Section 5 : Regression Discontinuity I

Andrew Bertoli

2 October 2013

Roadmap

1. Issues from the Last Homework
2. Permutation Inference
3. Regression Discontinuity
4. Questions

Last Homework

When estimating the variance, you are supposed to divide by $\frac{1}{n-1}$ instead of $\frac{1}{n}$.

Why: We are estimating $E[(X - \mu_x)^2]$, and we use the average of our sample ($\bar{X}$) for μ_x. However, our points will be slightly closer to their own average than the true population mean. So we want our estimate to be a little larger than it would be if we knew the true value of μ_x. Thus, we multiply $\sum_i (X_i - \bar{X})^2$ by $\frac{1}{n-1}$ instead of $\frac{1}{n}$.

So where is the $n - 1$ in this equation?

$$\widehat{SE}(\hat{\tau}) = \frac{\hat{\sigma}_T^2}{m} + \frac{\hat{\sigma}_C^2}{n-m}$$

Last Homework

We estimate $\hat{\sigma}^2_T$ and $\hat{\sigma}^2_C$ using the standard method of calculating the sample variance:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$$

Last Homework

There is only one design matrix.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \dots & x_{1p} \\ 1 & x_{22} & x_{23} & \dots & x_{2p} \\ 1 & x_{32} & x_{33} & \dots & x_{3p} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ . \\ . \\ \hat{\beta}_{p-1} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ . \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

Last Homework

Say we run an experiment where we look at one treatment and one outcome variable. If there is no treatment effect, what is the probability that our results will be significant at the 5% level?

5%

Last Homework

Now imagine that we look at k outcomes. Honesty is not our top priority, so we plan to only report one of the outcomes in our study and pretend like we did not look at any others. How many chances do we have to get a p-value below the 5% level? (Note: These chances will not be independent because of correlations in the data.)

k

Last Homework

Each control variable can be in the regression or not.

If the control variable is not in the regression, code it as 0, and if it is, code it as 1.

So we can write any possible control combination as sequences of 0's and 1's.

For instance, if we have 5 control variables, 00110 means that we control for just the third and fourth variables.

This means that the number of possible combinations of control variables is equal to the number of sequences of 0's and 1's of length j, which equals 2^j.

Last Homework

Now say we are doing an observational study like regression, and we have one treatment variable, j possible control variables, and k possible outcomes. How many chances will we have to get a p-value below the 5% level for our treatment variable?

$k \cdot 2^j$

Last Homework

How many chances do we get if we are also willing to control for 2 factor interactions?

Last Homework

So this problem is easy if we just think of each interaction as a new control variable. So there are $\binom{j}{2}$ possible interactions.

$$\text{Number of combinations} = k \cdot 2^j \cdot 2^{\binom{j}{2}}$$

$$= k \cdot 2^j \cdot 2^{\frac{j!}{(j-2)!2!}}$$

$$= k \cdot 2^j \cdot 2^{\frac{j(j-1)}{2}}$$

$$= k \cdot 2^{j+\frac{j(j-1)}{2}}$$

Last Homework

How many chances do we get if we are also willing to control for 2 factor interactions?

$k \cdot 2^{j+\frac{j(j-1)}{(2}}$

Last Homework

Some people counted the number of interactions using Guass's formula: $\frac{n(n+1)}{2}$

Their idea was that if we have j covariates, we can match the first to the $j-1$ others, the second to the $j-2$ others (excluding the first control), and so on until we get to the last control.

In other words, Covariate 1 gets matched to Covariate 2, Covariate 3, ... , Covariate j. Then Covariate 2 gets matched to Covariate 3, Covariate 4,..., Covariate j (we leave out Covariate 1 because it was already matched to it). We continue this process until we get down to Covariate j-1 and Covariate j.

Last Homework

This is correct, but the answer is not $1 + 2 + ... + j$. It is $1 + 2 + ... + j - 1$. We match Covariate 1 to j-1 units, not j units.

So the correct answer is

Number of interactions $= \frac{[j-1]([j-1]+1)}{2} = \frac{j(j-1)}{2}$

which agrees with our previous answer when we used $\binom{n}{j}$.

Last Homework

If today we are feeling especially creative and are willing to swap our treatment variable with any of our control variables, how many chances will we have to get a p-value below the 5% level?

Last Homework

If today we are feeling especially creative and are willing to swap our treatment variable with any of our control variables, how many chances will we have to get a p-value below the 5% level?

$$(j+1) \cdot k \cdot 2^{j+\frac{j(j-1)}{(2}}$$

Last Homework

For Part (e), how many chances do we have if we have 3 possible controls and 3 outcomes? What about 5 possible controls and 1 outcome? (Remember that these chances are not independent.)

Last Homework

For Part (e), how many chances do we have if we have 3 possible controls and 3 outcomes? What about 5 possible controls and 1 outcome? (Remember that these chances are not independent.)

$(3+1) \cdot 2^{3+3\cdot(3-1)/2} \cdot 3 = 768$

Last Homework

For Part (e), how many chances do we have if we have 3 possible controls and 3 outcomes? What about 5 possible controls and 1 outcome? (Remember that these chances are not independent.)

$(3+1) \cdot 2^{3+3\cdot(3-1)/2} \cdot 3 = 768$

$(5+1) \cdot 2^{5+5\cdot(5-1)/2} \cdot 3 = 196,608$

Permutation Inference

Imagine someone ran an experiment, and we want to do permutation inference to test their results.

We need three things:

(1) The vector of the treatment assignment $\mathbf{T} = (1, 0, 0, 1, ..., 0)$

(2) The vector of outcomes $\mathbf{Y} = (1.34, 2.31, 5.56, 1.14, ..., 0.55)$

(3) A description of how they randomized

Permutation Inference

Steps

(1) Compute the real test statistic (usually the difference in means between treatment and control group).

(2) Create a large number of new treatment assignments using their randomization technique.

(3) Use $\mathbf{Y}$ to calculate the test statistics for all these new randomizations.

(4) If the treatment had no effect whatsoever, the new test statistics should be similar to the real test statistic.

(5) The p-value is the proportion of the new test statistics that are as extreme or more extreme than the real test statistic.

Regression Discontinuity

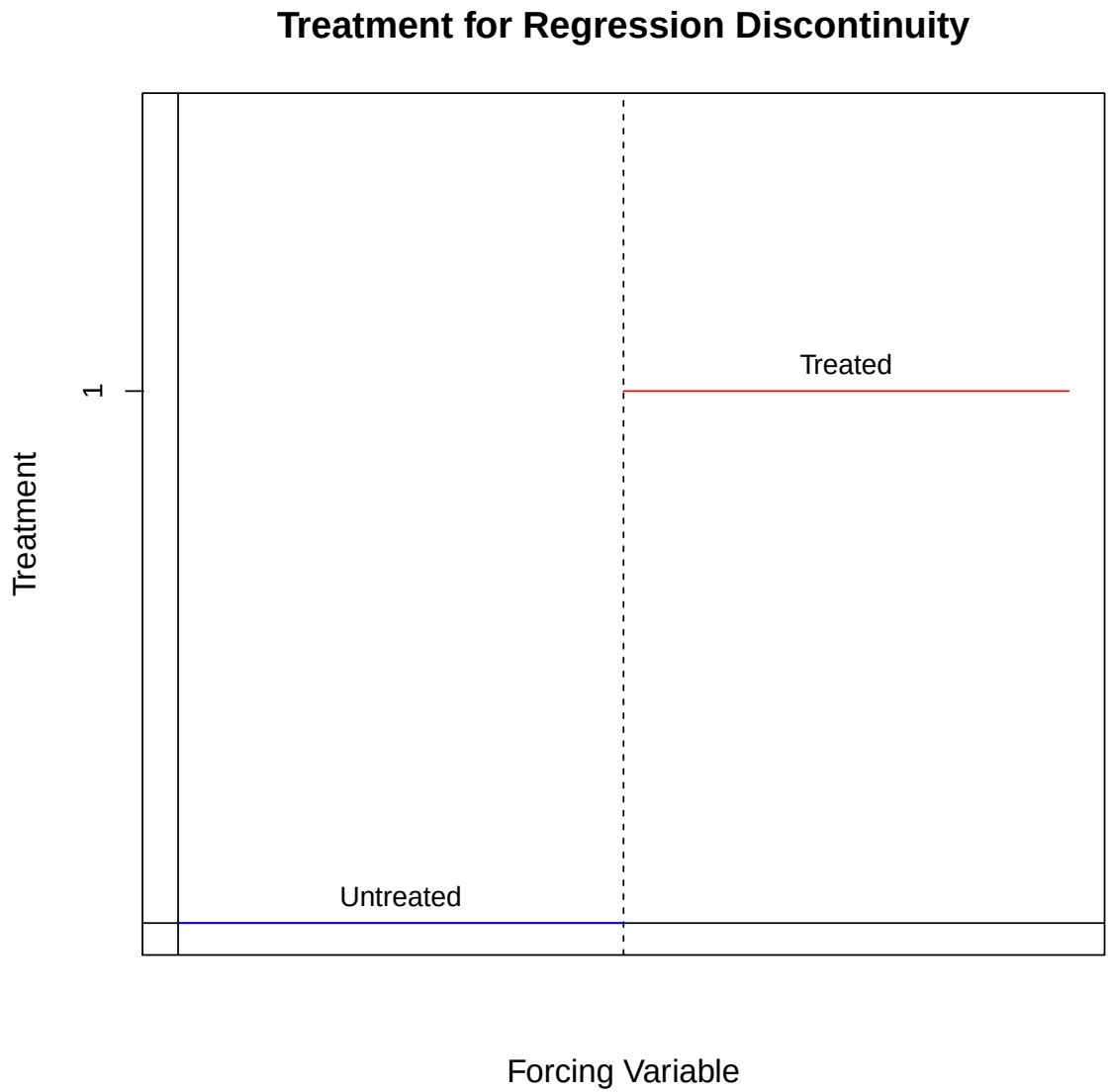
Basic Idea

1. Can be used when a treatment is given to units that score above a cut-point in a scoring system.

2. The idea is to estimate the treatment effect by comparing the units just above and just below the cut-point.

3. Allows researchers to estimate the Local Average Treatment Effect (LATE) at the cut-point.

Regression Discontinuity



Regression Discontinuity

Two Approaches to Inference

1. As-if randomness in a window around the cut-point
2. Continuity in potential outcomes at the cut-point

Regression Discontinuity

As-if randomness

1. Pick a window around the cut-point
2. Treat the data in the window like it is experimental data

Regression Discontinuity

Weaknesses of the as-if randomness approach

1. Because treatment assignment is not really random, the results will almost certainly be biased

2. Where you set the window can affect your results

Regression Discontinuity

Suggestions

1. You can often use a difference-in-differences estimator to help eliminate bias. This will also probably decrease the standard errors of your tests.

1. As a robustness check, use regression to control for a number of covariates.

2. Report the results for different windows around the cut-point.

Regression Discontinuity

Continuous Potential Outcomes

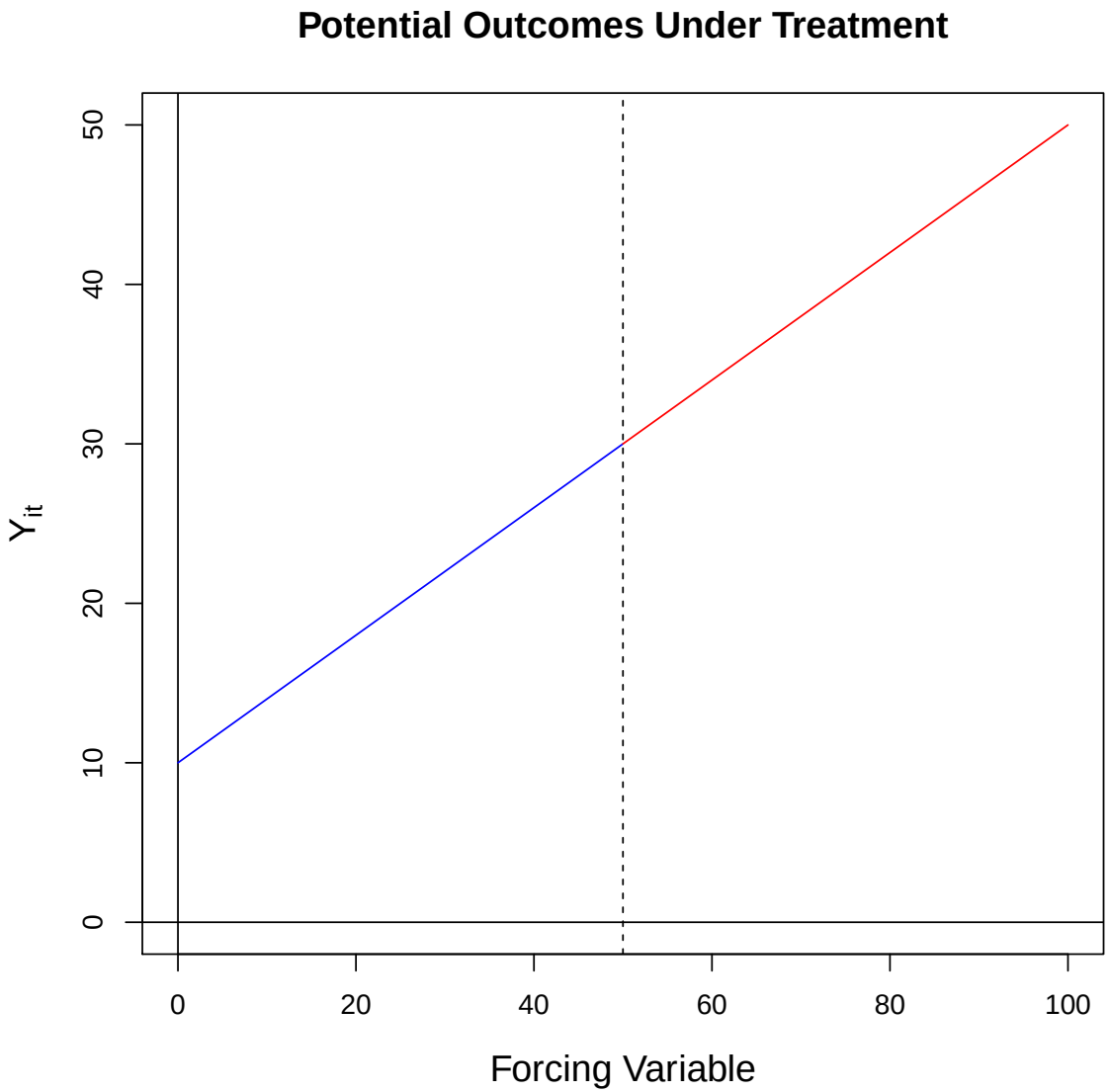
1. This approach does not require as-if randomness within an RD window.

2. Instead, it assumes that potential outcomes are continuous at the cut-point.

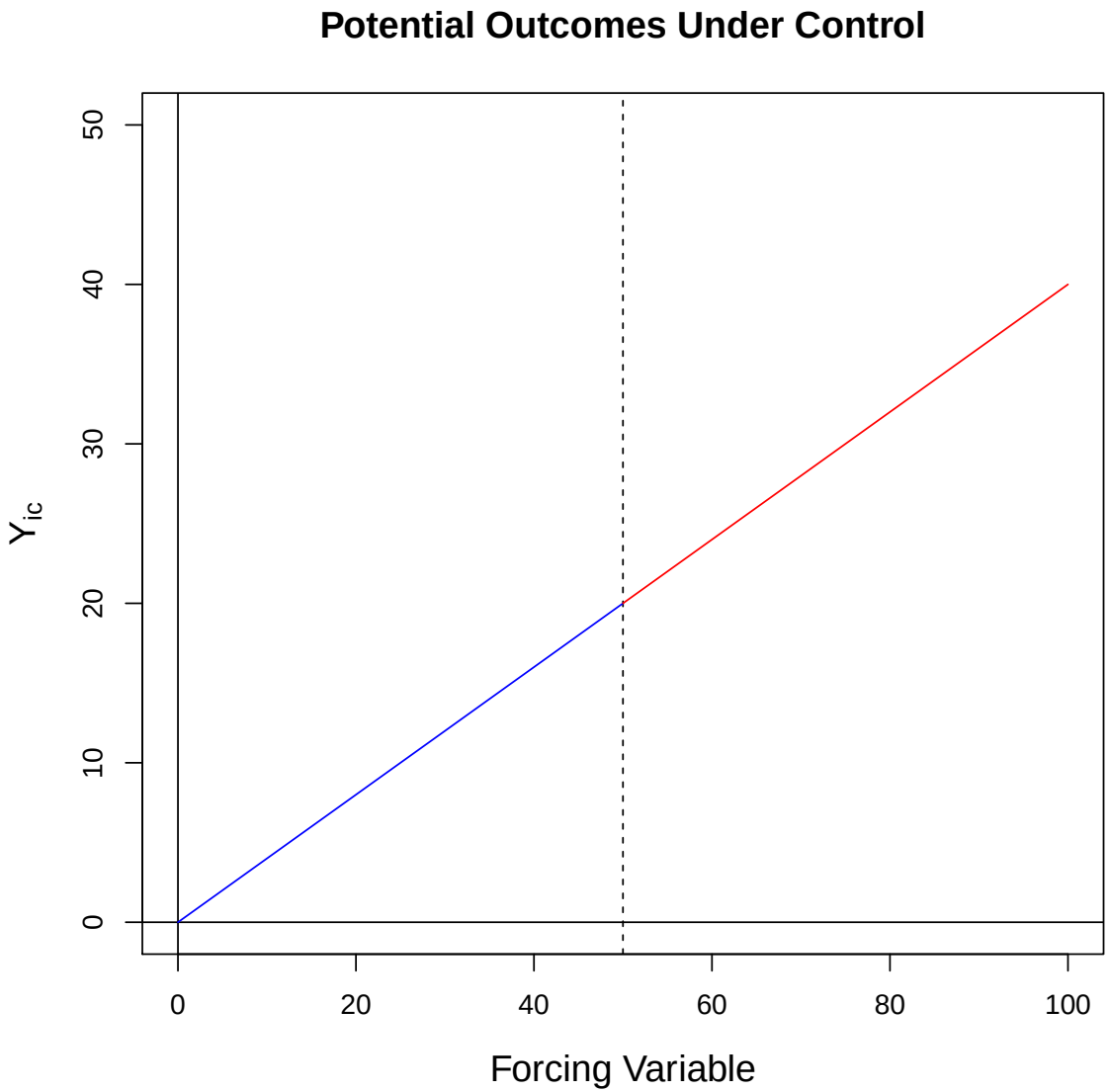
3. The data on the left is used to estimate Y_{ic} at the cut-point, and the data on the right is used to estimate Y_{it} at the cut-point.

4. The difference between these values is the estimated LATE at the cut-point.

Regression Discontinuity



Regression Discontinuity



Regression Discontinuity

Continuous Potential Outcomes

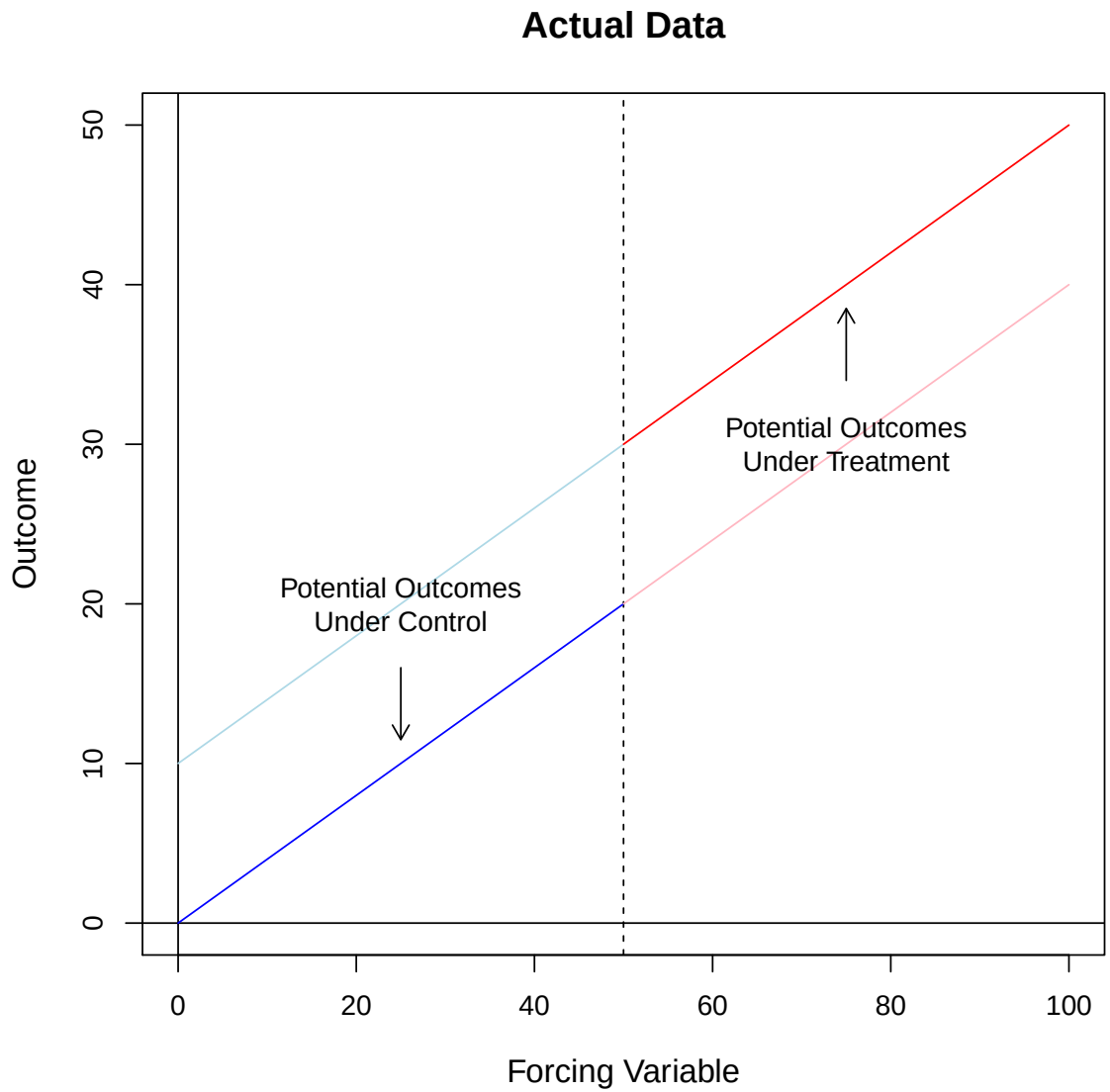
1. This approach does not require as-if randomness within an RD window.

2. Instead, it assumes that potential outcomes are continuous at the cut-point.

3. The data on the left is used to estimate Y_{ic} at the cut-point, and the data on the right is used to estimate Y_{it} at the cut-point.

4. The difference between these values is the estimated LATE at the cut-point.

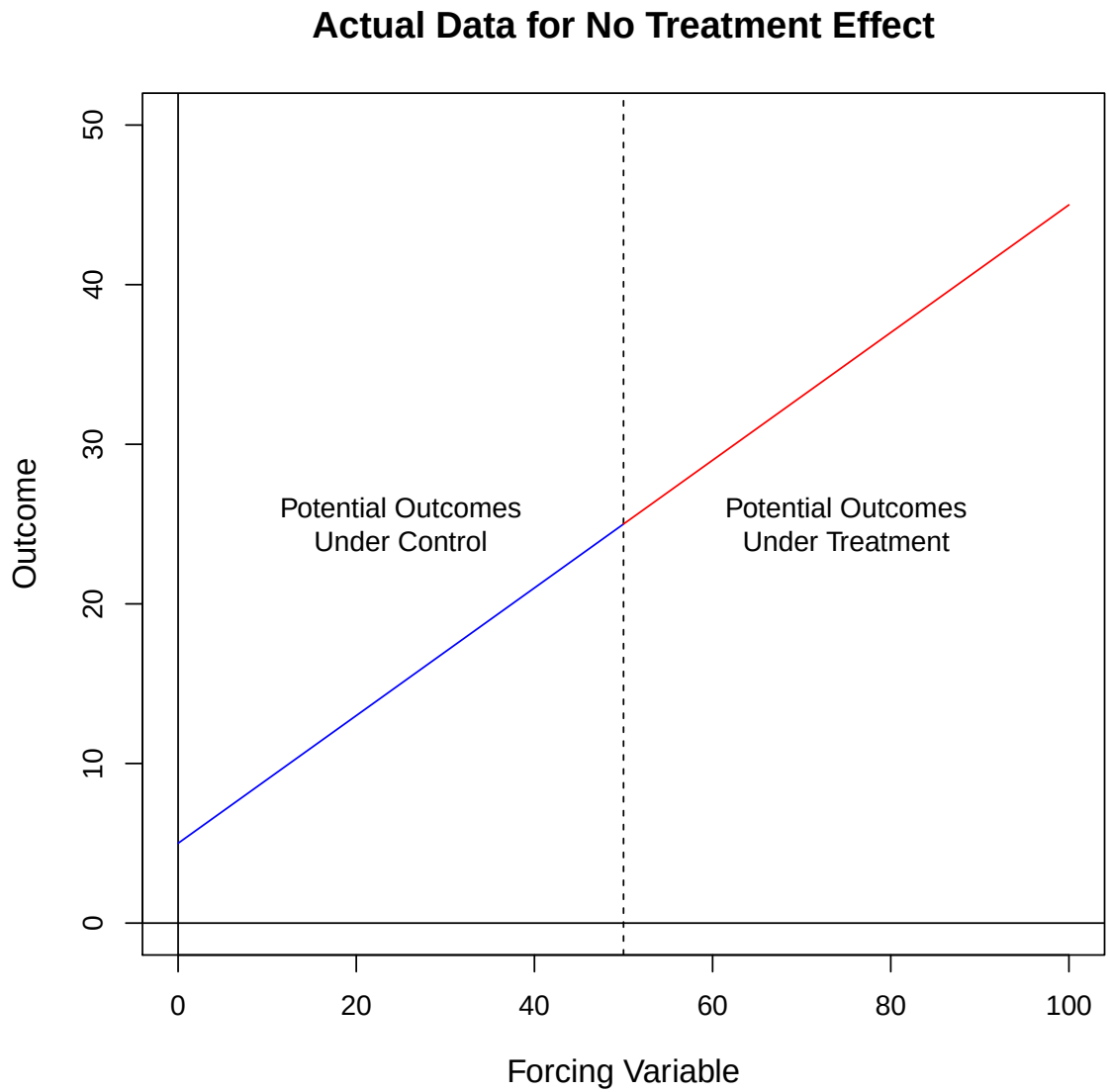
Regression Discontinuity



Regression Discontinuity

If there is no treatment effect, than each unit's potential outcome under treatment is equal to its potential outcome under control ($Y_{it} = Y_{ic}$). Since the Y_{it}'s and Y_{ic}'s are continuous at the cut-point, then the line in the actual data should also be continuous at the cut-piont.

Regression Discontinuity



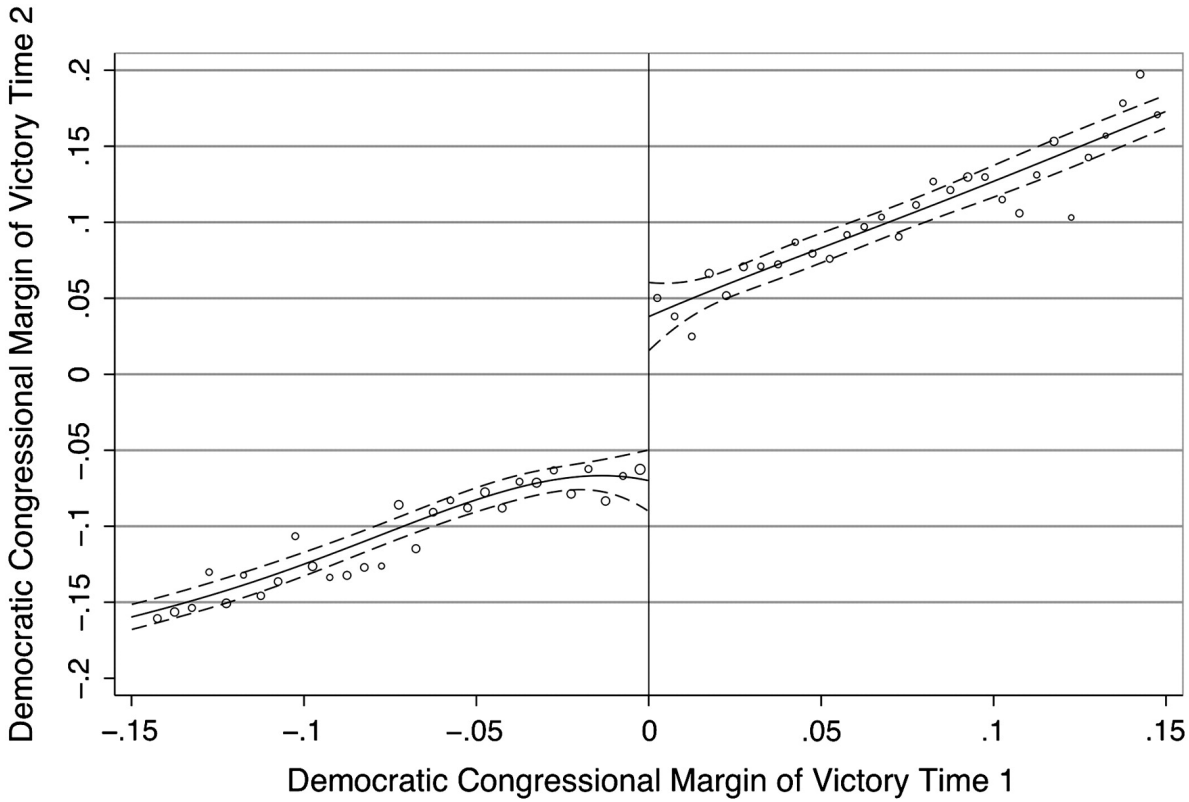
Regression Discontinuity

Weaknesses of Continuous Potential Outcomes

1. Requires modeling assumptions to construct the regression lines
2. These assumptions can influence the results

Regression Discontinuity

Creating the Graph



Regression Discontinuity

Constructing the Regression Lines

1. A common approach is to use local linear regression.
2. The bandwidth can be found using Caughey's code.

Regression Discontinuity

Estimating the Confidence Interval

1. The 95% confidence interval is found by bootstrapping

2. Starting with the left side of the cut-point, randomly sample (with replacement) from the points on the left. The size of the new sample should be the same as the number of points on the left.

3. Construct a regression line for this new sample

4. Repeat this process several thousand times. This will result in thousands of regression lines. Each regression line will be defined by a large number of (x, y) coordinates. Thus, each x will be associated with a large number of y's.

5. For each x, find the lower 2.5% and upper 97.5% quantile for the y's. These points will define the lower and upper bounds of the 95% confidence interval.

Regression Discontinuity

Common Forcing Variables

1. Score on a test
2. Vote share in an election
3. Age
4. Distance from border
5. Size or Population

Regression Discontinuity

Common Forcing Variables

1. **Score on a test**
2. Vote share in an election
3. Age
4. Distance from border
5. Size or Population

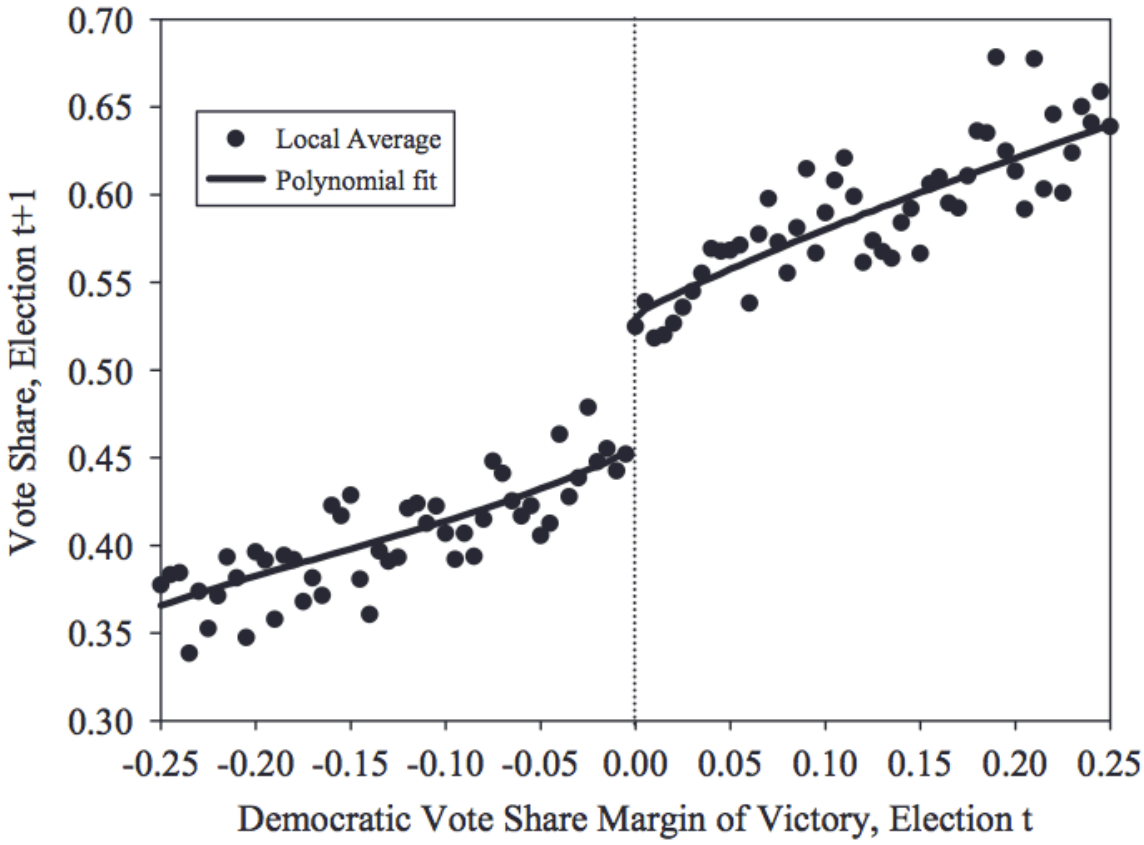
Regression Discontinuity

Lee (2008)

Question: How large is the incumbency advantage for the US Congress?

Design: Look at democratic vote share in cases where democratic candidates barely won and lost the previous election.

Regression Discontinuity



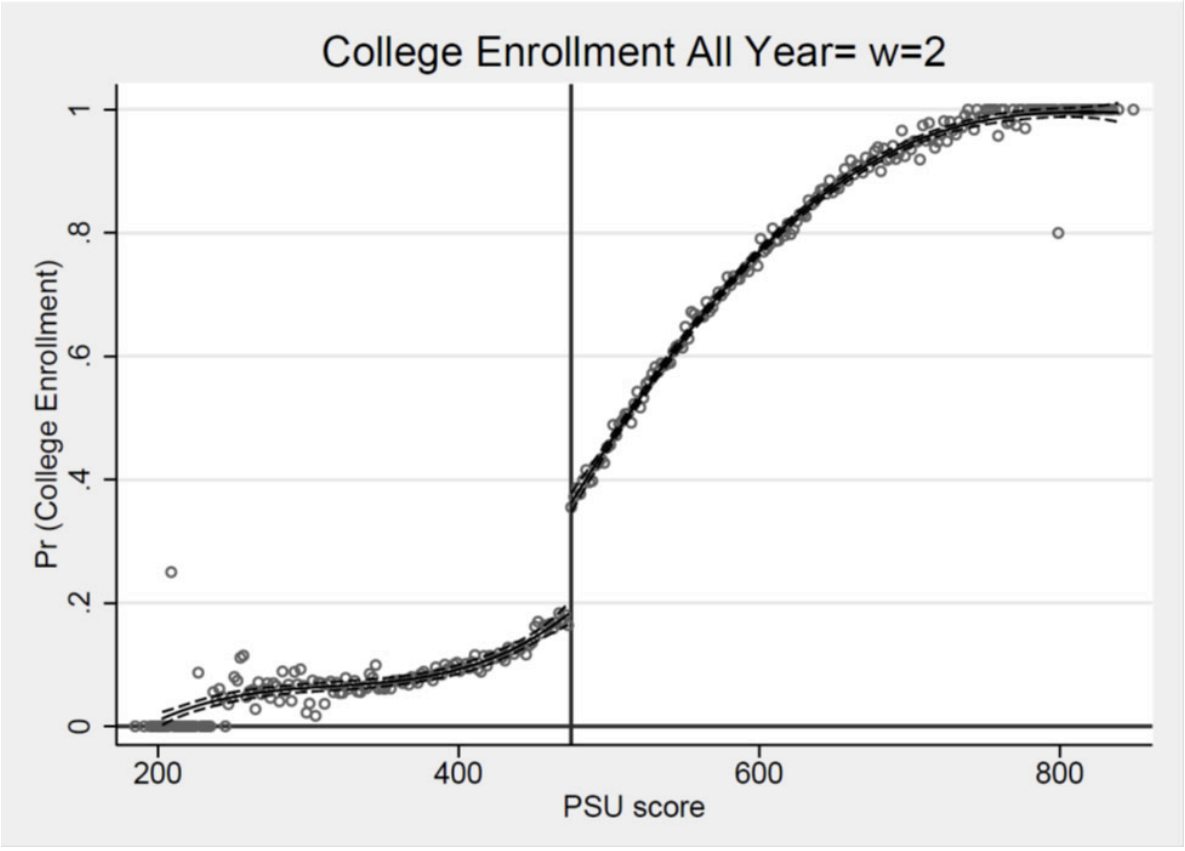
Regression Discontinuity

Solis (2011)

Question: How important are financial constraints in preventing people from going to college?

Design: Exploit a finical aid program in Chile that gave scholarships to students who achieved a certain score on a test.

Regression Discontinuity



Regression Discontinuity

Common Forcing Variables

1. Score on a test
2. Vote share in an election
3. **Age**
4. Distance from border
5. Size or Population

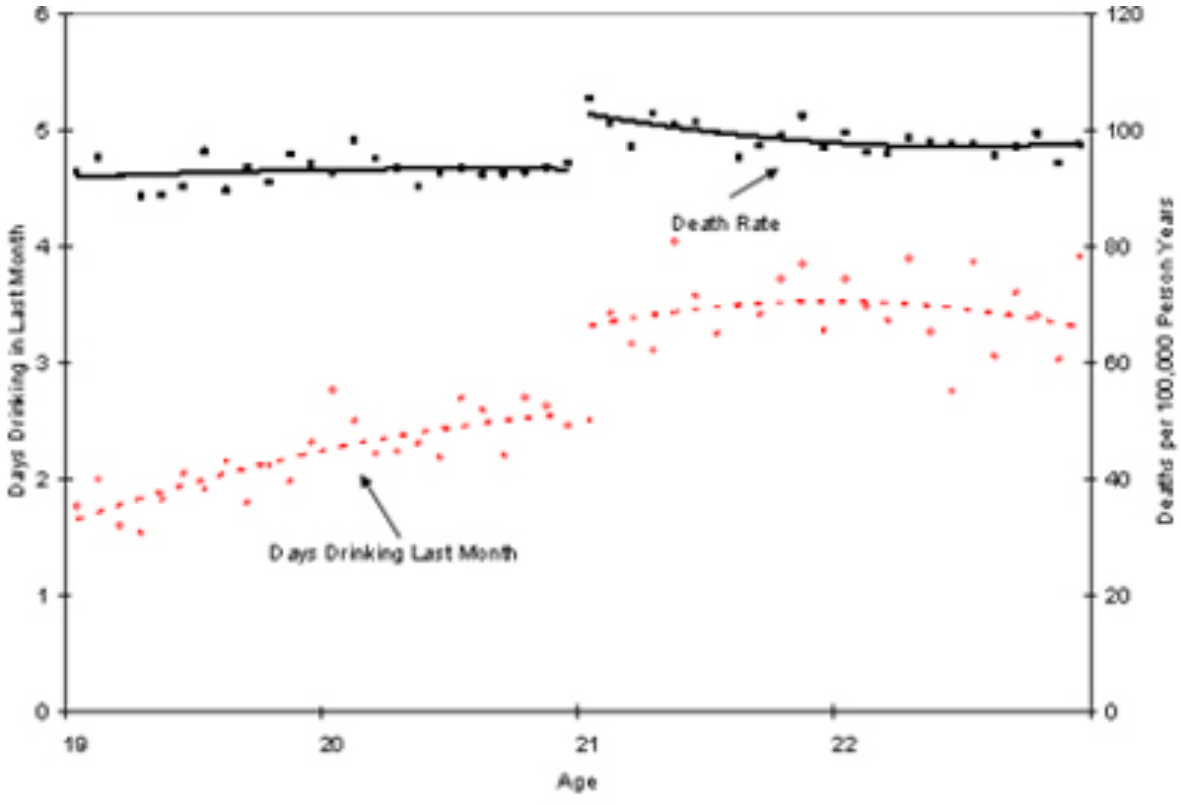
Regression Discontinuity

Carpenter and Dobkin (2009)

Question: Do laws against drinking prevent deaths from alcohol?

Design: Look at death rates from alcohol for people just above and just below the age of 21.

Regression Discontinuity



Regression Discontinuity

Common Forcing Variables

1. Score on a test
2. Vote share in an election
3. Age
4. **Distance from border**
5. Size or Population

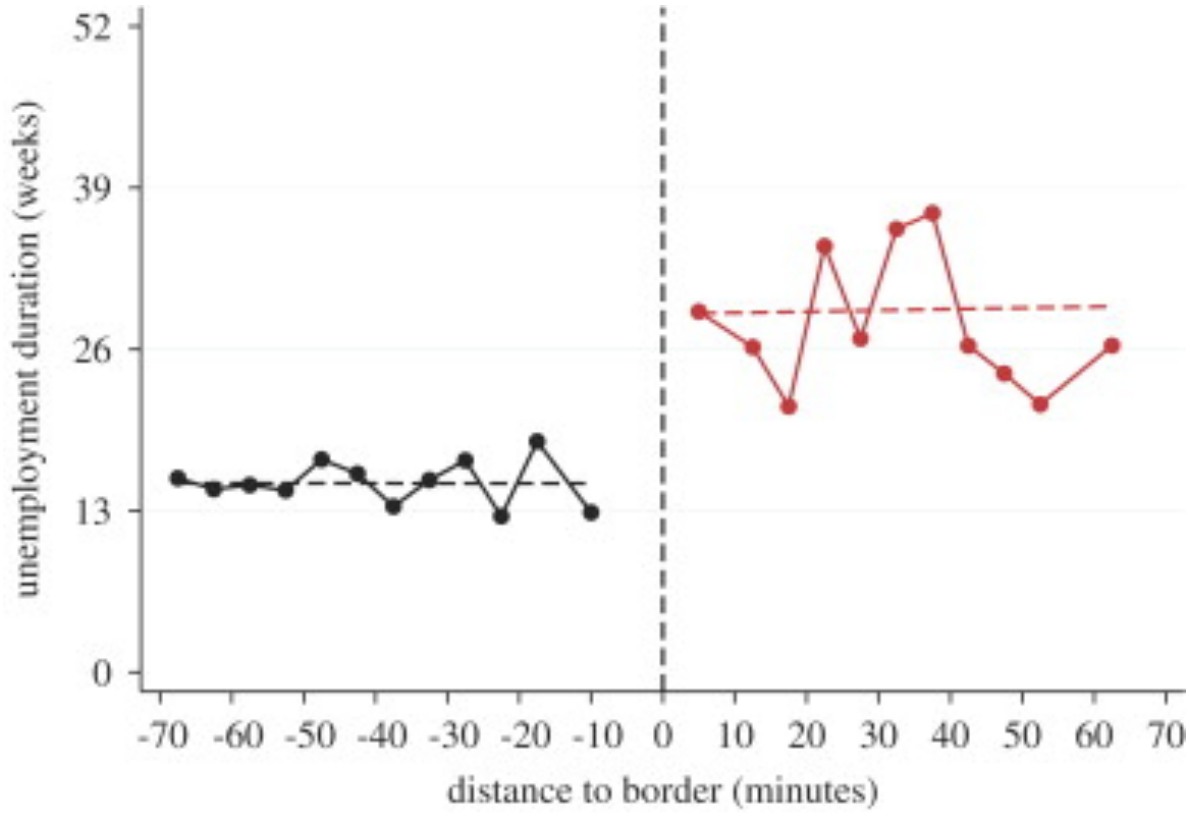
Regression Discontinuity

Lalive (2009)

Question: How do increased unemployment benefits affect the length of unemployment?

Design: Compare the average length of unemployment in two regions in Austria after one extends the maximum duration of unemployment benefits from 30 weeks to 209 weeks.

Regression Discontinuity



Discontinuity at threshold = 13.622; with std. err. = 2.988.

Regression Discontinuity

Common Forcing Variables

1. Score on a test
2. Vote share in an election
3. Age
4. Distance from border
5. **Size or Population**

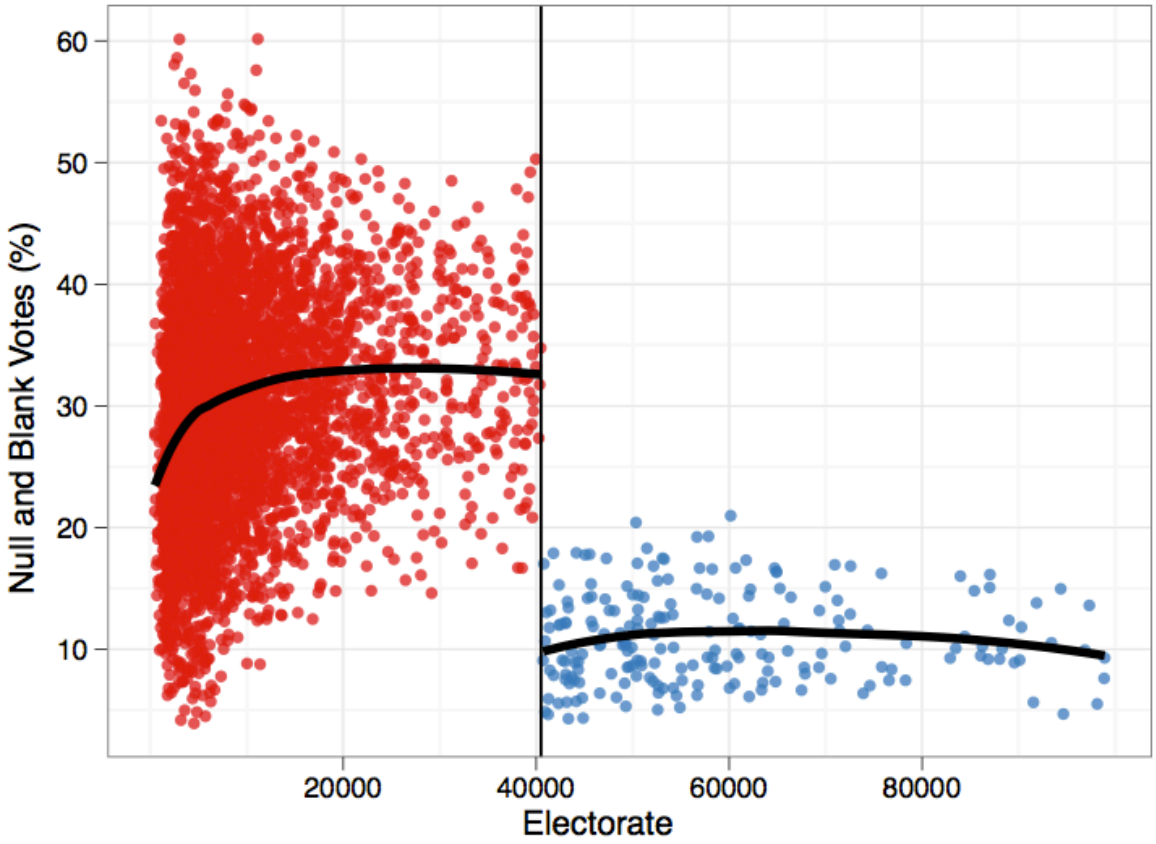
Regression Discontinuity

Hidalgo (2010)

Question: Can electronic voting machines make elections more fair by decreasing fraud and reducing the number of unreadable ballots?

Design: Exploit a 1998 policy in Brazil where municipalities with a 1996 electorate larger than 40,500 people were required to adopt electronic voting machines.

Regression Discontinuity



Regression Discontinuity

Potential Problems

1. Sorting at the cut-point
2. The LATE is very different than the ATE

Regression Discontinuity

Sorting at the Cut-Point

1. It is very important to show the results for the outcome variable before treatment, as well as other important covariates.

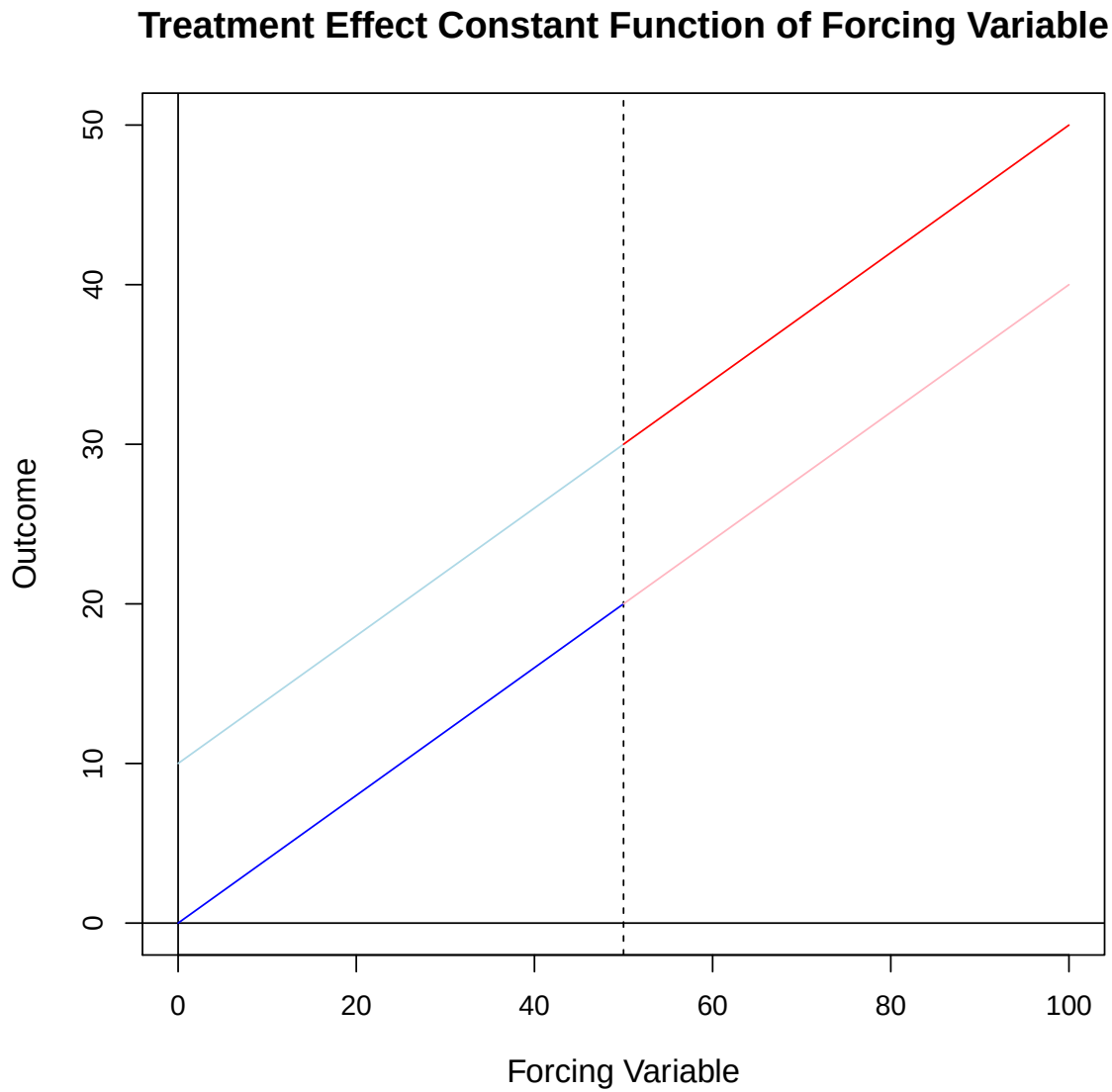
2. Remember that if sorting did not happen, the p-values for covariates should be distributed uniformly between 0 and 1 (roughly speaking).

Regression Discontinuity

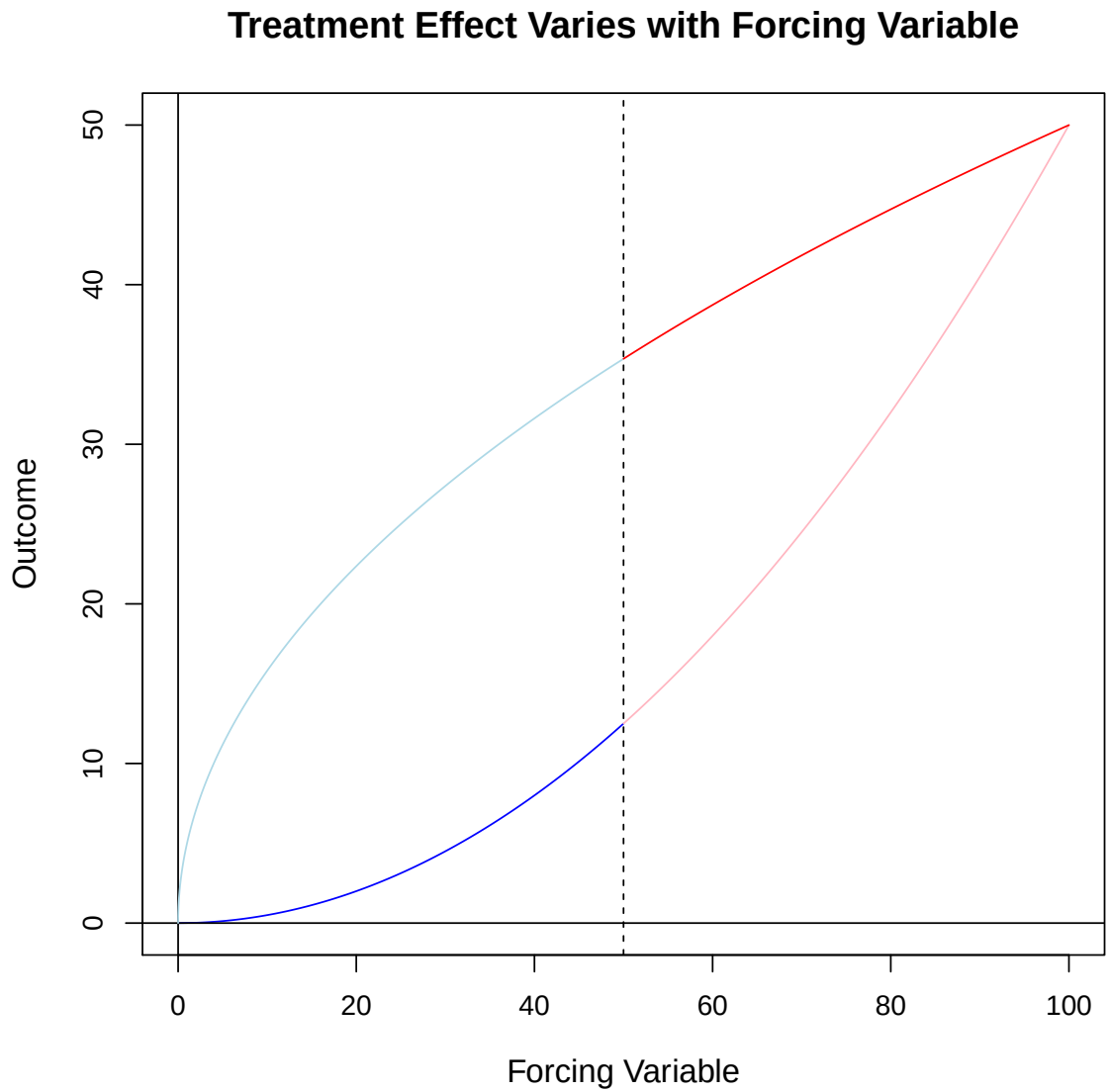
Potential Problems

1. Sorting at the cut-point
2. The LATE is very different than the ATE

Regression Discontinuity



Regression Discontinuity



Regression Discontinuity