Regression Discontinuity Design

Introduction

- Regression Discontinuity Design (RDD) is yet another advanced causal analysis technique.
- It works by exploiting arbitrary rules on how treatments are assigned.
- We will consider two types of RDD: **sharp** and **fuzzy**.
- These two types of RDD are characterised by what we know about the treatment rule.
- Note, I will typically refer to 'sharp RDD' as simply 'RDD'.

Sharp RDD

- Sharp RD is used when the decision to treat is a **deterministic** and **discontinuous** function of some regressor, say x_i.
- Consider the following treatment rule:

$$D_i = \begin{cases} 1 & \text{if } x_i \geq x_0 \\ 0 & \text{if } x_i < x_0. \end{cases}$$

  where x_0 is a known threshold. Note, x is normally referred to as the **forcing variable** or **running variable**.
- We say it is a deterministic rule because when we know x_i, we know D_i.

Some Examples

- Some examples will help make this clear.
- Suppose there is some scholarship that prospective university students are entitled to, but only if they score 80% or higher in their high school exams.
- There are many government policies which provide tax relief/benefits to those who earn under a certain income.

A Real Example and the First use of RDD

- US high school students are given a scholarship for university but only if they score high enough on a test taken in their final year of high school.
- People were interested in whether students who received theses scholarships were more likely to finish university.
- Of course, trying to estimate the causal effect of this scholarship is rife with OVB. The best and most motivated students are the ones who get the scholarship! We cannot naively compare the two groups.

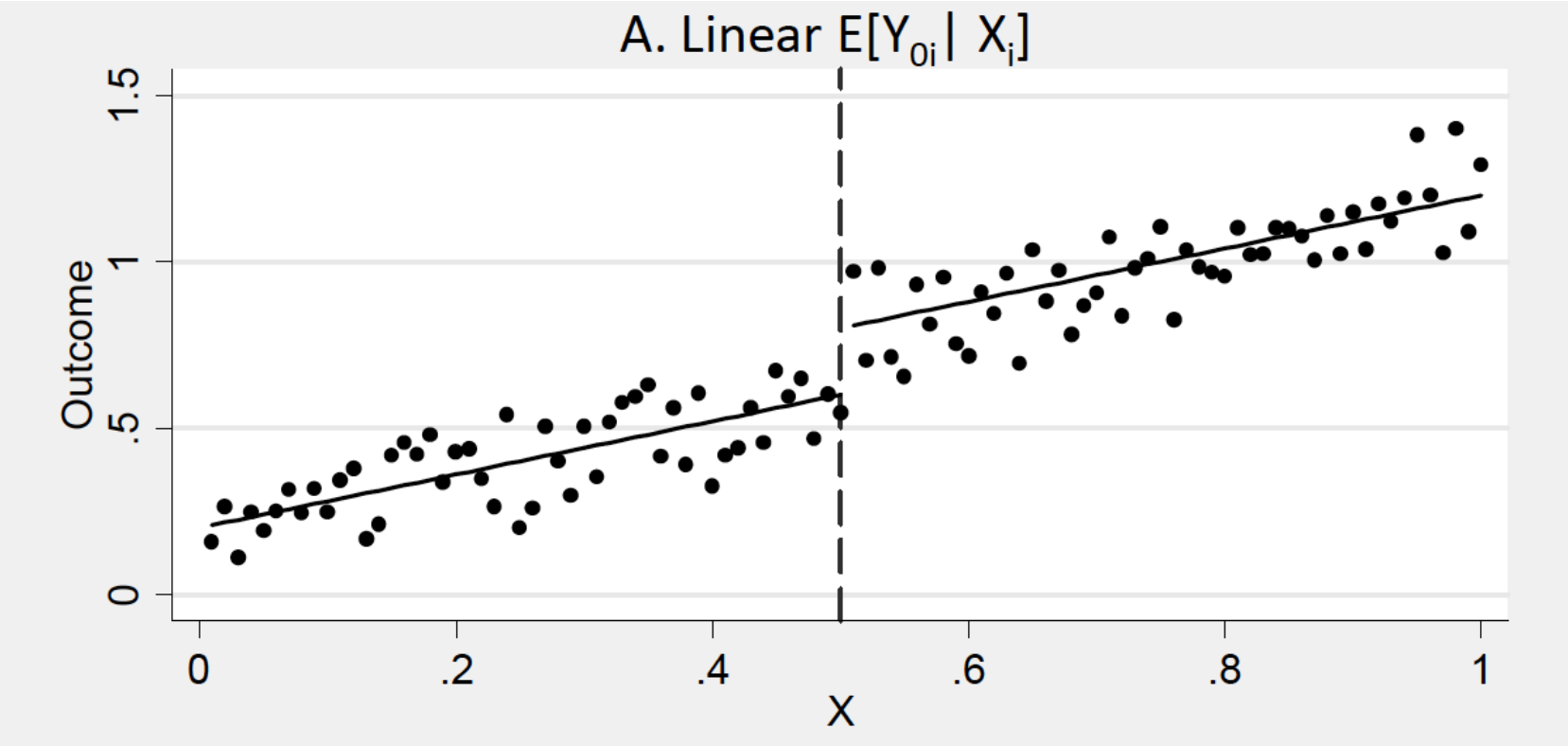
RDD for University Scholarships

- RDD was invented to answer this question.
- The idea is very simple, we should compare university completion rates for students who just missed out on the scholarship with those who just scraped into the scholarship program.
- Suppose the 'passing mark' is 80. Well, there is really no meaningful difference between those who score 79 and those scoring 80... apart from the scholarship!
- It is obvious then that this should give us a causal effect.

RDD and Regression

- If we make some additional assumptions, we can actually identify the causal effect with just a simple regression.
- Suppose we assume that the causal effect is additive and constant, and that the effect of x on y is constant across the discontinuity. (We will consider a less restrictive approach later on based on the idea on the previous slide)
- The idea is that without the treatment, there will be a normal regression line.
- But when you place the treatment at some point along this line you get a **discontinuity**. This is where the name comes from!
- It is this jump in the regression line that allows us to identify the causal effect. Check it out on the following graph:

Graphical Representation of RDD



(Credit: Mostly Harmless Econometrics, Angrist and Pischke)

The RDD Equation

- The actual regression to calculate the causal effect using RDD is very simple. Assuming a constant, additive effect, we have

$$Y_i = \beta_0 + \beta_1 x_i + \rho D_i + u_i.$$

- D_i is an indicator for whether you are to the left or right of the discontinuity. The causal effect is given by ρ.
- Notice that D_i is completely determined by x_i. Why is this not a problem for multicollinearity?
- The reason: it is a nonlinear function of x_i. Multicollinearity is only an issue of linear dependence.
- So RDD is distinguishing the nonlinear and discontinuous effect of D_i from the smooth, linear effect of x_i.

Example: Political Party Incumbency

- Lee (2008) studied the effect of party incumbency on the probability of re-election. He looks at whether the democratic candidate for a seat in the House of Representatives has an advantage if their party won the seat last time.
- House incumbents are known to have much greater success than non-incumbents, but is this because these incumbents are more popular or is it because they use their privileges and resources from being in office to gain an advantage?
- Lee looks at the likelihood of winning (y in the dataset) as a function of the difference in vote shares from the previous election (x in the dataset). So the cutoff is at $x_0 = 0$. (Note, the observations are grouped together for similar values of x)

Example: Political Party Incumbency ('house')

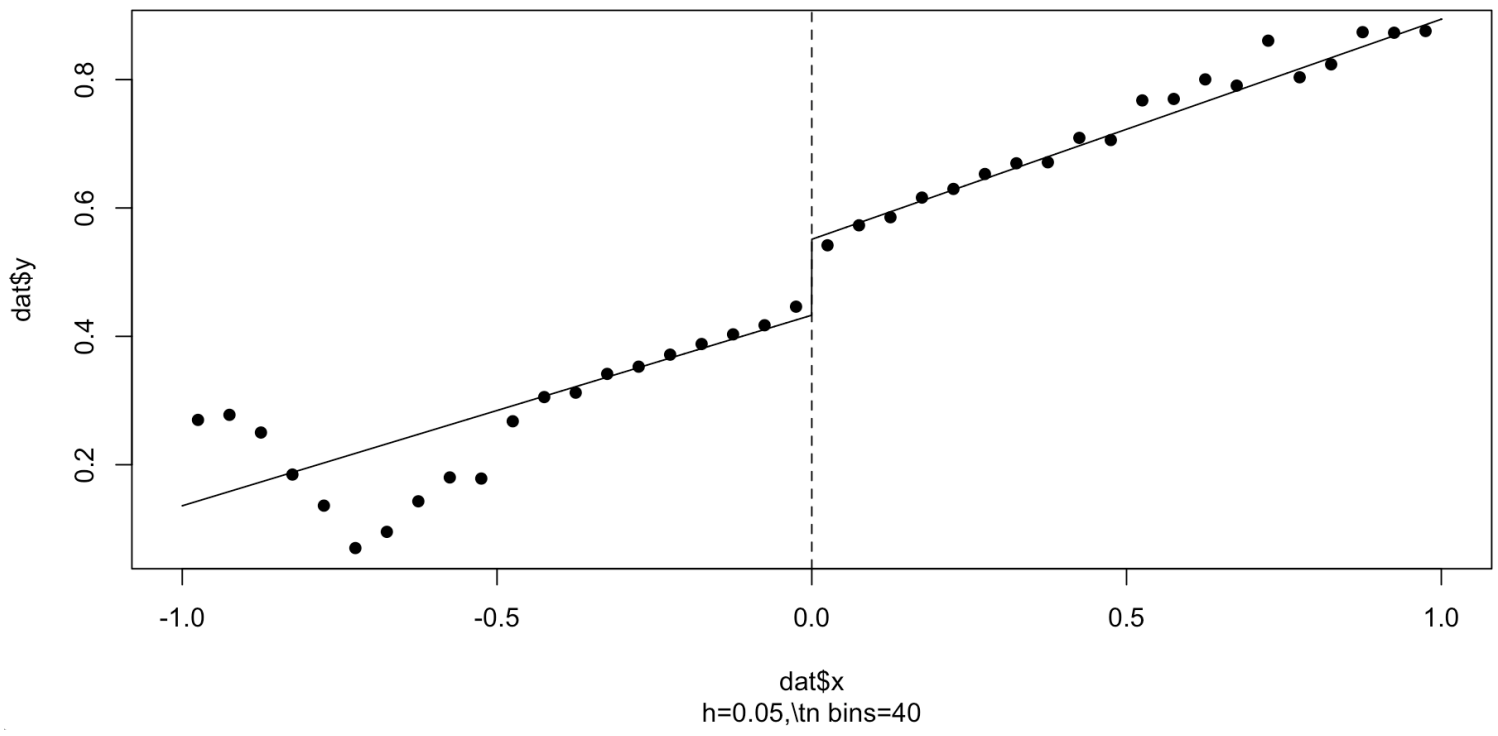
```
rm(list = ls())     # Clear workspace

library(rddtools)

# Just looking at the data
data(house)  # Load data
house_rdd = rdd_data(x=x, y=y, data=house, cutpoint=0)  # Automatically tic
summary(house_rdd)  # Summary of the data
plot(house_rdd)  # Plot of data with cutoff

# Run RDD linear regression
reg_para = rdd_reg_lm(rdd_object=house_rdd)
summary(reg_para)
plot(reg_para)
```

Example: Political Party Incumbency



Example: Political Party Incumbency

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.432948   0.004276 101.254  < 2e-16 ***
D           0.118231   0.005680  20.816  < 2e-16 ***
x           0.296906   0.011546  25.714  < 2e-16 ***
x_right     0.045978   0.013501   3.405 0.000665 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1384 on 6554 degrees of freedom
Multiple R-squared:  0.6707,    Adjusted R-squared:  0.6706
F-statistic:  4450 on 3 and 6554 DF,  p-value: < 2.2e-16
```

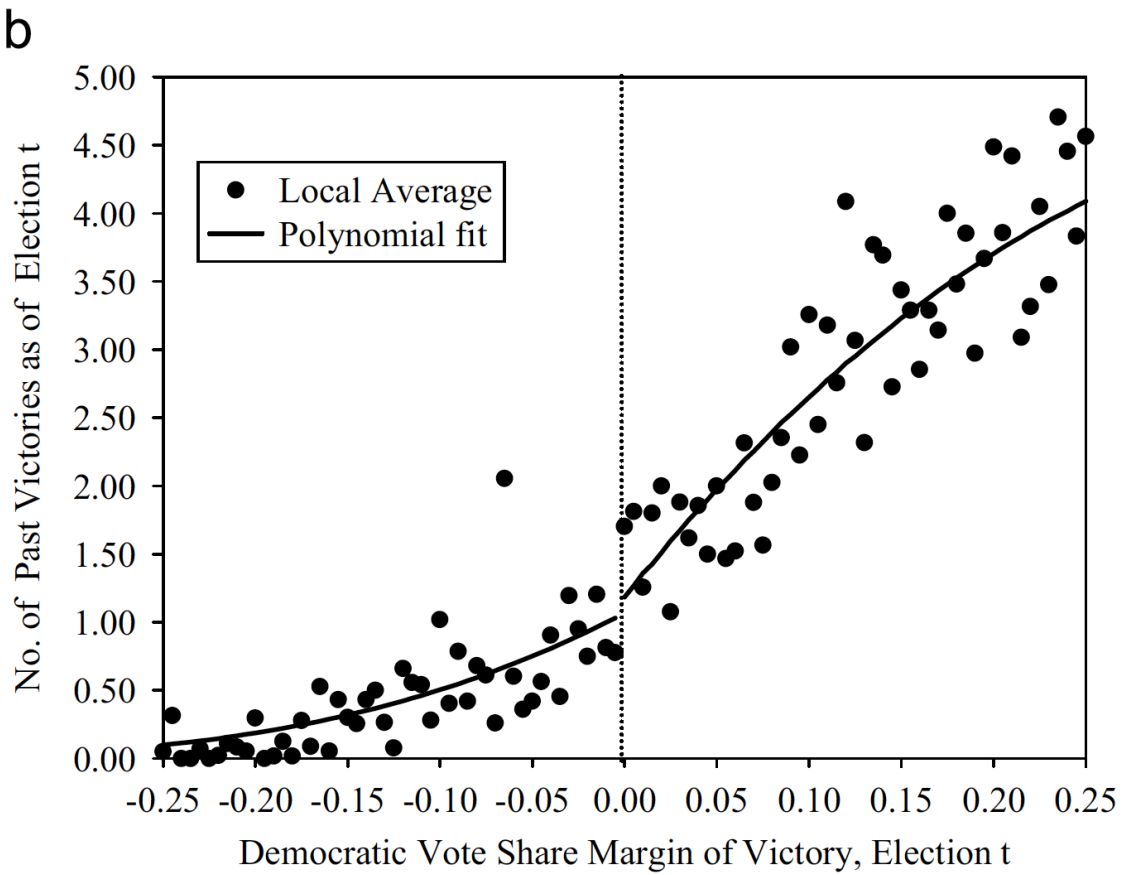
Example: Political Party Incumbency

- The probability of a democratic win is an increasing function of past vote share. As we would expect.
- The most important feature of the plot is the dramatic jump in win rates at the 0 percent mark, the point where a Democratic candidate gets more votes.
- Based on the size of the jump, incumbency appears to raise party re-election probability by about 12 percentage points.
- Lee goes on to test the validity of his analysis...

Example: Political Party Incumbency

- He performs the same analysis as before, but now his dependent variable is the number of victories **before** the previous election.
- The idea is that the current incumbency effect should clearly not have any impact on previous victories. If it does, then it indicates that there is a fundamental difference between those just over the threshold and those just under which has nothing to do with them actually being in office.
- If this turns out to be the case, then our RDD is flawed. We want to compare two identical people, one of whom is currently in office and one of whom is not.
- We do not have the full data he used, so cannot estimate it ourselves, but the following graph is taken from his paper.

Example: Political Party Incumbency

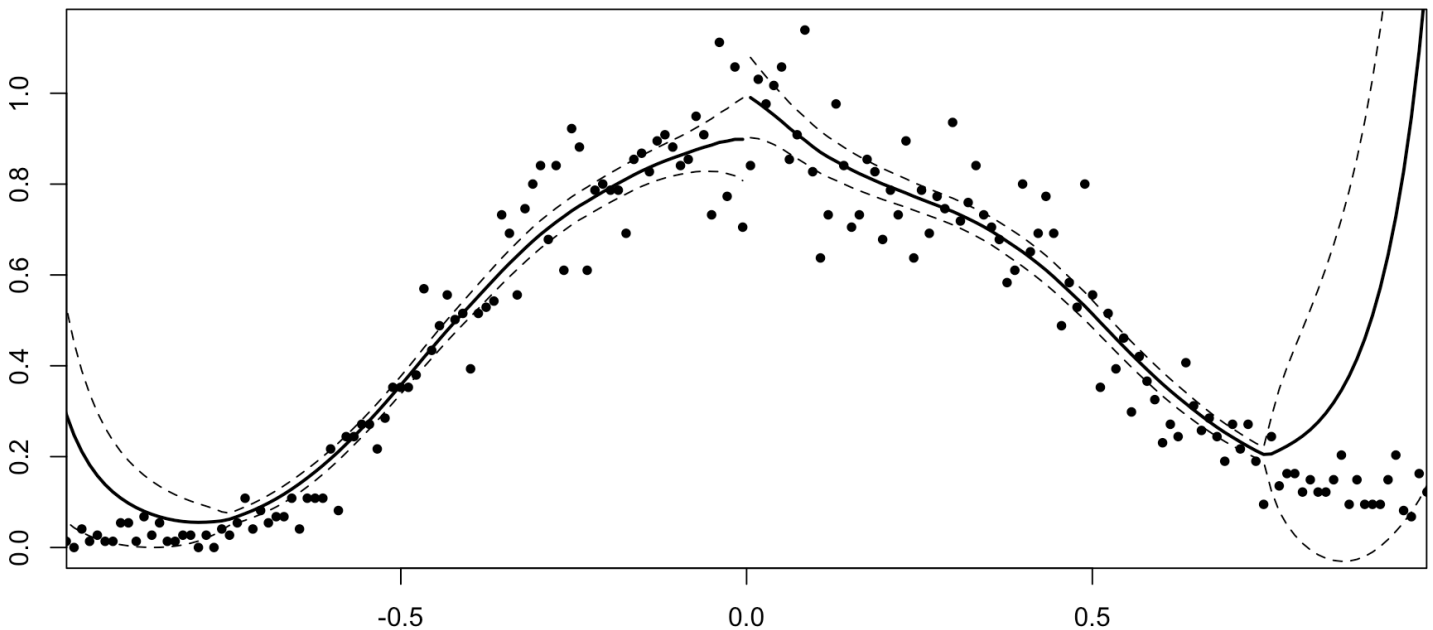


(Credit: Mostly Harmless Econometrics, Angrist and Pischke)

Checking Assumptions

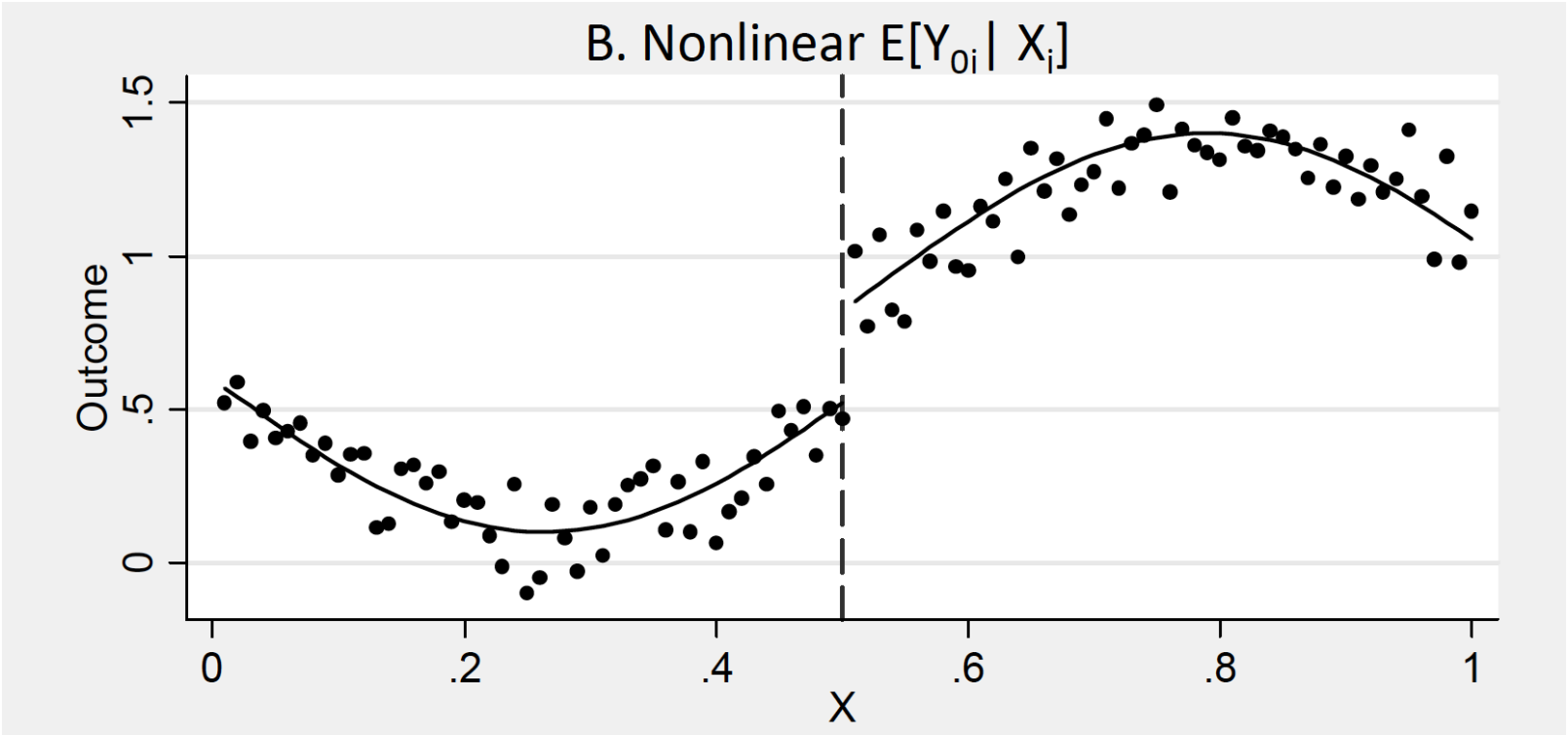
- This is basically just checking whether covariates are balanced in the 'treatment' and 'non-treatment' group. I.e. if we look at the characteristics of politicians who just missed out on election last time, and those who just got elected, there shouldn't be any systematic differences.
- Another test is the McCrary test. This looks at whether the density of the running variable is similar on either side of the cutoff. In R, you use the function dens_test() to test this.
- The idea is that if there are many more people on one side than the other (bunching of the running variable) then people can manipulate whether they are treated or not. This is not good for our analysis!

McCrary Test



Nonlinearity...

- What happens if the effects are actually nonlinear, like this:



(Credit: Mostly Harmless Econometrics, Angrist and Pischke)

Dealing with Nonlinearity

- So our model now looks like this

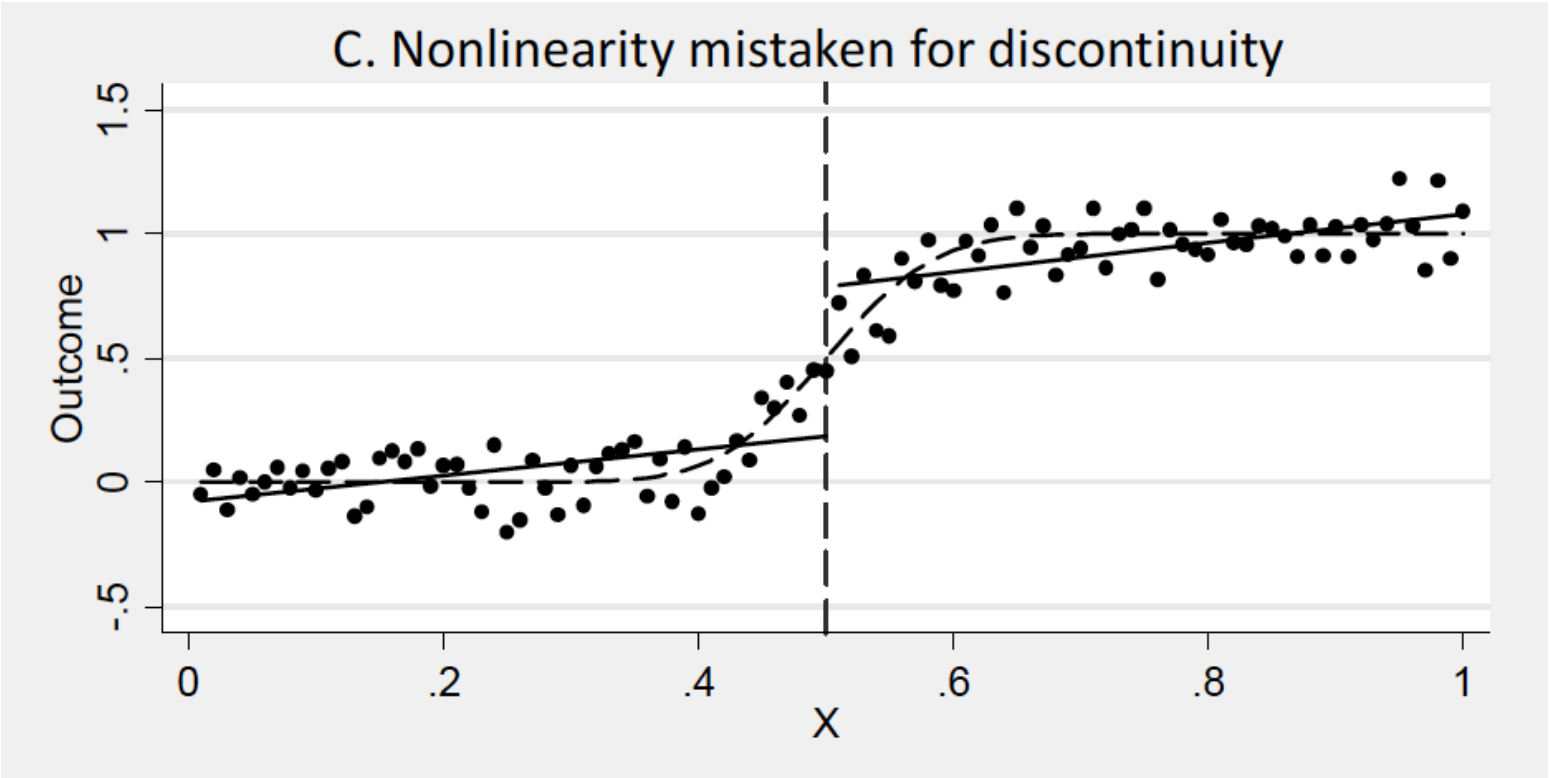
$$Y_i = g(x_i) + \rho D_i + u_i$$

  for some function $g(\cdot)$.

- We can appeal to the Weierstrass approximation theorem:
- *If* $g(\cdot)$ *is continuous on the real interval* $[a, b]$, *then for every* $\epsilon > 0$, *there exists a polynomial function,* $p(\cdot)$, *such that* $|f(x) - p(x)| < \epsilon \ \forall x$.
- For us, this means we can do something like this:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \rho D_i + u_i.$$

Importance of Allowing Nonlinearity



(Credit: Mostly Harmless Econometrics, Angrist and Pischke)

A More Robust Approach

- We have just seen that it is very important that we correctly specify the regression function.
- To reduce the chance of making a mistake like on the previous slide, we can appeal to the idea we had on slide 6: compare people just before and just after the cut-off.
- By doing this, we don't need to know what's going on with the rest of the regression line. We also don't even need to assume a constant causal effect. But we can then only say we've identified the causal effect for those with an x at the cut-off.
- This does introduce it's own problems though... because we only look at those close to the cut-off, we have very few observations. It can also be shown that the sample mean is biased at the boundary. Also, how do we even carry out this estimation?

Nonparametric Estimation

- We use **nonparametric** techniques.
- The most naive nonparametric estimator is to define the neighbourhood around the cutoff using a parameter δ, giving $[x_0 - \delta, x_0 + \delta]$. We then estimate the mean for all $x \in [x_0 - \delta, x_0)$, and the mean for all $x \in (x_0, x_0 + \delta]$. (We'll discuss choice of δ later)
- This is naive in two ways: (1) the sample mean is basic and suffers from bias at the boundary; (2) the weighting function is also very basic.
- The weighting function tells you which observations will be included in the estimate for the treated group and the untreated group, respectively. In this case it is 1 if in the neighbourhood and 0 everywhere else. (Illustrated on the board. If you missed this, check blackboard under 'Lecture Material'.)

Nonparametric Estimation - Weighting Function

- We first deal with the weighting function. Instead of a '**uniform kernel**', we could use a '**Gaussian kernel**' (or there's a bunch of other kernel functions).
- A uniform kernel places equal weight (specifically, 1) to every observation in the neighbourhood.But the observations that are closest to the cutoff are the most relevant, so they should be given most weight.
- Picture a Gaussian pdf. Now picture just the left side of it. Now align the straight right edge of this curve at the cutoff point, x_0. This function tells you what weight to assign an observation depending on where it falls on this curve. Observations at the cutoff receive the most weight, observations far to the left receive little weight. And those to the right of x_0 receive zero weight (i.e. are excluded).

Nonparametric Estimation - Estimator

- We mentioned that the sample mean is biased at the boundary. The problem is that you only have half the information. We want a more sophisticated approach.
- The idea is to use a linear regression (or you can even use a polynomial regression), but weighted using the kernel function discussed above - known as a **local linear estimator**.
- Imagine you want $E[y|X = \tilde{x}]$ for some value of $\tilde{x}$. We know how to do this using a simple linear regression. But now, instead of using the full sample of (y_i, x_i) to construct the estimator, we first weight each observation pair using a kernel function applied at point $\tilde{x}$ (illustrated on the board. On blackboard.)
- It turns out that by doing it this way, we do not get a bias at the boundary.

Nonparametric Estimation - Bandwidth Choice

- The parameter, δ, is known as a **bandwidth**.
- This bandwidth governs how many observations will be used in each estimate. (In the Gaussian kernel case, think of it like the variance of that Gaussian distribution)
- We have a tradeoff when choosing this parameter. If we pick a very small bandwidth, we will only look at observations that are very close to the point of interest, but we are using few observations. High variance but low bias.
- Conversely, if we choose δ to be large, we have low variance but potentially a high bias. In the extreme case of a huge δ, we use all observations and just end up with naive OLS.
- Luckily, some smart people (Imbens and Kalyanaraman, 2012) developed a way to choose δ optimally, so we don't have to worry.

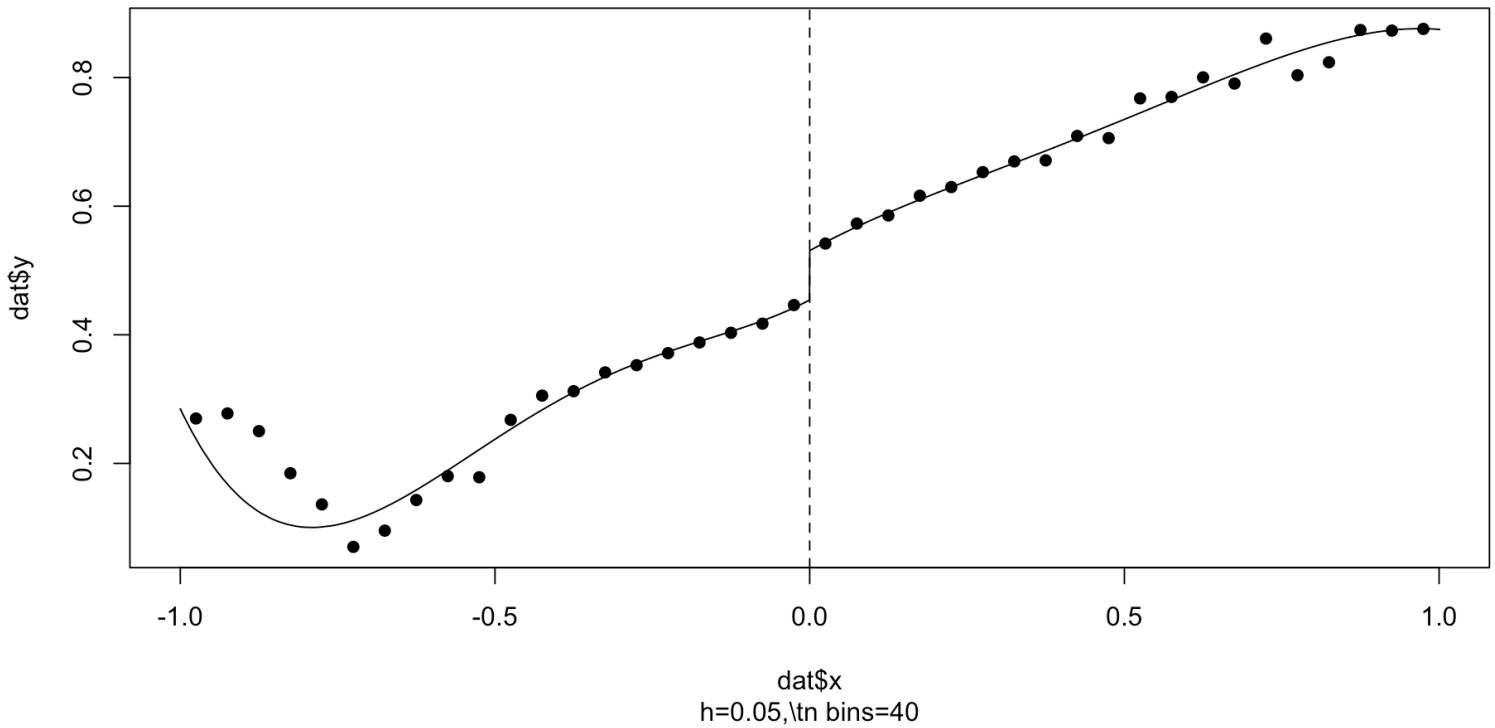
Example: Incumbency (Again)

- We're going to look at the same example, but now compare our previous results to those using nonlinear estimation and nonparametric techniques.
- The following R Code estimates a 4th order polynomial model, and a nonparametric (local linear) model.

```
# Run a parametric polynomial regression of order 4
reg_para4 = rdd_reg_lm(rdd_object=house_rdd, order=4)
summary(reg_para4)
plot(reg_para4)

# Run a nonparametric local linear regression
reg_nonpara = rdd_reg_np(rdd_object=house_rdd)
summary(reg_nonpara)
plot(reg_nonpara)
```

Example: Incumbency (Nonlinear)

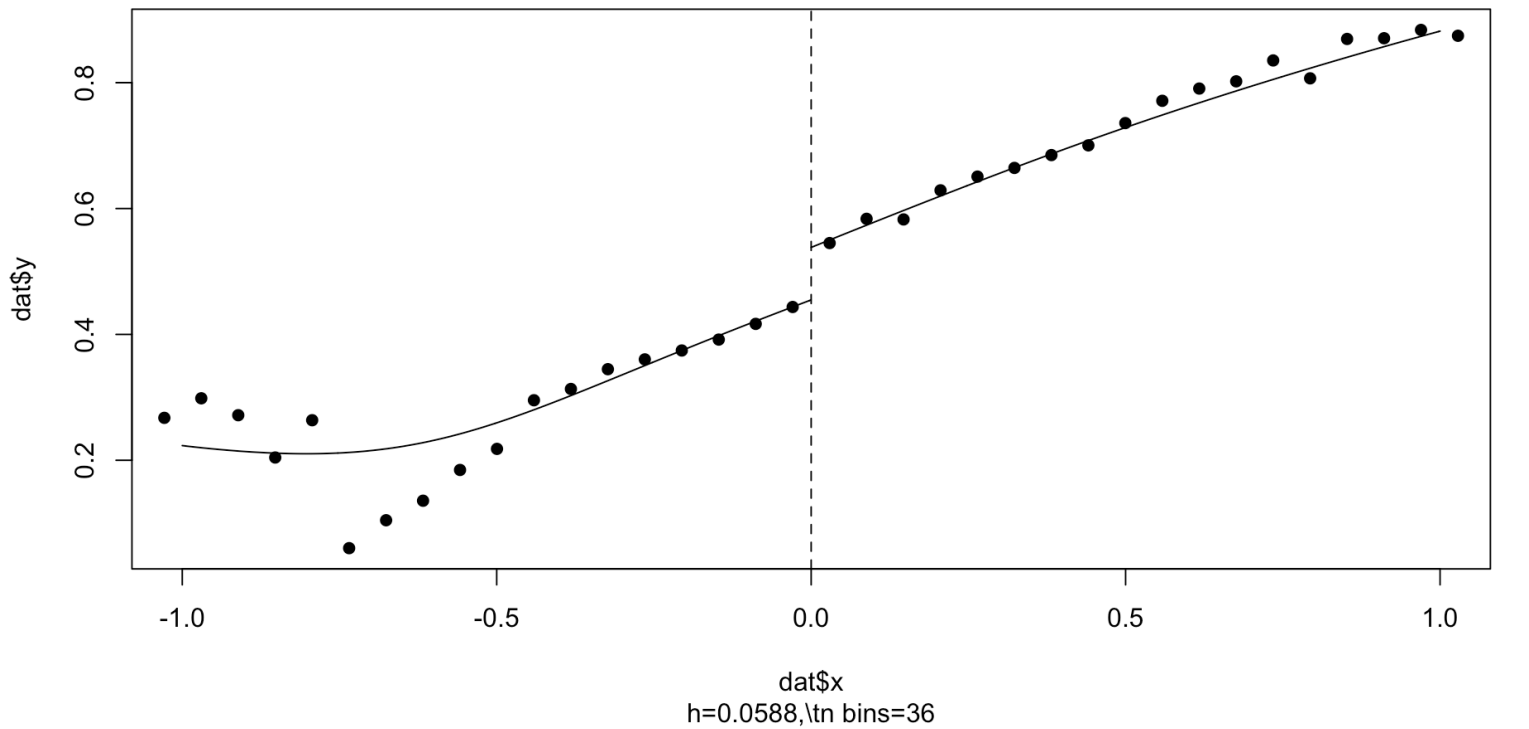


Example: Incumbency (Nonlinear)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.454167   0.009725  46.700  < 2e-16 ***
D            0.076590   0.013239   5.785 7.58e-09 ***
x            0.523595   0.158015   3.314 0.000926 ***
`x^2`        1.529216   0.740851   2.064 0.039044 *
`x^3`        4.220147   1.248511   3.380 0.000729 ***
`x^4`        3.045197   0.664059   4.586 4.61e-06 ***
x_right      0.019514   0.209112   0.093 0.925652
`x^2_right` -2.233991   0.948263  -2.356 0.018508 *
`x^3_right` -2.983991   1.551762  -1.923 0.054527 .
`x^4_right` -3.775626   0.808850  -4.668 3.10e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1356 on 6548 degrees of freedom
Multiple R-squared:  0.6838,    Adjusted R-squared:  0.6833
F-statistic:  1573 on 9 and 6548 DF,  p-value: < 2.2e-16
```

Example: Incumbency (Nonparametric)



Example: Incumbency (Nonparametric)

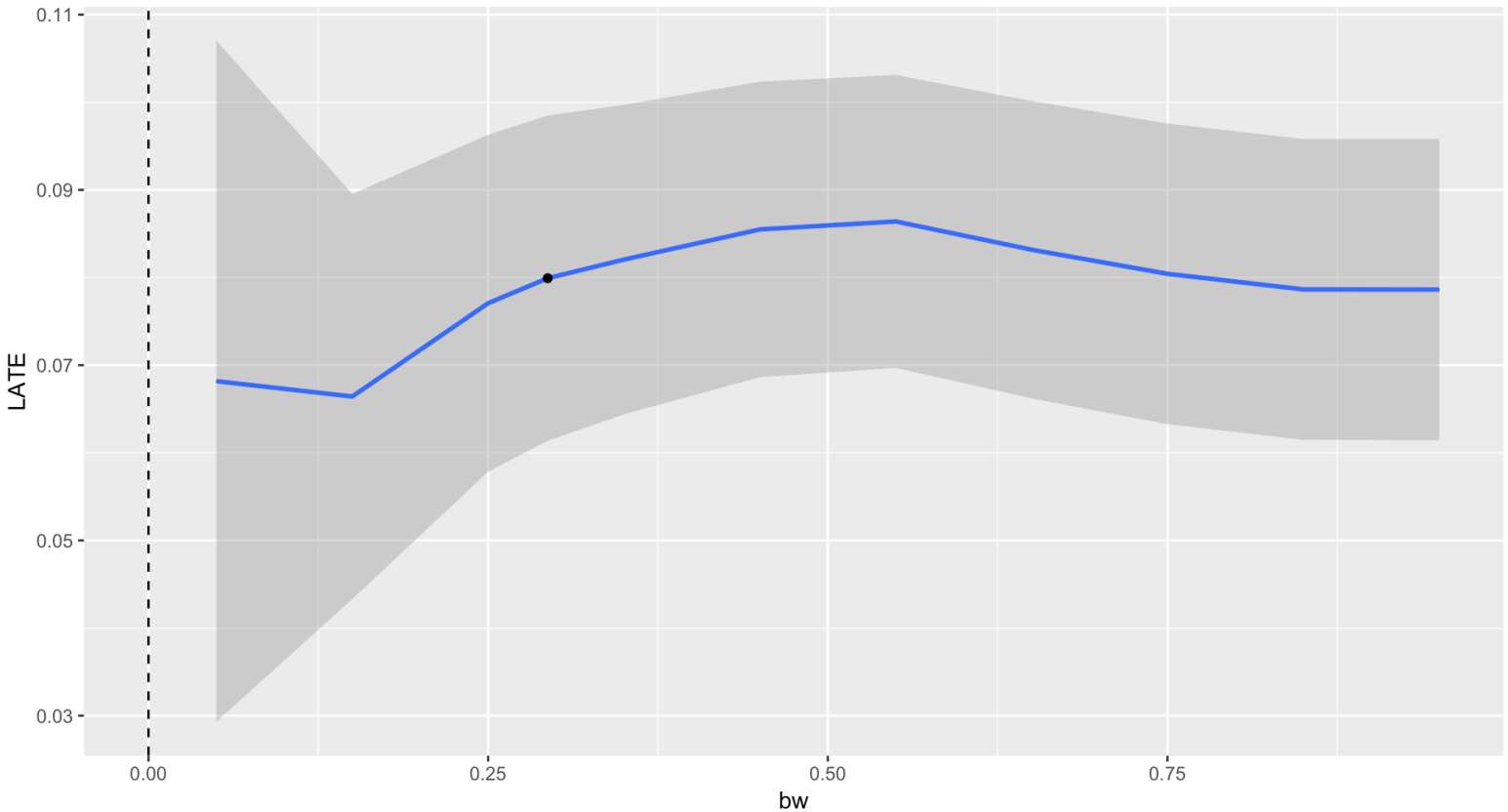
```
### RDD regression: nonparametric local linear###
        Bandwidth:  0.2938561
        Number of obs: 3200 (left: 1594, right: 1606)

        Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.97755 -0.06721 -0.00497  0.04504  0.93761

        Coefficient:
  Estimate Std. Error z value  Pr(>|z|)
D 0.079924   0.009465  8.4443 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        Local R squared: 0.3563
```

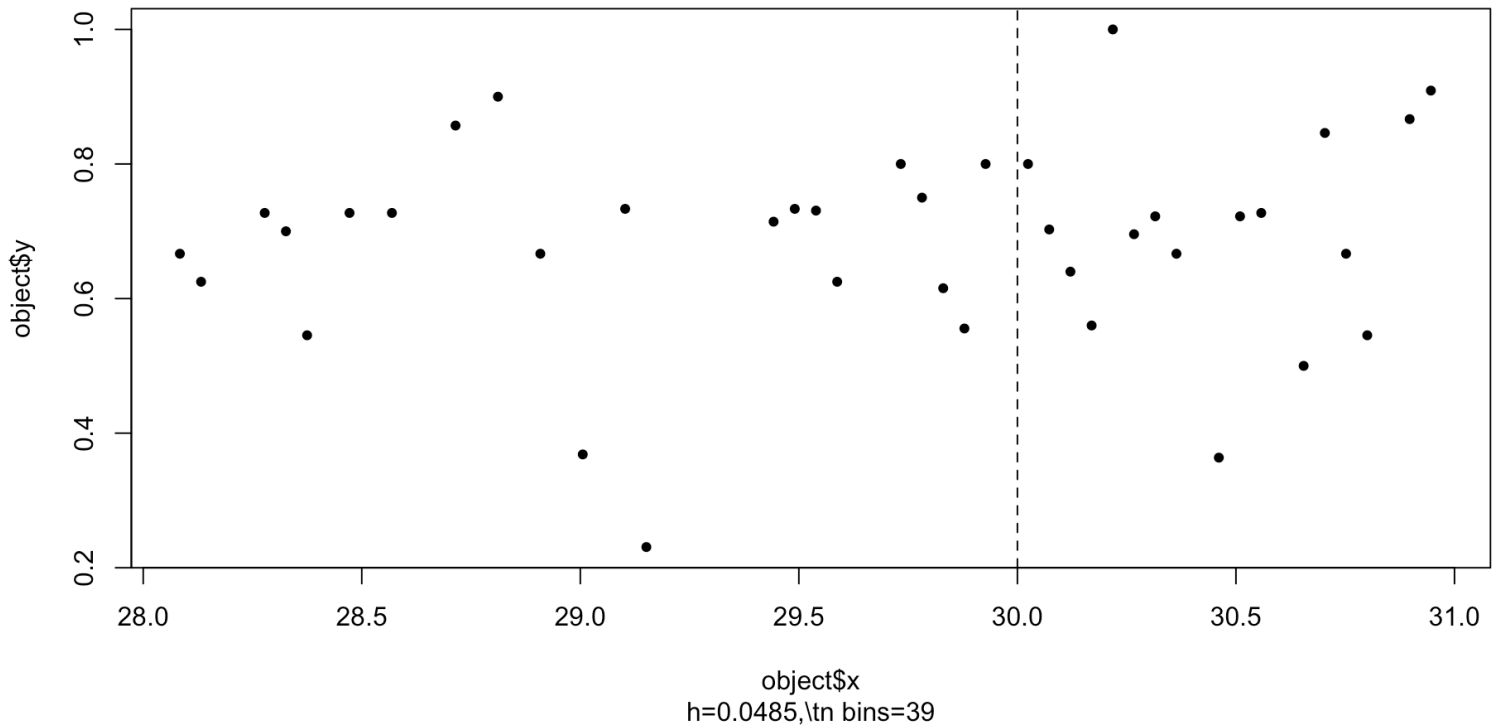
Example: Incumbency (Bandwidth Sensitivity)



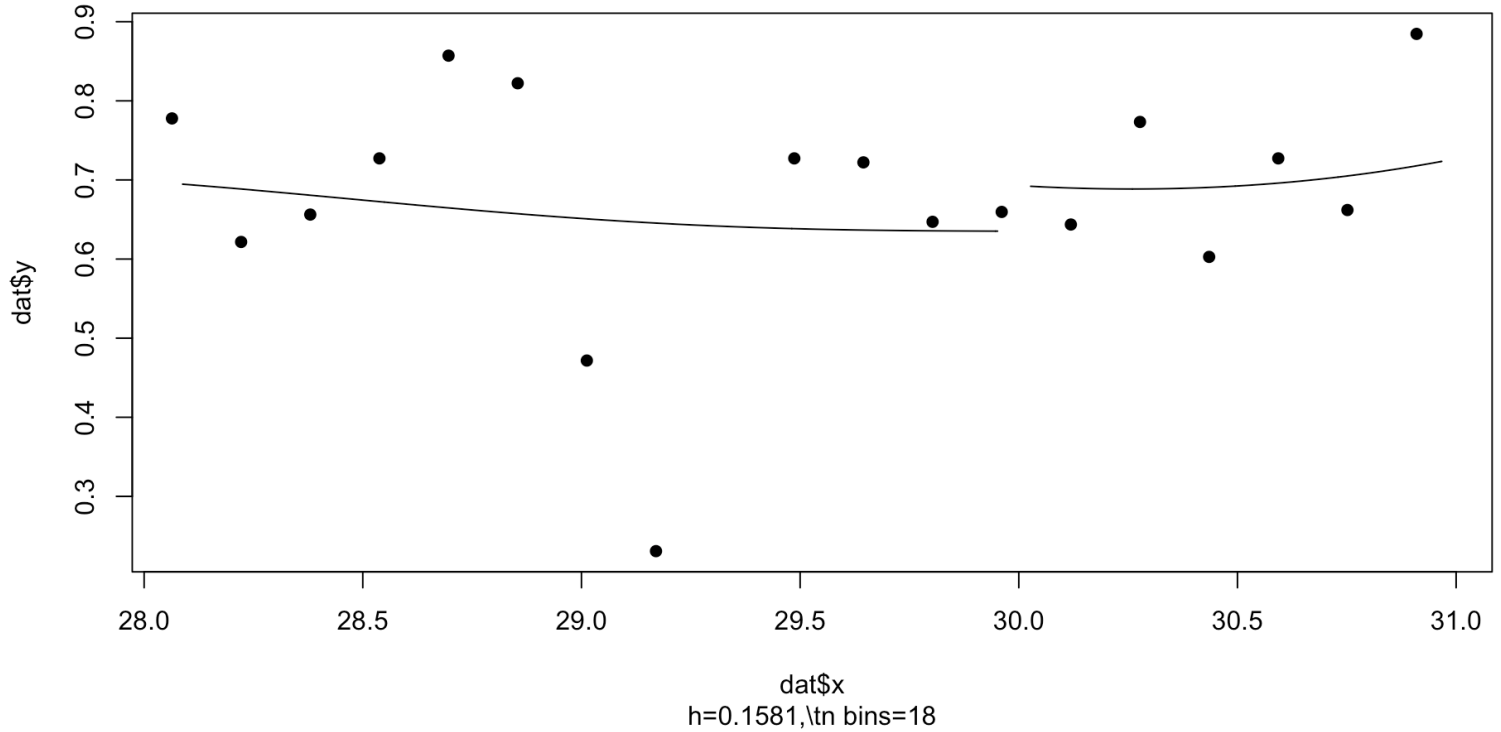
Things Aren't Always so Pretty (Example)

- The example we've just seen is a particularly well behaved example.
- I just want to quickly show you how the majority of real data applications actually look!
- We have data from a development project in Morocco. The variable of interest is the decision to contribute to a public good or not. The forcing variable is percentage in poverty, which represents the percentage of households in a commune living below the poverty threshold.
- As part of the project, communes with more than 30% of households below the poverty threshold were allowed a say in how the funds were distributed. The cutoff point for our analysis is therefore 30. (Note, observations are grouped by the poverty variable)

Example: Development ('indh')



Example: Development



Example: Development

```
### RDD regression: nonparametric local linear###
        Bandwidth:  0.790526
        Number of obs: 460 (left: 139, right: 321)

        Weighted Residuals:
     Min       1Q   Median       3Q      Max
-1.15126 -0.64666  0.12132  0.32261  0.43399

        Coefficient:
  Estimate Std. Error z value Pr(>|z|)
D 0.144775   0.095606  1.5143     0.13

        Local R squared: 0.007044
```

Fuzzy RDD

- When the treatment rule is not a deterministic function of x, but is instead a stochastic function of x, we have a **fuzzy RDD**.
- That is, if the **probability** of treatment increases at the cutoff point, then we must use fuzzy RDD techniques.
- In particular, the discontinuity becomes an instrumental variable for treatment status instead of deterministically switching treatment on or off.
- In comparison to the sharp RDD case, we now have the following treatment rule:

$$Pr\left[D_i = 1 \middle| x_i\right] = \begin{cases} g_1(x_i) & \text{if } x_i \geq x_0 \\ g_0(x_i) & \text{if } x_i < x_0, \end{cases}$$

where $g_0(x_0) \neq g_1(x_0)$. Typically, $g_0(\cdot)$ and $g_1(\cdot)$ are polynomials of some chosen order.

Fuzzy RDD - Probability of Treatment

- We can model the probability of treatment as

$$E\left[D_i|x_i\right] = P\left[D_i = 1|x_i\right] = g_0(x_i) + \left\{g_1(x_i) - g_0(x_i)\right\} T_i,$$

where $T_i = \mathcal{I}(x_i \geq x_0)$ is the point of discontinuity.

- Suppose

$$\begin{aligned} g_0(x_i) &= \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 \\ g_1(x_i) &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \end{aligned}$$

then we can write

$$g_1(x_i) - g_0(x_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2.$$

- This gives

$$E\left[D_i|x_i\right] = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \gamma_0 T_i + \gamma_1 x_i T_i + \gamma_2 x_i^2 T_i$$

Fuzzy RDD - 2SLS

- The previous equation is the first stage of a 2SLS estimator. We first predict the probability of treatment, given by the fitted value of $E[D_i|x_i]$.
- Note that we could just use T_i as a single instrument for fuzzy RDD.
- The rest of the estimation proceeds as with normal 2SLS where the fitted value (or just T_i), is used as an instrument for D_i in the previous sharp RDD regressions.
- The approach can be made nonparametric by conducting the estimation within a neighbourhood of the cutoff (we won't cover that here).

Fuzzy RDD - Example

- In Problem Set 8 you will go through the following example.
- Angrist and Lavy (1999) use Fuzzy RDD to estimate the effect of class size on student test scores.
- Class size in Israeli schools is capped at 40. Students in a grade with up to 40 students can expect to be in classes as large as 40, but grades with 41 students are split into two classes, grades with 81 students are split into three classes, and so on.

Summary

- We have looked in detail at the sharp regression discontinuity design approach to identifying causal effects.
- We briefly looked at a method to test the key assumption of the method.
- We looked at how to estimate the model in a nonlinear way.
- We also showed how nonparametric estimation works in this context.
- Finally, we considered the fuzzy RDD approach and how we can use 2SLS to estimate the model.