Chapter 8: Regression Discontinuity

Jonathan Roth

Mathematical Econometrics I
Brown University
Fall 2023

Motivation

- The key challenge in causal inference is finding a control group that is comparable to treated units in all ways except for treatment status

Motivation

- The key challenge in causal inference is finding a control group that is comparable to treated units in all ways except for treatment status
- We've seen so far how to estimate causal effects when:

Motivation

- The key challenge in causal inference is finding a control group that is comparable to treated units in all ways except for treatment status
- We've seen so far how to estimate causal effects when:
  - The treatment is literally randomly assigned

Motivation

- The key challenge in causal inference is finding a control group that is comparable to treated units in all ways except for treatment status
- We've seen so far how to estimate causal effects when:
  - The treatment is literally randomly assigned
  - The treatment is as-good-as-randomly assigned

Motivation

- The key challenge in causal inference is finding a control group that is comparable to treated units in all ways except for treatment status
- We've seen so far how to estimate causal effects when:
  - The treatment is literally randomly assigned
  - The treatment is as-good-as-randomly assigned
  - Selection bias is constant over time (parallel trends)

Motivation

- The key challenge in causal inference is finding a control group that is comparable to treated units in all ways except for treatment status
- We've seen so far how to estimate causal effects when:
  - The treatment is literally randomly assigned
  - The treatment is as-good-as-randomly assigned
  - Selection bias is constant over time (parallel trends)
  - We have an as-good-as-random and excludable instrument

Motivation

- The key challenge in causal inference is finding a control group that is comparable to treated units in all ways except for treatment status
- We've seen so far how to estimate causal effects when:
  - The treatment is literally randomly assigned
  - The treatment is as-good-as-randomly assigned
  - Selection bias is constant over time (parallel trends)
  - We have an as-good-as-random and excludable instrument
- Today we will see another strategy: **regression discontinuity (RD)**

Motivation

- The key challenge in causal inference is finding a control group that is comparable to treated units in all ways except for treatment status
- We've seen so far how to estimate causal effects when:
    - The treatment is literally randomly assigned
    - The treatment is as-good-as-randomly assigned
    - Selection bias is constant over time (parallel trends)
    - We have an as-good-as-random and excludable instrument
- Today we will see another strategy: **regression discontinuity (RD)**
- The idea behind RD is that sometimes treatment status is (partially) determined by whether a score is above/below a threshold

Motivation

- The key challenge in causal inference is finding a control group that is comparable to treated units in all ways except for treatment status
- We've seen so far how to estimate causal effects when:
  - The treatment is literally randomly assigned
  - The treatment is as-good-as-randomly assigned
  - Selection bias is constant over time (parallel trends)
  - We have an as-good-as-random and excludable instrument
- Today we will see another strategy: **regression discontinuity (RD)**
- The idea behind RD is that sometimes treatment status is (partially) determined by whether a score is above/below a threshold
  - For example, whether you are admitted to a school or university may depend on whether your test score exceeds a certain cutoff score

Motivation

- The key challenge in causal inference is finding a control group that is comparable to treated units in all ways except for treatment status
- We've seen so far how to estimate causal effects when:
  - The treatment is literally randomly assigned
  - The treatment is as-good-as-randomly assigned
  - Selection bias is constant over time (parallel trends)
  - We have an as-good-as-random and excludable instrument
- Today we will see another strategy: **regression discontinuity (RD)**
- The idea behind RD is that sometimes treatment status is (partially) determined by whether a score is above/below a threshold
  - For example, whether you are admitted to a school or university may depend on whether your test score exceeds a certain cutoff score
- In such cases, RD compares outcomes for people with scores just above the threshold to people with scores just below the threshold

Example - Hou and Chao (2008)

- Hu and Chao (2008) are interested in the question: does lack of health insurance prevent poor people from getting important medical care?

Example - Hou and Chao (2008)

- Hu and Chao (2008) are interested in the question: does lack of health insurance prevent poor people from getting important medical care?
- Why don't they just compare medical care utilization between people with and without health insurance?

Example - Hou and Chao (2008)

- Hu and Chao (2008) are interested in the question: does lack of health insurance prevent poor people from getting important medical care?

- Why don't they just compare medical care utilization between people with and without health insurance?
    - Confounding variables! People with health insurance may be richer, which also may affect health directly

Example - Hou and Chao (2008)

- Hu and Chao (2008) are interested in the question: does lack of health insurance prevent poor people from getting important medical care?
- Why don't they just compare medical care utilization between people with and without health insurance?
  - Confounding variables! People with health insurance may be richer, which also may affect health directly
- Context: the Republic of Georgia (not the state!) created a health insurance program in 2006. Each household received a "poverty score" derived from 80 household variables, and households with a score $\leq$100,000 received health insurance

Example - Hou and Chao (2008)

- Hu and Chao (2008) are interested in the question: does lack of health insurance prevent poor people from getting important medical care?
- Why don't they just compare medical care utilization between people with and without health insurance?
  - Confounding variables! People with health insurance may be richer, which also may affect health directly
- Context: the Republic of Georgia (not the state!) created a health insurance program in 2006. Each household received a "poverty score" derived from 80 household variables, and households with a score $\leq$100,000 received health insurance
- Key idea: people with scores just above 100,000 may be very similar to people with scores just below 100,000

Example - Hou and Chao (2008)

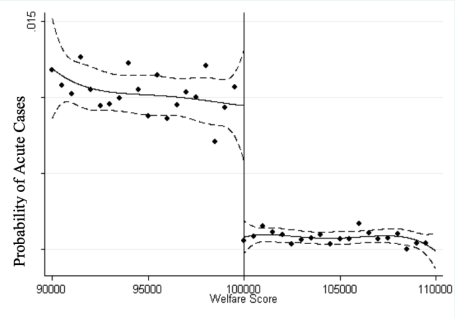
- Hu and Chao (2008) are interested in the question: does lack of health insurance prevent poor people from getting important medical care?

- Why don't they just compare medical care utilization between people with and without health insurance?
    - Confounding variables! People with health insurance may be richer, which also may affect health directly

- Context: the Republic of Georgia (not the state!) created a health insurance program in 2006. Each household received a "poverty score" derived from 80 household variables, and households with a score $\leq$100,000 received health insurance

- Key idea: people with scores just above 100,000 may be very similar to people with scores just below 100,000
    - Except those below 100,000 had the health insurance treatment!

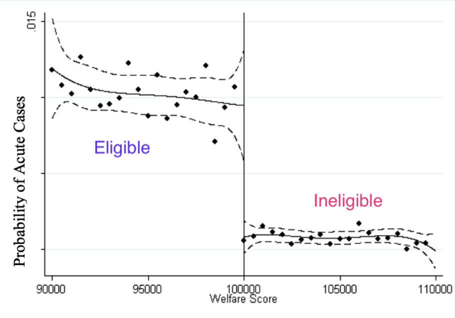
Figure 3: The Effect of MAP on Utilization of Acute Surgeries/In-Patient Services



Note: This figure plots probability of utilization of acute surgeries/inpatient services against welfare scores. Each dot is the average probability within 500 intervals of welfare scores. Solid lines are fitted values from 4^{th} order polynomial regressions on either side of the discontinuity. Dotted lines are 95% confidence intervals.

- This plot shows outcomes (acute surgeries) as a function of the score

Figure 3: The Effect of MAP on Utilization of Acute Surgeries/In-Patient Services

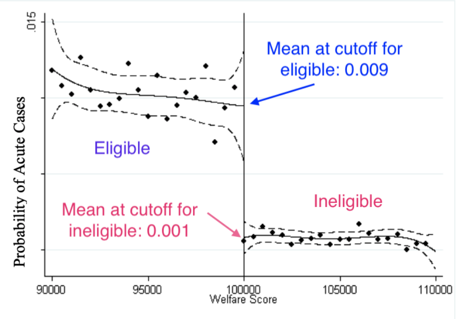


Note: This figure plots probability of utilization of acute surgeries/inpatient services against welfare scores. Each dot is the average probability within 500 intervals of welfare scores. Solid lines are fitted values from 4th order polynomial regressions on either side of the discontinuity. Dotted lines are 95% confidence intervals.

- People to the left of the vertical line (at 100,000) receive health insurance from the government

- We might expect people with scores just below 100,000 to be very similar to people just above 100,000 on all factors other than healthcare eligibility

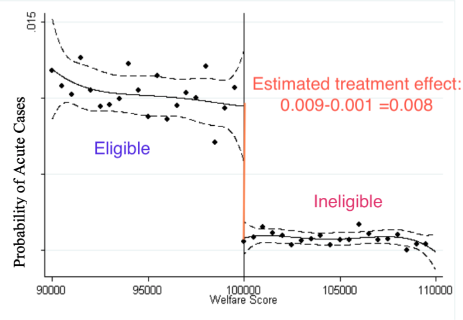
Figure 3: The Effect of MAP on Utilization of Acute Surgeries/In-Patient Services



Note: This figure plots probability of utilization of acute surgeries/inpatient services against welfare scores. Each dot is the average probability within 500 intervals of welfare scores. Solid lines are fitted values from 4^{th} order polynomial regressions on either side of the discontinuity. Dotted lines are 95% confidence intervals.

- We might expect people with scores just below 100,000 to be very similar to people just above 100,000 on all factors other than healthcare eligibility
- But people just below the threshold seem to get a lot more surgeries!

Figure 3: The Effect of MAP on Utilization of Acute Surgeries/In-Patient Services



Note: This figure plots probability of utilization of acute surgeries/inpatient services against welfare scores. Each dot is the average probability within 500 intervals of welfare scores. Solid lines are fitted values from 4th order polynomial regressions on either side of the discontinuity. Dotted lines are 95% confidence intervals.

- But people just below the threshold seem to get a lot more surgeries!
- If everything else is continuous at the threshold, the difference is the causal effect (for people at the threshold)!

Formalizing RD

- As usual, let D_i be a binary indicator for treatment (e.g. insurance status).

Formalizing RD

- As usual, let D_i be a binary indicator for treatment (e.g. insurance status).Let R_i be the *running variable* (e.g. the poverty score)

Formalizing RD

- As usual, let D_i be a binary indicator for treatment (e.g. insurance status).Let R_i be the *running variable* (e.g. the poverty score)

- In a *sharp* regression discontinuity design, unit i receives treatment if and only if R_i is above the threshold

$$D_i = 1[R_i \geq c]$$

(I've flipped the sign from the example to match usual convention)

Formalizing RD

- As usual, let D_i be a binary indicator for treatment (e.g. insurance status).Let R_i be the *running variable* (e.g. the poverty score)

- In a *sharp* regression discontinuity design, unit i receives treatment if and only if R_i is above the threshold

$$D_i = 1[R_i \geq c]$$

  (I've flipped the sign from the example to match usual convention)
  - Note unlike before, D_i is *deterministic* in a likely confounder – no room for a "selection-on-observables" story!

- The idea of RD is to compare the limits of the CEF $E[Y_i|R_i = r]$ around the cutoff c (we'll assume these limits exist!):

$$\tau_{RD} = \underbrace{\lim_{r \downarrow c} E[Y_i|R_i = r]}_{\text{Limit from above the cutoff}} - \underbrace{\lim_{r \uparrow c} E[Y_i|R_i = r]}_{\text{Limit from below the cutoff}}$$

Formalizing RD

- As usual, let D_i be a binary indicator for treatment (e.g. insurance status).Let R_i be the *running variable* (e.g. the poverty score)
- In a *sharp* regression discontinuity design, unit i receives treatment if and only if R_i is above the threshold

$$D_i = 1[R_i \geq c]$$

(I've flipped the sign from the example to match usual convention)

  - Note unlike before, D_i is *deterministic* in a likely confounder – no room for a "selection-on-observables" story!

- The idea of RD is to compare the limits of the CEF $E[Y_i|R_i = r]$ around the cutoff c (we'll assume these limits exist!):

$$\tau_{RD} = \underbrace{\lim_{r \downarrow c} E[Y_i|R_i = r]}_{\text{Limit from above the cutoff}} - \underbrace{\lim_{r \uparrow c} E[Y_i|R_i = r]}_{\text{Limit from below the cutoff}}$$

- When does τ_{RD} correspond to a causal effect?

Take it to the Limit

- By definition, people with $R_i < c$ get $D_i = 0$. Thus, the limit from below the cutoff is: $\lim_{r \uparrow c} E[Y_i | R_i = r] =$

Take it to the Limit

- By definition, people with $R_i < c$ get $D_i = 0$. Thus, the limit from below the cutoff is: $\lim_{r \uparrow c} E[Y_i | R_i = r] = \lim_{r \uparrow c} E[Y_i(0) | R_i = r]$

Take it to the Limit

- By definition, people with $R_i < c$ get $D_i = 0$. Thus, the limit from below the cutoff is: $\lim_{r \uparrow c} E[Y_i | R_i = r] = \lim_{r \uparrow c} E[Y_i(0) | R_i = r]$

- Similarly, people with $R_i \geq c$ get $D_i = 1$. Thus, the limit from above the cutoff is $\lim_{r \downarrow c} E[Y_i | R_i = r] =$

Take it to the Limit

- By definition, people with $R_i < c$ get $D_i = 0$. Thus, the limit from below the cutoff is: $\lim_{r \uparrow c} E[Y_i | R_i = r] = \lim_{r \uparrow c} E[Y_i(0) | R_i = r]$

- Similarly, people with $R_i \geq c$ get $D_i = 1$. Thus, the limit from above the cutoff is $\lim_{r \downarrow c} E[Y_i | R_i = r] = \lim_{r \downarrow c} E[Y_i(1) | R_i = r]$

Take it to the Limit

- By definition, people with $R_i < c$ get $D_i = 0$. Thus, the limit from below the cutoff is: $\lim_{r \uparrow c} E[Y_i | R_i = r] = \lim_{r \uparrow c} E[Y_i(0) | R_i = r]$

- Similarly, people with $R_i \geq c$ get $D_i = 1$. Thus, the limit from above the cutoff is $\lim_{r \downarrow c} E[Y_i | R_i = r] = \lim_{r \downarrow c} E[Y_i(1) | R_i = r]$

- **Key assumption:** Suppose that $f_d(r) = E[Y_i(d) | R_i = r]$ is *continuous* at c for $d = 0, 1$. Then:

Take it to the Limit

- By definition, people with $R_i < c$ get $D_i = 0$. Thus, the limit from below the cutoff is: $\lim_{r \uparrow c} E[Y_i | R_i = r] = \lim_{r \uparrow c} E[Y_i(0) | R_i = r]$

- Similarly, people with $R_i \geq c$ get $D_i = 1$. Thus, the limit from above the cutoff is $\lim_{r \downarrow c} E[Y_i | R_i = r] = \lim_{r \downarrow c} E[Y_i(1) | R_i = r]$

- **Key assumption:** Suppose that $f_d(r) = E[Y_i(d) | R_i = r]$ is *continuous* at c for $d = 0, 1$. Then:
    - $\lim_{r \uparrow c} f_0(r) = \lim_{r \uparrow c} E[Y_i(0) | R_i = r] = E[Y_i(0) | R_i = c]$

Take it to the Limit

- By definition, people with $R_i < c$ get $D_i = 0$. Thus, the limit from below the cutoff is: $\lim_{r \uparrow c} E[Y_i | R_i = r] = \lim_{r \uparrow c} E[Y_i(0) | R_i = r]$

- Similarly, people with $R_i \geq c$ get $D_i = 1$. Thus, the limit from above the cutoff is $\lim_{r \downarrow c} E[Y_i | R_i = r] = \lim_{r \downarrow c} E[Y_i(1) | R_i = r]$

- **Key assumption:** Suppose that $f_d(r) = E[Y_i(d) | R_i = r]$ is *continuous* at c for $d = 0, 1$. Then:
    - $\lim_{r \uparrow c} f_0(r) = \lim_{r \uparrow c} E[Y_i(0) | R_i = r] = E[Y_i(0) | R_i = c]$
    - $\lim_{r \downarrow c} f_1(r) = \lim_{r \downarrow c} E[Y_i(1) | R_i = r] = E[Y_i(1) | R_i = c]$

Take it to the Limit

- By definition, people with $R_i < c$ get $D_i = 0$. Thus, the limit from below the cutoff is: $\lim_{r\uparrow c} E[Y_i|R_i = r] = \lim_{r\uparrow c} E[Y_i(0)|R_i = r]$

- Similarly, people with $R_i \geq c$ get $D_i = 1$. Thus, the limit from above the cutoff is $\lim_{r\downarrow c} E[Y_i|R_i = r] = \lim_{r\downarrow c} E[Y_i(1)|R_i = r]$

- **Key assumption:** Suppose that $f_d(r) = E[Y_i(d)|R_i = r]$ is *continuous* at c for $d = 0, 1$. Then:
    - $\lim_{r\uparrow c} f_0(r) = \lim_{r\uparrow c} E[Y_i(0)|R_i = r] = E[Y_i(0)|R_i = c]$
    - $\lim_{r\downarrow c} f_1(r) = \lim_{r\downarrow c} E[Y_i(1)|R_i = r] = E[Y_i(1)|R_i = c]$

- And so

$$\tau_{RD} = \lim_{r\downarrow c} E[Y_i(1)|R_i = r] - \lim_{r\uparrow c} E[Y_i(0)|R_i = r]$$

Take it to the Limit

- By definition, people with $R_i < c$ get $D_i = 0$. Thus, the limit from below the cutoff is: $\lim_{r \uparrow c} E[Y_i | R_i = r] = \lim_{r \uparrow c} E[Y_i(0) | R_i = r]$

- Similarly, people with $R_i \geq c$ get $D_i = 1$. Thus, the limit from above the cutoff is $\lim_{r \downarrow c} E[Y_i | R_i = r] = \lim_{r \downarrow c} E[Y_i(1) | R_i = r]$

- **Key assumption:** Suppose that $f_d(r) = E[Y_i(d) | R_i = r]$ is *continuous* at c for $d = 0, 1$. Then:
  - $\lim_{r \uparrow c} f_0(r) = \lim_{r \uparrow c} E[Y_i(0) | R_i = r] = E[Y_i(0) | R_i = c]$
  - $\lim_{r \downarrow c} f_1(r) = \lim_{r \downarrow c} E[Y_i(1) | R_i = r] = E[Y_i(1) | R_i = c]$

- And so

$$\tau_{RD} = \lim_{r \downarrow c} E[Y_i(1) | R_i = r] - \lim_{r \uparrow c} E[Y_i(0) | R_i = r] = E[Y_i(1) - Y_i(0) | R_i = c]$$

Take it to the Limit

- By definition, people with $R_i < c$ get $D_i = 0$. Thus, the limit from below the cutoff is: $\lim_{r \uparrow c} E[Y_i | R_i = r] = \lim_{r \uparrow c} E[Y_i(0) | R_i = r]$

- Similarly, people with $R_i \geq c$ get $D_i = 1$. Thus, the limit from above the cutoff is $\lim_{r \downarrow c} E[Y_i | R_i = r] = \lim_{r \downarrow c} E[Y_i(1) | R_i = r]$

- **Key assumption:** Suppose that $f_d(r) = E[Y_i(d) | R_i = r]$ is *continuous* at c for $d = 0, 1$. Then:
  - $\lim_{r \uparrow c} f_0(r) = \lim_{r \uparrow c} E[Y_i(0) | R_i = r] = E[Y_i(0) | R_i = c]$
  - $\lim_{r \downarrow c} f_1(r) = \lim_{r \downarrow c} E[Y_i(1) | R_i = r] = E[Y_i(1) | R_i = c]$

- And so

$$\tau_{RD} = \lim_{r \downarrow c} E[Y_i(1) | R_i = r] - \lim_{r \uparrow c} E[Y_i(0) | R_i = r] = E[Y_i(1) - Y_i(0) | R_i = c]$$

the average treatment effect for people at the cutoff!

Evaluating Continuity

- The key assumption for RD is that the expectation of the potential outcomes, $E[Y_i(d)|R_i = r]$ is continuous at the cutoff (for $d = 0, 1$)

Evaluating Continuity

- The key assumption for RD is that the expectation of the potential outcomes, $E[Y_i(d)|R_i = r]$ is continuous at the cutoff (for $d = 0, 1$)
- Why might the continuity assumption be violated?

Evaluating Continuity

- The key assumption for RD is that the expectation of the potential outcomes, $E[Y_i(d)|R_i = r]$ is continuous at the cutoff (for $d = 0, 1$)
- Why might the continuity assumption be violated?
- #1: confounding factors change discontinuously at the cutoff

Evaluating Continuity

- The key assumption for RD is that the expectation of the potential outcomes, $E[Y_i(d)|R_i = r]$ is continuous at the cutoff (for $d = 0, 1$)
- Why might the continuity assumption be violated?
- #1: confounding factors change discontinuously at the cutoff
- #2: people can manipulate scores to get just above/below the cutoff

Example - Confounding Factors

- Policies often change discontinuously at state borders

Example - Confounding Factors

- Policies often change discontinuously at state borders
- Example: Holmes (1998) is interested in how "right-to-work laws," which weaken labor unions, affect businesses

Example - Confounding Factors

- Policies often change discontinuously at state borders
- Example: Holmes (1998) is interested in how “right-to-work laws,” which weaken labor unions, affect businesses
- He uses an RD to compare the density of manufacturing employment on both sides of borders between states that have/ don’t have right to work laws

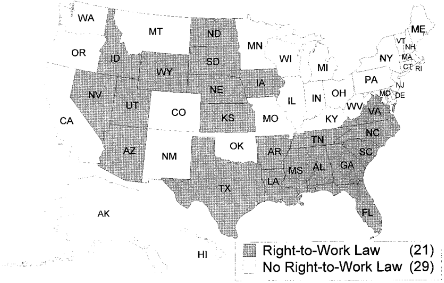


Fig. 1.—Geography of right-to-work laws

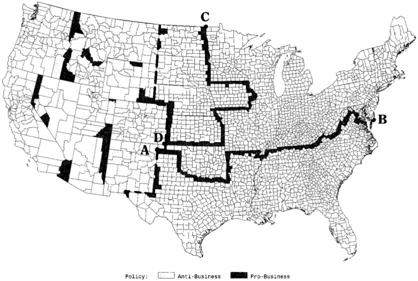


Fig. 3.—Counties within 25 miles of the policy change border

TABLE 1

Manufacturing Employment Shares and Growth Rates: Cross-County Averages by Distance from Border and Side of Border

Miles from Border	Coal Region Included: Share of 1992 Total (1)	Coal Region Included: Growth Rate, 1947–92 (2)	Coal Region Excluded: Share of 1992 Total (3)	Coal Region Excluded: Growth Rate, 1947–92 (4)
	A. Antibusiness Side of Border			
75–100	25.9	67.5	25.0	68.2
50–75	23.1	62.7	25.0	80.9
25–50	23.2	82.0	24.7	88.8
0–25	21.0	62.4	22.1	77.2
	B. Probusiness Side of Border			
0–25	28.6	100.7	27.9	104.2
25–50	26.7	89.1	25.5	88.3
50–75	26.7	92.9	24.5	90.1
75–100	25.4	91.8	23.1	93.5

- There does appear to be more manufacturing just to the RTW side of state borders

TABLE 1

MANUFACTURING EMPLOYMENT SHARES AND GROWTH RATES: CROSS-COUNTY AVERAGES BY DISTANCE FROM BORDER AND SIDE OF BORDER

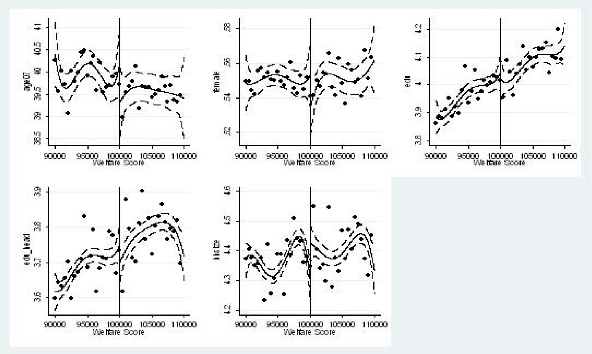
	Coal Region Included		Coal Region Excluded	
Miles from Border	Share of 1992 Total (1)	Growth Rate, 1947–92 (2)	Share of 1992 Total (3)	Growth Rate, 1947–92 (4)
	A. Antibusiness Side of Border			
75–100	25.9	67.5	25.0	68.2
50–75	23.1	62.7	25.0	80.9
25–50	23.2	82.0	24.7	88.8
0–25	21.0	62.4	22.1	77.2
	B. Probusiness Side of Border			
0–25	28.6	100.7	27.9	104.2
25–50	26.7	89.1	25.5	88.3
50–75	26.7	92.9	24.5	90.1
75–100	25.4	91.8	23.1	93.5

- There does appear to be more manufacturing just to the RTW side of state borders
- But are RTWs the only thing that vary at state borders?! Could it be that RTW laws are correlated with other policies that affect manufacturing?

- To partially address these types of concerns, it is common to show that observable features don't vary discontinuously at the cutoff

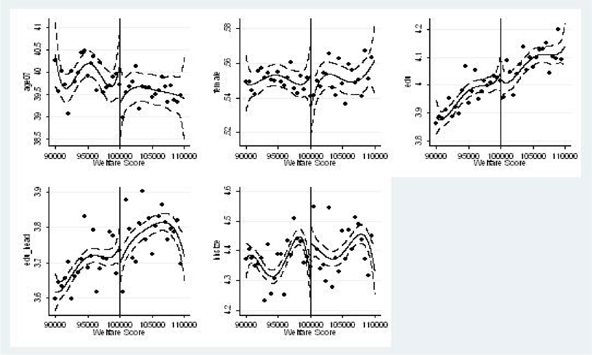
- To partially address these types of concerns, it is common to show that observable features don't vary discontinuously at the cutoff

- The figure below shows that there don't appear to be any discontinuities in age, sex, education, and household size in the Hou and Chao paper



- To partially address these types of concerns, it is common to show that observable features don't vary discontinuously at the cutoff

- The figure below shows that there don't appear to be any discontinuities in age, sex, education, and household size in the Hou and Chao paper



- But as usual, there may still be concern about other unobserved confounding factors varying at the cutoff

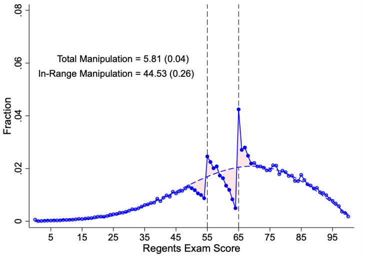
Example - Manipulation

- In NYC, students must take the Regents exams and get a score of at least 55 to get a diploma (and 65 to get a more prestigious diploma)

Example - Manipulation

- In NYC, students must take the Regents exams and get a score of at least 55 to get a diploma (and 65 to get a more prestigious diploma)
- Might be tempted to use this to study diploma effects... But it turns out there are way more students with scores just above the thresholds

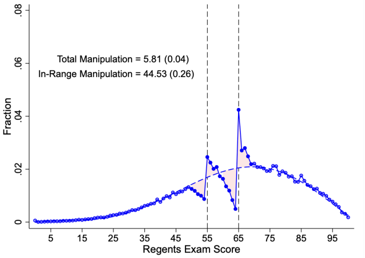
Figure 1
Test Score Distributions for Core Regents Exams, 2004-2010



Example - Manipulation

- In NYC, students must take the Regents exams and get a score of at least 55 to get a diploma (and 65 to get a more prestigious diploma)
- Might be tempted to use this to study diploma effects... But it turns out there are way more students with scores just above the thresholds

Figure 1
Test Score Distributions for Core Regents Exams, 2004-2010

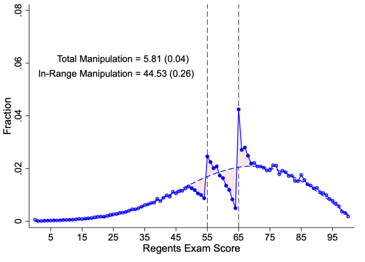


- Likely reason: teachers cheat to bump students over the threshold

Example - Manipulation

- In NYC, students must take the Regents exams and get a score of at least 55 to get a diploma (and 65 to get a more prestigious diploma)
- Might be tempted to use this to study diploma effects... But it turns out there are way more students with scores just above the thresholds

Figure 1
Test Score Distributions for Core Regents Exams, 2004-2010



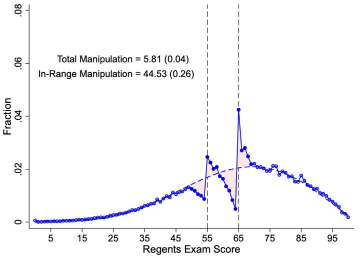
- Likely reason: teachers cheat to bump students over the threshold
- Do we think students who get bumped over the threshold versus not are similar?

Example - Manipulation

- In NYC, students must take the Regents exams and get a score of at least 55 to get a diploma (and 65 to get a more prestigious diploma)

- Might be tempted to use this to study diploma effects... But it turns out there are way more students with scores just above the thresholds

Figure 1
Test Score Distributions for Core Regents Exams, 2004-2010



- Likely reason: teachers cheat to bump students over the threshold

- Do we think students who get bumped over the threshold versus not are similar? Likely not, if teachers tend to bump better students...

Testing for Manipulation

- To test for such RD manipulation, it is common to check whether there are a similar number of units on both sides of the cutoff. This is often called a McCrary test.
- If there is bunching on one side of the cutoff, this is typically interpreted as evidence of manipulation
- The continuity assumption will usually be much more questionable if there is bunching.

Estimating RD

- Under continuity, we have identification:

$$\lim_{r \downarrow c} E[Y_i | R_i = r] - \lim_{r \uparrow c} E[Y_i | R_i = r] = E[Y_i(1) - Y_i(0) \mid R_i = c]$$

Estimating RD

- Under continuity, we have identification:

$$\lim_{r \downarrow c} E[Y_i | R_i = r] - \lim_{r \uparrow c} E[Y_i | R_i = r] = E[Y_i(1) - Y_i(0) \mid R_i = c]$$

- To estimate this CATE, all we need to do is estimate the CEF of the outcome Y_i given the running variable R_i and take it to the limit

Estimating RD

- Under continuity, we have identification:

$$\lim_{r \downarrow c} E[Y_i | R_i = r] - \lim_{r \uparrow c} E[Y_i | R_i = r] = E[Y_i(1) - Y_i(0) \mid R_i = c]$$

- To estimate this CATE, all we need to do is estimate the CEF of the outcome Y_i given the running variable R_i and take it to the limit

- How do we estimate CEFs?

Estimating RD

- Under continuity, we have identification:

$$\lim_{r \downarrow c} E[Y_i | R_i = r] - \lim_{r \uparrow c} E[Y_i | R_i = r] = E[Y_i(1) - Y_i(0) \mid R_i = c]$$

- To estimate this CATE, all we need to do is estimate the CEF of the outcome Y_i given the running variable R_i and take it to the limit

- How do we estimate CEFs? OLS!

RD with Linear Regression

- Suppose the CEF is piecewise linear:

$$
\begin{aligned}
E[Y_i|R_i = r] &\approx \alpha_0 + \alpha_1 r && \text{if } r < c \\
E[Y_i|R_i = r] &\approx \beta_0 + \beta_1 r && \text{if } r \geq c
\end{aligned}
$$

RD with Linear Regression

- Suppose the CEF is piecewise linear:

$$
\begin{aligned}
E[Y_i|R_i = r] &\approx \alpha_0 + \alpha_1 r && \text{if } r < c \\
E[Y_i|R_i = r] &\approx \beta_0 + \beta_1 r && \text{if } r \geq c
\end{aligned}
$$

- Then under the RD assumptions,

$$CATE = \lim_{r \downarrow c} E[Y_i|R_i = r] - \lim_{r \uparrow c} E[Y_i|R_i = r] \approx$$

RD with Linear Regression

- Suppose the CEF is piecewise linear:

$$E[Y_i|R_i = r] \approx \alpha_0 + \alpha_1 r \qquad \text{if } r < c$$
$$E[Y_i|R_i = r] \approx \beta_0 + \beta_1 r \qquad \text{if } r \geq c$$

- Then under the RD assumptions,

$$CATE = \lim_{r \downarrow c} E[Y_i|R_i = r] - \lim_{r \uparrow c} E[Y_i|R_i = r] \approx (\beta_0 + \beta_1 c) - (\alpha_0 + \alpha_1 c)$$

RD with Linear Regression

- Suppose the CEF is piecewise linear:

$$
\begin{aligned}
E[Y_i|R_i = r] &\approx \alpha_0 + \alpha_1 r && \text{if } r < c \\
E[Y_i|R_i = r] &\approx \beta_0 + \beta_1 r && \text{if } r \geq c
\end{aligned}
$$

- Then under the RD assumptions,

$$CATE = \lim_{r \downarrow c} E[Y_i|R_i = r] - \lim_{r \uparrow c} E[Y_i|R_i = r] \approx (\beta_0 + \beta_1 c) - (\alpha_0 + \alpha_1 c)$$

- We can estimate these regression coefficients via OLS to estimate the effect of the treatment

Doing it All in One Regression

- If we normalize $c = 0$ (without loss of generality), we can actually do it all in one regression

Doing it All in One Regression

- If we normalize $c = 0$ (without loss of generality), we can actually do it all in one regression

- Consider the regression:

$$Y_i = \alpha + \beta 1[R_i \geq 0] + \gamma R_i + \delta R_i 1[R_i \geq 0] + U_i$$

Doing it All in One Regression

- If we normalize $c = 0$ (without loss of generality), we can actually do it all in one regression

- Consider the regression:

$$Y_i = \alpha + \beta 1[R_i \geq 0] + \gamma R_i + \delta R_i 1[R_i \geq 0] + U_i$$

- Then

$$\text{if } r < 0, \qquad E[Y_i | R_i = r] \approx$$

Doing it All in One Regression

- If we normalize $c = 0$ (without loss of generality), we can actually do it all in one regression

- Consider the regression:

$$Y_i = \alpha + \beta 1[R_i \geq 0] + \gamma R_i + \delta R_i 1[R_i \geq 0] + U_i$$

- Then

$$\text{if } r < 0, \qquad E[Y_i | R_i = r] \approx \alpha + \gamma r$$

Doing it All in One Regression

- If we normalize $c = 0$ (without loss of generality), we can actually do it all in one regression

- Consider the regression:

$$Y_i = \alpha + \beta 1[R_i \geq 0] + \gamma R_i + \delta R_i 1[R_i \geq 0] + U_i$$

- Then

$$\begin{aligned} \text{if } r < 0, \qquad & E[Y_i | R_i = r] \approx \alpha + \gamma r \\ \text{if } r \geq 0, \qquad & E[Y_i | R_i = r] \approx \end{aligned}$$

Doing it All in One Regression

- If we normalize $c = 0$ (without loss of generality), we can actually do it all in one regression

- Consider the regression:

$$Y_i = \alpha + \beta 1[R_i \geq 0] + \gamma R_i + \delta R_i 1[R_i \geq 0] + U_i$$

- Then

$$\begin{aligned}
\text{if } r < 0, \qquad & E[Y_i | R_i = r] \approx \alpha + \gamma r \\
\text{if } r \geq 0, \qquad & E[Y_i | R_i = r] \approx \alpha + \beta + \gamma r + \delta r
\end{aligned}$$

Doing it All in One Regression

- If we normalize $c = 0$ (without loss of generality), we can actually do it all in one regression

- Consider the regression:

$$Y_i = \alpha + \beta 1[R_i \geq 0] + \gamma R_i + \delta R_i 1[R_i \geq 0] + U_i$$

- Then

$$\begin{aligned} &\text{if } r < 0, && E[Y_i|R_i = r] \approx \alpha + \gamma r \\ &\text{if } r \geq 0, && E[Y_i|R_i = r] \approx \alpha + \beta + \gamma r + \delta r \end{aligned}$$

- So

$$\begin{aligned} ATE &= \lim_{r \downarrow 0} E[Y_i|R_i = r] - \lim_{r \uparrow 0} E[Y_i|R_i = r] \\ &\approx \end{aligned}$$

Doing it All in One Regression

- If we normalize $c = 0$ (without loss of generality), we can actually do it all in one regression

- Consider the regression:

$$Y_i = \alpha + \beta 1[R_i \geq 0] + \gamma R_i + \delta R_i 1[R_i \geq 0] + U_i$$

- Then

$$\begin{aligned} &\text{if } r < 0, && E[Y_i|R_i = r] \approx \alpha + \gamma r \\ &\text{if } r \geq 0, && E[Y_i|R_i = r] \approx \alpha + \beta + \gamma r + \delta r \end{aligned}$$

- So

$$\begin{aligned} ATE &= \lim_{r \downarrow 0} E[Y_i|R_i = r] - \lim_{r \uparrow 0} E[Y_i|R_i = r] \\ &\approx (\alpha + \beta + (\gamma + \delta) \times 0) - (\alpha + \gamma \times 0) = \end{aligned}$$

Doing it All in One Regression

- If we normalize $c = 0$ (without loss of generality), we can actually do it all in one regression

- Consider the regression:

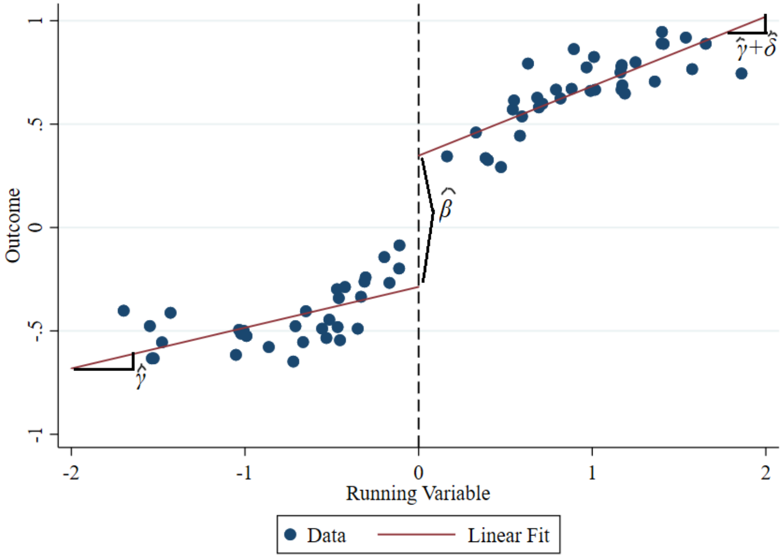
$$Y_i = \alpha + \beta 1[R_i \geq 0] + \gamma R_i + \delta R_i 1[R_i \geq 0] + U_i$$

- Then

$$\begin{aligned} \text{if } r < 0, \qquad & E[Y_i|R_i = r] \approx \alpha + \gamma r \\ \text{if } r \geq 0, \qquad & E[Y_i|R_i = r] \approx \alpha + \beta + \gamma r + \delta r \end{aligned}$$

- So

$$\begin{aligned} ATE &= \lim_{r \downarrow 0} E[Y_i|R_i = r] - \lim_{r \uparrow 0} E[Y_i|R_i = r] \\ &\approx (\alpha + \beta + (\gamma + \delta) \times 0) - (\alpha + \gamma \times 0) = \beta \end{aligned}$$


Outcome
1
.5
0
-.5
-1
-2
-1
0
1
2
Running Variable
$\hat{\gamma}$
$\hat{\beta}$
$\hat{\gamma}+\hat{\delta}$
Data
Linear Fit

Higher Order Terms

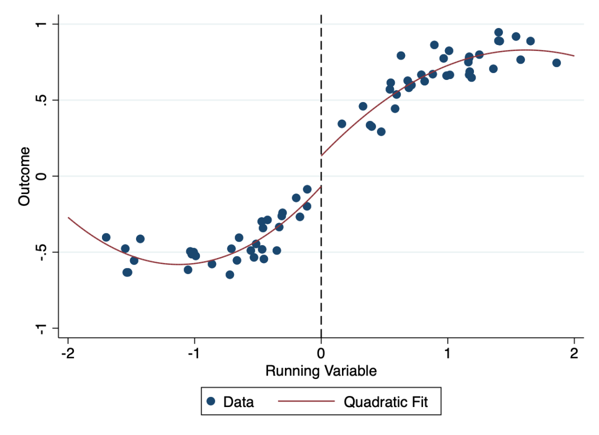
- This will only work well if the linear approximation to the CEF is good

Higher Order Terms

- This will only work well if the linear approximation to the CEF is good
- As before, we can also add higher-order terms to the regression

Higher Order Terms

- This will only work well if the linear approximation to the CEF is good
- As before, we can also add higher-order terms to the regression
- Here, for example, we fit a quadratic on each side:



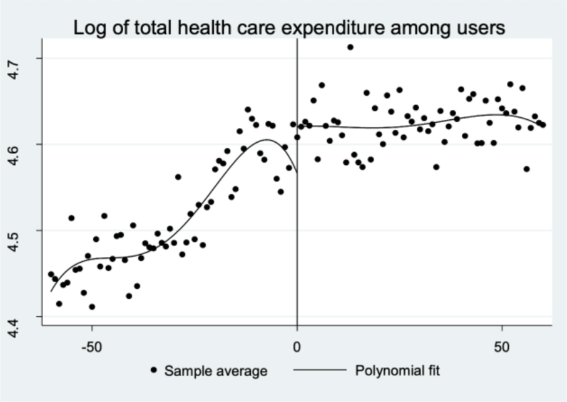
Problems with Higher-Order Polynomials

- The issue with higher-order polynomials is they might overfit the data
  - Particularly bad when extrapolating to the “boundary” of the data

Problems with Higher-Order Polynomials

- The issue with higher-order polynomials is they might overfit the data
  - Particularly bad when extrapolating to the "boundary" of the data

- Is there really a discontinuity in this picture?



Stick to Linear?

The overfitting problem has led a certain Nobel laureate (and Brown graduate) to take a strong stand!

Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs

Andrew Gelman
Department of Statistics and Department of Political Science, Columbia University, New York, NY, 10027 (*gelman@stat.columbia.edu*)

Guido Imbens
Graduate School of Business, Stanford University, Stanford, CA 94305, and NBER, Stanford University, Stanford, CA 94305 (*imbens@stanford.edu*)

What to Do in Practice?

- Clearly, there are tradeoffs between having a rich enough approximation to the CEF and over-fitting the data

What to Do in Practice?

- Clearly, there are tradeoffs between having a rich enough approximation to the CEF and over-fitting the data
- In practice, the usual approach is to use **local linear regression**

What to Do in Practice?

- Clearly, there are tradeoffs between having a rich enough approximation to the CEF and over-fitting the data
- In practice, the usual approach is to use **local linear regression**
- The basic idea is to fit a linear regression but to only use points that are "close" to the boundary

What to Do in Practice?

- Clearly, there are tradeoffs between having a rich enough approximation to the CEF and over-fitting the data
- In practice, the usual approach is to use **local linear regression**
- The basic idea is to fit a linear regression but to only use points that are "close" to the boundary
  - In the simplest case, just use points within a "bandwidth" of the cutoff

What to Do in Practice?

- Clearly, there are tradeoffs between having a rich enough approximation to the CEF and over-fitting the data
- In practice, the usual approach is to use **local linear regression**
- The basic idea is to fit a linear regression but to only use points that are "close" to the boundary
  - In the simplest case, just use points within a "bandwidth" of the cutoff
  - More complicated versions put higher weight on observations closer to the cutoff

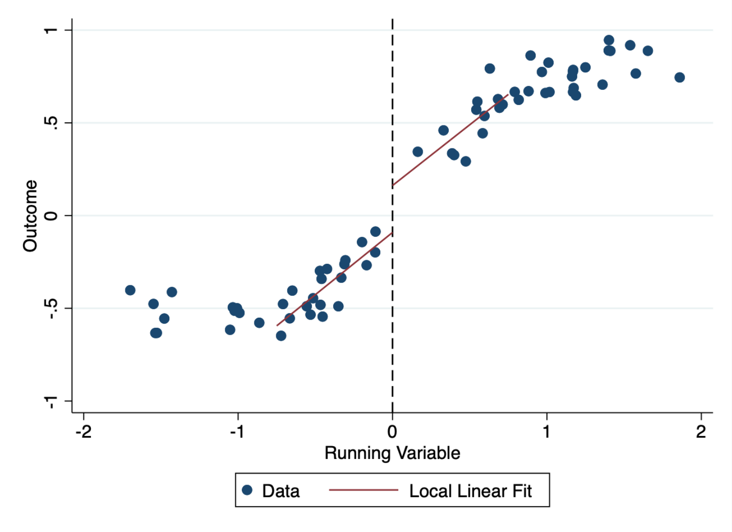
What to Do in Practice?

- Clearly, there are tradeoffs between having a rich enough approximation to the CEF and over-fitting the data
- In practice, the usual approach is to use **local linear regression**
- The basic idea is to fit a linear regression but to only use points that are "close" to the boundary
  - In the simplest case, just use points within a "bandwidth" of the cutoff
  - More complicated versions put higher weight on observations closer to the cutoff
- There are some data-driven ways to choose the bandwidth (e.g. use a smaller bandwidth when there are more observations)

What to Do in Practice?

- Clearly, there are tradeoffs between having a rich enough approximation to the CEF and over-fitting the data
- In practice, the usual approach is to use **local linear regression**
- The basic idea is to fit a linear regression but to only use points that are "close" to the boundary
  - In the simplest case, just use points within a "bandwidth" of the cutoff
  - More complicated versions put higher weight on observations closer to the cutoff
- There are some data-driven ways to choose the bandwidth (e.g. use a smaller bandwidth when there are more observations)
- This is easy to implement with the `rd` package in Stata

Local Linear Regression



RD Estimation is Tricky!

- Local linear regression is a relatively good medium between being rich enough and over-fitting
- But it's not perfect! In general, difficult to distinguish between a very non-linear CEF and a discontinuity

RD Estimation is Tricky!

- Local linear regression is a relatively good medium between being rich enough and over-fitting
- But it's not perfect! In general, difficult to distinguish between a very non-linear CEF and a discontinuity
- Good to trust your eyes – does it look like there's a discontinuity on the plot?
- The most convincing RDs are obvious from the plot and don't need any fancy econometrics

Fuzzy RD

- Sometimes crossing a threshold doesn't completely determine treatment status, but discontinuously increases treatment probability
- This is called a *fuzzy* RD

Fuzzy RD

- Sometimes crossing a threshold doesn't completely determine treatment status, but discontinuously increases treatment probability
- This is called a *fuzzy* RD
- The basic idea is similar to IV — we estimate the effect of being above the threshold on the outcome, then divide this by the effect on the treatment (i.e. the change treatment probability)

Fuzzy RD

- Sometimes crossing a threshold doesn't completely determine treatment status, but discontinuously increases treatment probability
- This is called a *fuzzy* RD
- The basic idea is similar to IV — we estimate the effect of being above the threshold on the outcome, then divide this by the effect on the treatment (i.e. the change treatment probability)
- Under certain conditions, this will identify the LATE at the cutoff

Example – Bleemer and Mehta (2022)

- Bleemer and Mehta ask a very important question (for you): does majoring in economics make you rich?

Example – Bleemer and Mehta (2022)

- Bleemer and Mehta ask a very important question (for you): does majoring in economics make you rich?
- They study this Q in the context of UC Santa Cruz (UCSC), where the econ department only allowed people with GPA below 2.8 in intro classes to major in econ "at the discretion of the department"

First Stage

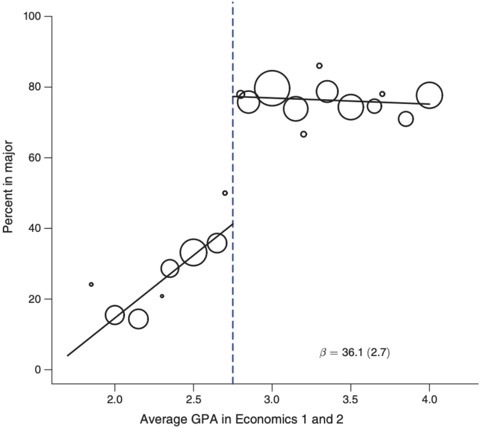


FIGURE 1. THE EFFECT OF THE UCSC ECONOMICS GPA THRESHOLD ON MAJORING IN ECONOMICS

Notes: Each circle represents the percent of economics majors (y-axis) among 2008–2012 UCSC students who earned a given *EGPA* in Economics 1 and 2 (x-axis). The size of each circle corresponds to the proportion of students who earned that *EGPA*. *EGPA*s below 1.8 are omitted, leaving 2,839 students in the sample. Fit lines and beta estimate (at the 2.8 GPA threshold) from linear RD specification; standard error (clustered by *EGPA*) in parentheses.

Students above the threshold about 36 pp more likely to major in econ

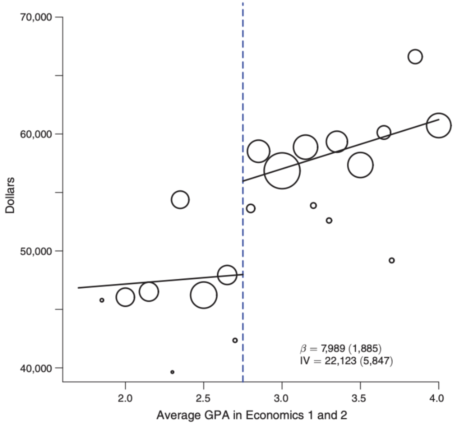


FIGURE 2. THE EFFECT OF THE UCSC ECONOMICS GPA THRESHOLD ON ANNUAL WAGES

- Students about the threshold earn about $8K more

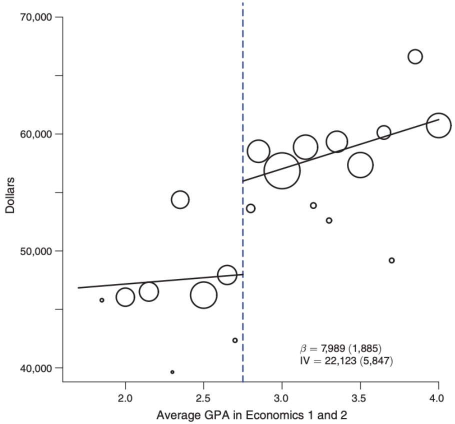


FIGURE 2. THE EFFECT OF THE UCSC ECONOMICS GPA THRESHOLD ON ANNUAL WAGES

- Students about the threshold earn about $8K more
- Aren't you glad you took this course?!

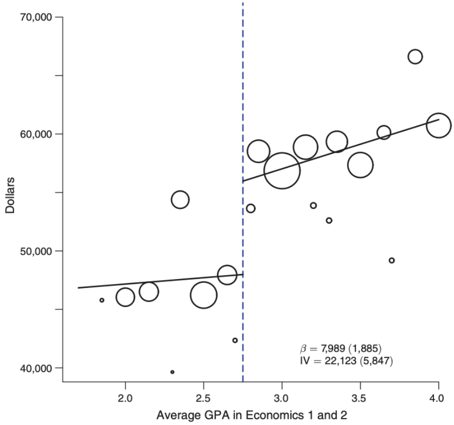


Figure 2. The Effect of the UCSC Economics GPA Threshold on Annual Wages

- Students about the threshold earn about \$8K more
- Aren't you glad you took this course?!
- The fuzzy RD estimate of the effect of majoring in econ is then $\$8K/0.36 \approx 22K$, or 40% of mean earnings

When does Fuzzy RD Give a LATE?

- Under similar assumptions to those for IV, fuzzy RD gives a LATE for compliers at the cutoff

When does Fuzzy RD Give a LATE?

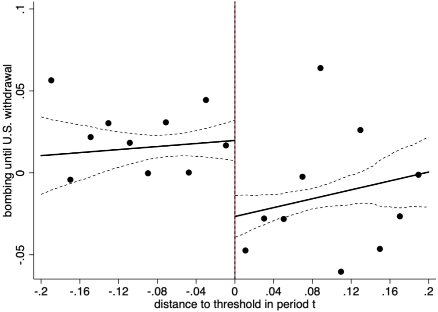
- Under similar assumptions to those for IV, fuzzy RD gives a LATE for compliers at the cutoff
- **Continuity**. Need that $E[Y_i(d)|R_i]$ is continuous at the cutoff for $d = 0, 1$
- **Exclusion**. Being just above/below the cutoff affects outcomes only through its effect on treatment
- **Relevance**. There is a discontinuity in treatment takeup at the cutoff.
- **“Local” Monotonicity**. No defiers who only take treatment if below the cutoff.

Dell and Querubin (2017)

- Dell and Querubin exploit the fact that during the Vietnam War, the US airforce selected bombing targets based on a risk score formed using 169 security/political/economic characteristics of villages

- The algorithm produced a continuous score which was rounded to the nearest integer before being given to generals → Dell and Querubin exploit the discontinuity from rounding

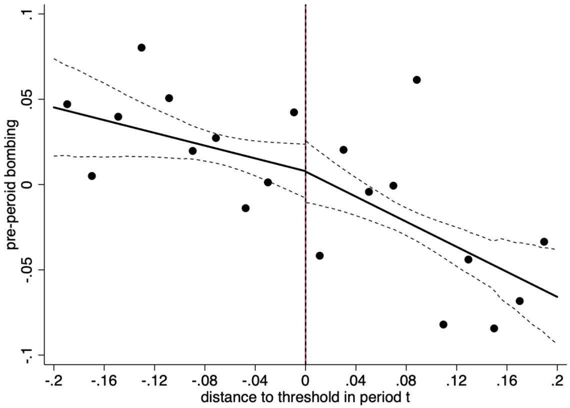
(b) Cumulative First Stage



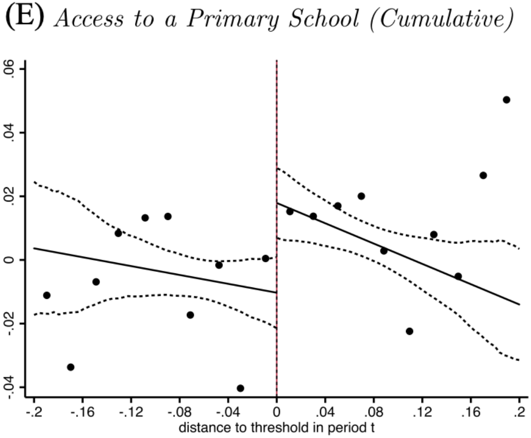
Checking for Balance

- Areas above/below the threshold have similar characteristics prior to bombing decision
- E.g., they have similar # of bombings prior to when score was used

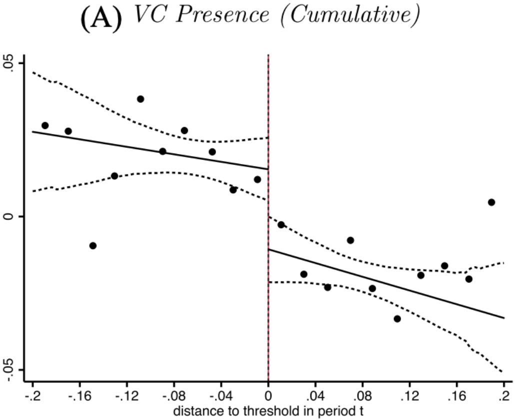
(c) All Prior Quarters Bombing



- Bombing appears to be bad for villages → fewer schools



- Bombing also appears to be bad for military objectives $\rightarrow$ more long-run Vietcong (VC) activity



So to Conclude...

- Bombing is bad ✓
- Econometrics is good ✓