text{and} \quad \hat{\mu}_r(c) = \hat{\alpha}_r + \hat{\beta}_r \cdot (c - c) = \hat{\alpha}_r$$

Finally, estimate $\hat{\tau}_{SRD} = \hat{\alpha}_r - \hat{\alpha}_l$.

In practice, it is often easier to implement this estimator using a single regression. Specifically, in the SRD design, run the following regression:

$$Y_i = \alpha + \tau D_i + \beta(X_i - c) + \gamma(X_i - c) \cdot D_i + u_i \text{ for the sample with } c - h < X_i < c + h$$

The coefficient $\hat{\tau}$ will be numerically identical to $\hat{\tau}_{SRD}$ above. The advantage to the single regression, besides being simpler, is that we can use the least squares (robust) standard errors for statistical inference. As with kernel regression, the two factors that the researcher must choose are the kernel function and the bandwidth, h. The kernel choice we implicitly ignored by choosing the uniform kernel, and we know that kernel choice is not too important anyway – the important choice is bandwidth. Imbens suggests a bandwidth proportional to $N^{-\delta}$,

where $1/5 < \delta < 2/5$, but as always it's best to check that the results are not sensitive to doubling or halving the bandwidth.

Covariates could also be added to the regression above to improve precision. Unlike with the running variable, it is not necessary to interact the covariates with the treatment indicator (which is effectively an indicator for whether the unit lies above or below c, given that this is the SRD design).

3.2 Estimation in the Fuzzy RD

In the FRD design, we have two effects to estimate: the effect of crossing the threshold on the treatment (the "first stage") and the effect of crossing the threshold on the outcome (the "reduced form"). We again use local linear regression. As you might expect, we apply the same methodology as in Section 3.1 to estimate the effect of crossing the threshold on Y_i and the effect of crossing the threshold on D_i. Specifically, we run the regressions:

$$Y_i = \pi_0 + \pi_1 Z_i + \pi_2(X_i - c) + \pi_3(X_i - c) \cdot Z_i + u_i \text{ for the sample with } c - h < X_i < c + h$$

and

$$D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2(X_i - c) + \gamma_3(X_i - c) \cdot Z_i + v_i \text{ for the sample with } c - h < X_i < c + h$$

where $Z_i = \mathbf{1}(X_i \geq c)$. In other words, we regress Y_i and D_i on an indicator for whether an observation falls above or below the discontinuity threshold (and also controlling for the running variable and an interaction of the running variable and the above/below indicator) using the sample with $c - h < X_i < c + h$. The fuzzy RD estimator is:

$$\hat{\tau}_{FRD} = \frac{\hat{\pi}_1}{\hat{\gamma}_1}$$

In other words, the FRD estimator is simply the ratio of the reduced form and first stage estimates, i.e. the effect of crossing the discontinuity threshold on the outcome divided by

the effect of crossing the discontinuity threshold on the treatment. Again, it is trivial to add covariates to the regressions above, and the formula $\hat{\tau}_{FRD} = \hat{\pi}_1/\hat{\gamma}_1$ will still apply.

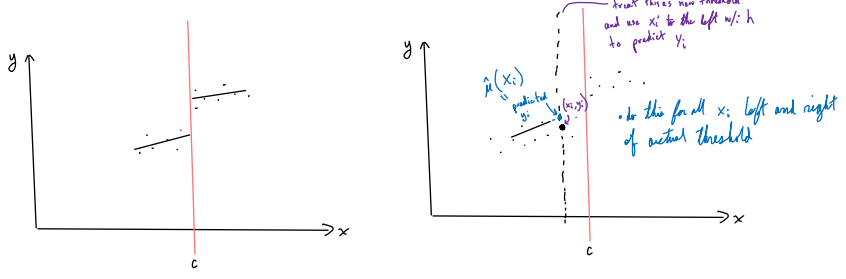
Given the discussion above, it should be obvious that we can estimate $\hat{\tau}_{FRD}$ using a TSLS regression in which Y_i is the outcome, D_i is the treatment, Z_i is the instrument, and $(X_i - c)$ and $(X_i - c) \cdot Z_i$ are covariates (this regression would of course be limited to the sample with $c - h < X_i < c + h$). The advantage of this approach, besides ease of implementation, is that we can use the 2SLS (robust) standard errors for statistical inference.

3.3 Optimal Bandwidth

As with kernel density estimation, choosing the bandwidth (h) is more an art than a science. If you really love doing math, check out the section in Imbens and Lemieux (2008) on bandwidth selection. They discuss a cross-validation method in the context of the SRD. Recall that the optimal bandwidth is generally a function of the regression function that we are trying to estimate (which itself depends on the bandwidth). The cross-validation method minimizes the mean squared error between Y_i and predicted Y_i with respect to h. The setup here is actually simpler than it was with kernel density estimators in that we only care about minimizing the MSE near a single point (c) rather than over the entire support of X. We do this by minimizing

$$CV_Y^{\delta}(h) = \frac{1}{N} \sum_{i \ni q_{X,\delta,l} \leq X_i \leq q_{X,1-\delta,r}} (Y_i - \hat{\mu}(X_i))^2$$

where $q_{X,\delta,l}$ and $q_{X,1-\delta,r}$ are the δ and $1 - \delta$ quantiles (respectively) of the empirical distributions of the samples with $X_i < c$ and $X_i \geq c$ (respectively). We limit the sample in this manner because we're really interested in the optimal bandwidth in some region around c. Literally what the CV procedure does is finds the bandwidth h that minimizes the mean squared error (i.e., the difference between the actual Y_i and the predicted Y_i) in the trimmed data set.



In the CV criterion formula above, $\hat{\mu}(X_i)$ is the intercept from a local linear regression using observations falling in $(X_i - h, X_i)$ if $X_i < c$ or observations falling in $(X_i, X_i + h)$ if $X_i \geq c$. These regressions are estimated as described in the beginning of Section 3.1, but now X_i replaces c. Why do we proceed in this manner? Recall that the goal of our left side local linear regression is to estimate the conditional expectation of $Y_i(0)$ at $X_i = c$. Because $E[Y_i|X_i]$ jumps discontinuously at $X_i = c$ (and because there are virtually no observations with $X_i = c$), we cannot test our regression's performance by comparing its prediction to actual observations of Y_i when $X_i = c$. Instead, in our data we observe pairs (Y_i, X_i) at many values of X_i other than c. For each observation we thus pretend that X_i is the threshold value c, estimate a local linear regression at X_i using other nearby observations, and measure how well the local linear regression does at predicting the observed value of Y_i for the observation (Y_i, X_i).[9]

Specifically, for each point X_i that is to the left of c (but right of $q_{X,\delta,l}$), we run the regression:

$$Y_j = \alpha_l(X_i) + \beta_l(X_i) \cdot (X_j - X_i) + u_j \quad \text{for the sample with } X_i - h < X_j < X_i$$

We limit the sample to the left side of X_i to mimic the local linear regression that we estimate on the left side of c.[10] Note that the parameters $\alpha_l(X_i)$ and $\beta_l(X_i)$ are functions of X_i because we are running a different local linear regression for each point X_i in the left subsample. We set $\hat{\mu}(X_i) = \hat{\alpha}_l(X_i)$ for each point X_i in the left subsample.[11]

For each point X_i that is to the right of c (but left of $q_{X,1-\delta,r}$), we run the regression:

$$Y_j = \alpha_r(X_i) + \beta_r(X_i) \cdot (X_j - X_i) + u_j \quad \text{for the sample with } X_i < X_j < X_i + h$$

[9] Stepping back to see the forest rather than the trees, our ultimate goal is to choose the optimal bandwidth h^* that produces the set of local linear regressions which best predict the observed values of Y_i in the trimmed data set.

[10] In principle we could estimate separate bandwidths for either side of c, but in practice we use the same bandwidth on both sides of c.

[11] $\hat{\mu}(X_i)$ is not a function of $\hat{\beta}_l(X_i)$ because the regressor $X_j - X_i$ approaches zero as X_j approaches the simulated threshold X_i. If we specified the regressor as X_j instead of $X_j - X_i$, then we would need to set $\mu(X_i) = \alpha_l(X_i) + \beta_l(X_i)X_i$ in order to estimate the conditional expectation of Y_i at X_i.

We limit the sample to the right side of X_i to mimic the local linear regression that we estimate on the right side of c. We set $\hat{\mu}(X_i) = \hat{\alpha}_r(X_i)$ for each point X_i in the right subsample.

The procedure is computationally intensive – for each potential value of h, you are running $(1-\delta) \cdot N$ local linear regressions. In practice, you might start with $\delta = 0.50$ and assess the sensitivity of h to using larger values of δ (e.g., 0.8, 0.9).

For the FRD design, one could in principle choose separate bandwidths for the reduced form and the first stage. In practice, however, we generally choose the same bandwidth for both regressions. One option is to use the optimal bandwidth for the reduced form for both regressions. Alternatively, we could choose the minimum of the two optimal bandwidths (the reduced form optimal bandwidth and the first stage optimal bandwidth).

Regardless of how you choose the bandwidth, it always a good idea to test the sensitivity of your results to choice of bandwidth by doubling/halving the bandwidth.

3.4 Alternative Estimators

An alternative to the SRD and FRD estimators above is to use all of the data when estimating the treatment effect but to control for the conditional expectation of the outcome as a function of the running variable. For example, in the SRD design, we could express the outcome as a function of the running variable and the treatment:

$$Y_i = \alpha + m(X_i) + \tau D_i + v_i$$

If D_i is a treatment indicator, then the coefficient $\hat{\tau}$ should estimate the average treatment effect for units at the discontinuity. In practice, however, we do not know the function $m(.)$. We therefore approximate $m(.)$ using a low order polynomial of the running variable (e.g., a cubic or quartic), fully interacted with the treatment indicator. We can then estimate the treatment effect by simply regressing Y_i on D_i, a polynomial of X_i, and that polynomial

interacted with D_i. The coefficient on D_i gives the estimate of the treatment effect. This is the type of estimator applied in Card and Lee (2008) (see Section 3.6).

In the FRD design, we can use the same methodology as above, but apply it to the 2SLS estimator instead of the OLS estimator. Specifically, replace D_i with $Z_i = \mathbf{1}(X_i \geq c)$, and run a 2SLS regression in which the endogenous regressor is D_i, the instrument is Z_i, and the covariates are a polynomial in X_i and the interaction between that polynomial and Z_i.

However, the general trend in the applied literature has been to use local linear regression (LLR) estimators over these "global polynomial" estimators. There appear to be sound theoretical reasons for this choice. Gelman and Imbens (2019) argue *against* using global polynomial estimators (over LLR estimators) on several grounds.

First, like any regression estimate, we may interpret the global polynomial estimator as a weighted difference in means between treated and untreated units, with the weights determined by the values of X_i and higher order terms of X_i. Inspection of the weights reveals that in some cases they are highly variable (with respect to the order of the polynomial) and place large amounts of weight on observations very far from the RD threshold. This seems undesirable. Second, choice of polynomial order for the global polynomial regression is analogous to the choice of bandwidth for the LLR. However, we do not have any good methods for choosing the polynomial order with global polynomial regressions. Finally, statistical inference based on global polynomial regressions is often poor, with confidence intervals that fail to include zero at a rate higher than the Type I error rate (5%) even when there is no discontinuity in the regression function.

3.5 Specification Tests

As in the graphical analysis, we can run several sets of specification tests. One type of specification check tests for discontinuities in covariates at the threshold, c. To perform this test, simply replace Y_i in the regressions above with the covariate of interest, Z_i. Another specification check, presented in McCrary (2008), tests whether there is a discontinuous

jump in the running variable density at the threshold. As argued above, such a jump would be evidence that individuals are manipulating their values of the running variable in order to select into/out of treatment. Implementing this test requires us to revisit kernel density estimation.

The first step in the test involves estimating a histogram using a similar methodology to that in Section 2.2. Note that the histogram bins are defined such that no bin crosses the threshold value, c. The goal here is to plot the frequency of observations, rather than the average outcome. For each bin we therefore plot $\hat{F}_k = N_k/(N \cdot b)$, where b is the binwidth and

$$N_k = \sum_{i=1}^{N} \mathbf{1}(b_k < X_i \leq b_{k+1}).$$

We are therefore plotting the number of observations that fall in each histogram bin, normalized by $N \cdot b$ (we normalize by b so that the binwidth does not impact the height of the histogram bars). We then estimate local linear regressions using the histogram bin midpoints as data – this is effectively smoothing the histogram. As in previous sections, we estimate this regression separately for the left and right sides of the threshold value – we do not want to smooth across the threshold point. Let $\{X_1^*, ..., X_{K_0}^*\}$ be the set of histogram bin midpoints that fall below c. At every point $x < c$, run the following weighted regression:

$$\hat{F}_k = \alpha_l(x) + \beta_l(x)(X_k^* - x) \text{ using weights } \sqrt{K((X_k^* - x)/h)}$$

where $K(t) = \max\{0, 1 - |t|\}$ is the triangle kernel and h is the bandwidth (keep in mind that the bandwidth h is no longer the same thing as the binwidth b!). Note that within each regression, x is a constant – the parameters are written as functions of x to reflect the fact that we are estimating a separate regression at each point $x < c$. The density estimate at point x is then $\hat{f}(x) = \hat{\alpha}_l(x)$. Repeat the same analysis using only the set of histogram bin midpoints that lie above c, $\{X_1^*, ..., X_{K_1}^*\}$. From this analysis, recover the density estimate at all points $x \geq c$, $\hat{f}(x) = \hat{\alpha}_r(x)$.

The test statistic will be $\hat{\theta} = \ln \hat{f}(x)^+ - \ln \hat{f}(x)^-$, where $\hat{f}(x)^+$ and $\hat{f}(x)^-$ are estimated just to the right and left of c respectively. We normalize this statistic by its standard error, $\hat{\sigma}_\theta = \sqrt{4.8 \cdot (\frac{1}{\hat{f}^+} + \frac{1}{\hat{f}^-})/(Nh)}$. We can then evaluate $\hat{\theta}/\hat{\sigma}_\theta$ using a standard t-distribution.

In order to implement this estimator, you need to choose the binwidth, b, and the bandwidth, h. McCrary's simulations indicate that the choice of b is not too important, but the choice of h can be important. His recommendation is to choose $b = 2\hat{\sigma}/\sqrt{N}$, where $\hat{\sigma}$ is the standard deviation of the running variable. For the bandwidth, his recommendation is to choose it via visual inspection, experimenting with different bandwidths. If you prefer to use a plug-in bandwidth formula, however, or if you need an initial value for the bandwidth, McCrary offers the following plug-in bandwidth estimator:

1. Using the first-step histogram, estimate a global 4th order polynomial (in the running variable) on each side of the threshold, c.

2. On each side of c, compute $3.348 \cdot [\hat{\sigma}^2(b-a)/\sum_k \hat{f}''(X_k)^2]^{0.2}$, where $\hat{\sigma}^2$ is the mean squared error of the regression, $b-a$ is $X_K - c$ on the right-side regression and $c - X_1$ on the left-side regression, and $\hat{f}''(X_k)$ is the estimated second derivative from the global polynomial model. Set the estimated bandwidth, $\hat{h}$, to be the average of two quantities that you have computed.

As always, no matter how you choose your bandwidth, you will want to check the sensitivity of your results to choice of bandwidth.

McCrary (2008) applies the test to Lee's Congressional election data and to House roll call votes. In the former example, we expect no discontinuity in the running variable density at c, and we find none, as evidenced in Figure 3 and Table 2. In the roll call votes, however, there are relatively few voters (only the members of the House of Representatives), and the votes are public knowledge. Close votes thus may involve intense lobbying of swing members, and votes may be more likely to barely pass than they are to barely fail. Indeed, that is what appears in Table 2 and Figure 4.

Figure 3: Democratic vote share relative to cutoff, popular elections to the House of Representatives, 1900-1990. Source: McCrary (2008)

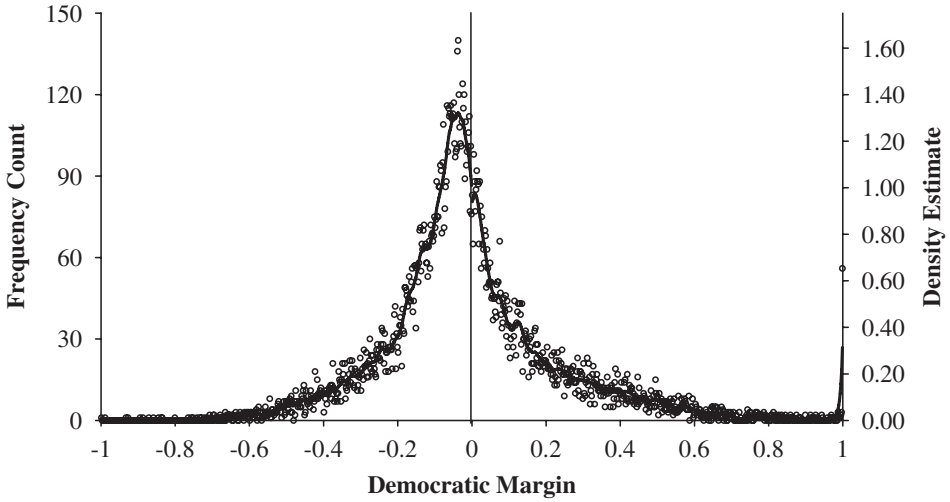
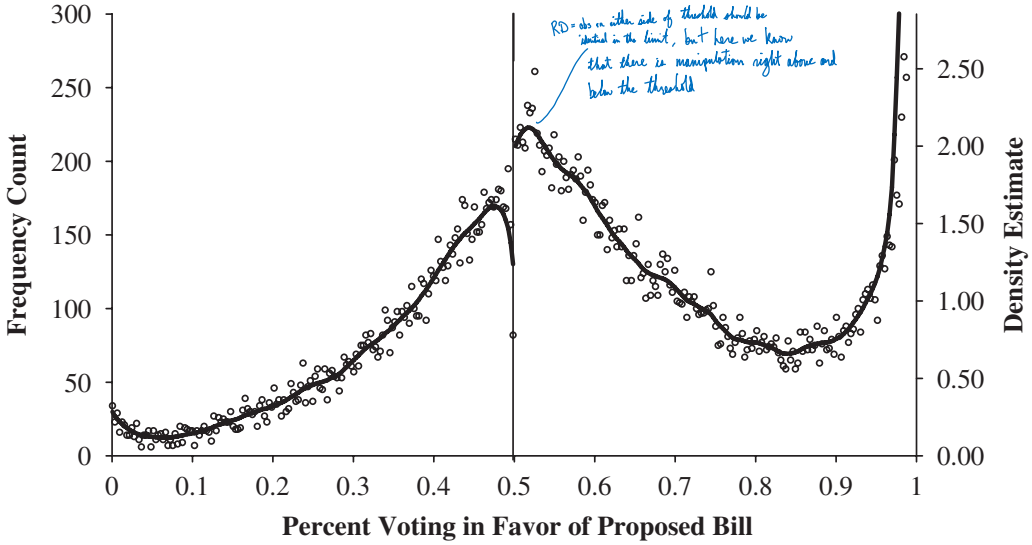


Table 2
Log discontinuity estimates

	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.

Figure 4: Percent voting yes, roll call votes, US House of Representatives, 1857-2004. Source: McCrary (2008)



3.6 Discrete Running Variables

In some cases, the running variable is discrete (technically, this is true in all cases, but in some cases the discreteness is non-trivial, e.g., if the running variable is age and you only have the data available in months or quarters). In these cases, it is impossible to estimate the conditional expectation function arbitrarily close to the discontinuity threshold, generating a form of model misspecification (Card and Lee 2008). This specification error can lead to group structure in the variance-covariance matrix — non-zero covariances between observations with the same value of the discretized running variable, X_i, may arise because there is a deviation between the true conditional expectation and the predicted conditional expectation (using the coarse running variable). In extreme (i.e., very discrete) cases this can create issues. Nevertheless, results in Kolesar and Rothe (2018) suggest that a discrete running variable does not create problems *per se* as long as there remain a sufficient number of points of support in the RD estimation sample (i.e., the sample on which you run the regression). In practice this means that you should be fine as long as you can tighten the bandwidth to a reasonable value without reducing the number of points of support too far (e.g., below 10).

3.7 Bias Correction

While local linear regression yields asymptotically consistent estimates of τ_{SRD} (or τ_{FRD}), in any finite sample there may be some residual bias. Calonico et al. (2014) suggest using a bias-corrected local regression estimator, but the bias correction term itself introduces additional variance that should be accounted for. In practice they find that a simple procedure in which you choose the RD bandwidth, h_1, using a local linear regression but then compute t-statistics using a local quadratic regression (with bandwidth h_1) appears to work well. They also have software available for alternative bias-correction procedures and associated standard errors.

4 Additional References

Calonico, S., M. Cattaneo, and R. Titiunik. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica*, 2014, 82, 2295-2326.

Card, David and David Lee. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics*, 2008, 142, 655-674.

Gelman, Andrew and Guido Imbens. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business and Economic Statistics*, 2019, 37, 447-456.

Imbens, Guido, and Karthik Kalyanaraman. "Optimal bandwidth choice for the regression discontinuity estimator." *The Review of Economic Studies*, 2012, 79, 933-959.

Kolesar, M. and C. Rothe. "Inference in Regression Discontinuity Designs with a Discrete Running Variable." *American Economic Review*, 2018, 108(8): 2277-2304.

Lee, D. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics*, 2008, 142, 675-697.

Serial Correlation:

BDM Model: $y_t = \alpha + \beta x_t + \varepsilon_t$ $\quad \varepsilon_{t+1} = \rho \varepsilon_t + u_{t+1}$ (AR1)

3 Takeaways

① If ε_t or x_t independent across t i.e. $E[\varepsilon_t \varepsilon_{t+1}] = 0$
⇒ no serial correlation

② Positive serial correlation in both ε_t & x_t
⇒ OLS SEs are too small

③ Degree of bias ↑ w/ T.

---

① Draw year from uniform