

# STA 235H - Regression Discontinuity Design

Fall 2022

McCombs School of Business, UT Austin

# Announcements

- Feedback for the **project's idea pitch** was posted!
  - Review the comments and incorporate them moving forward (e.g. switch from causal inference to prediction).
  - If you have questions, reach out!
  - Next submission is on Nov 10th (check the website for details)
- Remember that you can **drop an assignment**:
  - That is meant to be used in any situation you can't submit on time.
- Office hours on Thursday will start at **4.00pm**

# Last class

- **Natural Experiments**
  - RCTs in the wild.
  - Importance of the context and checking potential confounding.
- **Difference-in-Differences (DD):**
  - How we can use two wrong estimates to get a right one.
  - Assumptions behind DD.



# Today



- **Regression Discontinuity Design (RDD):**
  - How can we use discontinuities to recover causal effects?
  - Assumptions behind RD designs.
- Structure for today:
  - Start: Material + Examples
  - Finish: Exercise

Mind the gap

# Another identification strategy

- We have seen:

RCTs

Selection on observables

Natural experiments

Difference-in-Differences

**Regression Discontinuity Designs**

**Tell me something about the readings/videos you had to watch for this week**

# Introduction to Regression Discontinuity Designs

## Regression Discontinuity (RD) Designs

Arbitrary rules determine treatment assignment

E.g.: If you are above a threshold, you are assigned to treatment, and if your below, you are not (or vice versa)



# Geographic discontinuities

Turnout • 0.2 • 0.4 • 0.6  
Treatment Status (Eastern Side of Time Zone Border) • No • Yes

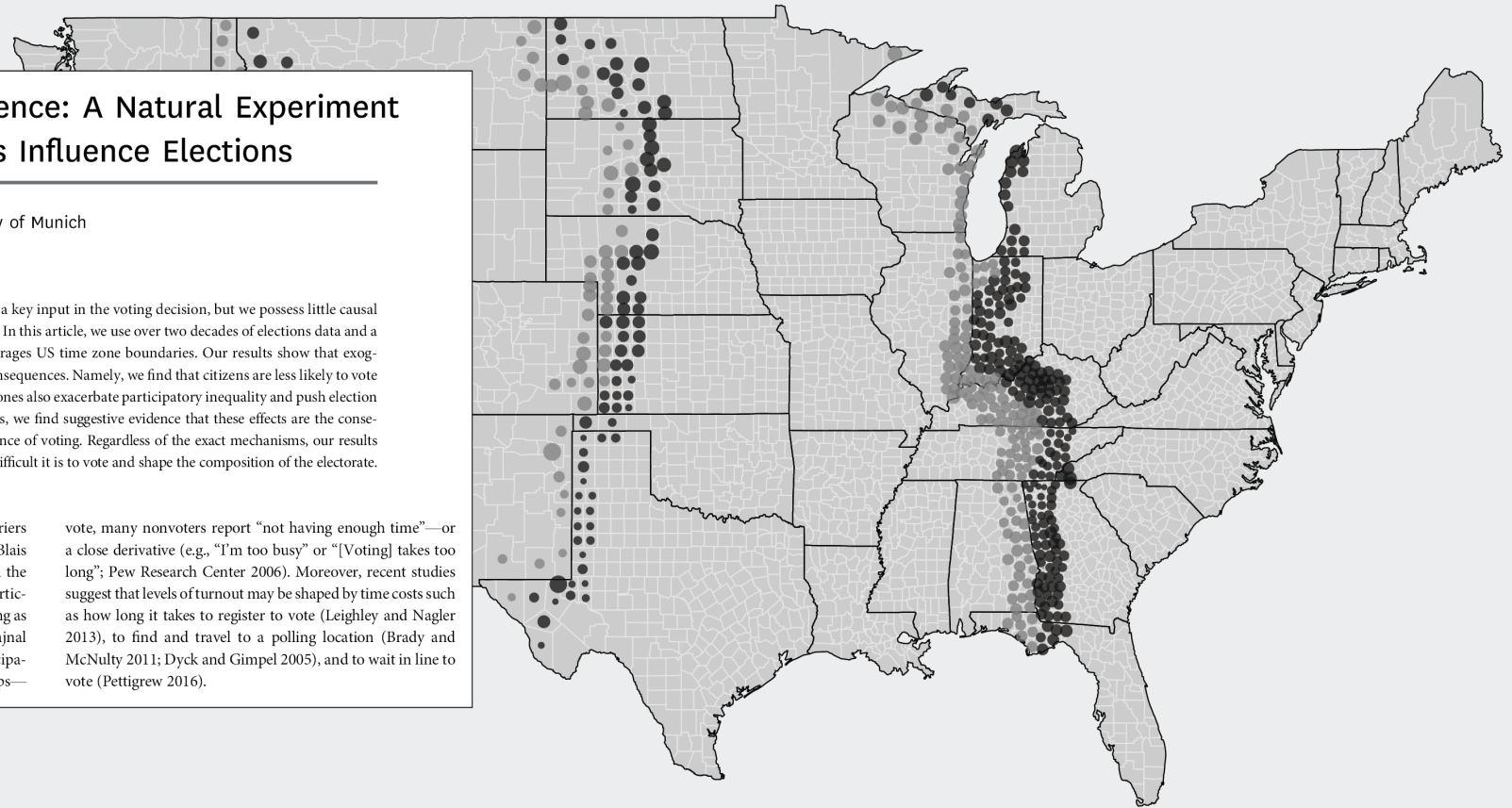
## When Time Is of the Essence: A Natural Experiment on How Time Constraints Influence Elections

**Jerome Schafer**, Ludwig Maximilian University of Munich  
**John B. Holbein**, University of Virginia

Foundational theories of voter turnout suggest that time is a key input in the voting decision, but we possess little causal evidence about how this resource affects electoral behavior. In this article, we use over two decades of elections data and a novel geographic regression discontinuity design that leverages US time zone boundaries. Our results show that exogenous shifts in time allocations have significant political consequences. Namely, we find that citizens are less likely to vote if they live on the eastern side of a time zone border. Time zones also exacerbate participatory inequality and push election results toward Republicans. Exploring potential mechanisms, we find suggestive evidence that these effects are the consequence of insufficient sleep and moderated by the convenience of voting. Regardless of the exact mechanisms, our results indicate that local differences in daily schedules affect how difficult it is to vote and shape the composition of the electorate.

Although in recent years the administrative barriers to voting have declined in many democracies (Blais 2010), many eligible citizens still fail to vote. In the United States, about 40% of registered voters do not participate in presidential elections, with abstention rates soaring as high as 60% in midterms and 70% in local elections (Hajnal and Trounstein 2016). Moreover, rates of political participation have remained stubbornly low among vulnerable groups—

vote, many nonvoters report “not having enough time”—or a close derivative (e.g., “I’m too busy” or “[Voting] takes too long”; Pew Research Center 2006). Moreover, recent studies suggest that levels of turnout may be shaped by time costs such as how long it takes to register to vote (Leighley and Nagler 2013), to find and travel to a polling location (Brady and McNulty 2011; Dyck and Gimpel 2005), and to wait in line to vote (Pettigrew 2016).



# Time discontinuities

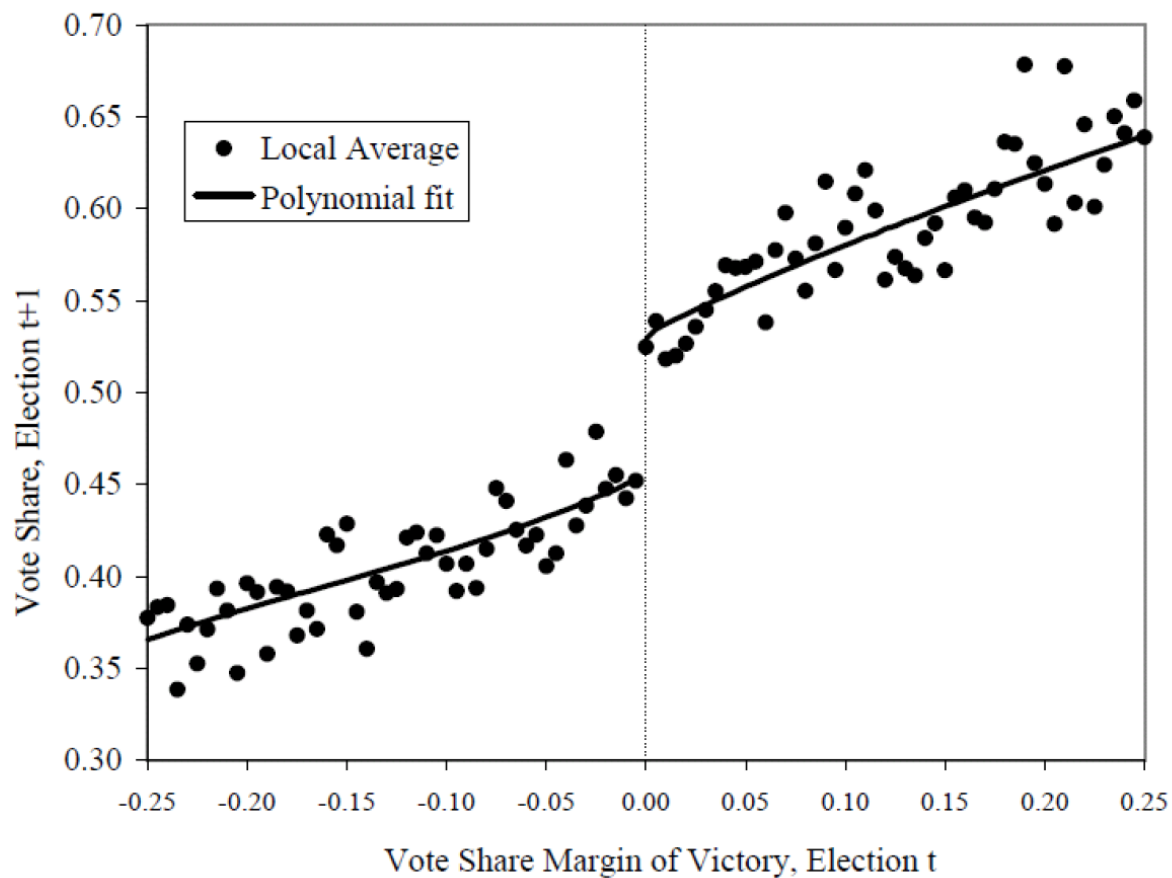
## After Midnight: A Regression Discontinuity Design in Length of Postpartum Hospital Stays<sup>†</sup>

By DOUGLAS ALMOND AND JOSEPH J. DOYLE JR.\*

*Estimates of moral hazard in health insurance markets can be confounded by adverse selection. This paper considers a plausibly exogenous source of variation in insurance coverage for childbirth in California. We find that additional health insurance coverage induces substantial extensions in length of hospital stay for mother and newborn. However, remaining in the hospital longer has no effect on readmissions or mortality, and the estimates are precise. Our results suggest that for uncomplicated births, minimum insurance mandates incur substantial costs without detectable health benefits. (JEL D82, G22, I12, I18, J13)*

# Voting discontinuities

Figure IVa: Democrat Party's Vote Share in Election  $t+1$ , by Margin of Victory in Election  $t$ : local averages and parametric fit



**You can find discontinuities  
everywhere!**

# Key Terms

**Running/ forcing variable**

Index or measure that determines eligibility

**Cutoff/ cutpoint/ threshold**

Number that formally assigns you to a program or treatment

**Let's look at an example**

# Hypothetical tutoring program

**Students take an entrance exam**

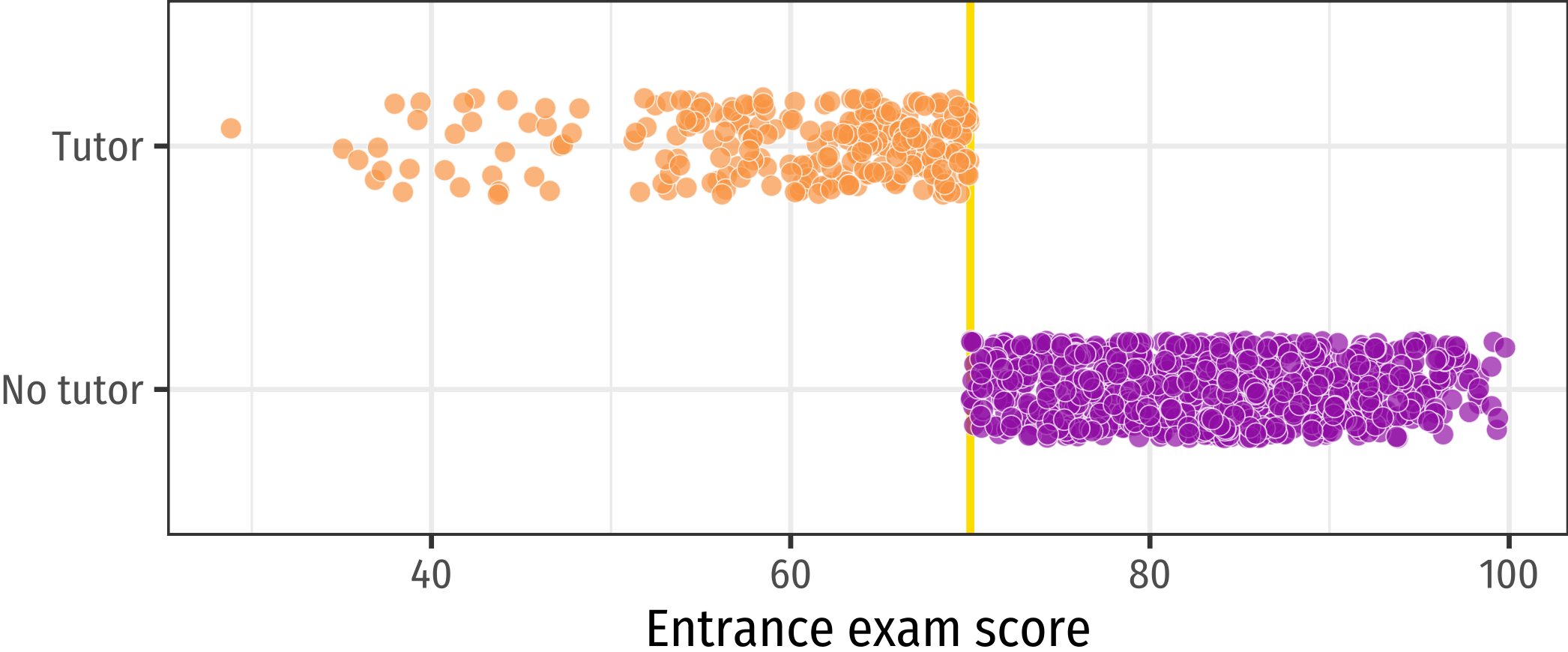
**Those who score 70 or lower  
get a free tutor for the year**

**Students then take an exit exam  
at the end of the year**

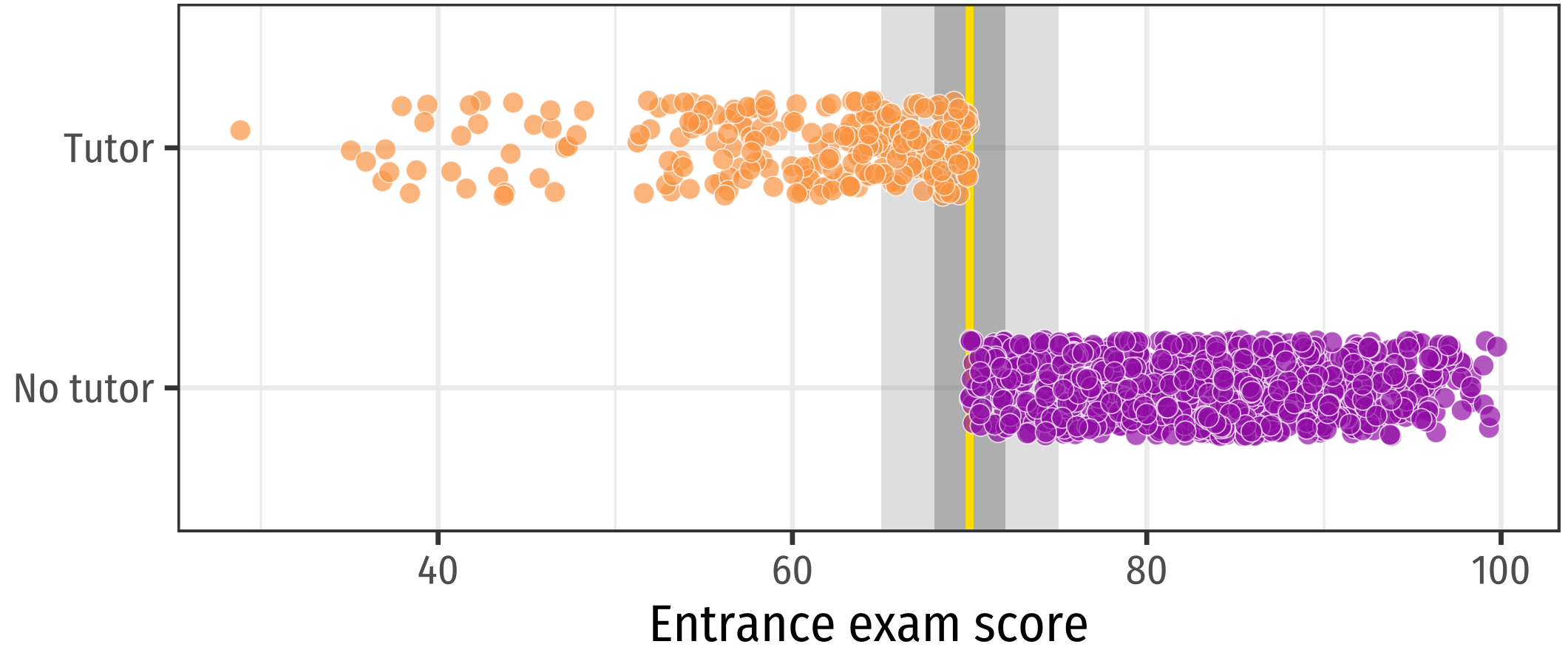
**Can we compare students who got a tutor vs those that did not to capture the effect of having a tutor on their exit exam?**



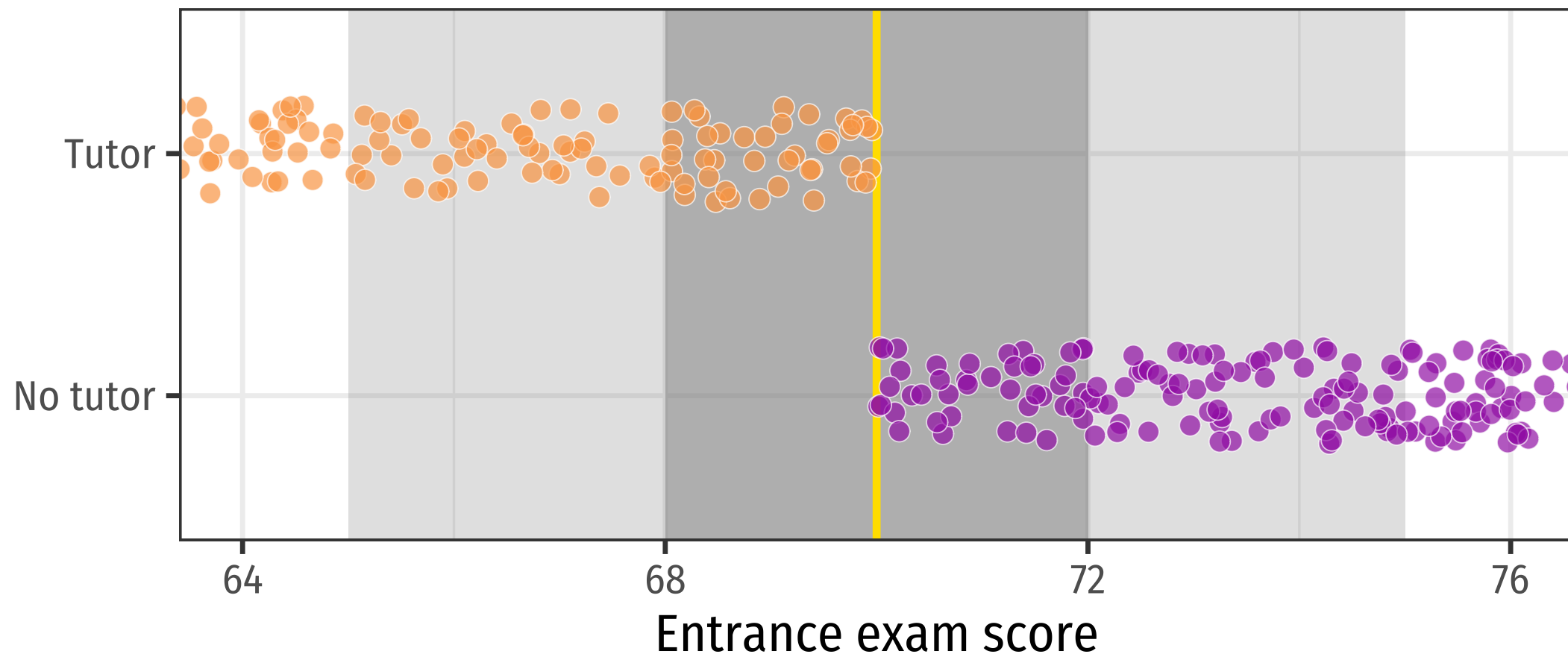
# Assignment based on entrance score



# Let's look at the area close to the cutoff



# Let's get closer



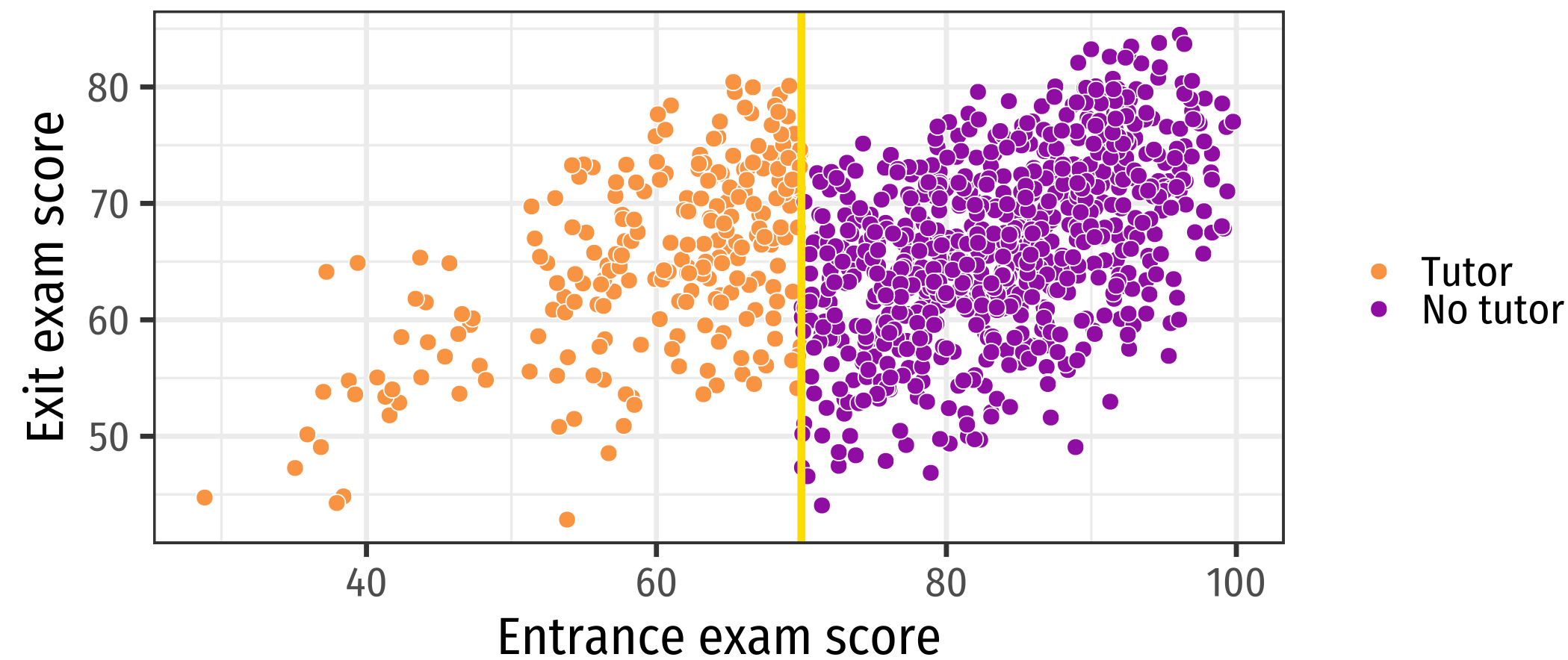
# Causal inference intuition

Observations right before and after the threshold are essentially the same

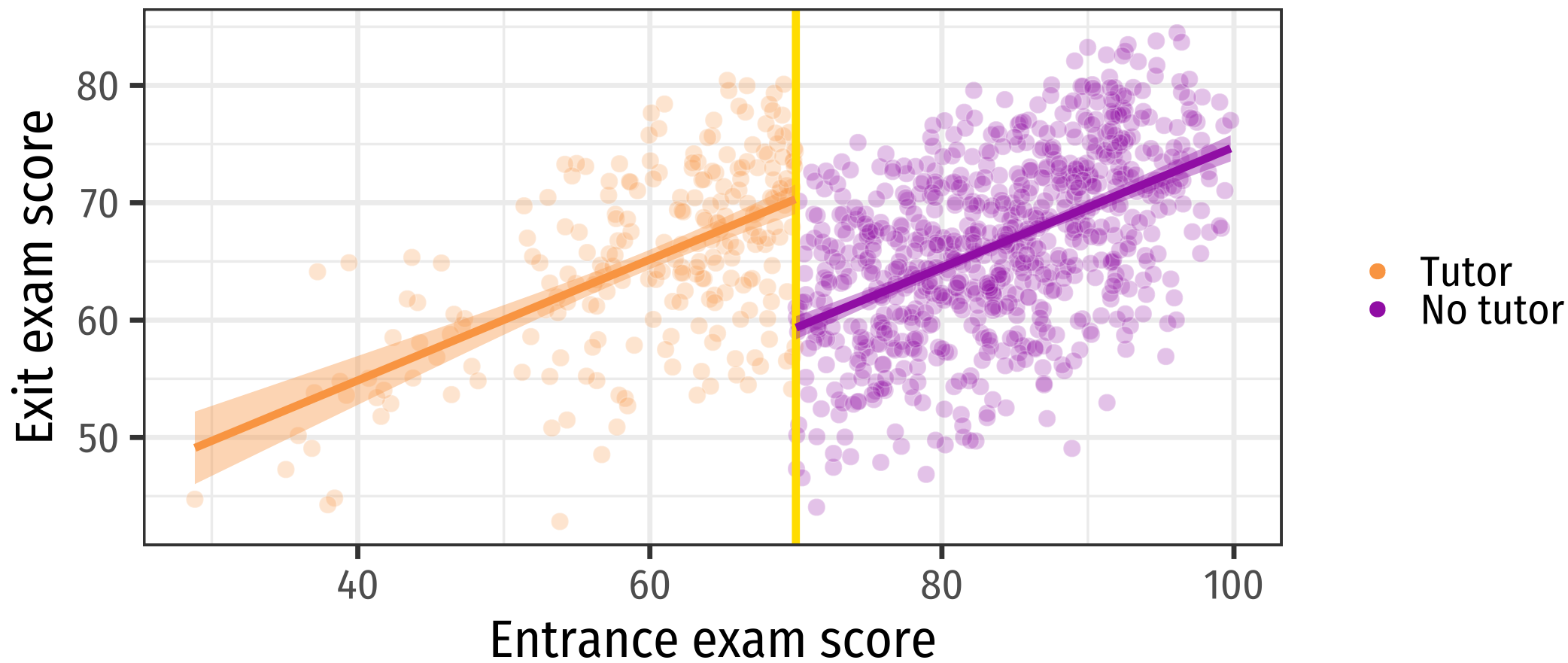
Pseudo treatment and control groups!

Compare outcomes right at the cutoff

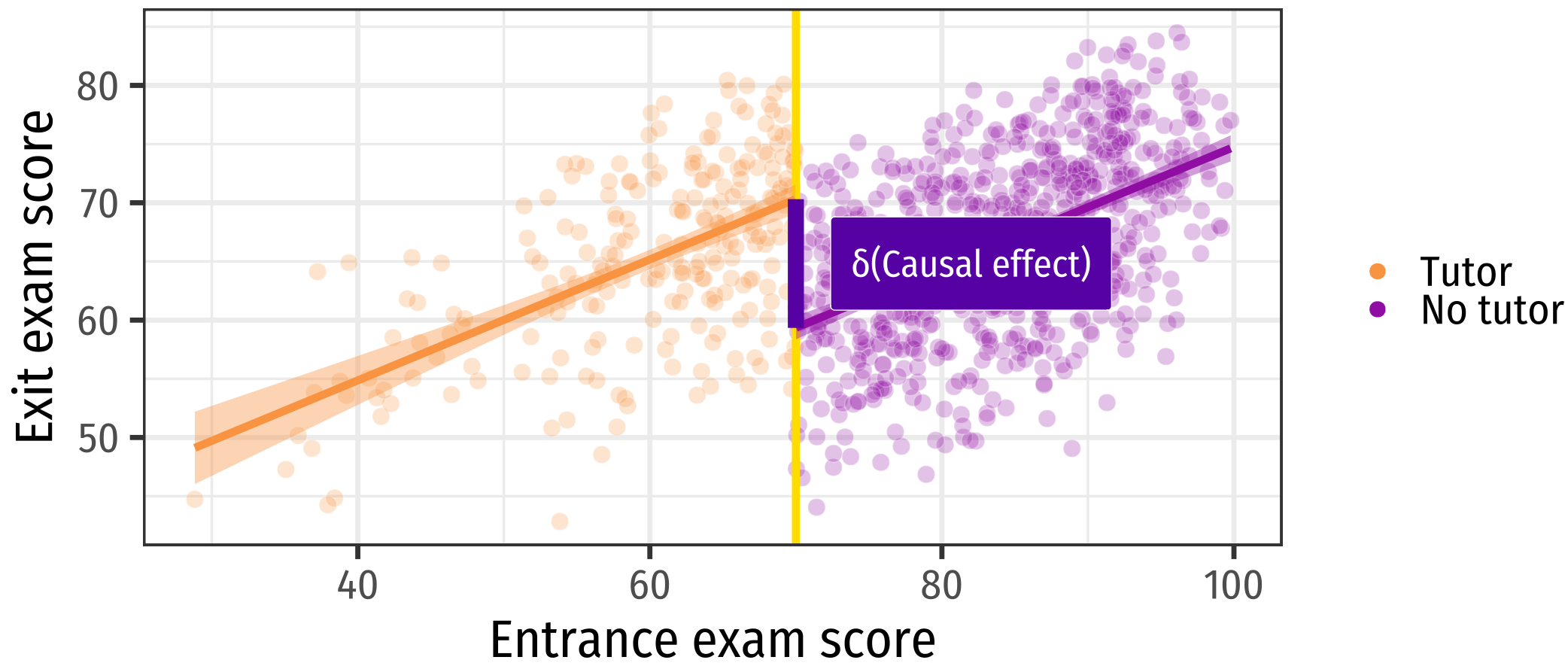
# Exit exam results according to running variable



# Fit a regression at the right and left side of the cutoff



# Fit a regression at the right and left side of the cutoff



**What population within my  
sample am I comparing?**



**My estimand is the  
Local Average Treatment Effect  
(LATE) for units at  $R=c$**

Is that what we want?

Probably not ideal, there may not be *any* units with  $R=c$

... but better LATE than nothing!

# Conditions required for identification

- Threshold rule **exists** and cutoff point is **known**
- The running variable  $R_i$  is **continuous** near  $c$ .
- Key assumption:

Continuity of  $E[Y(1)|R]$  and  $E[Y(0)|R]$  at  $R=c$

That's the math-y way to say that the only thing that changes right at the cutoff is the treatment assignment!

**How can I check if this assumption holds?**

You can't! (it's an assumption)

... but there are robustness checks (e.g. check covariates at the threshold)

Estimation in practice

# How do we actually estimate an RD?

- The simplest way to do this is to fit a regression using **an interaction of the treatment variable and the running variable**:

$$Y = \beta_0 + \beta_1(R - c) + \beta_2 I[R > c] + \beta_3 (R - c) I[R > c] + \varepsilon$$

# How do we actually estimate an RD?

- The simplest way to do this is to fit a regression using **an interaction of the treatment variable and the running variable**:

$$Y = \beta_0 + \beta_1 \underbrace{(R - c)}_{\text{Distance to the cutoff}} + \beta_2 \underbrace{\mathbb{I}[R > c]}_{\text{Treatment}} + \beta_3 \underbrace{(R - c)}_{\text{Distance to the cutoff}} \underbrace{\mathbb{I}[R > c]}_{\text{Treatment}} + \varepsilon$$

- We can simplify this with new notation:

$$Y_i = \beta_0 + \beta_1 R' + \beta_2 Z + \beta_3 R' \times Z$$

where  $Z$  is a binary treatment variable and  $R'$  is the running variable centered around the cutoff

**Can you identify these parameters in a plot?**

**Let's identify coefficients**



**Let's see an example**

# Discount and sales

- You are managing a retail store and notice that sales are low in the mornings, so you want to improve those numbers.
- You decide to give the first 1,000 customers that show up **10% off**



# Discounts and sales

- 1) Can you compare sales for the first 1,000 customers with sales for 1,001-2,000 to get a causal effect?
- 2) Can you compare sales for the first 1,000 customers today vs 1,000 customers yesterday to get a causal effect?
- 3) What could you do to get a causal effect?

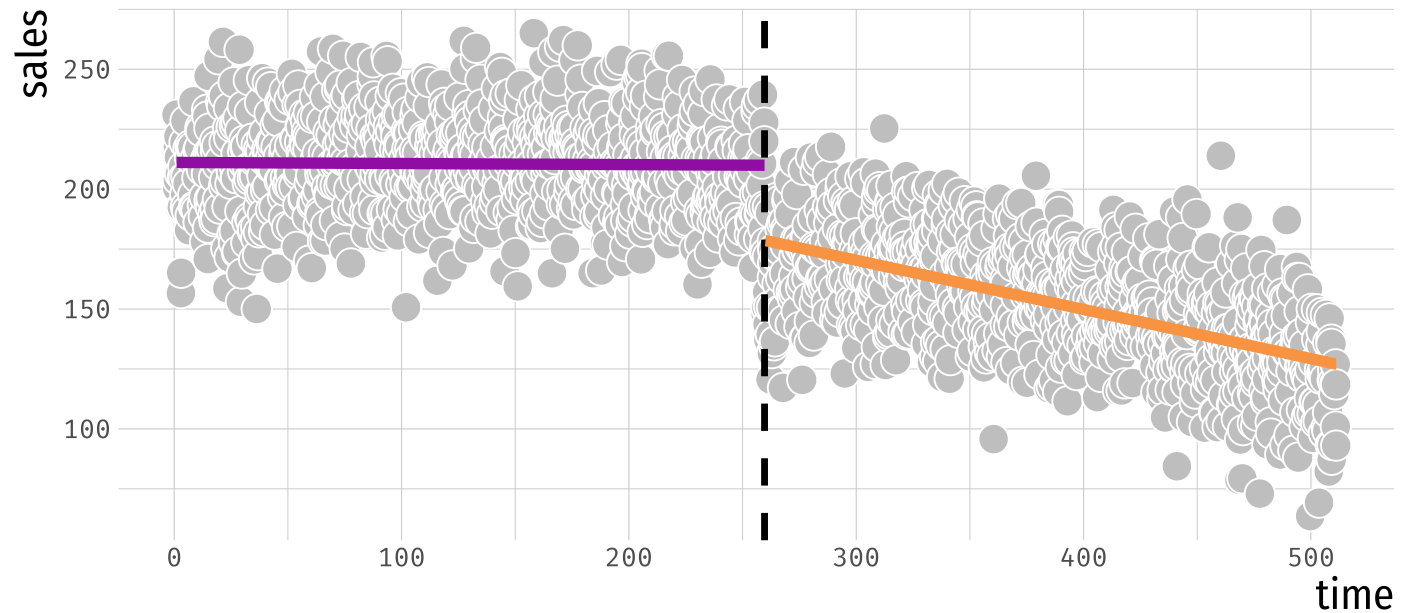
# Discounts and sales: Can we use an RDD?

- In RDD, we need to check that there are **unbalances in covariates across the threshold**.

```
sales <- sales %>% mutate(dist = c-time)
lm(income ~ dist*treat, data = sales)
```

# RDD on sales using linear models

```
lm(sales ~ dist*treat, data = sales)
```



# RDD on sales using linear models

```
summary(lm(sales ~ dist*treat, data = sales))
```

```
##
## Call:
## lm(formula = sales ~ dist * treat, data = sales)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -65.738 -13.940   0.051  13.538  76.515
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 178.640954   1.300314  137.38  <2e-16 ***
## dist         0.205355   0.008882   23.12  <2e-16 ***
## treat        31.333952   1.842338   17.01  <2e-16 ***
## dist:treat   -0.200845   0.012438  -16.15  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.52 on 1996 degrees of freedom
## Multiple R-squared:  0.6939,    Adjusted R-squared:  0.6934
## F-statistic: 1508 on 3 and 1996 DF,  p-value: < 2.2e-16
```

# We can be more flexible

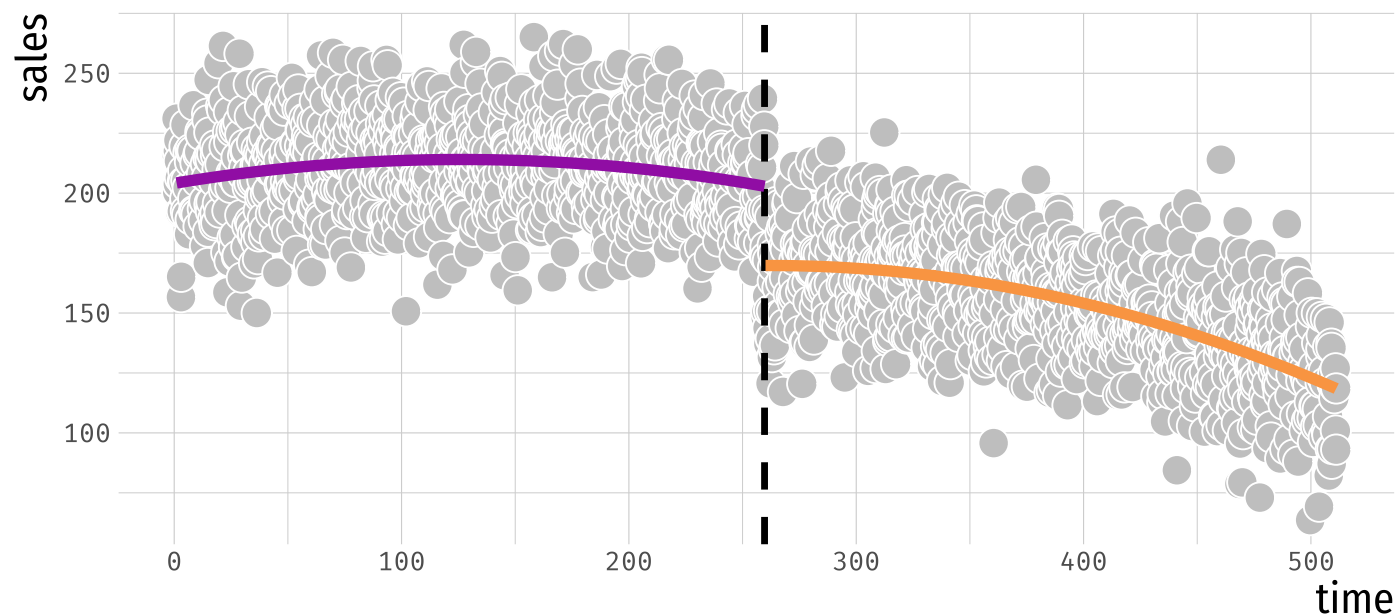
- The previous example just included linear terms, but you can also be more flexible:

$$Y = \beta_0 + \beta_1 f(R') + \beta_2 Z + \beta_3 f(R') \times Z + \varepsilon$$

- Where  $f$  is any function you want.

# What happens if we fit a quadratic model?

```
lm(sales ~ dist*treat + treat*I(dist^2), data = sales)
```





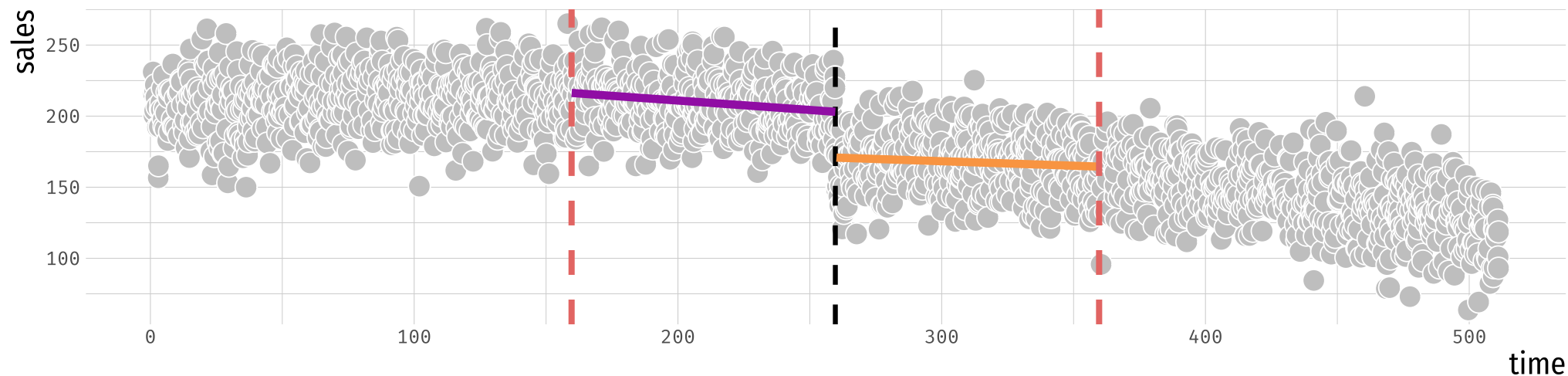
# What happens if we fit a quadratic model?

```
summary(lm(sales ~ dist*treat + treat*I(dist^2), data = sales))
```

```
##
## Call:
## lm(formula = sales ~ dist * treat + treat * I(dist^2), data = sales)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -66.090 -13.979   0.239  13.154  76.656
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.698e+02  1.937e+00  87.665 < 2e-16 ***
## dist         -4.302e-03  3.556e-02  -0.121  0.903725
## treat         3.308e+01  2.747e+00  12.041 < 2e-16 ***
## I(dist^2)     -8.288e-04  1.363e-04  -6.083  1.41e-09 ***
## dist:treat     1.713e-01  4.964e-02   3.452  0.000569 ***
## treat:I(dist^2) 2.034e-04  1.877e-04   1.084  0.278554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.23 on 1994 degrees of freedom
## Multiple R-squared:  0.7029,    Adjusted R-squared:  0.7021
## F-statistic: 943.5 on 5 and 1994 DF,  p-value: < 2.2e-16
```

# What happens if we only look at observations close to $c$ ?

```
sales_close <- sales %>% filter(dist>-100 & dist<100)  
lm(sales ~ dist*treat, data = sales_close)
```



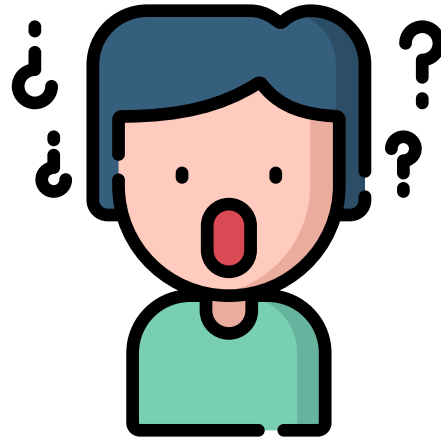
# How do they compare?

```
summary(lm(sales ~ dist*treat, data = sales_close))
```

```
##
## Call:
## lm(formula = sales ~ dist * treat, data = sales_close)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -53.241 -14.764   0.268  12.938  57.811
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 170.84457    2.05528   83.125  <2e-16 ***
## dist         0.06345    0.03542    1.791   0.0736 .
## treat        32.21243    2.93614   10.971  <2e-16 ***
## dist:treat   0.06909    0.05047    1.369   0.1714
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.25 on 782 degrees of freedom
## Multiple R-squared:  0.5261,    Adjusted R-squared:  0.5243
## F-statistic: 289.4 on 3 and 782 DF,  p-value: < 2.2e-16
```

# Potential problems

- There are **many potential problems** with the previous examples:
  - Which polynomial function should we choose? Linear, quadratic, other?
  - What bandwidth should we choose? Whole sample?  $[-100, 100]$ ?



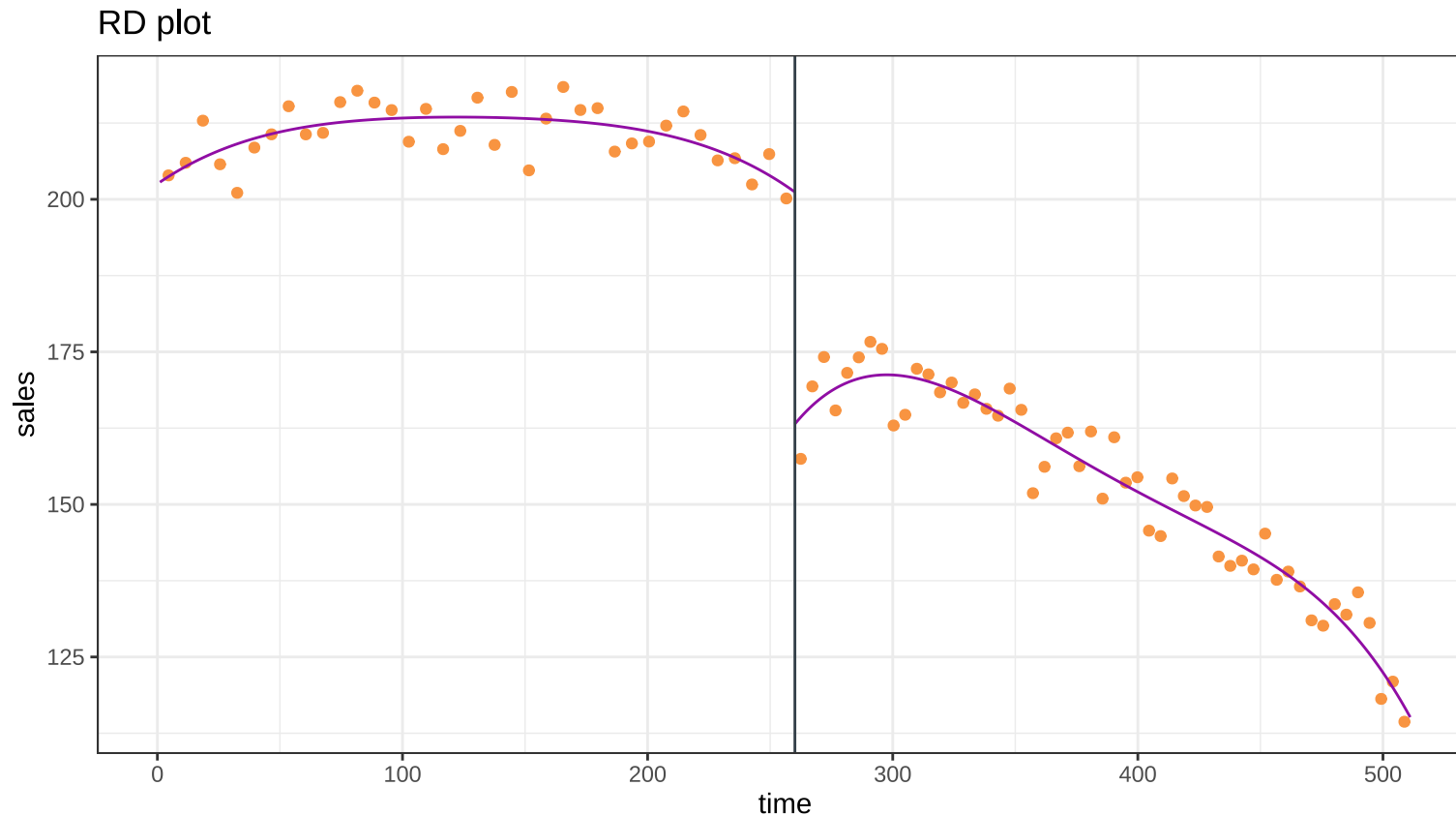
- There are some ways to address these concerns.

# Package `rdrobust`

- Robust Regression Discontinuity introduced by Cattaneo, Calonico, Farrell & Titiunik (2014).
- Use of **local polynomial** for fit.
- **Data-driven optimal bandwidth** (bias vs variance).
- `rdrobust`: Estimation of LATE and opt. bandwidth
- `rdplot`: Plotting RD with nonparametric local polynomial.
- `rddensity`: Check density across the cutoff (from the `rddensity` package)

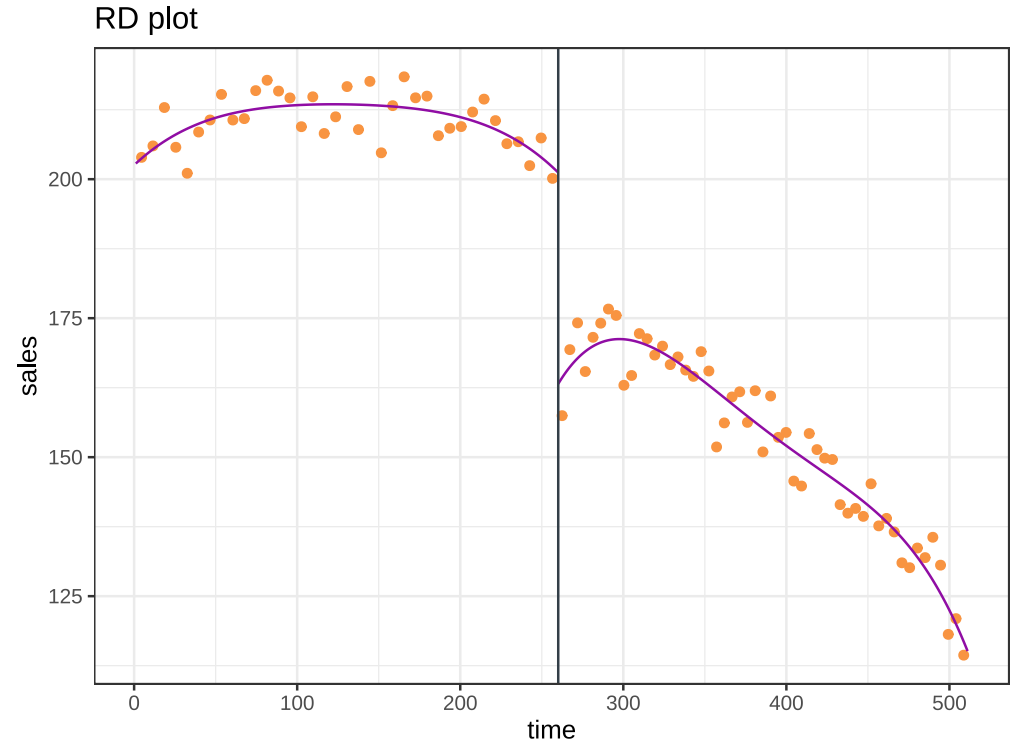
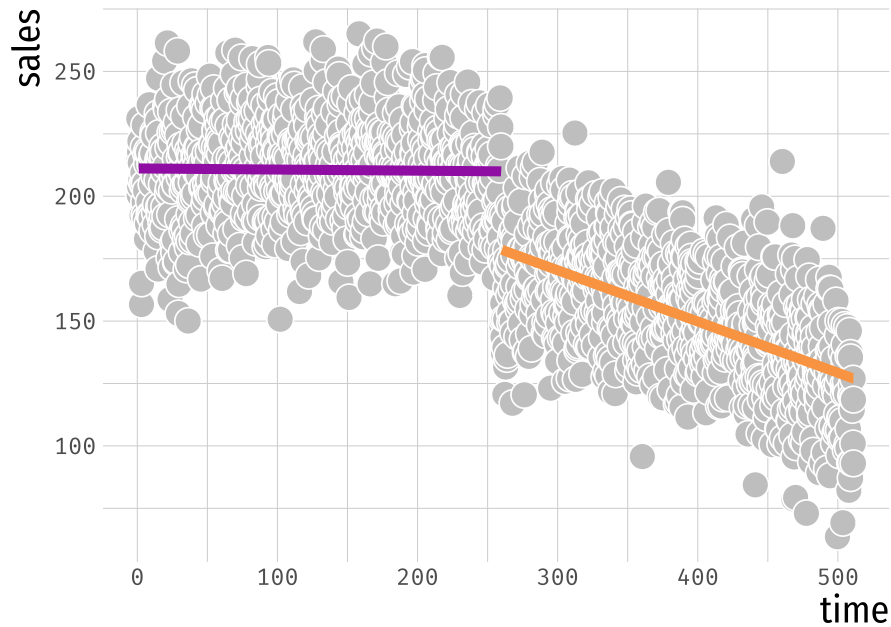
# Let's compare with previous parametric results

```
rdplot(y = sales$sales, x = sales$time, c = c,  
       title = "RD plot", x.label = "time", y.label = "sales")
```



# Let's compare with previous parametric results

```
rdplot(y = sales$sales, x = sales$time, c = c,  
       title = "RD plot", x.label = "time", y.label = "sales")
```



# Let's compare with previous parametric results

```
summary(rdrobust(y = sales$sales, x = sales$time, c = c))
```

```
## Sharp RD estimates using local polynomial regression.
```

```
##
```

```
## Number of Obs.          2000
```

```
## BW type                  mserd
```

```
## Kernel                   Triangular
```

```
## VCE method              NN
```

```
##
```

```
## Number of Obs.          1000      1000
```

```
## Eff. Number of Obs.      202       213
```

```
## Order est. (p)           1         1
```

```
## Order bias (q)           2         2
```

```
## BW est. (h)              54.304    54.304
```

```
## BW bias (b)              87.787    87.787
```

```
## rho (h/b)                0.619     0.619
```

```
## Unique Obs.             1000      1000
```

```
##
```

```
## =====
```

```
##      Method      Coef. Std. Err.      z    P>|z|      [ 95% C.I. ]
```

```
## =====
```

```
## Conventional  -37.434    4.344   -8.618   0.000  [-45.948 , -28.921]
```

```
## Robust        -         -    -7.610   0.000  [-48.596 , -28.691]
```

```
## =====
```



# Summary

For conducting an RDD:

- 1) Check that there is a **discontinuity in assignment** (e.g. `plot treat vs r`)
- 2) Check that **covariates are smooth across the cutoff** (e.g. RDD for `x` (baseline covariate))
- 3) Run a **RDD** using either a linear model (`lm( )`) or a nonparametric model (`rdrobust( )`)

# Takeaway points

- RD designs are **great** for causal inference!
  - Strong internal validity
  - Number of robustness checks
- **Limited** external validity.
- Make sure to check your data:
  - Discontinuity in treatment assignment
  - Density across the cutoff
  - Smoothness of covariates



# References

- Angrist, J. and S. Pischke. (2015). "Mastering Metrics". *Chapter 4*.
- Social Science Research Institute at Duke University. (2015). "Regression Discontinuity: Looking at People on the Edge: Causal Inference Bootcamp"