# STA 235 - Causal Inference: Regression Discontinuity Design

## Spring 2021

McCombs School of Business, UT Austin

# Another identification strategy

- We have seen:

RCTs

Selection on observables

Natural experiments

Differences-in-Differences

Regression Discontinuity Designs

# I'm on the edge [of glory?]

# 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)

# Key Terms

**Running/ forcing variable**

Index or measure that determines eligibility

**Cutoff/ cutpoint/ threshold**

Number that formally assigns you to a program or treatment

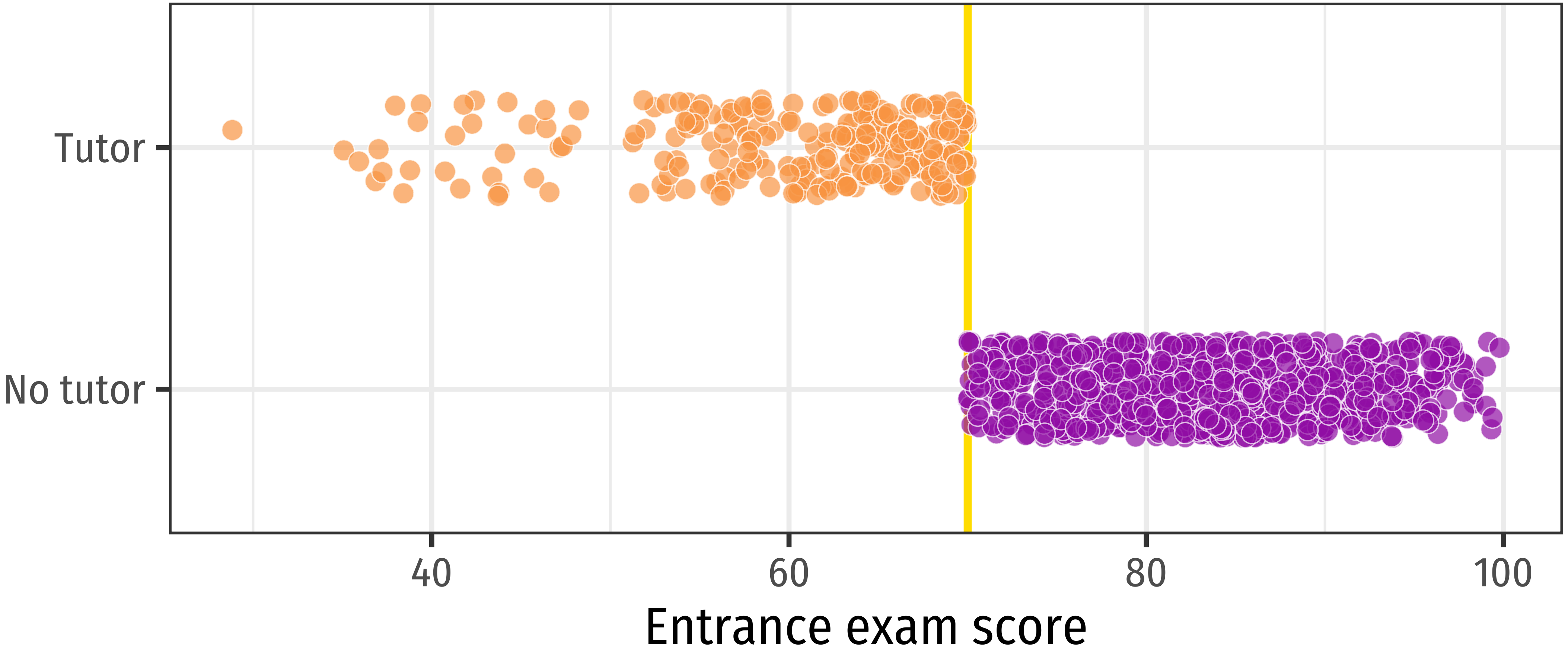
# Hypothetical tutoring program

Students take an entrance exam

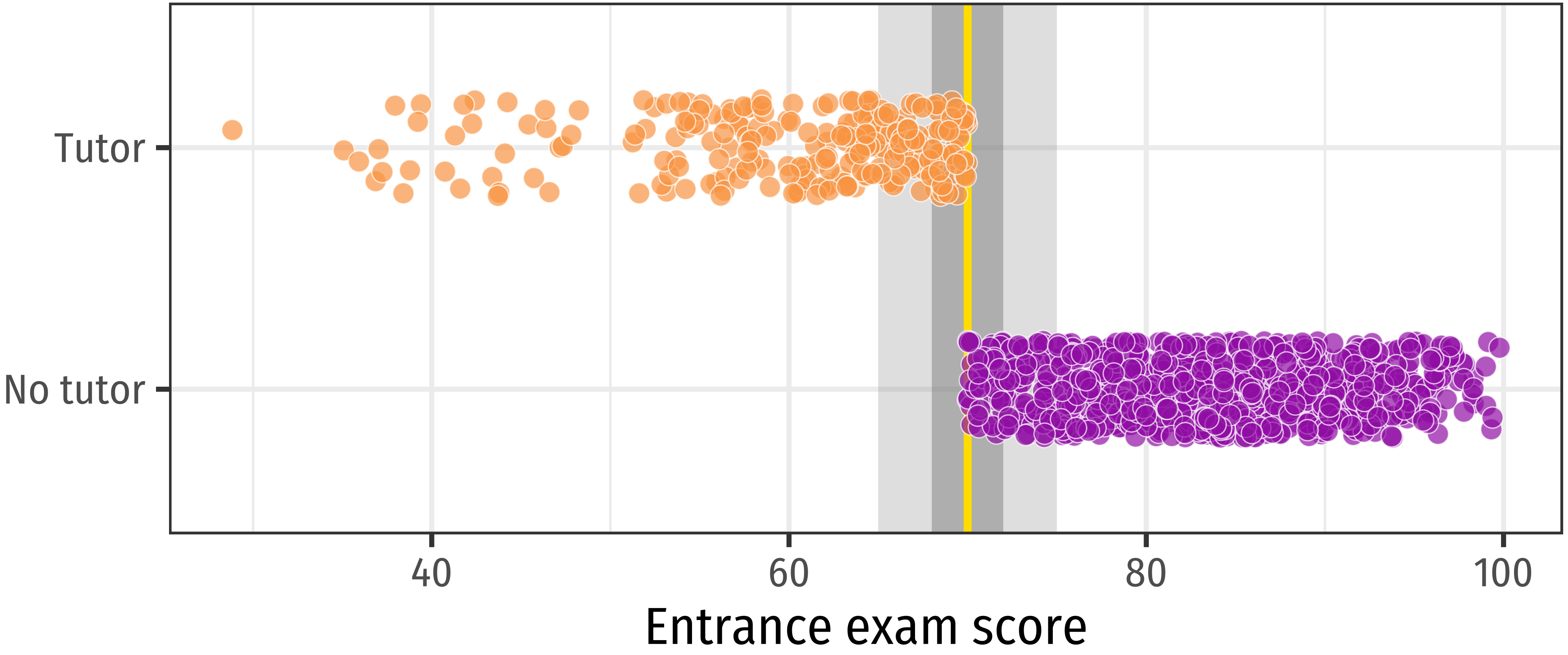
Those who score 70 or lower get a free tutor for the year

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

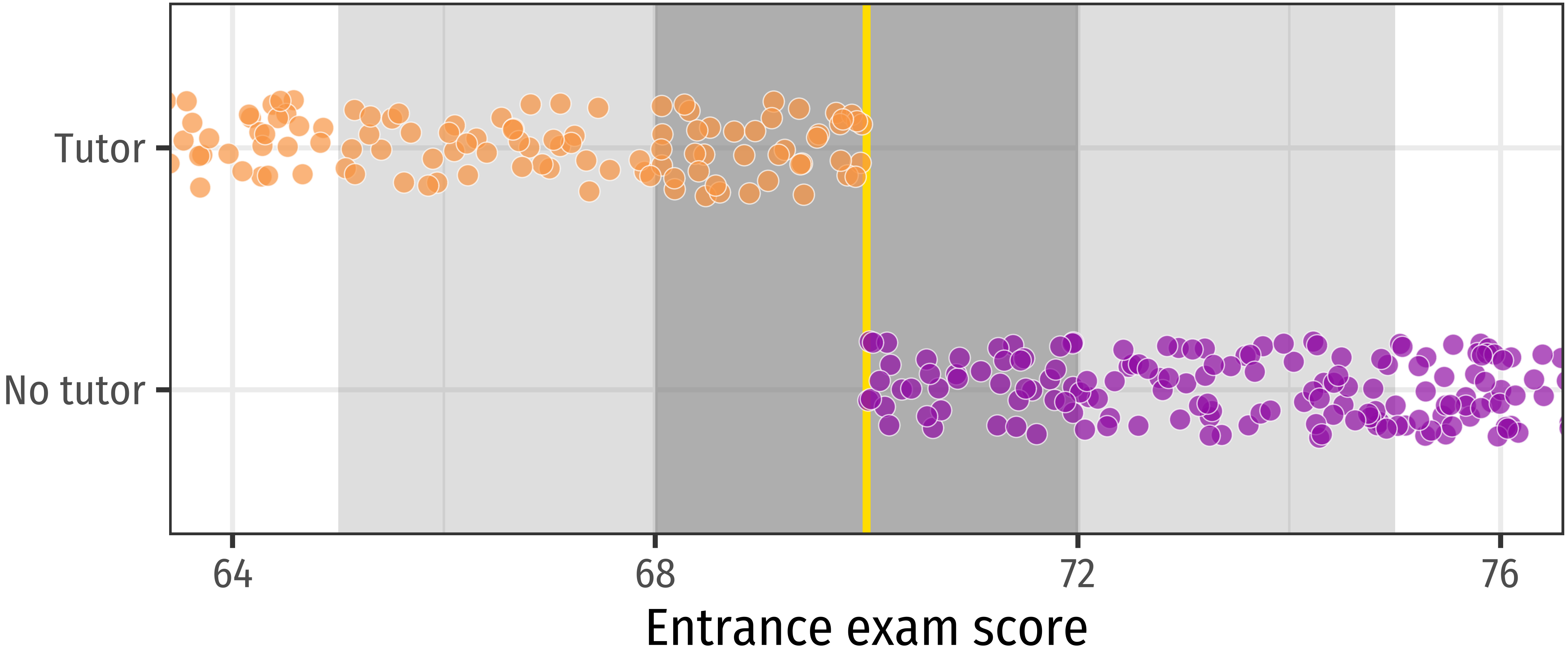
# Assignment based on entrance score



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



# Let's get closer



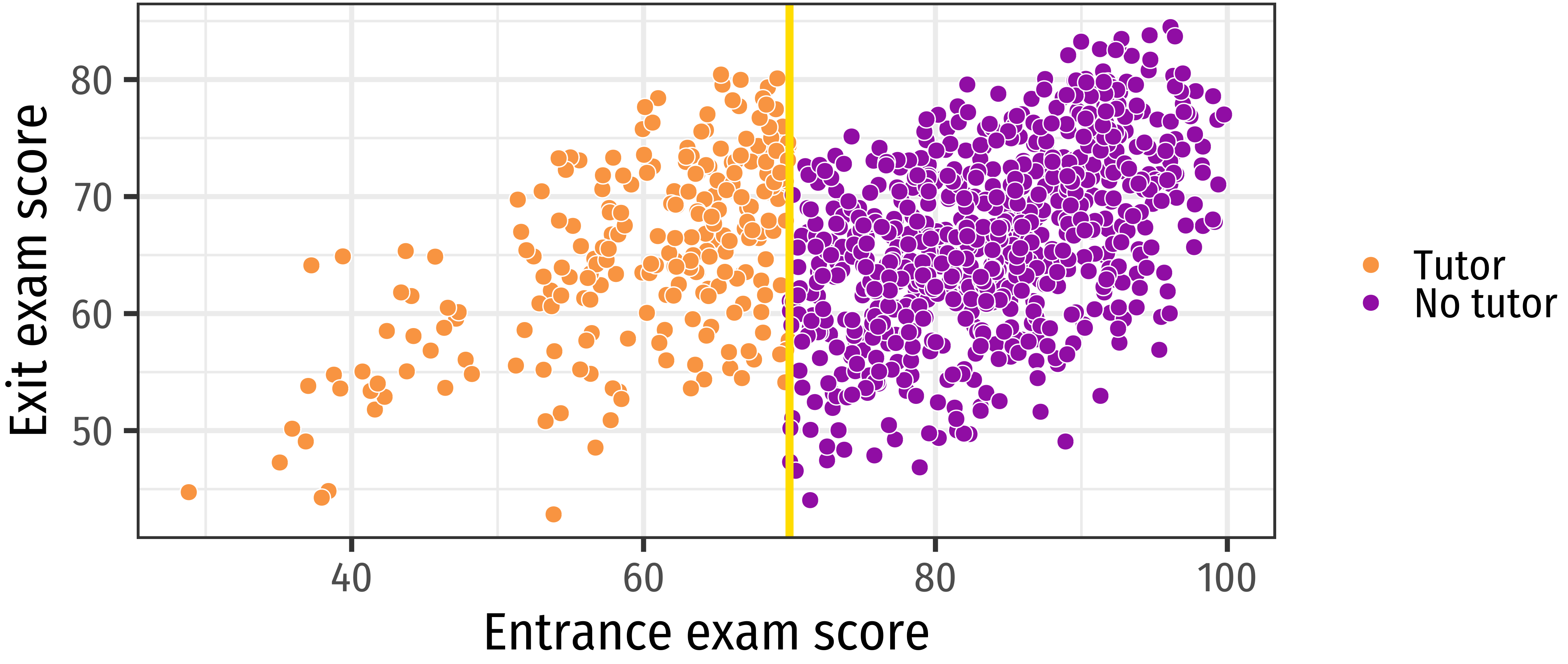
# Causal inference intuition

Observations right before and after the threshold are essentially the same

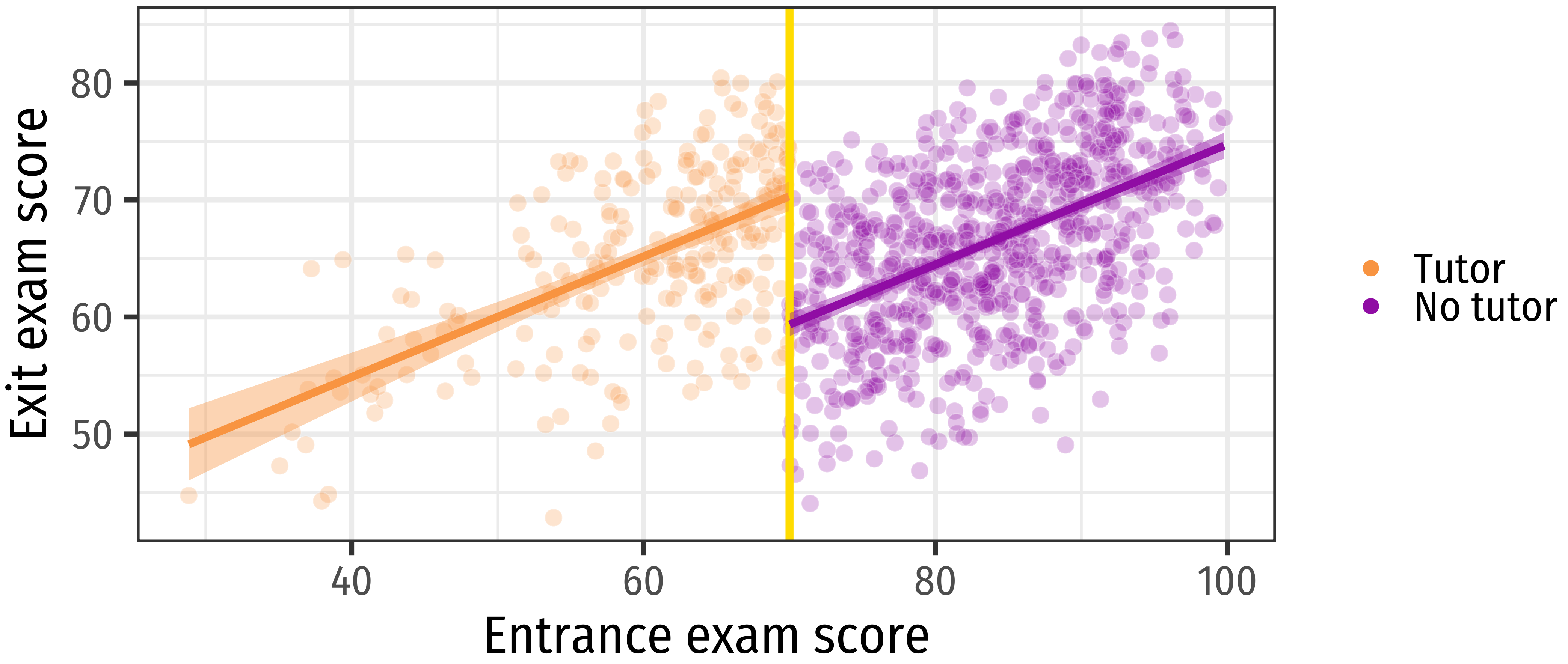
Pseudo treatment and control groups!

Compare outcomes right at the cutoff

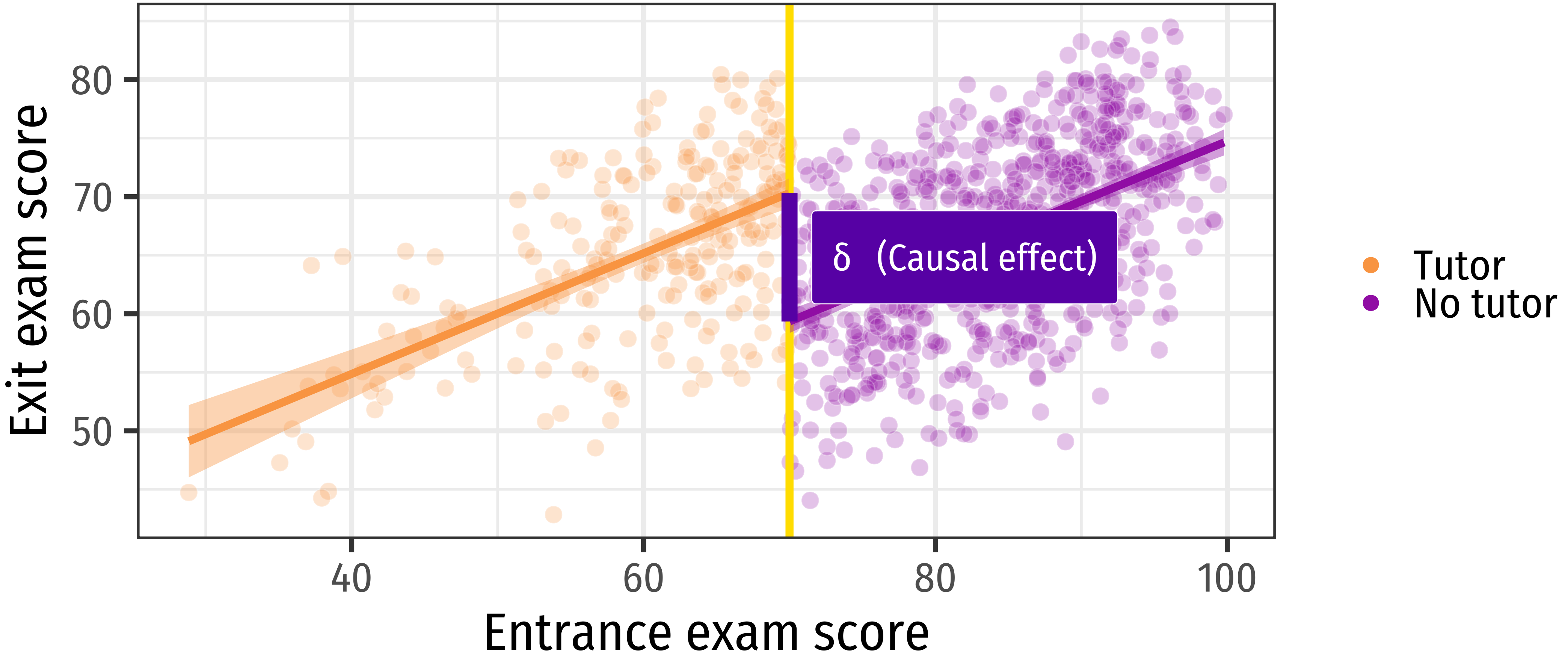
# Exit exam results according to running variable



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

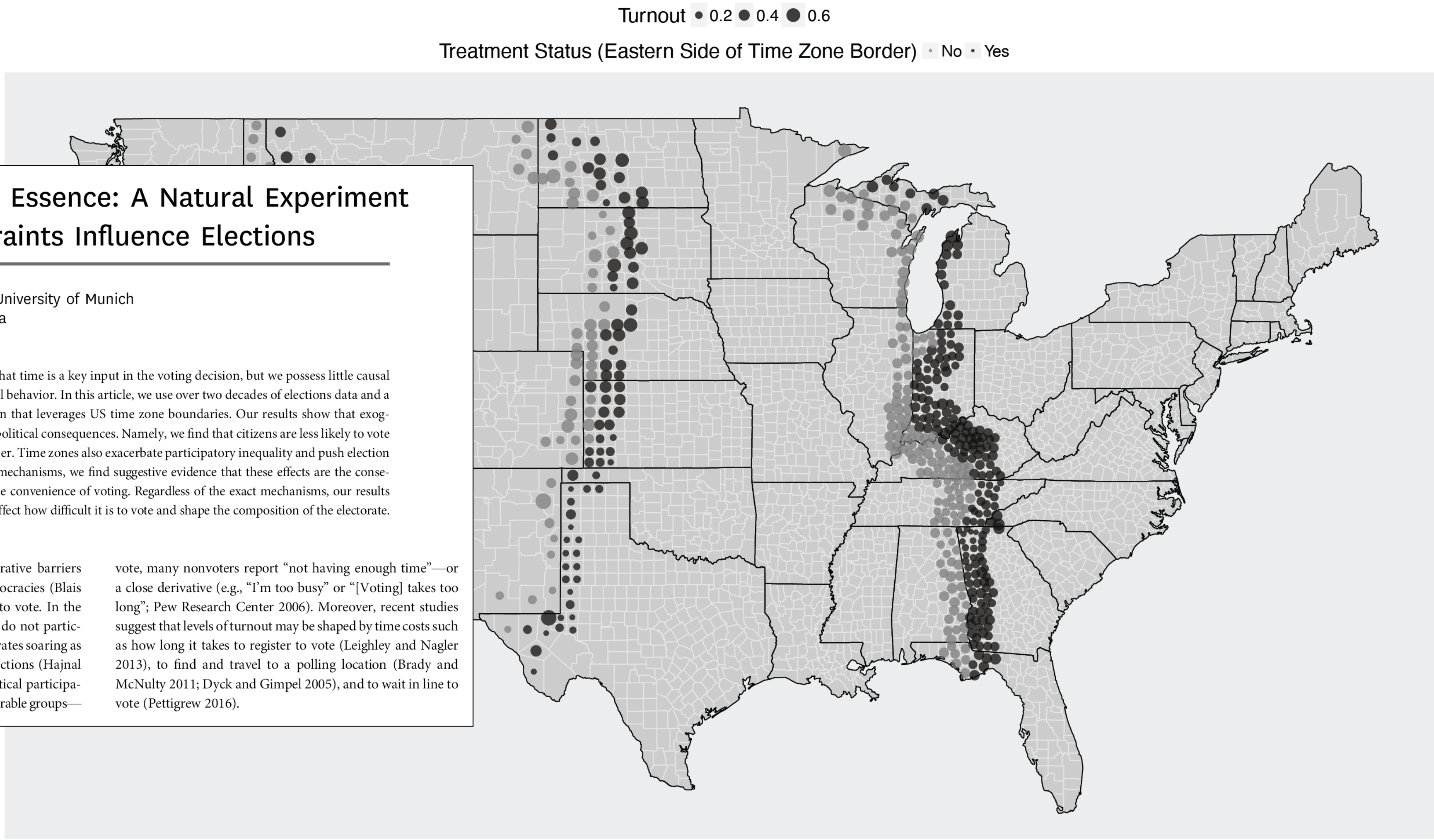


# Fit a regression at the right and left side of the cutoff



You can find discontinuities everywhere!

# Geographic discontinuities




## 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 Trounstine 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).

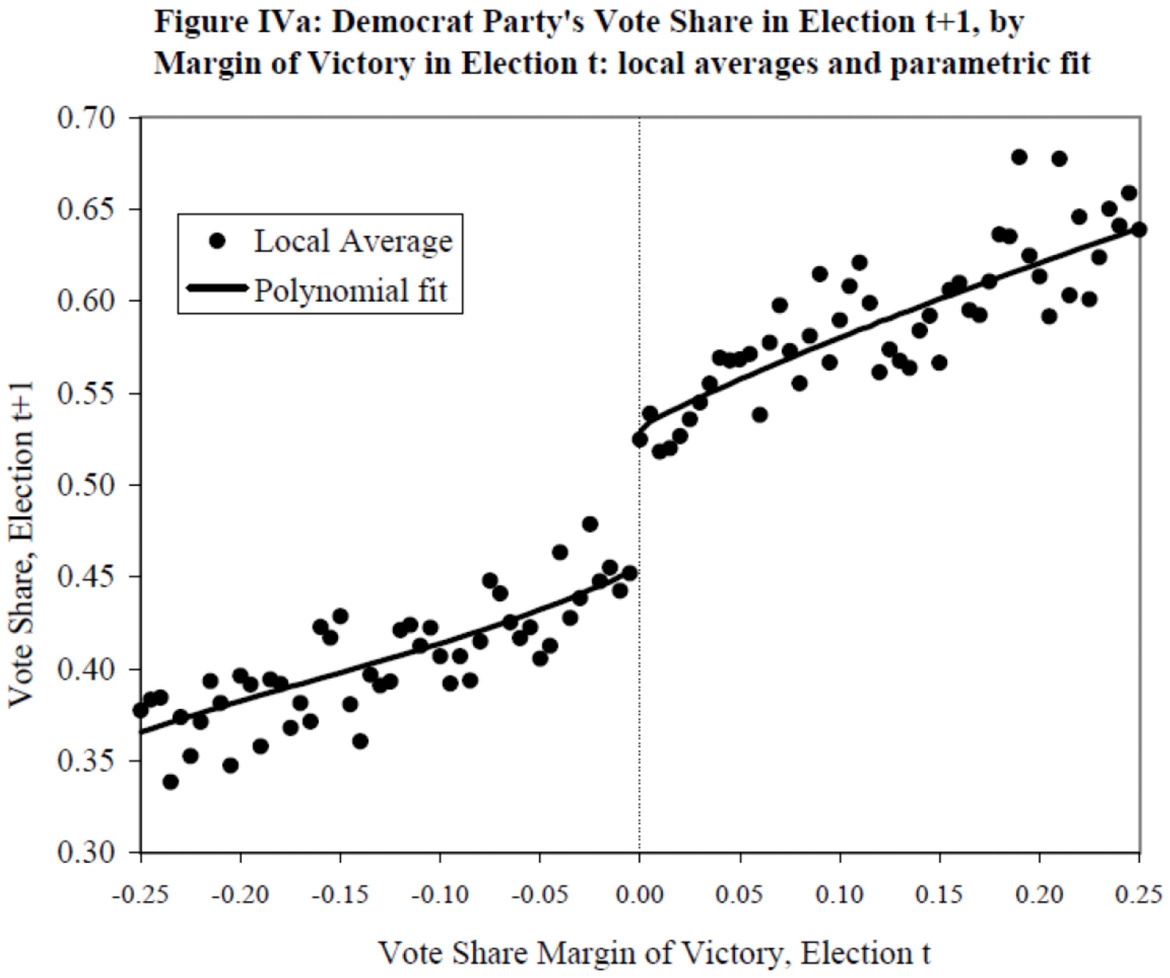
# Time discontinuities

## After Midnight: A Regression Discontinuity Design in Length of Postpartum Hospital Stays†

*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



# How do we do RDs in practice?

# Behind the scenes of RDs

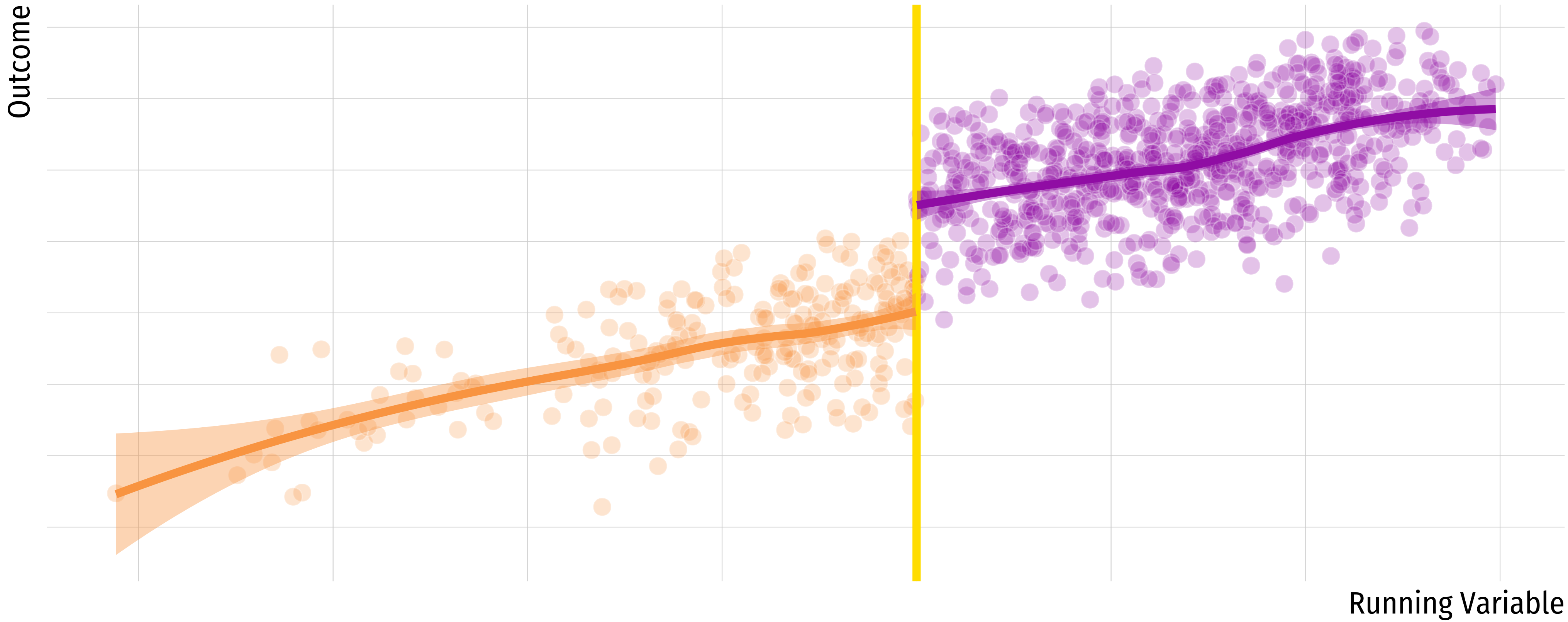
- Basically, regression discontinuities work under an **asymptotic assumption**:
- Let $Y_i$ be the outcome of interest, $Z_i$ the treatment assignment, $R_i$ the running variable, and $c$ the cutoff score:

$$Z_i = \begin{cases} 0 & R_i \leq c \\ 1 & R_i > c \end{cases}$$

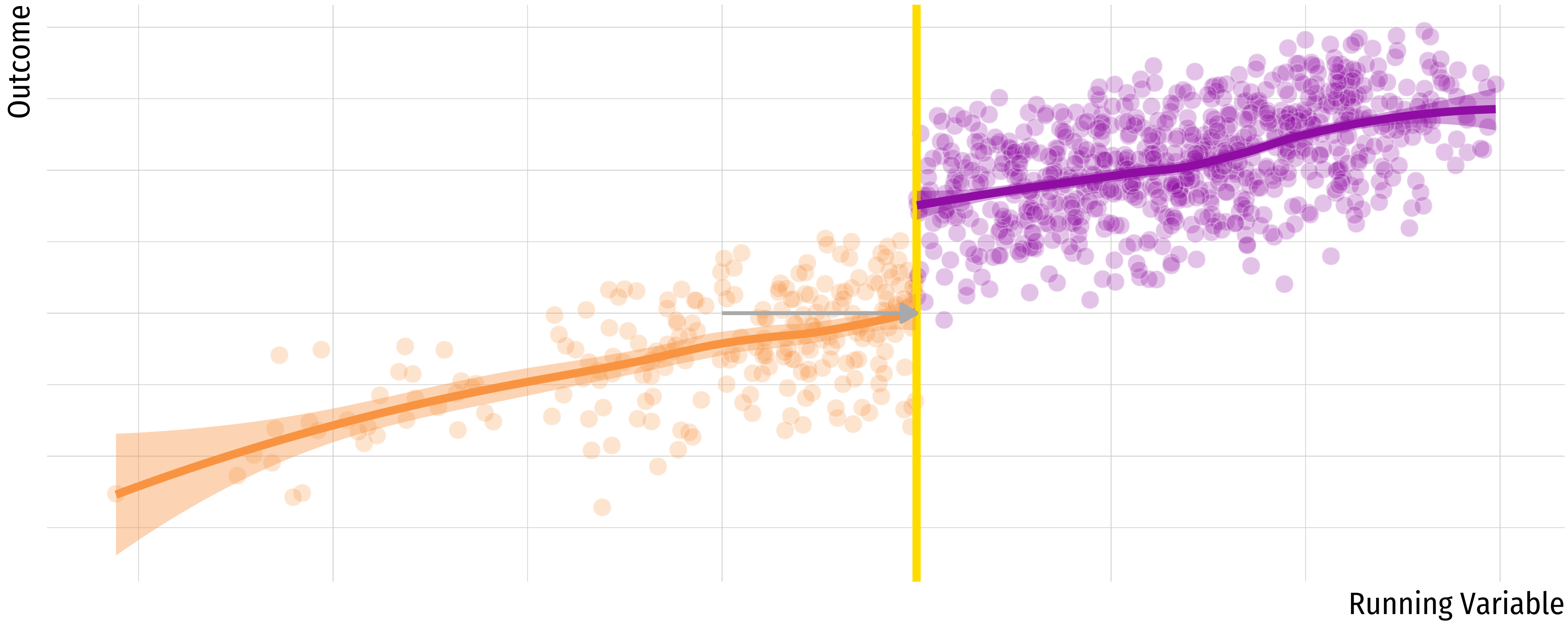
- Then, we can define the treatment effect $\delta$ as:

$$\delta = \lim_{\epsilon \to 0^+} E[Y_i | R_i = c + \epsilon] - \lim_{\epsilon \to 0^-} E[Y_i | R_i = c + \epsilon]$$

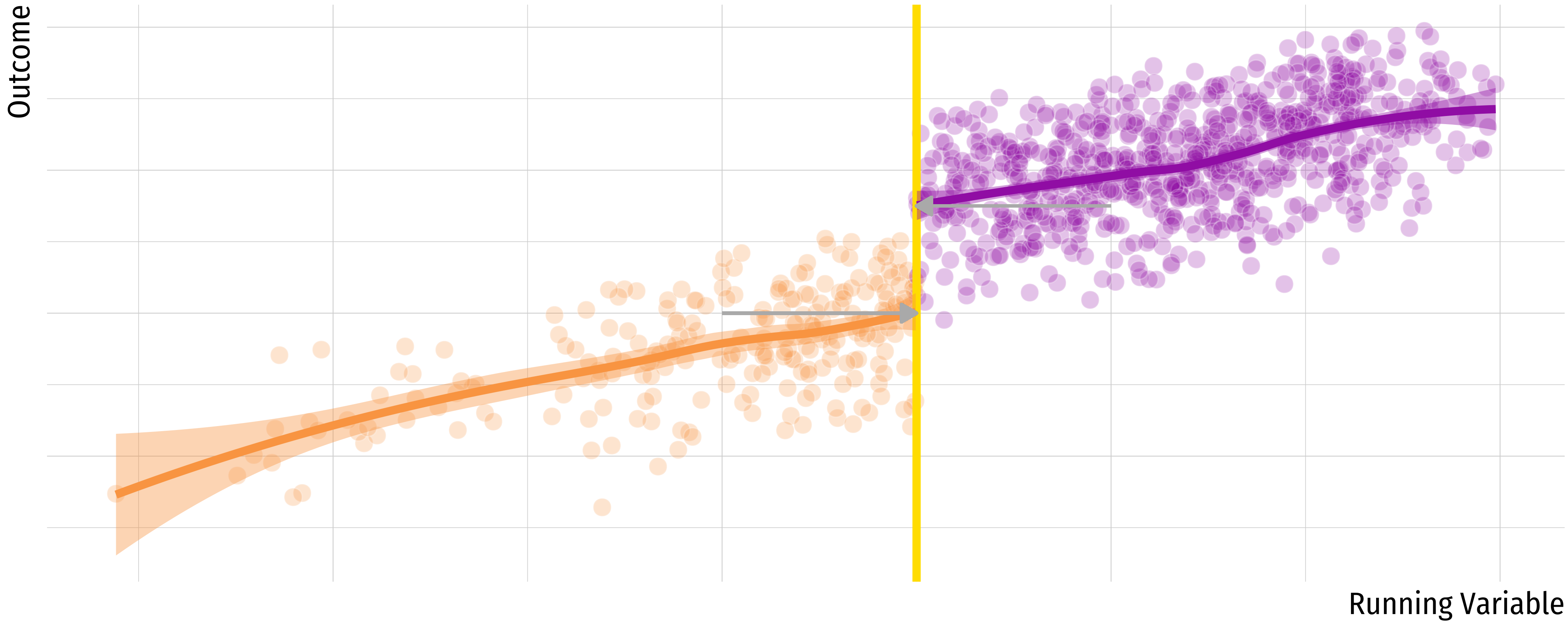
# What does the limit expression mean?



# What does the limit expression mean?



# What does the limit expression mean?



# What is the estimand we are estimating?

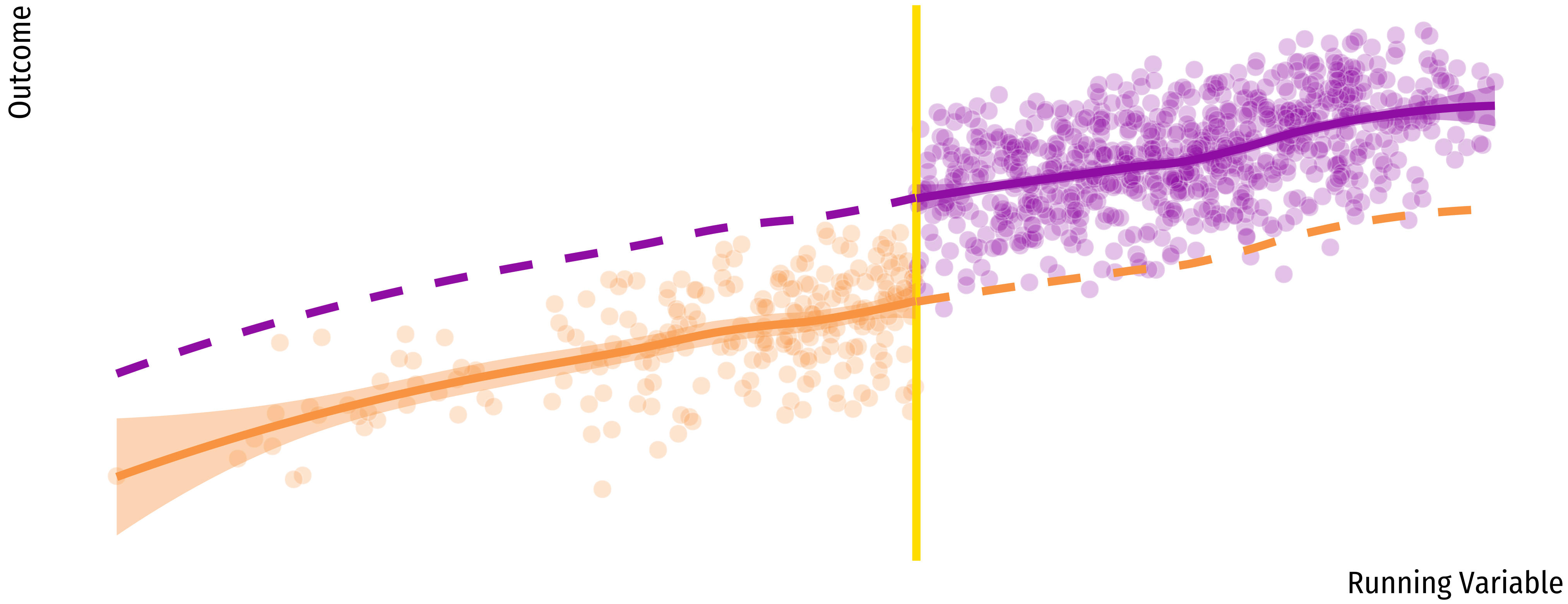
# Local Average Treatment Effect (LATE) for units at R=c

# Conditions required for identificaton

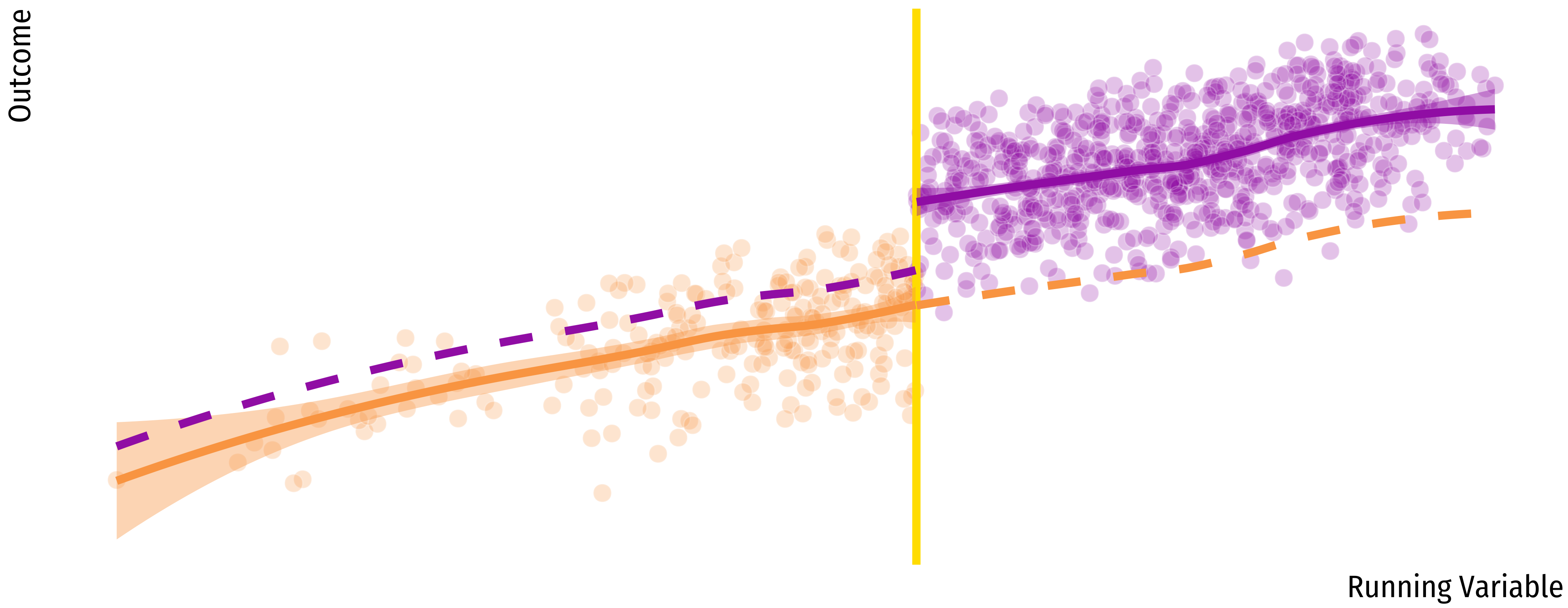
- Threshold rule **exists** and cutoff point is **known**
- The running variable $R_i$ is **continuous** near $c$.
- Key assumption:

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

# Potential outcomes need to be smooth across the threshold



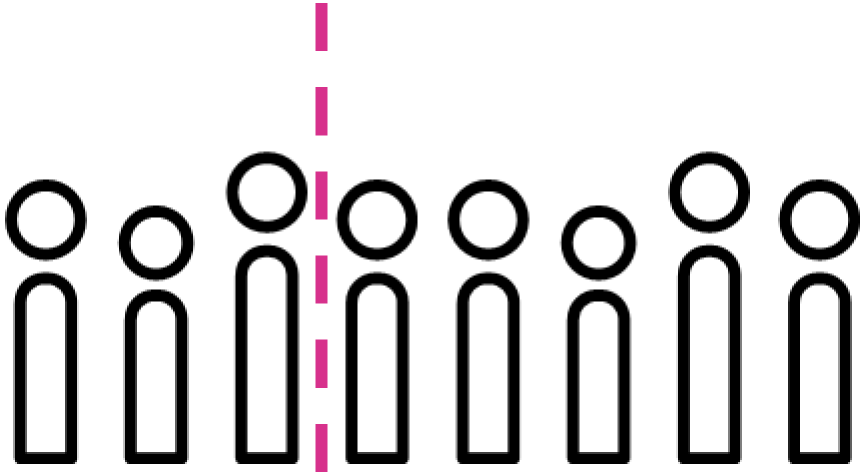
# Potential outcomes need to be smooth across the threshold



Can you think situations where that could happen?

# Let's go back to our discount example

- Customers are given discounts based on their **order of arrival**



- We could think of this as an **RD in time**, where $c$ is the time of arrival of customer 1,000.

# Work in groups

1) Each group will be given a task and some code

2) You need to complete the code and discuss the results

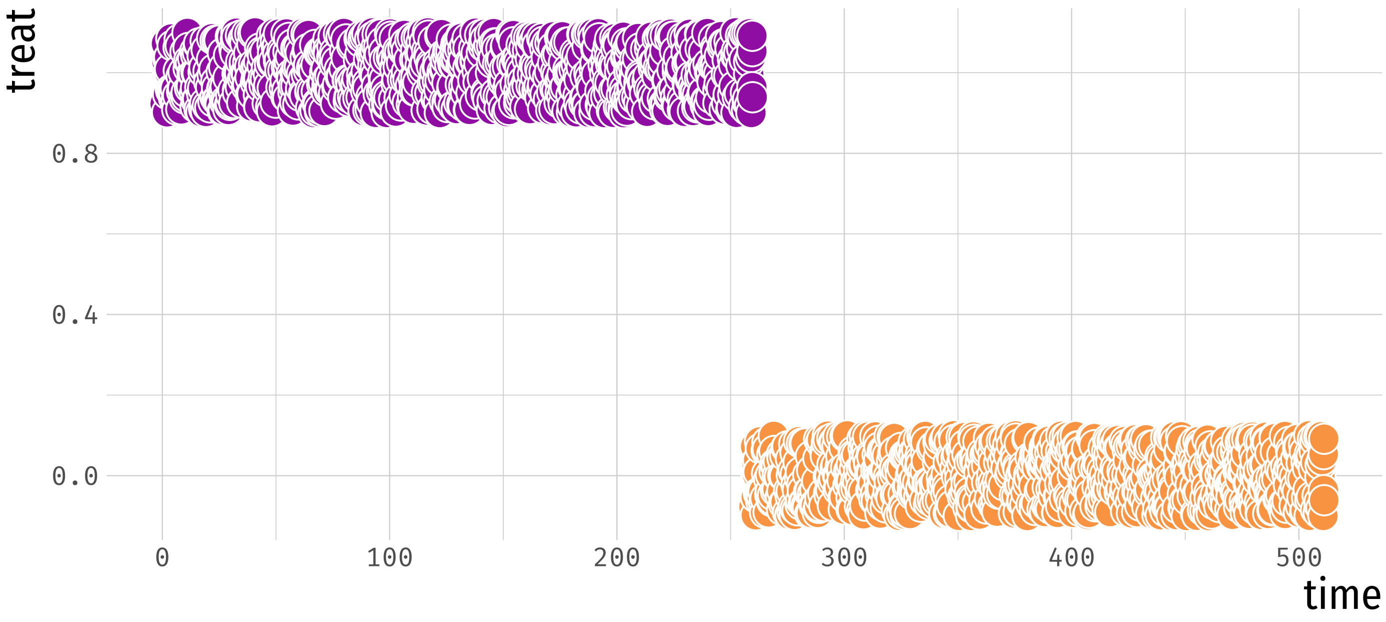
# Group 1

**What did you have to do?**

# Group 1: Check the treatment assignment

```
c = max(sales$time[sales$treat==1])

ggplot(data = sales, aes(x = time, y = treat)) +
  geom_point(data = filter(sales, time<=c), pch = 21, color = "white", fill="#900DA4", position = po
  geom_point(data = filter(sales, time>c), pch = 21, color = "white", fill="#F89441", position = pos
```

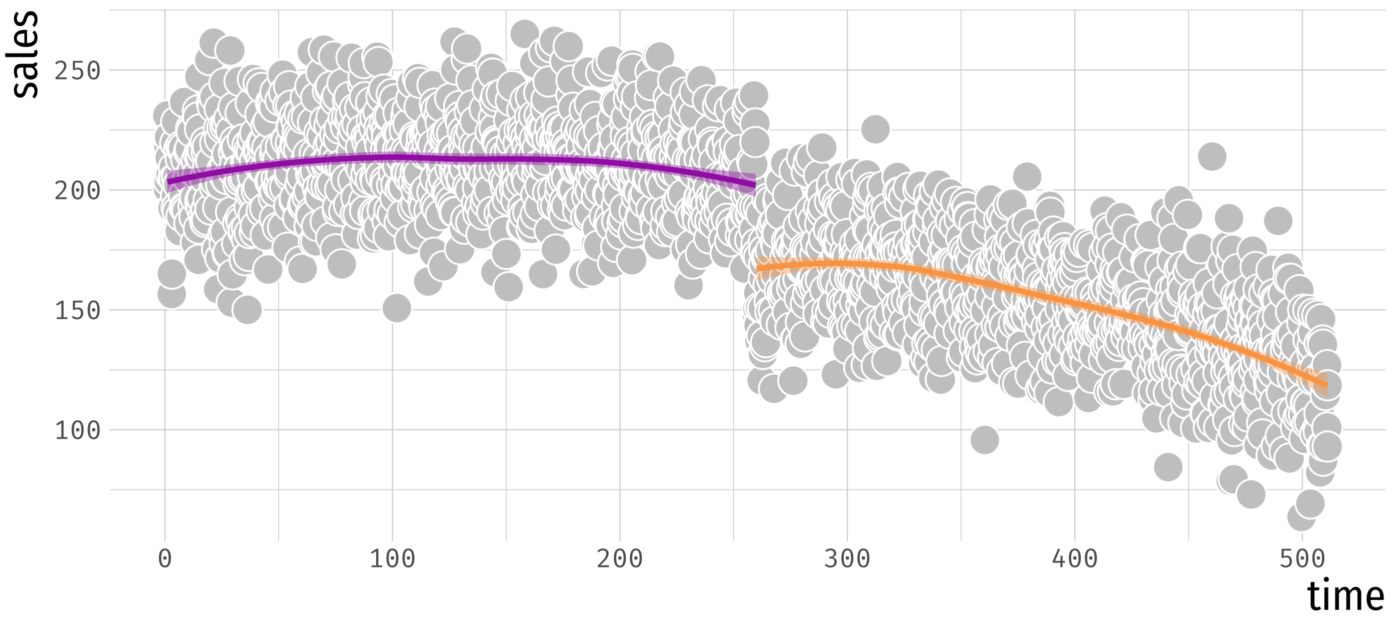


# Group 2

**What did you have to do?**

# Group 2: Check the regression discontinuity on the outcome

```
ggplot(data = sales, aes(x = time, y = sales)) +
  geom_point(pch = 21, color = "white", fill="grey") +
  geom_smooth(data = filter(sales, time>c), method = "loess", se=TRUE, color = "#F89441", fill = "#
  geom_smooth(data = filter(sales, time<=c), method = "loess", se=TRUE, color = "#F89441", fill = "#
```

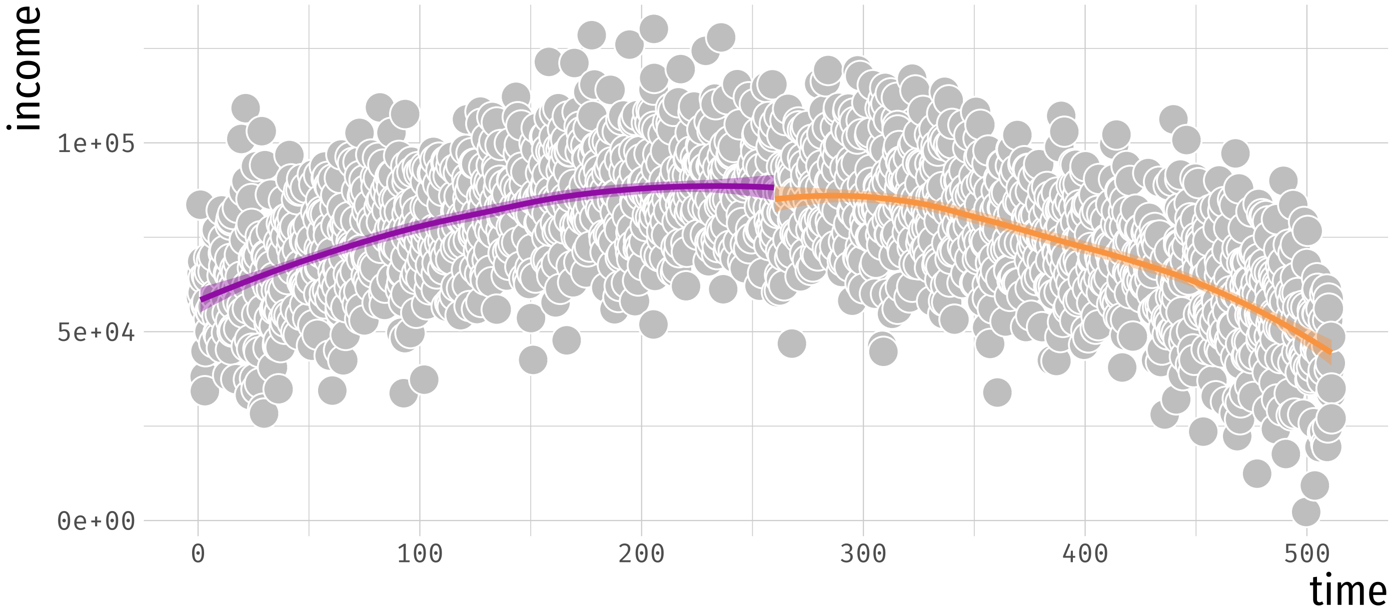


# Group 3

**What did you have to do?**

# Group 3: Check smoothness of income

```
ggplot(data = sales, aes(x = time, y = income)) +
  geom_point(pch = 21, color = "white", fill="grey") +
  geom_smooth(data = filter(sales, time>c), method = "loess", se=TRUE, color = "#F89441", fill = "#F
  geom_smooth(data = filter(sales, time<=c), method = "loess", se=TRUE, color = "#F89441", fill = "#
```

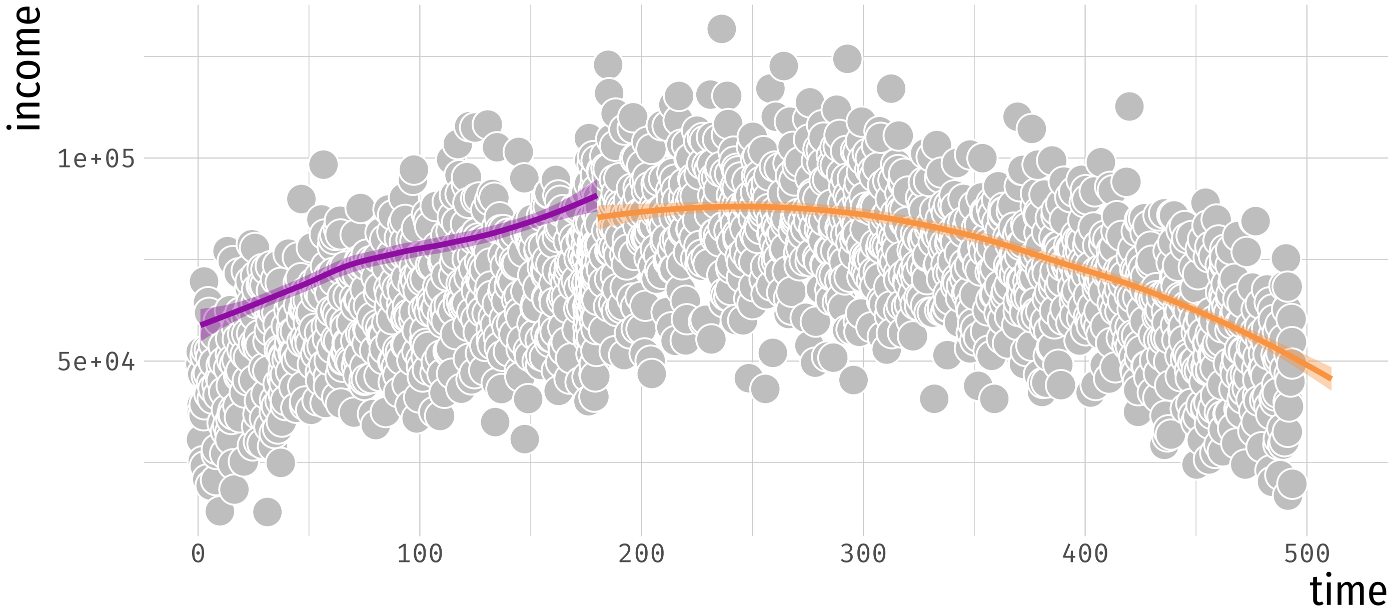


# Group 4

What did you have to do?

# Group 4: Check smoothness of income under different assignment

```
ggplot(data = sales_mod, aes(x = time, y = income)) +
  geom_point(pch = 21, color = "white", fill="grey") +
  geom_smooth(data = filter(sales, time>c), method = "loess", se=TRUE, color = "#F89441", fill = "#I
  geom_smooth(data = filter(sales, time<=c), method = "loess", se=TRUE, color = "#F89441", fill = "#
```



# How do we actually estimate an RD?

- The simplest way to do this is to fit a regression:

$$Y_i = \beta_0 + \beta_1(R_i - c) + \beta_2 \mathrm{I}[R_i > c] + \beta_3(R_i - c)\mathrm{I}[R_i > c]$$

# How do we actually estimate an RD?

- The simplest way to do this is to fit a regression:

$$Y_i = \beta_0 + \beta_1 \underbrace{(R_i - c)}_{\text{Distance to the cutoff}} + \beta_2 \mathbf{I}[R_i > c] + \beta_3 \overbrace{(R_i - c)}^{\text{Distance to the cutoff}} \mathbf{I}[R_i > c]$$

# How do we actually estimate an RD?

- The simplest way to do this is to fit a regression:

$$Y_i = \beta_0 + \beta_1(R_i - c) + \beta_2\underbrace{\mathrm{I}[R_i > c]}_{\text{Treatment}} + \beta_3(R_i - c)\overbrace{\mathrm{I}[R_i > c]}^{\text{Treatment}}$$

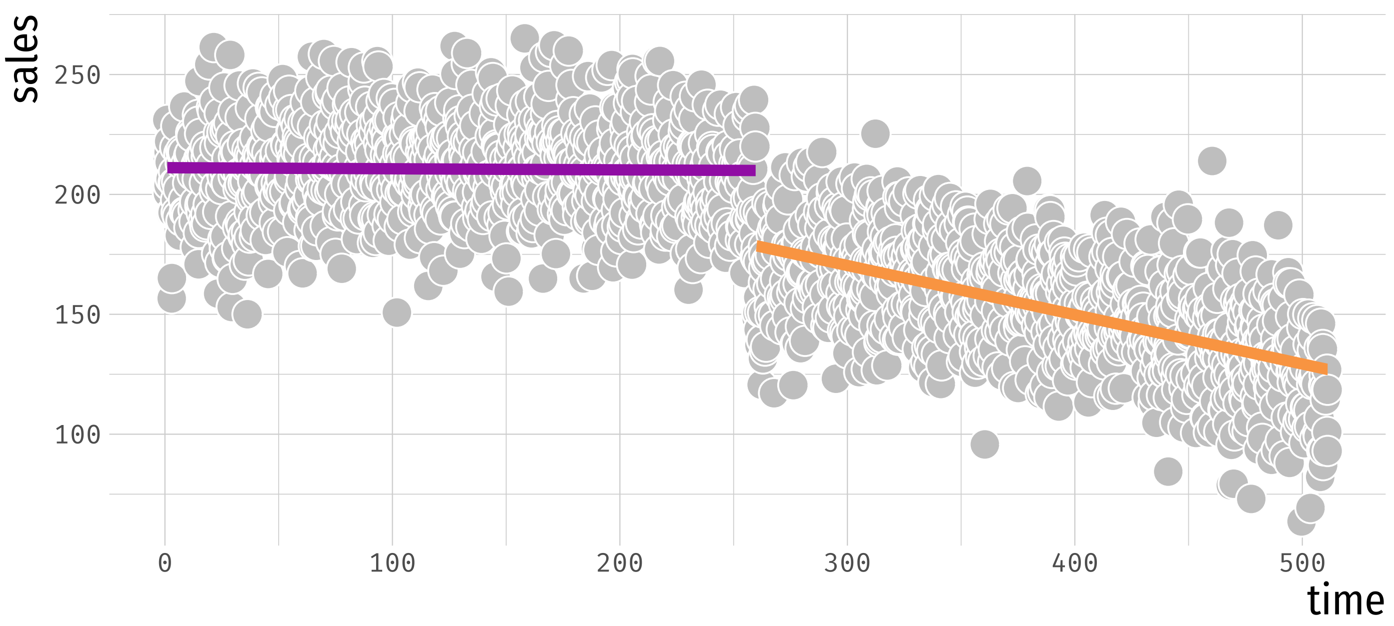
- You want to add **flexibility** for each side of the cutoff.

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

# Let's see some examples: Sales using a linear model

```
sales <- sales %>% mutate(dist = c-time)

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



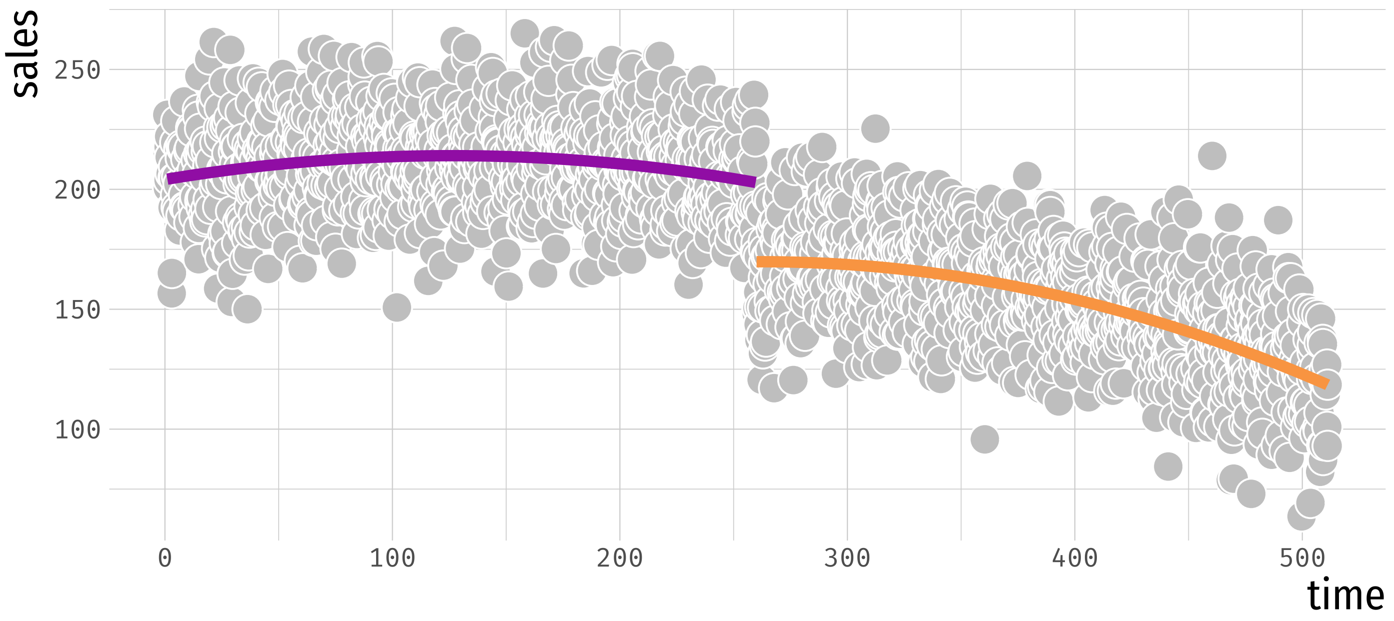
# Let's see some examples: Sales using a linear model

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

```
##
## Call:
## lm(formula = sales ~ dist + treat + 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
```

# What happens if we fit a quadratic model?

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

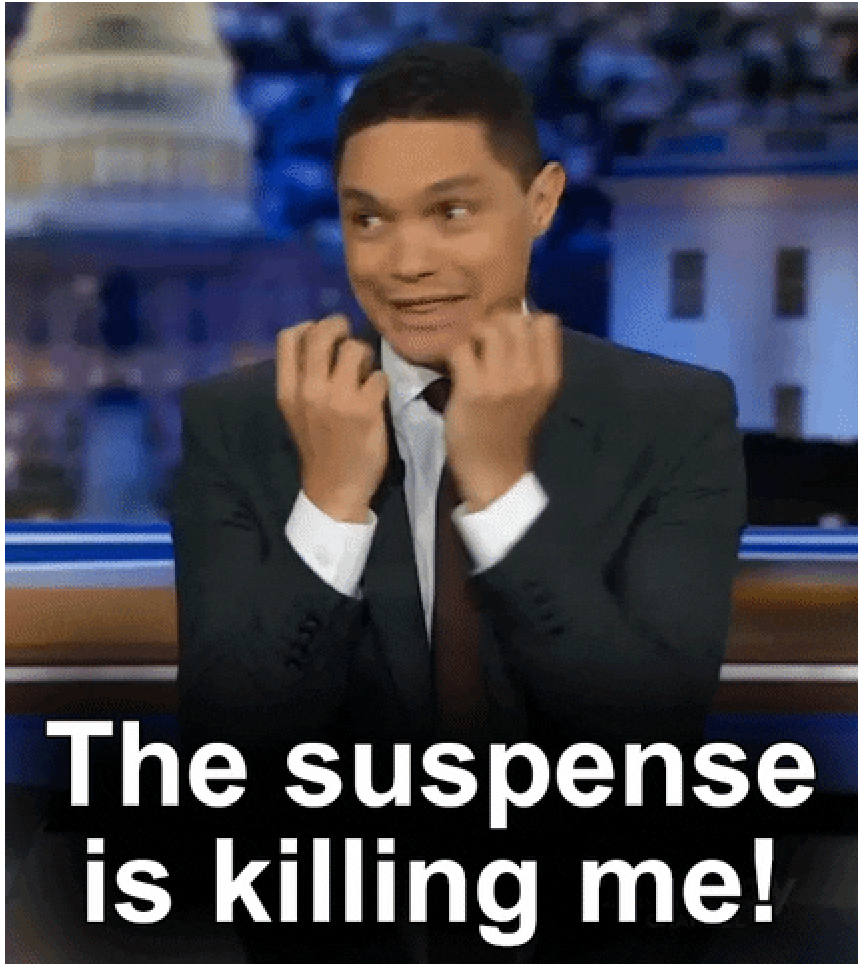


# What happens if we fit a quadratic model?

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

```
##
## Call:
## lm(formula = sales ~ dist + I(dist^2) + treat + 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
## I(dist^2)       -8.288e-04  1.363e-04  -6.083 1.41e-09 ***
## treat            3.308e+01  2.747e+00  12.041  < 2e-16 ***
## dist:treat       1.713e-01  4.964e-02   3.452 0.000569 ***
## I(dist^2):treat  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
```

# Next class




- Check how to rely less on **parametric assumptions**
- What is the **optimal bandwidth** to estimate our RD?
- Talk about **fuzzy regression discontinuities**

# Have a good Spring Break!

# References


- Angrist, J. and S. Pischke. (2015). "Mastering Metrics". *Chapter 4*.
- Heiss, A. (2020). "Program Evaluation for Public Policy". *Class 10: Regression Discontinuity I, Course at BYU*.
- Lee, D. and T. Lemieux. (2010). "Regression Discontinuity in Economics". *Journal of Economic Literature 48, pp 281-355*.