A Brief Introduction to Popular Identification Strategies for Observational Data

POLS 602
Fall 2025

Dr. Mike Burnham
Texas A&M Political Science

Review

- Causal inference basics
  - Understand its significance in the scientific process
  - Counterfactuals and the fundamental problem of causal inference
  - Randomization
  - Experimental research design

Review

- Descriptive statistics
  - Central tendency
  - Spread
  - Correlation and covariance
  - Basic plots in R

Review

- Predictive modeling with regression
  - The purpose of predictive modeling
  - An intuitive understanding for what linear regression is
  - What OLS is
  - How to interpret a regression table
  - Interpret goodness of fit metrics

Review

- Multiple regression
  - A conceptual understanding of endogeneity
  - Controlling for confounders with multiple regression
  - Model fitting with R
  - Controlling for categorical variables and interaction terms
  - Gauss Markov Theorem
  - Bias and potential sources

Identification

The ability to determine the true value of a model's parameters based on the distribution of the observable data

Causal Identification

The conditions or assumptions under which a causal effect can be determined from observed data

Identification Strategy

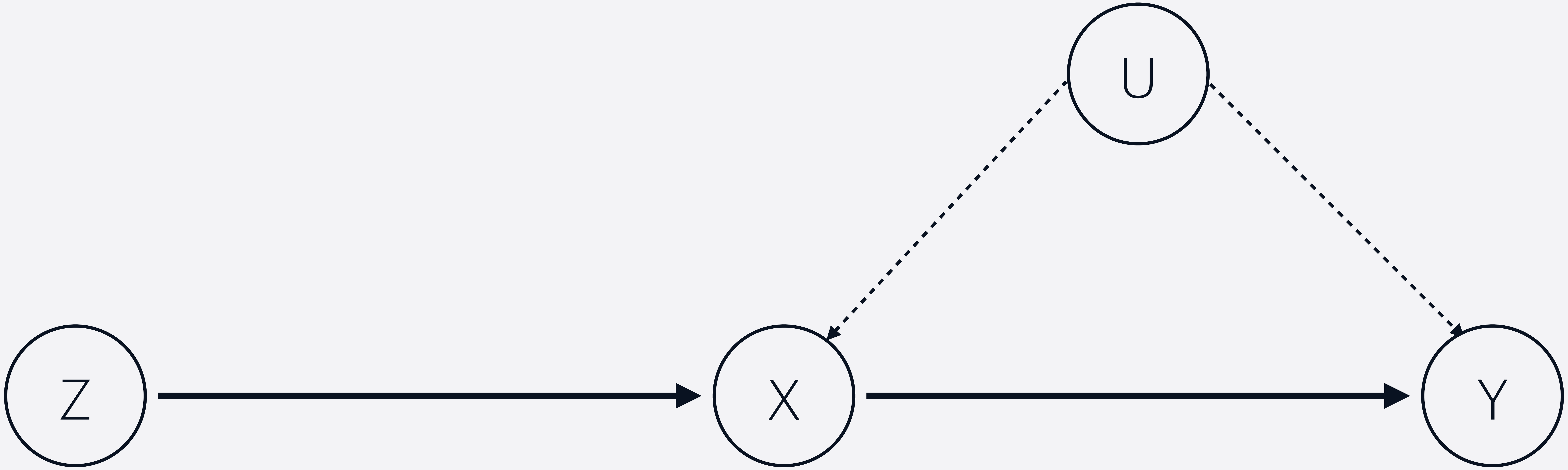
The research design, assumptions, and modeling strategy used to support your empirical claim

Today

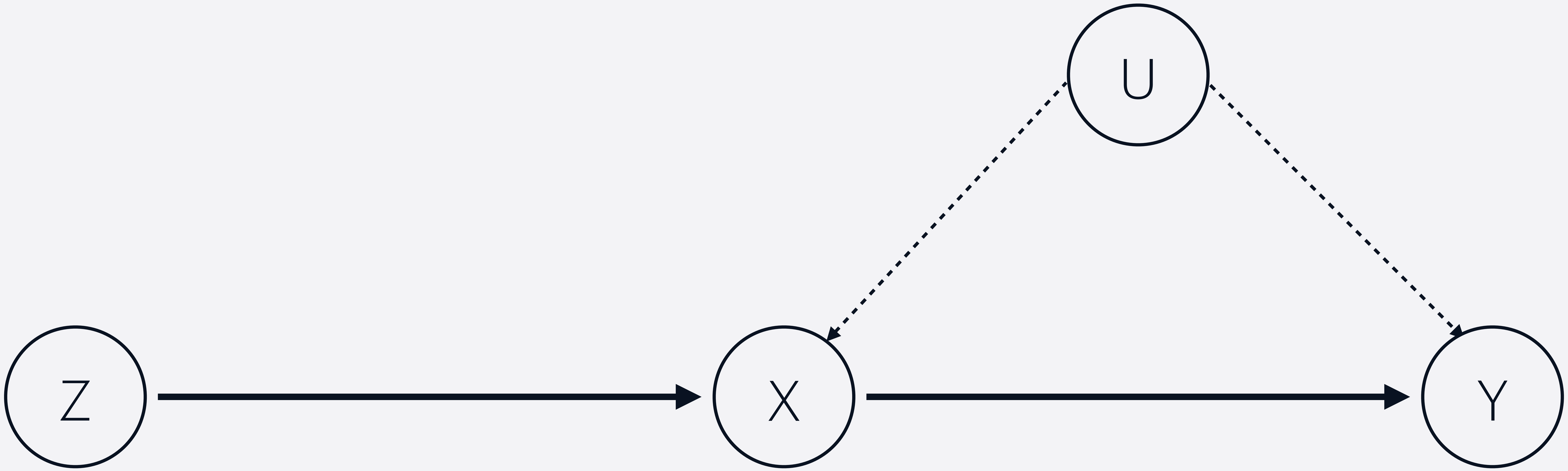
- A brief introduction to three of the most popular identification strategies for observational data
- Learn how these these research designs exploit **exogenous** variation to make credible causal estimates
- Give you a framework to think about exogenous variation, and point you in the right direction for future work

Instrumental Variables

IV DAG



IV DAG



Core idea: Use variation in X caused by Z, to estimate the causal effect on Y

IV

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 U_i + \epsilon_i$$ U is an unobserved confounder, thus:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$ Where $$e_i = \beta_2 U_i + \epsilon_i$$

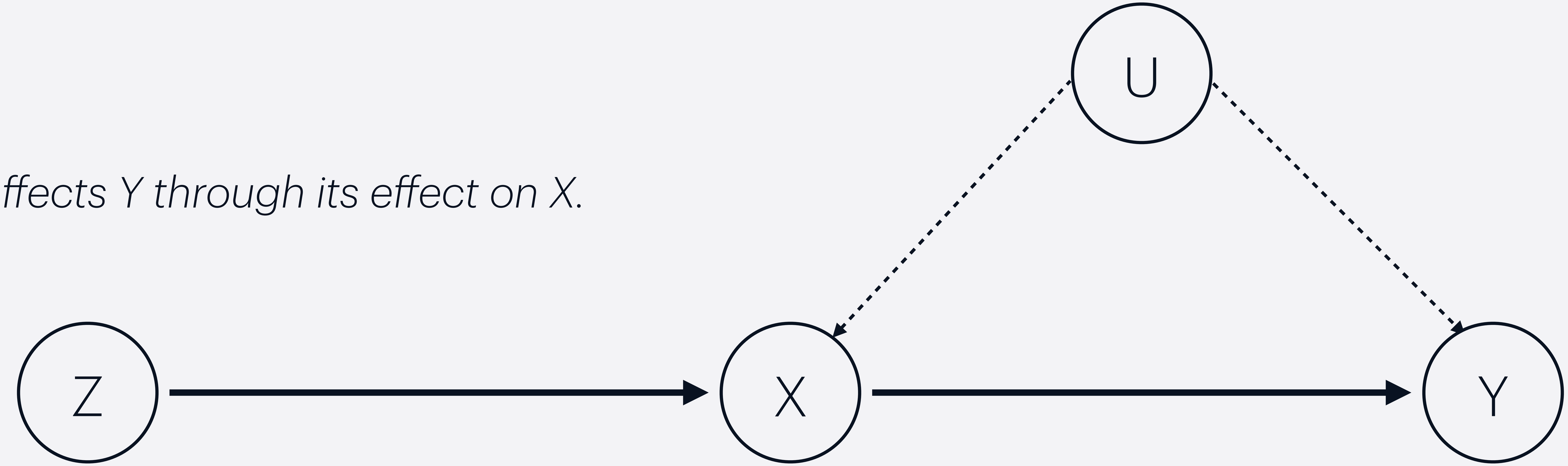
$$X_i = \alpha + \gamma Z_i + u_i$$

IV Assumptions

1. **Exclusion Restriction:** The instrument (Z) is independent of other variables that determine Y except for X.

2. **Non-zero First Stage:** Z is correlated with X, and therefore correlated with Y through its effect on X.

Z only affects Y through its effect on X.



The two-stage least squares estimator

```{r}
n <- 1000
# instrument Z
Z <- rnorm(n)

# error terms
e <- rnorm(n) # affects Y
u <- rnorm(n) # affects X
u <- 0.7*e + sqrt(1 - 0.7^2)*u  # u as a function of e to induce correlation

# generate endogenous regressor X and outcome Y
X <- 0.8*Z + u # X as a function of Z
Y <- 1 + 2*X + e # Y as a function of X, true causal effect on Y is 2
```

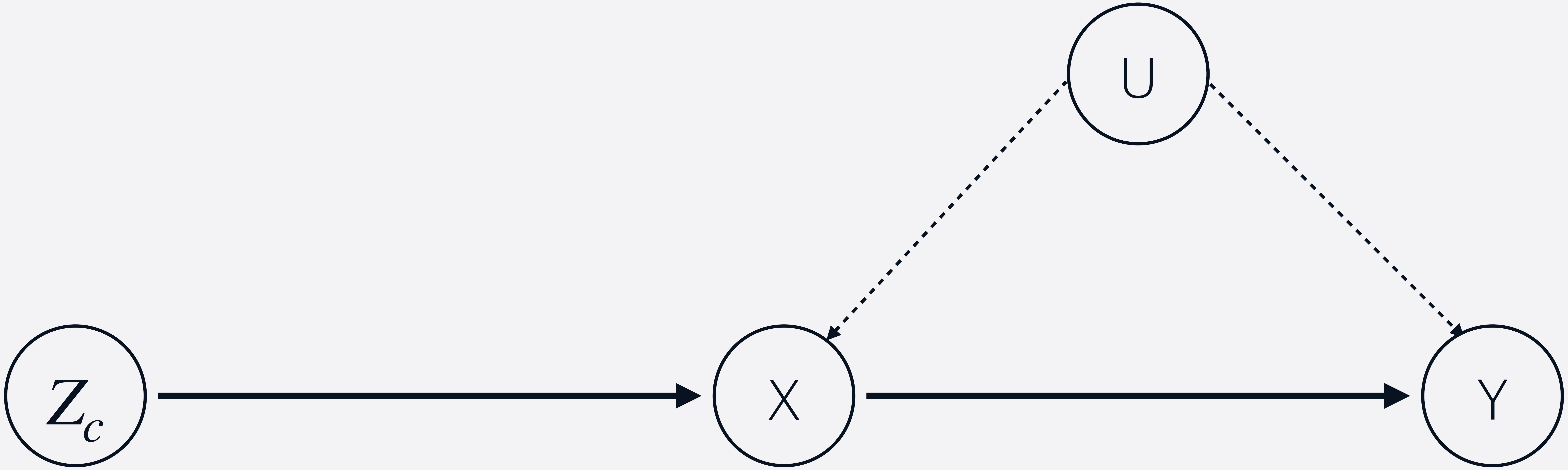
```{r}
# basic OLS
ols_model <- lm(Y ~ X)
summary(ols_model) # estimate is biased upwards
```

```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.20271 -0.56251 -0.00755  0.56948  2.69888 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.94952    0.02574   36.89   <2e-16 ***
X            2.45094    0.01943  126.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8139 on 998 degrees of freedom
Multiple R-squared:  0.941,	Adjusted R-squared:  0.9409 
F-statistic: 1.591e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

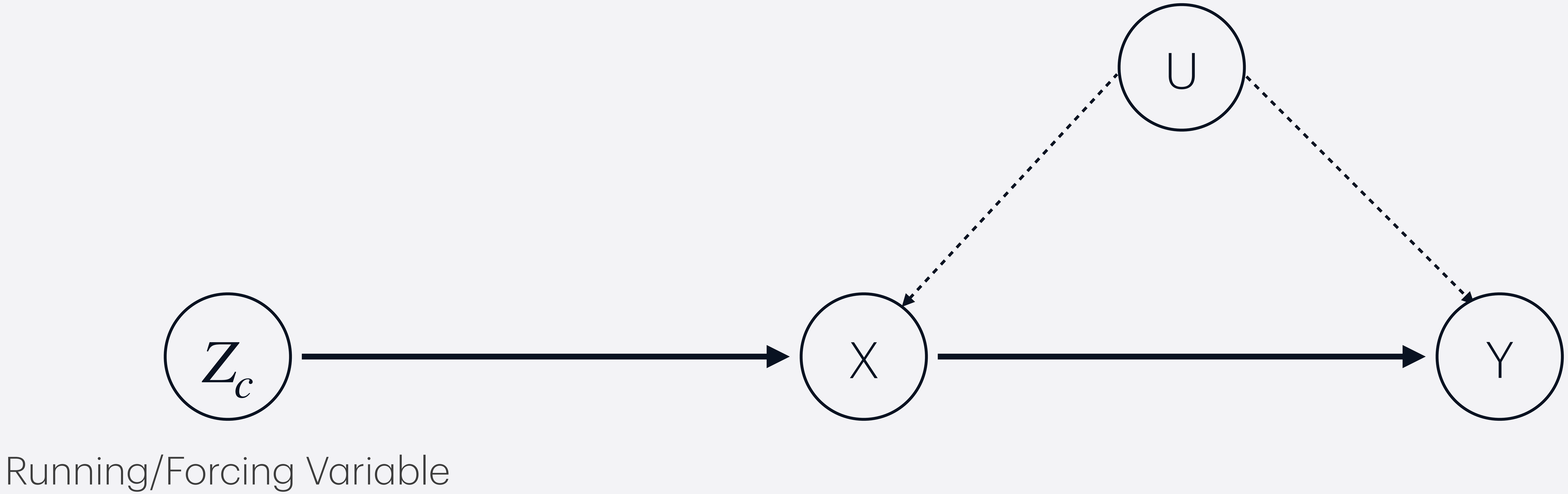
```r
# IV regression (2SLS)
# first stage: regress X on Z
first_stage <- lm(X ~ Z)
X_hat <- fitted(first_stage)

# second stage: regress Y on predicted X
iv_model <- lm(Y ~ X_hat)
summary(iv_model)
```

```

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.20271 -0.56251 -0.00755  0.56948  2.69888 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.94952    0.02574   36.89   <2e-16 ***
X            2.45094    0.01943  126.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8139 on 998 degrees of freedom
Multiple R-squared:  0.941,	Adjusted R-squared:  0.9409 
F-statistic: 1.591e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

Regression Discontinuity



Z_c
X
U
Y

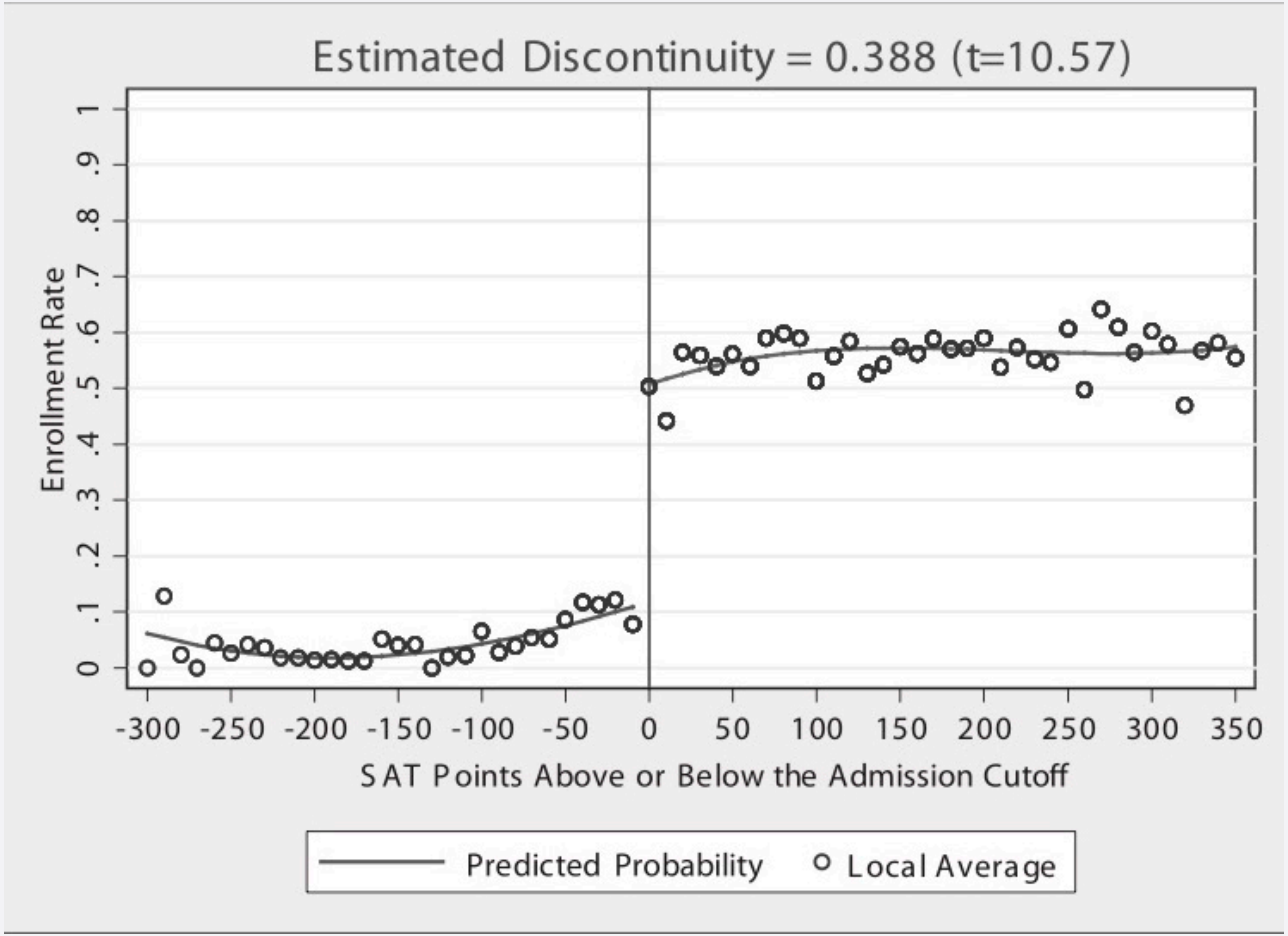

U
Z_c
X
Y
Running/Forcing Variable

RDD key assumptions

- **Continuity**: Potential outcomes are continuous functions of the forcing variable at the threshold. The outcome variable would be a smooth continuous function if not for the treatment.

- **No manipulation**: Observations cannot precisely manipulate the forcing variable around the cutoff.

- **Local randomization**: Units above and below the cutoff are similar in all respects, and thus the treatment is “as-if” random. Note: This does not mean treatment around the forcing variable is randomly assigned.


Estimated Discontinuity = 0.388 (t=10.57)
Enrollment Rate
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
-300
-250
-200
-150
-100
-50
0
50
100
150
200
250
300
350
SAT Points Above or Below the Admission Cutoff
Predicted Probability
Local Average

RDD

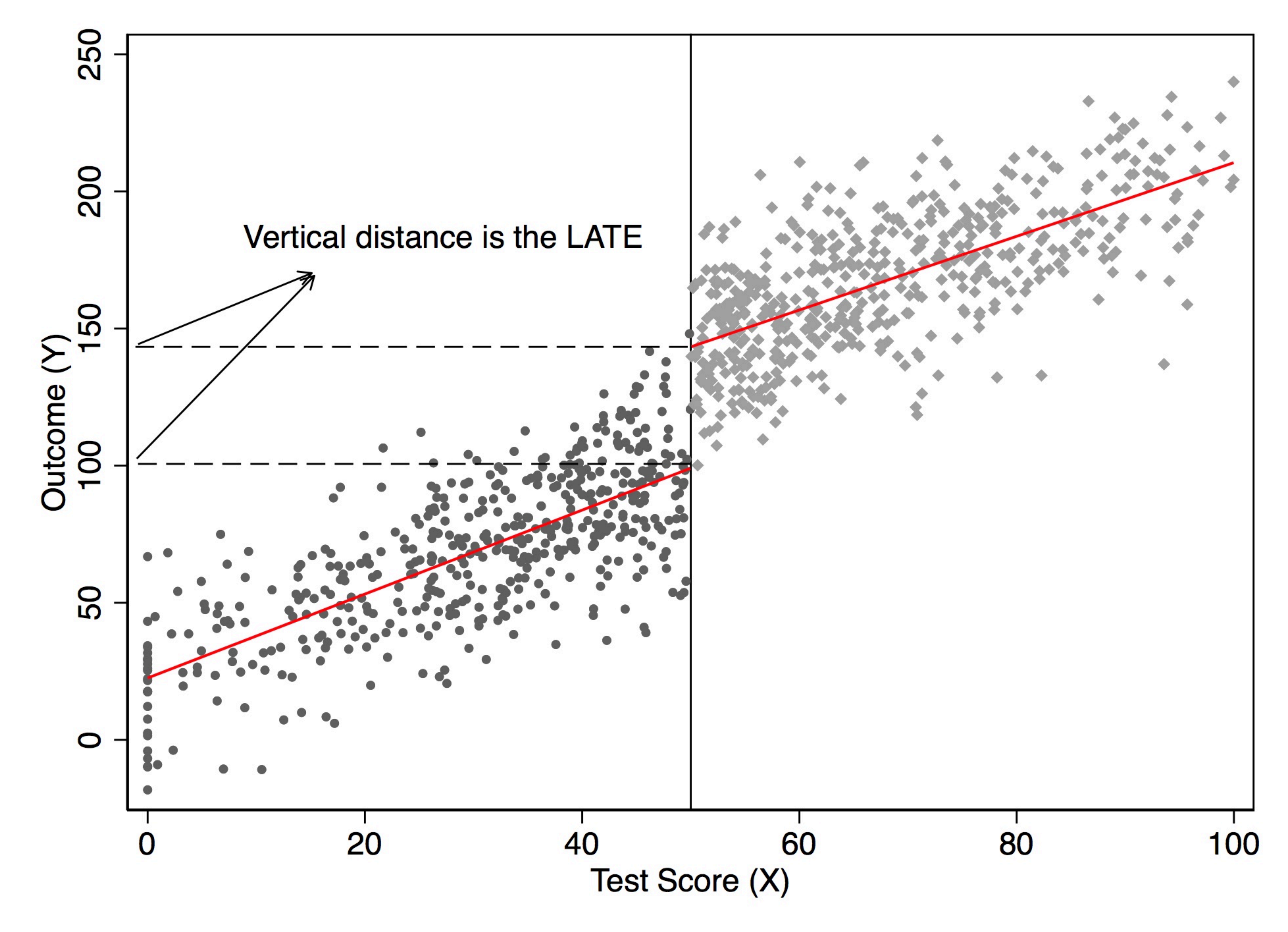
$$X_i = \begin{cases} 1 \text{ if } Z_i \geq C \\ 0 \text{ if } Z_i \leq C \end{cases}$$

$Y_i^0 = \beta_o + \beta_1 Z_i$

$Y_i^1 = Y_i^0 + \gamma$ where $\gamma =$ the treatment effect parameter

$Y_i = \beta_o + \beta_1 Z_i + \gamma X_i + \epsilon_i$

Local Average Treatment Effect (LATE)



Estimate the LATE in R

```r
library(tidyverse)
library(haven)
library(estimatr)

read_data <- function(df)
{
  full_path <- paste("https://github.com/scunning1975/mixtape/raw/master/",
                     df, sep = "")
  df <- read_dta(full_path)
  return(df)
}

lmb_data <- read_data("lmb-data.dta")

lmb_subset <- lmb_data %>%
  filter(lagdemvoteshare>.48 & lagdemvoteshare<.52)

lm_1 <- lm_robust(score ~ lagdemocrat, data = lmb_subset, clusters = id)
lm_2 <- lm_robust(score ~ democrat, data = lmb_subset, clusters = id)
lm_3 <- lm_robust(democrat ~ lagdemocrat, data = lmb_subset, clusters = id)

summary(lm_1)
summary(lm_2)
summary(lm_3)
```

Difference in Differences

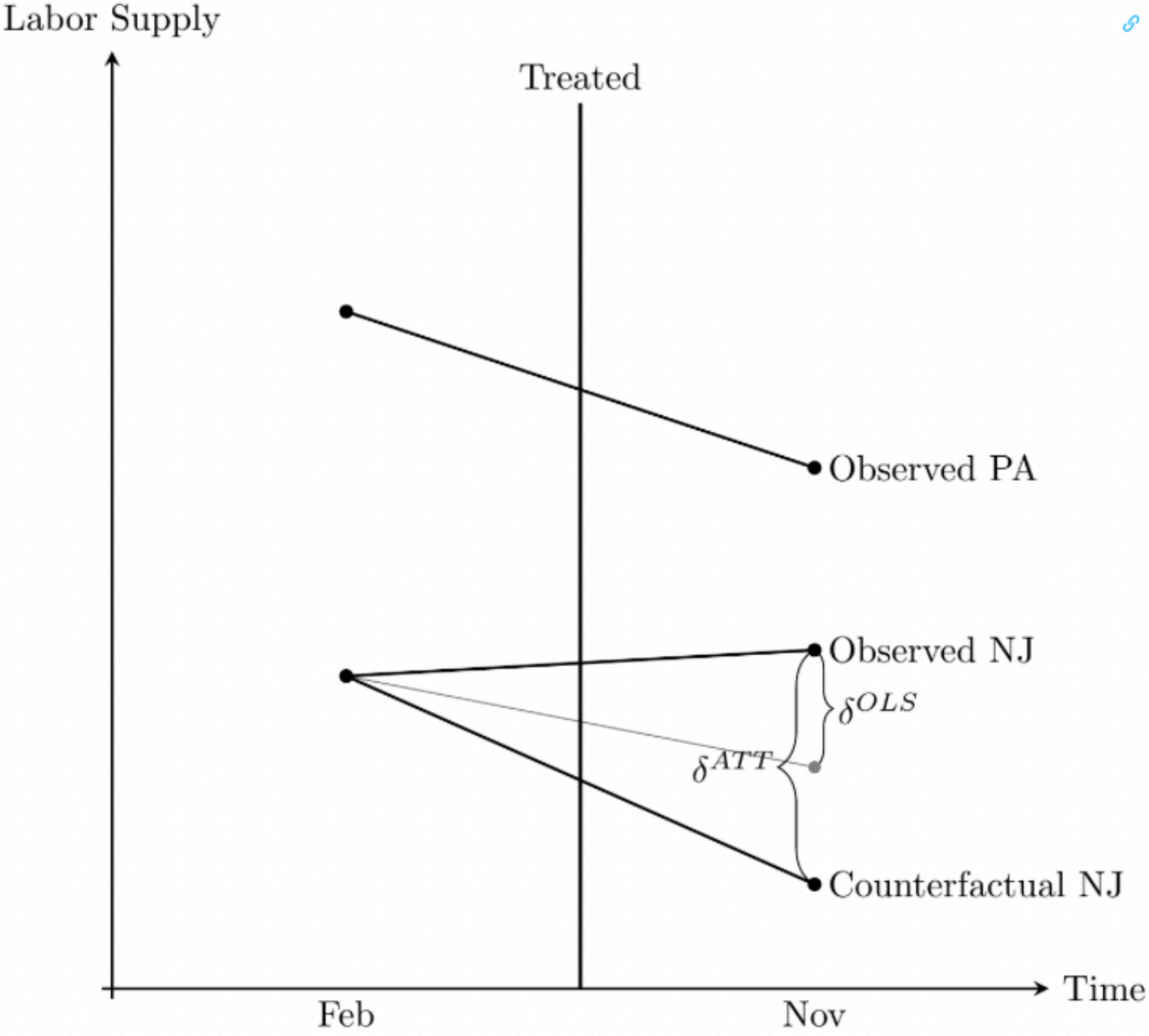
John Snow and Cholera

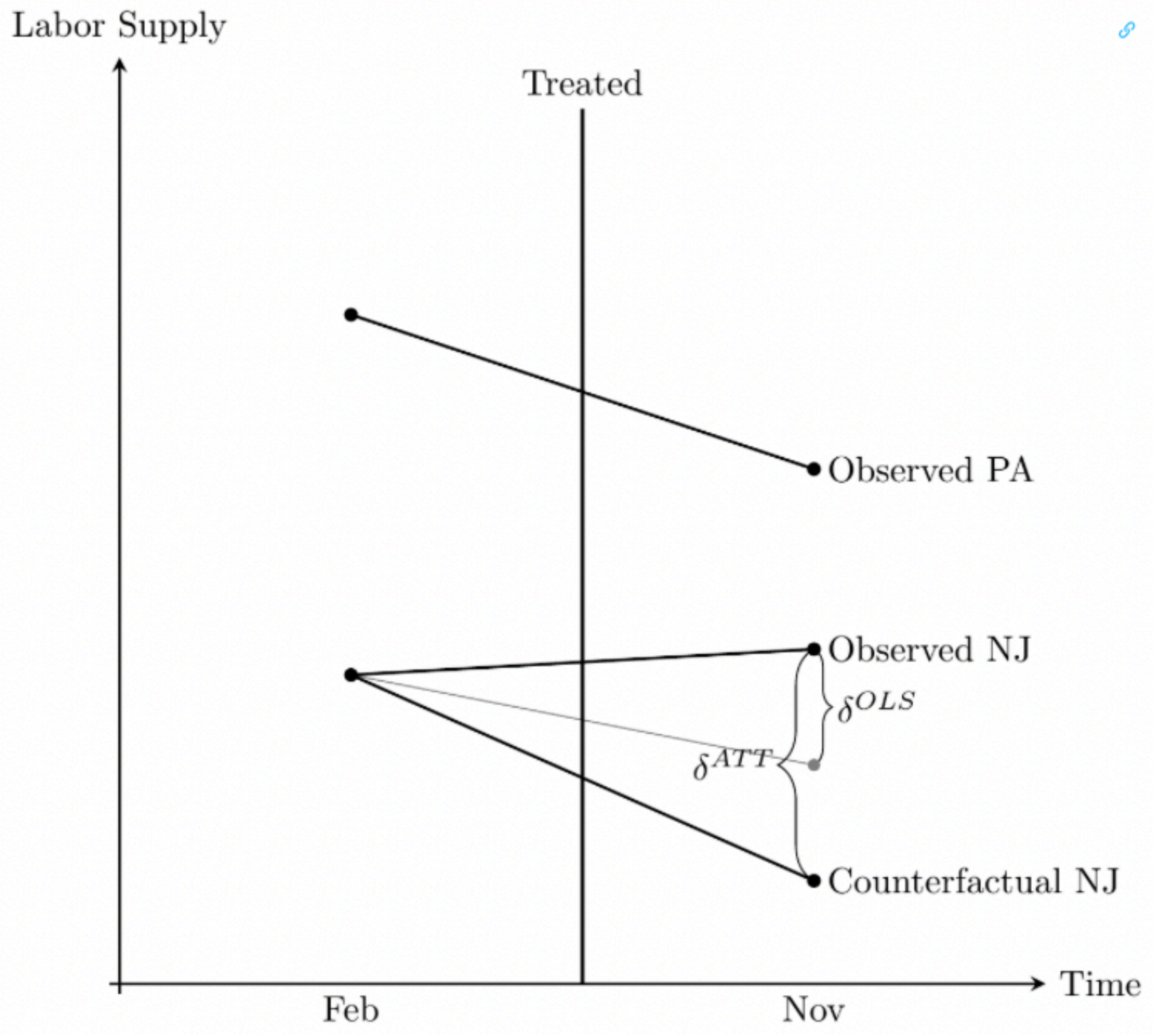
DiD assumptions

- **(Conditonal) Parallel trends:** In the absence of treatment, the outcome trends in the treatment and control groups would have been parallel over time.
- **Exogenous treatment assignment:** Treatment assignment is exogenous
- **No anticipation effects:** Treatment and control groups should not anticipate the treatment and change their behavior.
- **No spillover effects:** Effects from the treatment group do not spillover into the control group.

The 2x2 DiD

$$\hat{\gamma} = (E[Y_k | \text{Post}] - E[Y_k | \text{Pre}]) - (E[Y_u | \text{Post}] - E[Y_u | \text{Pre}])$$


Labor Supply
Treated
Observed PA
Observed NJ
δ^{OLS}
δ^{ATT}
Counterfactual NJ
Time
Feb
Nov



$$Y_{it} = \beta_0 + \beta_1 NJ_i + \beta_3 X_i + \beta_4 (NJ \times X_i) + \epsilon_i$$

```r
#-- DD estimate of 15-19 year olds in repeal states vs Roe states
library(tidyverse)
library(haven)
library(estimatr)

read_data <- function(df)
{
  full_path <- paste("https://github.com/scunning1975/mixtape/raw/master/",
                     df, sep = "")
  df <- read_dta(full_path)
  return(df)
}

abortion <- read_data("abortion.dta") %>%
  mutate(
    repeal = as_factor(repeal),
    year   = as_factor(year),
    fip    = as_factor(fip),
    fa     = as_factor(fa),
  )

reg <- abortion %>%
  filter(bf15 == 1) %>%
  lm_robust(lnr ~ repeal*year + fip + acc + ir + pi + alcohol+ crack + poverty
            data = ., weights = totpop, clusters = fip)
```