LECTURE 8. REGRESSION DISCONTINUITY

Dr. Tom O'Grady

Note: the diagrams in this lecture originally come from a lecture given by Teppei Yamamoto at MIT

Regression discontinuity (RD) analysis refers to situations where a rule arbitrarily divides groups into treatment and control without the intervention of the analyst. RD is therefore another example of a natural experiment. It has become enormously popular in recent years, and is often thought to have the highest internal validity of any non-experimental technique, but the lowest external validity. We'll consider why that is.

1. The Basic Idea

Regression discontinuities often emerge in a rule-based world with resource constraints, where access to some treatment is rationed by an eligibility criterion at a cutoff. Examples include:

1. **Educational Programs** are sometimes available only to students who exceed some threshold score. This is actually where RD analysis was invented (by Donald Campbell and Donald Thistelthwaite in 1960), who studied the impact of an American college scholarship program that was available only to students exceeding a threshold on their SAT scores (akin to A Levels).

2. **Voting** is legal for anyone aged 18 years and 0 days or more, but illegal for anyone aged 17 years and 364 days or less

3. **New-born babies** born in the US just below 1500 grams are put into specialist neo-natal care but babies just above it are not

4. **Elections** are won (in a two-candidate contest) by any candidate obtaining 50.0000001% of the vote and lost by any candidate obtaining 49.999999% of the vote

5. **Administrative borders** divide countries or provinces, separating populations into two regimes. Anyone living one meter inside the boundary for country A is a citizen of country A, and anyone one meter inside the boundary for country B is a citizen of country B.

In each case the rule is chosen non-randomly, so we cannot simply study all units that fall on each side of the threshold. They are likely to have very different potential outcomes. Instead, we can study *only those units that fall just below and just above the threshold*. These units *are* likely to have equivalent potential outcomes. People aged 12 and people aged 24 are on average very different to each other, but people aged 17 and 364 days and aged 18 and 0 days are not. Likewise, candidates who obtain 30% of the vote and candidates who obtain 70% are

very different to each other, but candidates who obtain 49.9% and 50.1% are not. Winning a constituency election by just a handful of votes is just like winning thanks to a coin toss. The election could have gone the other way if just a tiny handful of voters had changed their minds. This randomness opens up the possibility of exploring difficult-to-answer questions:

1. **What is the impact of education on earnings?** A comparison of earnings for the educated and uneducated would be subject to selection bias, but those who narrowly win a place at university should have the same average potential outcomes as those who narrowly lose

2. **Do politicians use their office to enrich themselves?** A comparison of earnings or wealth for former politicians and those from other careers is likely to be subject to selection bias. Instead, Eggers and Hainmueller (in the paper on this week's reading list) compare wealth at death for British people who narrowly won their election to the House of Commons to those who narrowly lost. Those groups probably have very similar potential outcomes. They find that Conservatives who narrowly won their elections died on average almost £250,000 wealthier than candidates who narrowly failed.

3. **Is voting habit-forming?** People who voted in past elections will have very different potential outcomes to people who did not, but those who were aged just under 18 at the last election should be near-identical to those just over 18. Daniel de Kadt (2017)[1] uses this to study the impact of participating in South Africa's first democratic election after Apartheid in 1994, when Nelson Mandela was elected president. He finds that turnout in future elections was around 6 percentage points higher among those who were just old enough to take part in the election compared to those who were just too young

4. **Did colonialism affect future economic development?** Countries that were colonised and not colonised are very different. Places that western powers chose to colonise are likely to have been better-suited for agriculture and industry to begin with. But within some countries or areas, borders can be drawn in a way that is arbitrary, sometimes literally by drawing a straight line on a map. Mattingly (2017)[2] argues that this is how China drew its border with Mongolia in the early twentieth century. In 1932, Japan colonised much of China and occupied the Chinese side of the border but did not invade Mongolia. The Chinese side was therefore subjected to Japanese colonial rule for fifteen years while the Mongolian side was not. Communities on either side of the border should have the same average potential outcomes. By comparing these two sets of communitiies, Mattingly argues that Japanese colonisation had positive effects on development and growth because Japan invested heavily in public services and bureaucracy, leaving a long-term legacy.

There is an important distinction between two different types of RD setups, "sharp" and "fuzzy":

[1]Daniel de Kadt (2017). "Voting then, voting now: The long term consequences of participation in South Africa's first democratic election." *Journal of Politics* 79 (2)

[2]Daniel Mattingly (2017). "Colonial Legacies and State Institutions in China: Evidence from a Natural Experiment." *Comparative Political Studies* 50 (4): 434-463

- **Sharp Regression Discontinuity:** Treatment status is wholly determined by the threshold. All units above it are treated, and all units below it are not treated. Two-candidate elections and colonial borders are examples. All towns on one side of the border must belong to country A and all towns on the other side of the border to country B
- **Fuzzy Regression Discontinuity:** The probability of being treated increases at the threshold but is not wholly determined by it. Scholarships and voting are examples. Being 18 or over makes citizens *eligible* to vote but does not force them to do so, except in countries with compulsory voting.

Fuzzy regression discontinuities are just like a situation with encouragement and one-sided non-compliance in randomised experiments. The rule at the discontinuity encourages some units to take the treatment but does not compel them to do so. Therefore fuzzy regression discontinuities are in effect just another example of an instrumental variable. In this lecture we only consider sharp regression discontinuities, which require a somewhat different approach compared to instrumental variables.

2. Calculating the LATE

RD estimation allows us to compare groups that are just above and just below the threshold, because at the threshold the treatment is as-if randomly assigned. As in instrumental variables, we have found something akin to a randomised experiment in a situation where the treatment is not in general randomly assigned. We have:

- **Treatment** $D_i \in \{0, 1\}$
- **Threshold** c above which the treatment applies
- **Potential Outcomes** under Treatment and Control: $Y_i(1)$ and $Y_i(0)$
- **"Running" (or "Forcing") Variable** X_i that determines whether i is above or below c, i.e.

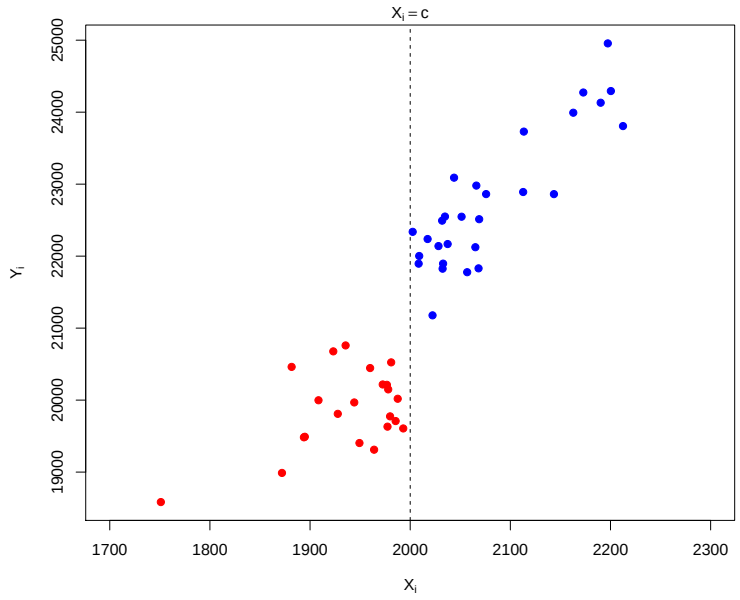
$$D_i = \begin{cases} 1 & \text{if } X_i > c \\ 0 & \text{if } X_i \leq c \end{cases}$$

In de Kadt's study of South African voters, D is voting in 1994, c is 18 years, and X is age. In Eggers and Hainmueller's study of British MPs, D is being a politician, c is 0, and X is the difference in vote share between the candidate and either the winner (for losing candidates) or the nearest challenger (for winners). We can therefore compute the Local Average Treatment Effect at the threshold:

$$\tau \;=\; E[Y_i(1) - Y_i(0) \mid X_i = c]$$

This LATE cannot be estimated for any i because we do not observe i in both treatment and control (the fundamental problem of causal inference). As always, we replace the potential outcomes with the difference in means between treated and control units, in this case only at the threshold where $x_i = c$. This is shown graphically in Figure 1, where there appears to be a positive treatment effect at the discontinuity. Figure 2 shows mean potential outcomes for treated and untreated units at different values of X. Note that the blue line is higher than the

Figure 1: **Graphical Illustration of Regression Discontinuity**



red: treated units are systematically different to untreated units, a sign of selection bias. But as shown in Figure 3, only half of these mean potential outcomes are actually observed. Above c, all units are treated and below c, all units are untreated. At the discontinuity, treatment is as-if randomly assigned and the estimated treatment effect is simply the vertical difference (in green) between the treated and untreated units, i.e. the difference in observed means.

Figure 2: **Regression Discontinuity: True Potential Outcomes**

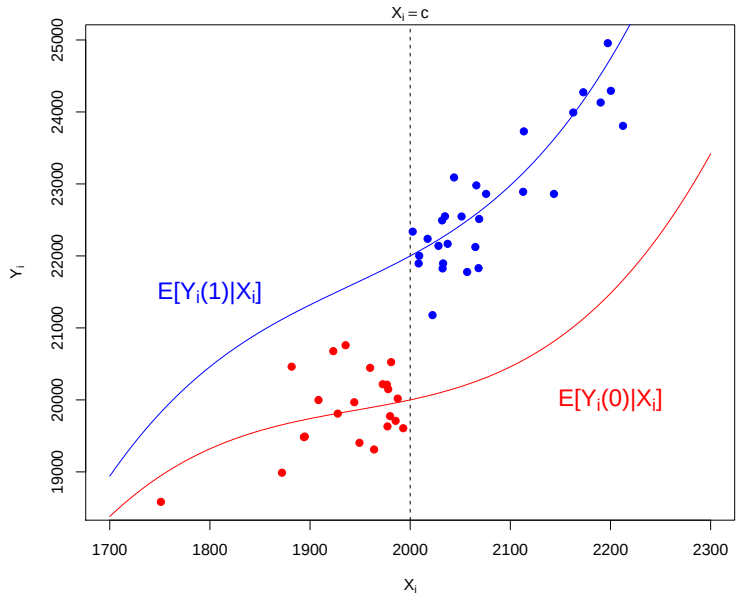
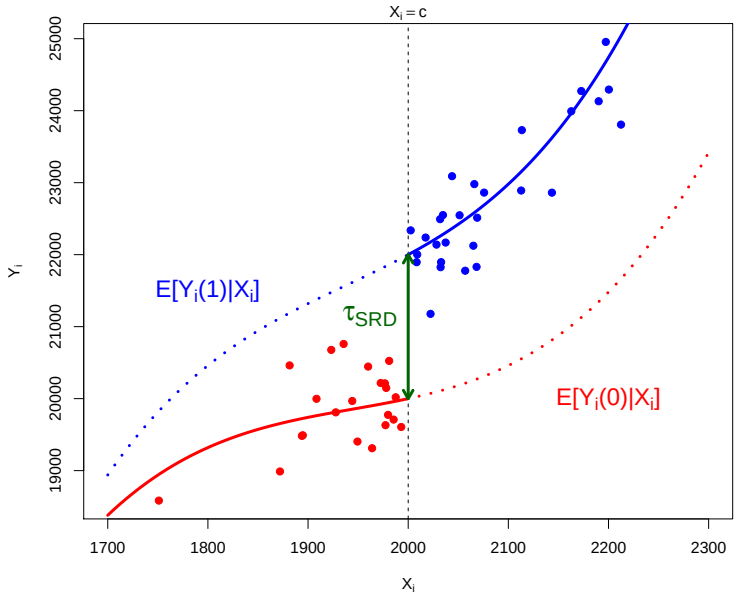


Figure 3: **Regression Discontinuity: Observed Potential Outcomes and Local Average Treatment Effect**



3. Modeling the LATE

Actual estimation of the treatment effect requires accurately modeling the curves shown in Figure 3. In particular, we need an accurate estimate of where exactly the curves "hit" the cutoff, or our results will be biased. We can capture everything necessary using regression models. To do this, we must first recode the forcing variable to deviations from threshold. Thus $\tilde{X}_i = X_i - c$, and:

- $\tilde{X}_i = 0$ if $X_i = c$
- $\tilde{X}_i > 0$ if $X_i > c$ and thus $D_i = 1$
- $\tilde{X}_i < 0$ if $X_i < c$ and thus $D_i = 0$

We must then decide on an appropriate form for a regression model. The most simple possible model just uses a linear model with a common slope, giving the plot in Figure 4:

$$E[Y_i \mid D_i, X_i] = \alpha + \beta\tilde{X}_i + \tau D_i$$

The equation demonstrates why it is necessary to re-center the running variable X around c as $\tilde{X}$ and control for it in the regression, because when $X = c$, $E(Y) = \alpha + \tau D$ and therefore τ measures the treatment effect of D at the threshold. This regression specification is only one possibility amongst many, though. For instance, it is often the case that there are different slopes on either side of the discontinuity. In Figure 4, the relationship may be steeper above c. How can we guard against this possibility, or other potential mis-specifications of the model?

One important option is to try different model specifications and see how sensitive the results are to different modelling choices. Visual inspection can be used to determine which

Figure 4: **Linear Regression with Common Slopes**

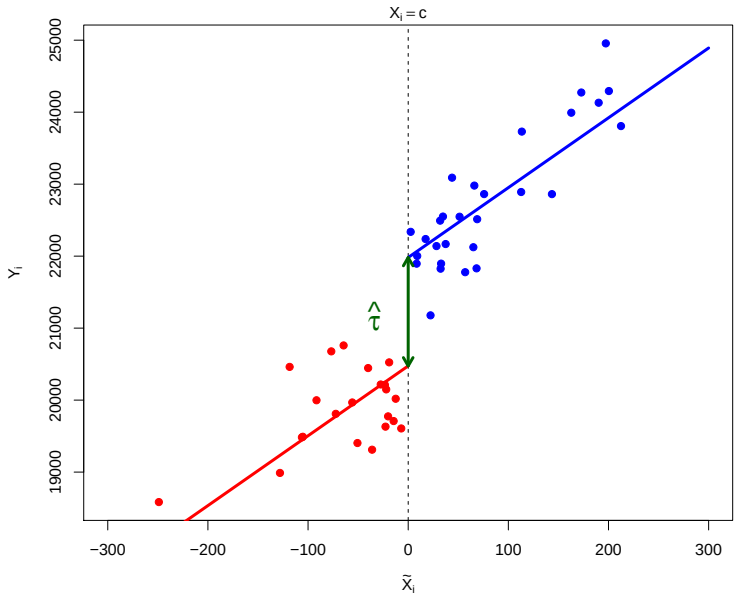
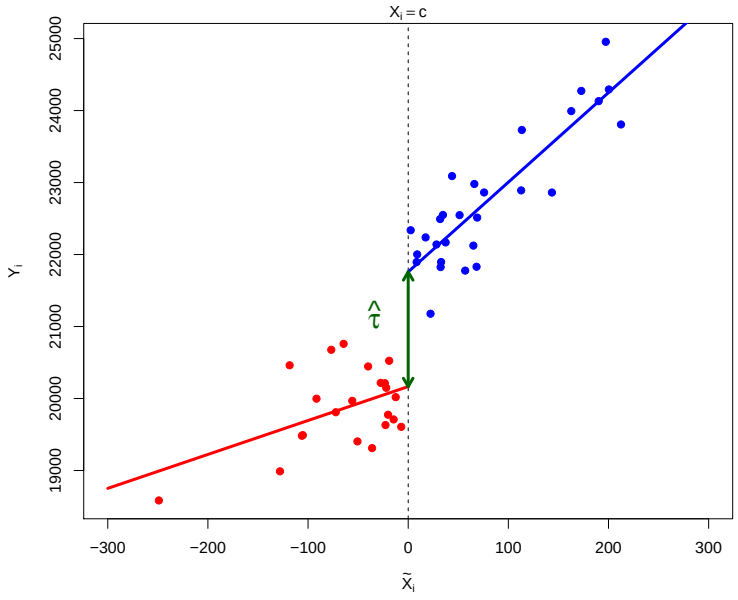


Figure 5: **Linear Regression with Different Slopes**



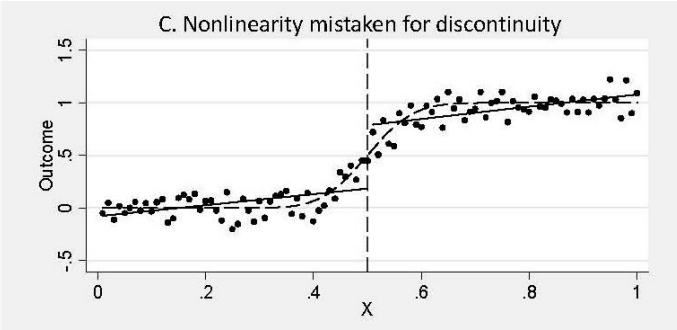
models seem to fit the data best. Different slopes can simply be accommodated by using an interaction term:

$$E[Y_i \mid D_i, X_i] \;=\; \alpha + \beta\tilde{X}_i + \tau D_i + \gamma D_i\tilde{X}_i$$

This is shown in Figure 5, which appears to give a much better fit to the data. However, this still risks us making a serious error. What if the relationship between $\tilde{X}$ and Y doesn't feature

a discontinuity at all, but simply a non-linearity around the cutoff window? Such a possibility is shown in Figure 6, taken from Angrist and Pischke. What looks like a discontinuity is nothing more than a standard non-linearity that happens to coincide with the threshold $X = c$.

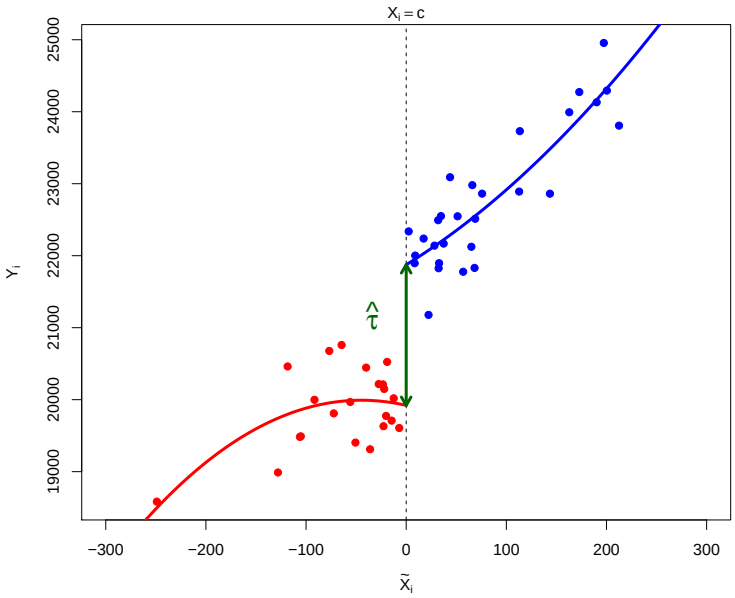
Figure 6: **Mistaking a Non-Linearity for a Discontinuity**



As we know from basic regression theory, while the regression function itself must be linear, it is perfectly possible to model non-linearities with linear regression by including quadratic terms such as squares and cubes. For example, we could fit the following model to the same data shown in Figures 4 and 5, which allows for both a curvilinear relationship and different slopes either side of c, shown in Figure 7:

$$E[Y_i \mid D_i, X_i] \;=\; \alpha + \beta_1 \tilde{X}_i + \beta_2 \tilde{X}_i^2 + \tau D_i + \gamma_1 D_i \tilde{X}_i + \gamma_2 D_i \tilde{X}_i^2$$

Figure 7: **Linear Regression with Non-Linearities**



This appears to give the best fit to the data of all. While there is no hard and fast rule about how to choose the right model, it is always important in RD estimation to plot the data

and carefully compare different model choices with a range of different flexible models. The aim is to produce a good fit to the data such that the result we obtain is determined by it, rather than by the model. This also highights a condition for effective use of RD estimation: there should be a lot of data points in the neighbourhood of the discontinuity. This allows the data to speak for itself in detemining exactly where the regression lines hit the discontinuity, reducing model dependency.

In fact, standard practice now in RD estimation is to use the most flexible possible types of non-linear regression, known variously as 'local linear regression', 'locally-weighted polynomial regression' or 'LOESS', which allow for a very large amount of non-linearity, leaving the data to speak for itself. We won't go into the details, but essentially these techniques carry out a series of 'mini-regressions' at many levels of X and then stitch them together into a single regression line. This is the type of regression that is used by default in R's `rdd` package, which we'll use in this course, and is also what Eggers and Hainmueller use in their paper.

They also carry out a common robustness check for this functional form of the regression known as a **placebo test** on page 14. Placebo tests in general are of the form: "did something that we do not expect to happen actually not happen?" We are looking for the absence of an effect. Here, the idea is that it is possible that the relationship between wealth at death and the vote share margin is fundamentally discontinuous and marked by a lot of discontinuities, in which case the discontinuity observed at $X = c$ is merely one of many such 'jumps' that occur commonly, and have nothing to do with a genuine treatment effect. They test for this by choosing a range of "placebo thresholds" and examining whether or not the data are discontinuous in each of those placebo regions. If the running variable is fundamentally continuous everywhere except the discontinuity, then we should find a series of null results (effects indistinguishable from zero). And indeed, this is exactly what they do find (Table 5), which gives us confidence that there is indeed something special about the discontinuity at c. This placebo test is very commonly carried out in RD studies.

No matter how we ultimately specify the model, the τ that we eventually obtain is called a Local Average Treatment Effect because it is local to only those units that are at the threshold, whose potential outcomes are comparable. That means that this is an extremely "local" effect, limiting the external validity of RD estimation. We cannot, for instance, make inferences about incumbency effects for politicians in general, but only for those who narrowly win their elections. In the same vein, we can only make inferences about the impact of participation for people who were just over 18 at the last election, rather than for voters in general. Nonetheless, internal validity is likely to be very high, provided that we have an accurate model of the relationship between Y and X and that treatment is genuinely randomly assigned at the threshold. Having considered the first those two problems, we'll now look at the second.

4. Violations of Assumptions: Composite Treatments and Sorting

As with everything else we have studies in this course, regression discontinuity analysis requires that the treatment – and only the treatment – be genuinely as-if randomly assigned at the threshold. There are two common ways in which this can be violated

4.1. Composite Treatments

Composite treatments are akin to the exclusion restriction in instrumental variables analysis. They occur when, at the threshold, other rules or policies other than the treatment of interest also change. One common scenario involves rules that are based on the size of administrative units like cities. For example, Andy Eggers shows in a recent study that French municipalities of 3,500 people or more must use a PR electoral system for their local councils, whereas municipalities of less than 3,500 must use a plurality rule (winner-takes-all).[3] He uses this discontinuity to estimate the effect of electoral systems on voter turnout at the threshold. The idea is that PR systems may increase turnout because people's votes are more likely to determine the outcome of the election. A vote for the second-placed party or candidate may still 'count.'

However, a common feature of these population-size thresholds is that many other policies also change at the threshold that might plausibly affect the outcome, including things like the size of the council or the policies over which it has control. If towns of 3,500+ have many more powers, turnout may go up because the election is more important in determining policy outcomes, rather than being due to the electoral system. Instead of estimating the treatment effect that we care about – the impact of electoral systems – we actually end up estimating the effect of a whole set of varied policies. Hence it is always important in RD studies to carefully check, using knowledge of the case at hand, what else might change at the cutoff.

4.2. Sorting Around the Threshold

Another possibility is a direct violation of randomisation, whereby units are able to cheat their way into the treatment or control group by manipulating their value for the running variable. For example, candidates might commit electoral fraud to win close elections or voters might fake their identification to vote while too young. If these groups of 'cheaters' have different potential outcomes to non-cheaters, then random assignment is violated and the treatment and control groups are no longer comparable at the threshold. As always, this can be tested using balance tests for observable characteristics at the threshold. Imbalance is suggestive of violations of randomisation.

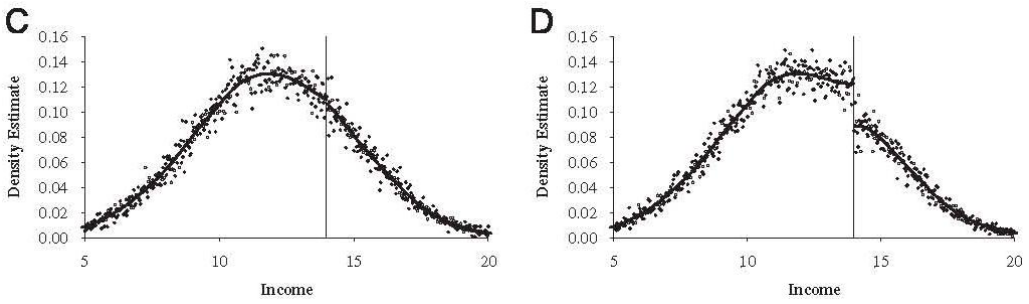
One common way to carry out these balance tests is with another type of placebo test: placebo regression discontinuities. Here, we simply replace the outcome variable with each of the background covariates in turn, and carry out the same regression discontinuity estimation. If units are genuinely randomly assigned at the threshold, then we should find no evidence of discontinuities there for the background characteristics. Eggers and Hainmueller do this in Table 6 of their paper.

Another test of this can be obtained by examining the density of the running variable, which should be continuous around the threshold. If there is a 'bump' in the histogram just above c, that would suggest that units might be pushing themselves into the treatment group. A high incidence of electoral fraud, for example, could be indicated by a bunching of results at just over the 50% mark. Candidates are unlikely to try to fix elections so egregiously that what should have been a close race turns into a landslide victory. But they may well commit 'just enough' fraud to push themselves over the winning line. Similarly, voters aged 17 are unlikely to falsely pose as forty year-olds, but instead as eighteen or nineteen year-olds.

[3] Eggers (2015). "Proportionality and Turnout: Evidence from French Municipalities." *Comparative Political Studies* 48 (2): 135-167.

In fact, the method for detecting this sort of sorting behaviour has many similarities to statistical methods that are used to detect electoral fraud. The **McRary Sorting Test** looks for evidence of discontinuous jumps in the running variable at the threshold. Remember that in an RD setting, a treatment effect occurs when there is a discontinuous jump in the *outcome variable*, but the running variable should be continuous around the threshold. If fraud or cheating is happening, it would be indicated by a 'bunching' of units at just over the threshold value, trying to cheat in a subtle way. To look for this, we can examine the density of the running variable to look for such jumps.

Figure 8: **Non-Violation and Violation of Sorting**



The left-hand panel of Figure 8 shows a situation with no evidence of sorting, whereas the right-hand panel is a case where sorting appears to have occurred. The McRary test in R will produce this plot for you. It also produces a test statistic for testing the null hypothesis of no jump in the density at the discontinuity. A statistically significant result means that we can reject this null, providing evidence in favour of the alternative hypothesis of a genuine difference, indicating the possibility of sorting. The code for running this in R is simply:

```
library(rdd)
DCdensity(running.var,verbose=T)
```

5. Bandwidth Selection

A final issue in RD estimation is the selection of an appropriate *bandwidth* around the cutoff. Essentially this asks: how much of the data for the running variable should we use to construct our regressions? In RD estimation, all that we really care about is accurately modeling patterns at the cutoff. Should we bother to include data that may be very far from the cutoff? In answering this question, we face a *bias-variance tradeoff*. The more data that we use, the lower will be the variance (precision is inversely related to sample size). But at the same time, the regression lines will be influenced by data points that are further and further away from the cutpoint. Particularly in cases where there are outliers a long way from the cutpoint, the estimated regression lines could provide an inaccurate picture of patterns directly at the cutpoints themselves. Therefore bias may increase, the more data that we use. There is a tradeoff: achieving a lower variance leads to greater potential for bias, and vice versa.

The bandwidth for RD estimation is defined by points h and $-h$. We use data for observations only within the bandwidth $c - h \leq X_i \leq c + h$, where h is a point on the scale of the

running variable X. The **Imbens-Kalyanaram Procedure** solves an optimisation problem to find the optimal h that balances the twin goals of variance reduction and bias reduction. Before doing RD estimation, it is standard practice to find the optimal bandwidth and restrict the sample to within that bandwidth. The R code to do this is given by:

```
library(rdd)
bandwidth <- IKbandwidth(running.variable, outcome.variable)
```

Having found this bandwidth, the rdd estimate of the LATE can be computed as follows, using highly flexible local non-linear regression as standard:

```
library(rdd)
RDestimate(outcome~running.variable, bw = bandwidth, data=)
```

These days, an even more common procedure is simply to estimate the LATE at a very wide range of different bandwidths, allowing the reader to see how sensitive the results are to different bandwidths. Such a plot is usually found in modern RD papers. Ideally, the size and significance of our main result should not change much as the threshold changes: we'd like the result to not be sensitive to the choice of bandwidth. Otherwise, if the result held only at one particular choice of bandwidth, then it could have arisen due to chance alone.