Lecture 14: Regression Discontinuity 1

Tyler Stoeger

18 November 2025

University of Illinois Chicago

Announcements

- DataCamp: *Intermediate Regression in R*
  - Due Sunday (11/23) by 11:59 pm
- PS4 now posted
  - Due 11/30 by 11:59 pm
  - No coding portion!
- Final Exam will be Monday, December 8th from 10:30 am to 12:30 pm

Motivation

- There are countless policies that govern our lives, and all these policies are associated with various rules.
  - A winner of an election is the candidate that receives at least $50+\epsilon$ percent of the vote.
  - A child is allowed to enroll in kindergarten this year if they are born before September 1st; otherwise, they must wait a year.
  - A person is allowed to buy alcohol when they turn 21 years old.
  - People become eligible for Social Security when they turn 62.

Motivation

- There are countless policies that govern our lives, and all these policies are associated with various rules.
  - Medical interventions occur when a diagnostic test passes a certain threshold.
    - People become eligible for cholesterol-lowering drugs when their LDL levels go above 160 mg/dL.
    - Babies born at 1,500 grams or less a subject to extra neonatal care.

Motivation

- The thresholds for these rules are often chosen arbitrarily.
  - For kindergarten enrollment, why is the cutoff September 1^{st}? Why not August 31^{st}?
  - For neonatal intervention, what is special about 1,500 grams. Is this really some sort of “magic number” where extra intervention is needed? What about babies a little heavier, or a little lighter.

Regression Discontinuity (RD) Design

- We can use these rather arbitrary cutoffs to estimate the effect of being on one side of the cutoff versus the other.
- Idea: If we get close enough to the cutoff, groups on either side of the cutoff are about the same, and thus good counterfactuals for each other.
  - Children born on August 31st are probably close to the same on average as children born on September 2nd, so we can use these groups as counterfactuals to see what the effect of going to kindergarten a year early is.
  - Babies born weighing 1,499 grams are about the same babies born weighing 1,501 grams, so we can see what is the effect of the extra neonatal care.

Example - GPA

- When you graduate, depending on your GPA, you might be awarded Latin Honors.

Summa Cum Laude	Highest Distinction	3.90 to 4.00
Magna Cum Laude	High Distinction	3.75 to 3.89
Cum Laude	Distinction	3.50 to 3.74

Example - GPA

- We can use these rules to estimate the effect of being on one side of the threshold versus the other.
- What is the effect of graduating with distinction on earnings after graduation.
    - This could be a difficult question to answer. Those who graduate with distinction have higher GPAs. So, there are presumably many other differences between those with and without honor awards.
    - Lots of selection = lots of endogeneity!

Example - GPA

- The idea of RD is to compare the earnings of people who were just barely above the 3.5 GPA cutoff to those just below the cutoff.
- The GPAs at this cutoff is nearly identical.
  - 3.49 is essentially the same as 3.51
  - Because these two GPAs are nearly the same, their unobservables are probably very similar.
    - The only difference is that the people with the 3.51 get the award, while the people with the 3.49 do not.

Basic RD Model

$$y_i = \beta_0 + \beta_1 Above_i + \beta_2 S_i + u_i$$

- Here:
  - y_i is the outcome variable.
  - $Above_i$ is a binary variable that switches on if we are at or above the cutoff.
  - S_i is the **running variable**.
    - The running variable represents the X_i variable that determines who gets treatment.
    - Typically, we want to center the running variable on zero, so $S_i = X_i - \text{cutoff}$, where cutoff is the threshold that determines treatment.
      - We are subtracting the same value from all of our observations, so even though our numbers change, the underlying information contained in the data does not.

Basic RD Model Interpretation

$$y_i = \beta_0 + \beta_1 Above_i + \beta_2 S_i + u_i$$

- How do we interpret our coefficients?
  - β_0 is the intercept.
    - In RD, this will almost always be meaningless.
  - β_1 gives us the discontinuity.
    - A discontinuity is a point at which a graph suddenly jumps up or down.
    - **This is our variable of interest!**
    - β_1 gives us the effect of being just on the other side of the threshold.
  - β_2 gives us the slope, that is, the average change in y_i for every unit increase in S_i.

Basic RD Model with Slope Changes

- The previous model assumes that the relationship between S_i and y_i is the same on either side of the cutoff.
  - The slope on either side of the cutoff is the same.
- Oftentimes, there is good reason to believe that the slope on either side of the cutoff is not the same.
  - Even if it is, it can be nice to add flexibility into the model.

Basic RD Model with Slope Changes

$$y_i = \beta_0 + \beta_1 Above_i + \beta_2 S_i + \beta_3 Above_i * S_i + u_i$$

- Here:
  - $Above_i * S_i$ is an interaction term that allows for a difference in slope above the cutoff.
  - β_3 is the difference between the slope below the cutoff and the slope above the cutoff
  - $\beta_2 + \beta_3$ is the slope above the cutoff.
  - **Our variable of interest is still $\boldsymbol{\beta_1}$.**

Example - GPA

- We want to know the effect of graduating with distinction on earnings.
  - To graduate with distinction, you need at least a 3.50 GPA.

- What is the outcome variable?
- What is the X_i variable?
- How do we construct the running variable?

Example - GPA

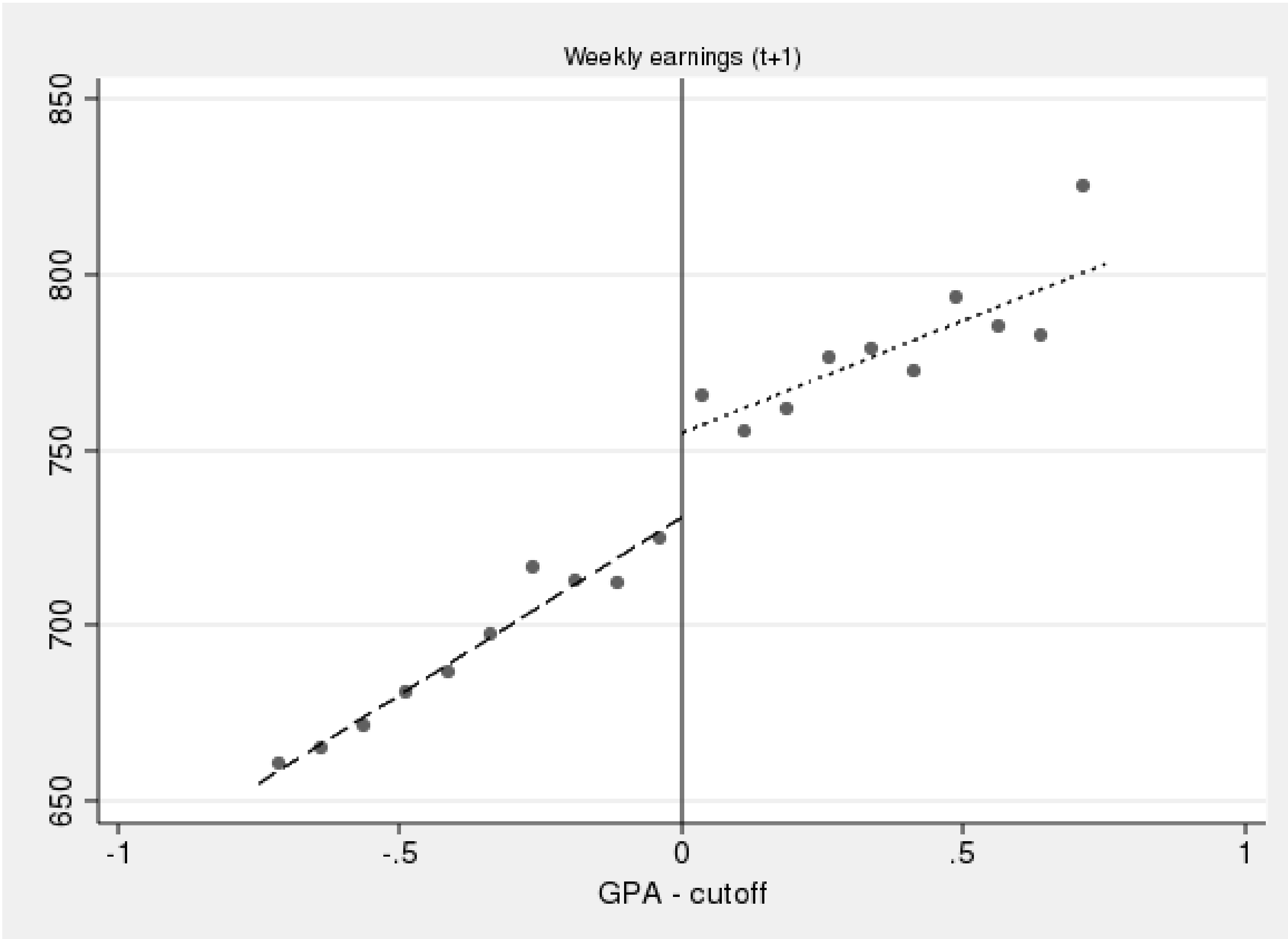
- We want to know the effect of graduating with distinction on earnings.
  - To graduate with distinction, you need at least a 3.50 GPA.

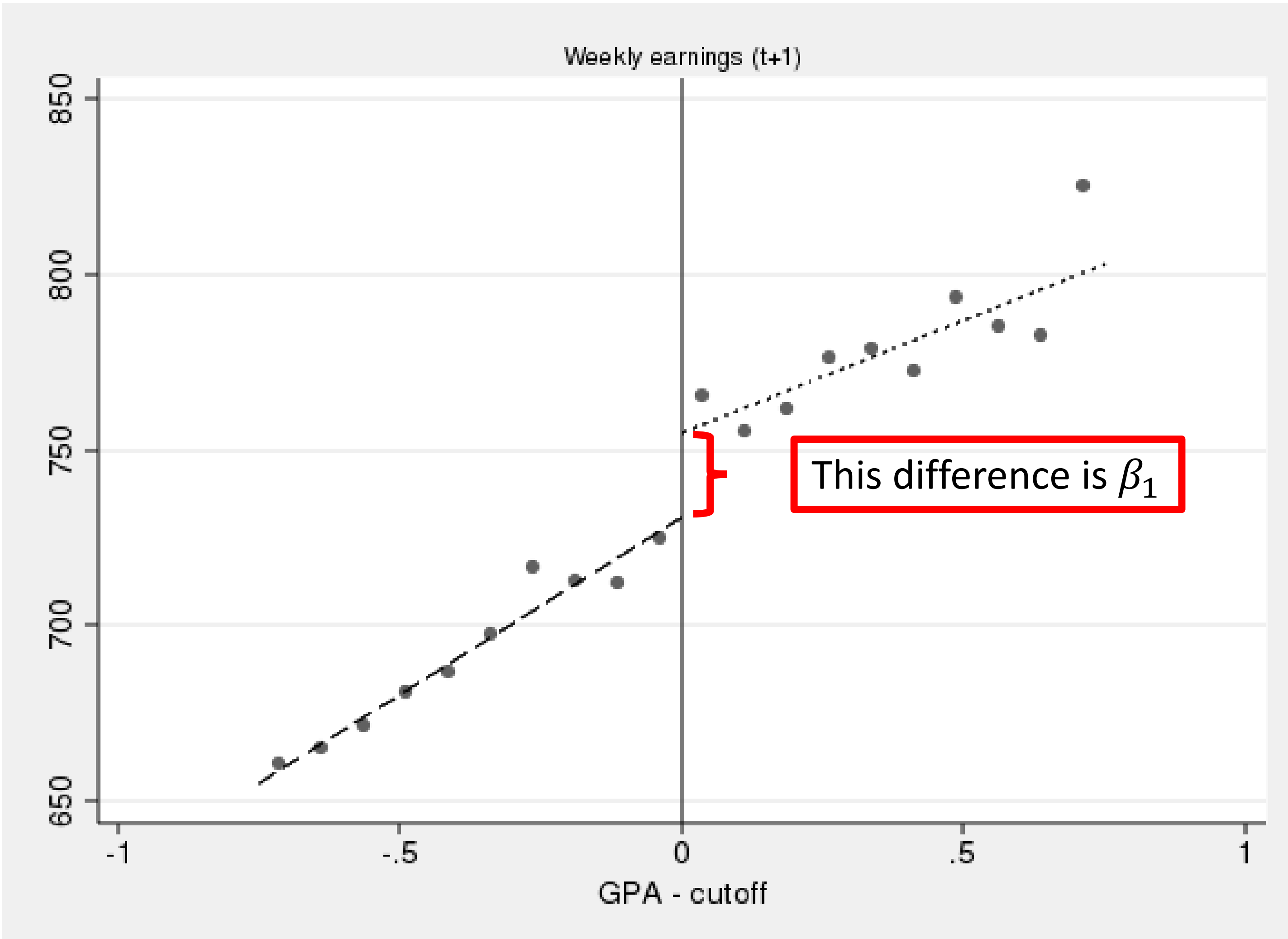
- What is the outcome variable?
  - Earnings
- What is the X_i variable?
  - GPA
- How do we construct the running variable?
  - GPA – 3.50

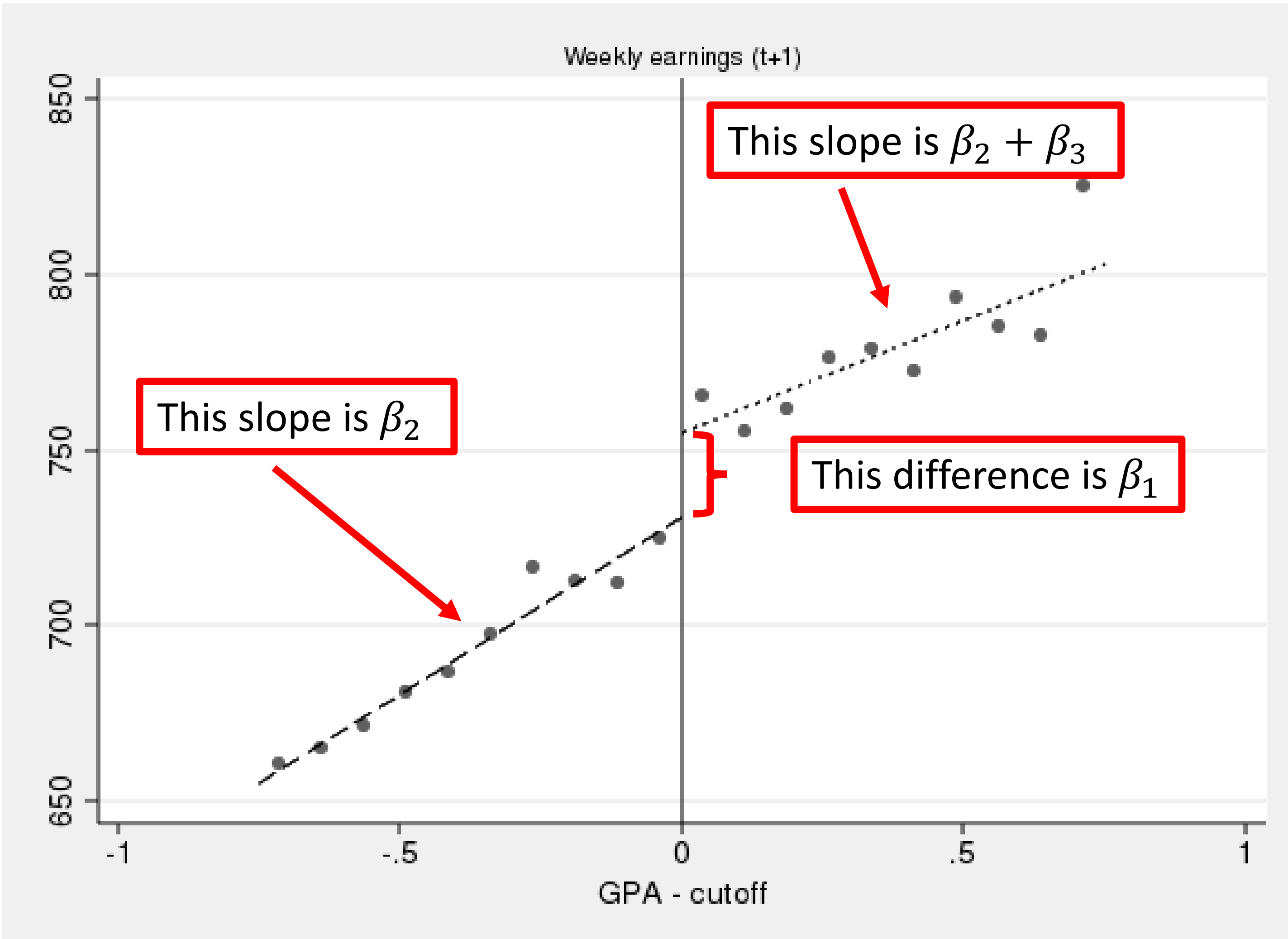
Example - GPA

$$Earnings_i = \beta_0 + \beta_1 Above_i + \beta_2 (GPA - 3.50)_i + \beta_3 Above_i * (GPA - 3.50)_i + u_i$$

- Here, $(GPA - 3.50)_i = S_i$, the running variable.
- $Above_i = 1$ if GPA is greater than or equal to 3.50, and zero otherwise
- β_1 is the variable of interest and estimates the size of the discontinuity.


Weekly earnings (t+1)
850
800
750
700
650
-1
-.5
0
.5
1
GPA - cutoff


Weekly earnings (t+1)
850
800
750
700
650
-1
-.5
0
.5
1
GPA - cutoff
This difference is β_1



Weekly earnings (t+1)
850
800
750
700
650
-1
-.5
0
.5
1
GPA - cutoff
This slope is $\beta_2 + \beta_3$
This slope is β_2
This difference is β_1

Example - GPA

$$Earnings_i = \beta_0 + \beta_1 Above_i + \beta_2 (GPA - 3.50)_i + \beta_3 Above_i * (GPA - 3.50)_i + u_i$$

- Normally, we worry that our treatment variable is correlated with unobservables in the error term.
  - If we were to run a regression of earnings on GPA, this would be a significant concern!
- Here, we assume that people close to the cutoff are about the same, and if this assumption holds, those endogeneity concerns go away.

Example - GPA

- So, what is the intuition here?
  - We've established that GPA is not random.
  - However, we can assume that people cannot perfectly control their GPA.
    - Clearly this is true, otherwise people would get perfect GPAs.
    - People do better or worse than expected in classes, leading to different grades.
  - Then, there is an element of luck in whether a person ends up just above the cutoff.

Example - GPA

- If we were to zoom in on people sufficiently close to the 3.50 cutoff:
  - We would expect there to be roughly the same number of people on either side of the cutoff.
  - We would expect them to have similar observables.
  - Since we expect them to have similar observables, we could assume that they have similar unobservables, and so the only difference between them was the award.

RD Assumptions

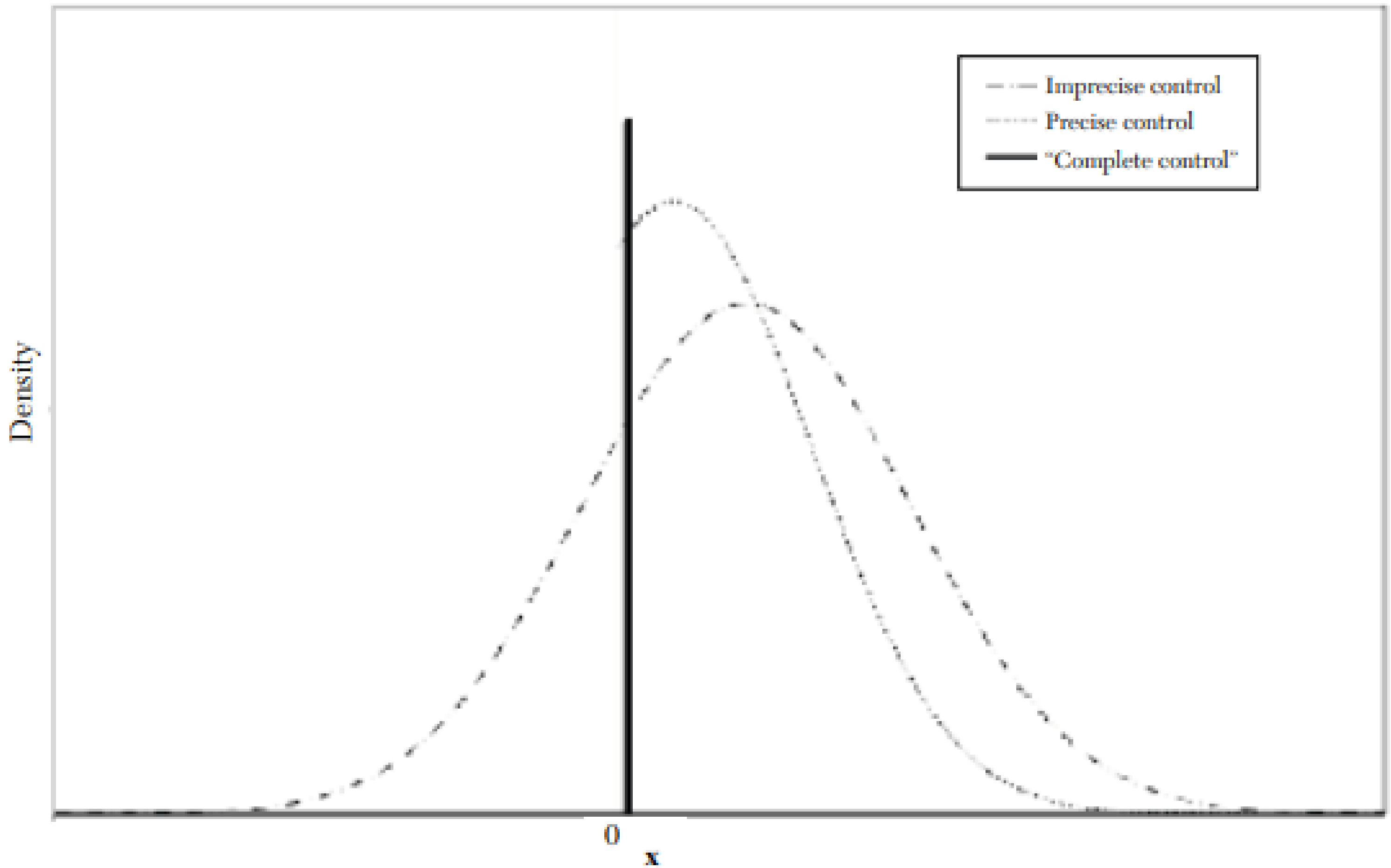
- The RD design relies on the following key assumptions.
  - Observations on either side of the cutoff are good counterfactuals for each other.
    - “Potential outcomes are continuous at the cutoff.”
  - There is no precise manipulation of the running variable.
    - Individuals cannot finely change their running variable to land on whichever side of the cutoff.

- In the context of our GPA example, students just above the a 3.50 need to be about the same as students just below a 3.50, and students shouldn’t be able to precisely determine what their GPA is.

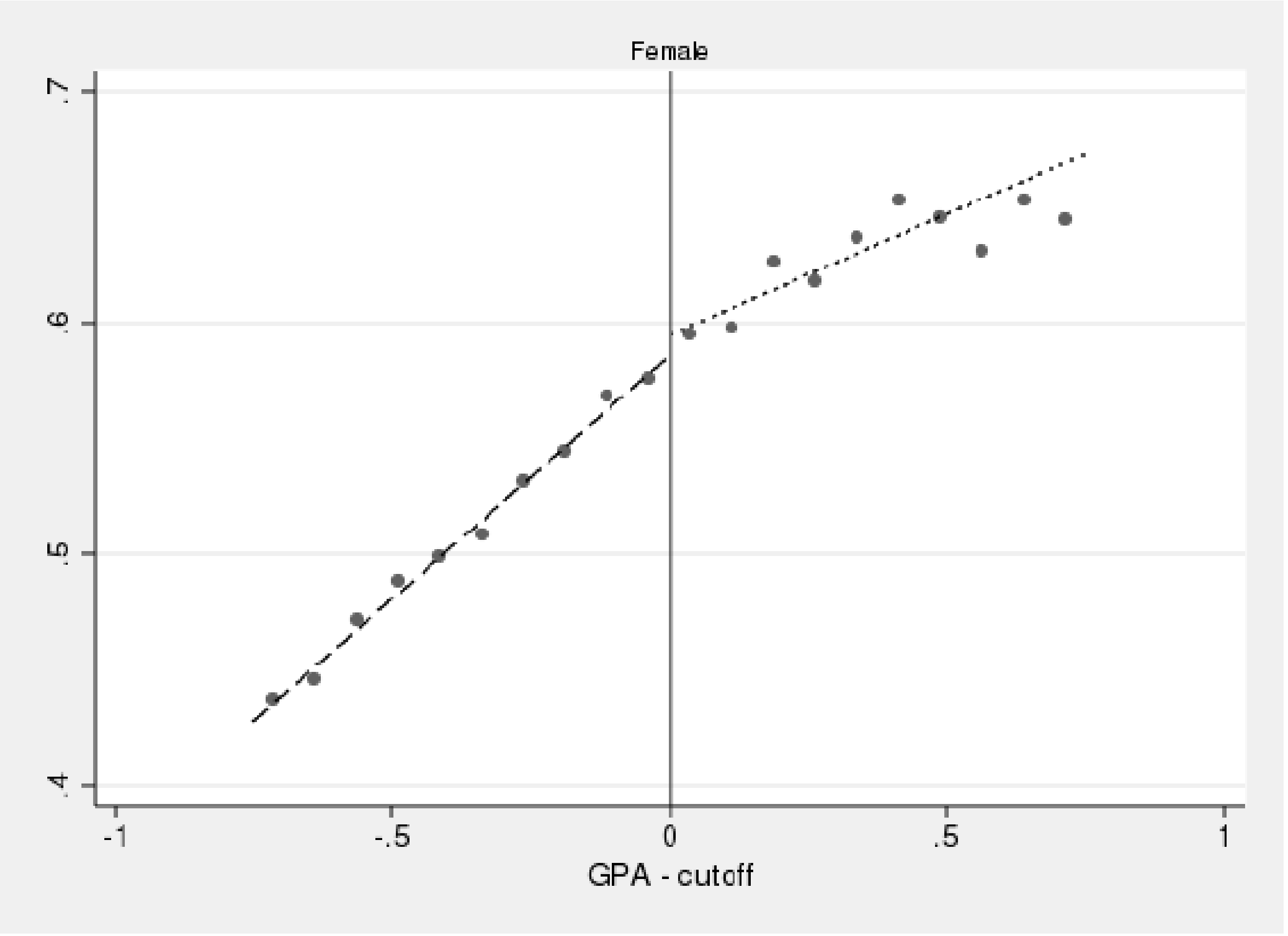
Testing RD Assumptions

- One of the appeals of RD is that there are very convincing tests to determine whether the RD assumptions are valid.
  - Histogram test
    - Plot a histogram of the running variable. If there is no manipulation, we would expect the histogram to be smooth through the cutoff. If there is manipulation, we would expect to see bunching on either side of the cutoff.
  - Covariate smoothness test.
    - Run an RD regression on the other observable covariates. If we find smoothness (no discontinuity) through the cutoff for these other observables, this is good evidence that unobservables are also balanced through the cutoff.
      - This suggests that observations on either side of the cutoff are good counterfactuals for each other!

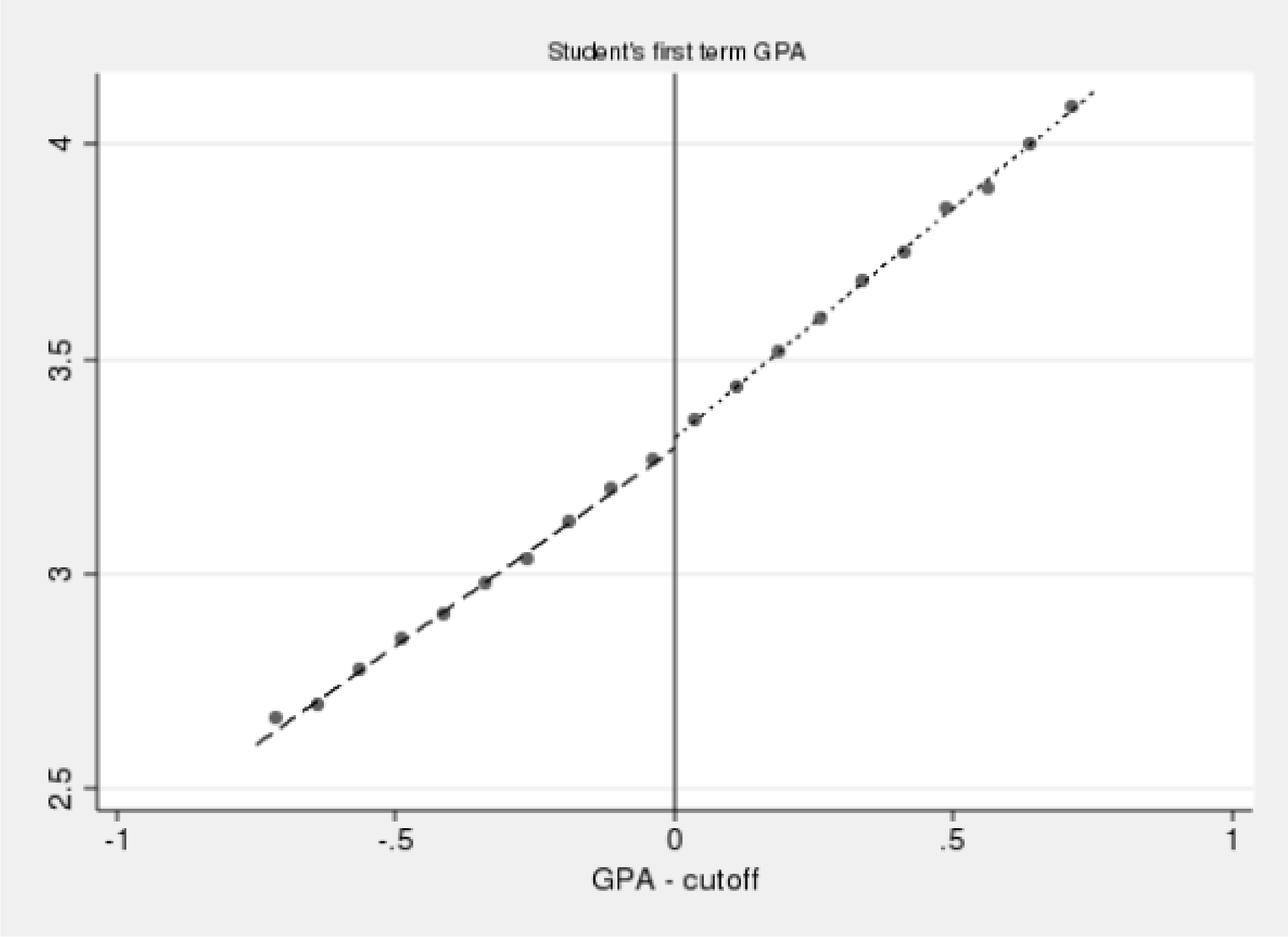
Histogram Test



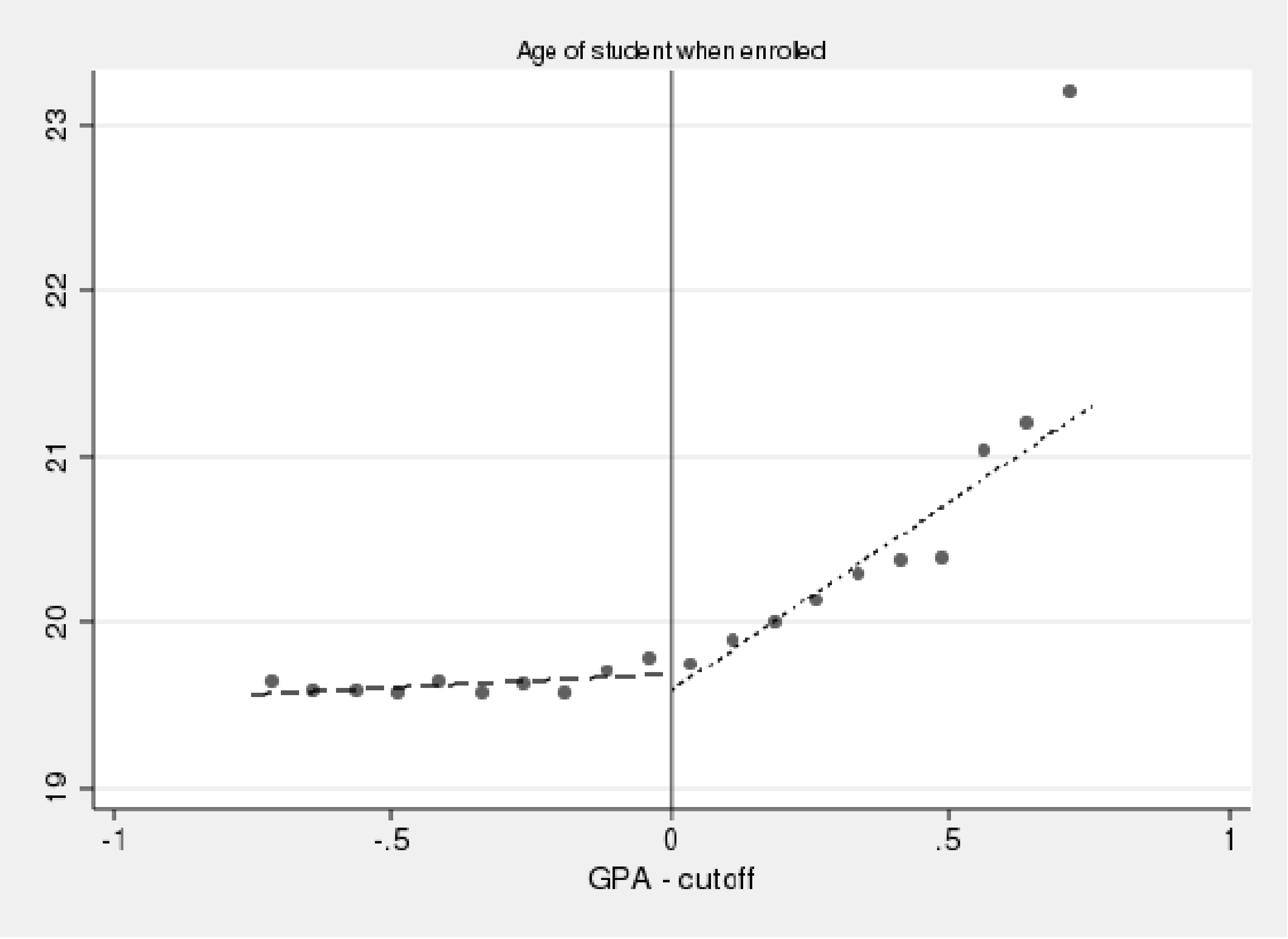
Covariate Smoothness Test



Covariate Smoothness Test



Covariate Smoothness Test



When is an RD Design Appropriate?

- There needs to be a threshold that determines treatment.
- You need some knowledge about how assignment is determined,
  - Can people precisely manipulate the running variable?
  - Are people on either side of the cutoff good counterfactuals for each other?

When is an RD Design Appropriate?

- If individuals cannot precisely manipulate the running variable, then we have "local randomization" around the cutoff.
  - Treatment is as good as randomly assigned around the cutoff.
- The RD estimate captures the Local Average Treatment Effect (LATE) around the cutoff.
  - This estimate is only valid for observations sufficiently close to the cutoff.
  - In the context of the GPA example, the coefficient we estimate is valid for people right around a 3.50; it is probably not valid for someone with a 4.00.