# Econometrics III Final

Owen Jetton

06/10/2022

---

## Question 01 For each of the following identification strategies, define the main assumption(s) that are required for unbiased (or consistent) estimates of a treatment effect:

### [01a] A selection-on-observables design

- Conditional on covariates $X_i$, the treatment variable ($D$) is independent of the outcome variable for treated and untreated observations ($Y_{0i}, Y_{1i}$)

### [01b] Instrumental variables

- The outcome variable for treated and untreated observations ($Y_{0i}, Y_{1i}$) is independent of the instrument ($Z$).
- The covariance between the instrument ($Z$) and the treatment variable ($D$) is not 0
- The covariance between the instrument ($Z$) and the disturbance term ($u_i$) from a regression of the treatment on the outcome variable, is 0

### [01c] Sharp regression discontinuity

- The probability of being assigned to treatment jumps from 0 to 1 as the running variable crosses the cutoff.
- The treatment effect doesn't depend on the running variable.
- The expected value of the outcome variable (whether treated or untreated) is continuous in the running variable, and all covariates.

## Question 02 Many identification strategies try create a situation where the potential outcomes $Y_0$ and $Y_1$ are uncorrelated with treatment (possibly conditional on some controls). Do regression discontinuities rely upon the potential outcomes being uncorrelated with treatment? Explain your answer.

Regression discontinuities don't rely upon assuming potential outcomes are uncorrelated with treatment, instead they rely on 2 similar assumptions: that the difference in the potential outcomes are independent of the running variable, and that the probability of treatment changes (in a fuzzy or sharp way) with the running variable. These two assumptions tell us that *the difference in the potential outcomes is independent of the running variable* in regression discontinuity.

### Question 03 You are interesting in estimating the effect of some treatment D on an outcome Y, but you are quite confident that D is "selected" (i.e., correlates with potential outcomes). You know of a variable W that is very predictive of treatment status and uncorrelated with the potential outcomes. For some individuals, W greatly increases their likelihood of treatment; for other individuals, W reduces their likelihood of treatment. Explain whether is or is not be a good instrument—or if it depends.

To ensure the coefficient on the IV yields a valid Local Average Treatment Effect interpretation among compliers, we need our instrument to be such that there is uniformity among the instrument's effect on likelihood of being treated. Since this potential instrument, $W$, increases the likelihood of treatment for some, and decreases the likelihood of treamtnet for others, $W$ violates this monotonicity assumption. Since it violates this assumption, it is unlikely that any result would be usefully interpretable, therefore I would not advise using $W$ as an instrument.

### Question 04 When people talk about "clustering their errors", are they talking about correlated disturbances or correlated treatment assignment? Explain.

Clustering standard errors refers to correlated disturbances *and* correlated treatment assignment. Clustering is used if treatment is assigned to individuals in a group, or if it is suspected that disturbances within groups are correlated, or possibly if both are present.

### Question 05 Linear regression is clearly restrictive—it's hard to expect the real world to follow

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

**Explain why we (economists and related social scientists) still rely so much on linear, least squares regression.**

Linear regression is incredibly useful for identifying causal relations and for yielding understandable interpretations of variables' effects on each other. We also care about linear regression because it is a helpful way to identify and estimate the **Conditional Expectation Function**, which tells us the expected value of an outcome variable holding covariates constant.

---

**Which type of treatment effect you would be estimating (Average Treatment Effect, Treatment On Treated, Intent To Treat, etc.) and explain your answer in one short sentence.**

### Question 06 You wish to estimate the effect of a college degree on earnings. First, you randomly distribute a set of scholarships to high-school seniors. Then, using the individuals' earnings—along with the outcome of the scholarship lottery—you estimate the returns to education by regressing earnings on an indicator for whether individuals received the random scholarships.

This is intention to treat (**ITT**) because we are including all individuals in their assigned treatment group, whether or not they completed college (i.e. complied with treatment).

### Question 07 Slight update to [06]: You still want to estimate the effect of a college degree on earnings. You randomly distribute a set of scholarships to high-school seniors. Then, using the individuals' earnings (and your information on the scholarship lottery), you estimate the returns to education as the ratio A/B where

### - A = the scholarship's effect on earnings

### - B = the scholarship's effect on the probability of receiving a college degree

This is estimating the treatment on the treated (**TOT**) because we're weighing for the compliance rate (the probability of receiving a college degree)

### Question 08 Continuing [07]: For this question only: What if the scholarship increases graduation rates from lower-income students and decreases graduation rates for higher-income students?

This is estimating the Local Average Treatment Effect (**LATE**) because we are looking at changes to income, and there are non-uniform effects on the treatment along values of this variable.

### Question 09 Continuing [07]: What if the effect of college degrees is the same for all individuals?

This is the average treatment effect (**ATE**) because if we know $Y_{1i} - Y_{0i} = 0$ then the estimate from this regression will be the effect an individual randomly drawn from our sample experiences.

### Question 10 To estimate the effect of access to banking on consumption, you match neighborhoods with banks to similar neighborhoods without banks. You then compare consumption in each neighborhood that has a bank to its closest-matched neighborhods that do not have banks (taking the difference). Finally, you take the average of each of these differences.

This is the average treatment effect (**ATE**) because we're averaging the treated and untreated, then taking the difference. However, we should not assume that the treatment was randomly assigned, shining skepticism on any inference from our estimate.

### Question 11 For each treated individual i, you calculate $\tau_i = Y_i - Y_{j(i)}$, where individual j(i) is a randomly selected person from the control group (sampled with replacement). You then estimate the causal effect of treatment as the mean of the $\tau_i$'s.

This is the **bootstrap estimate** because we're randomly sampling with replacement.

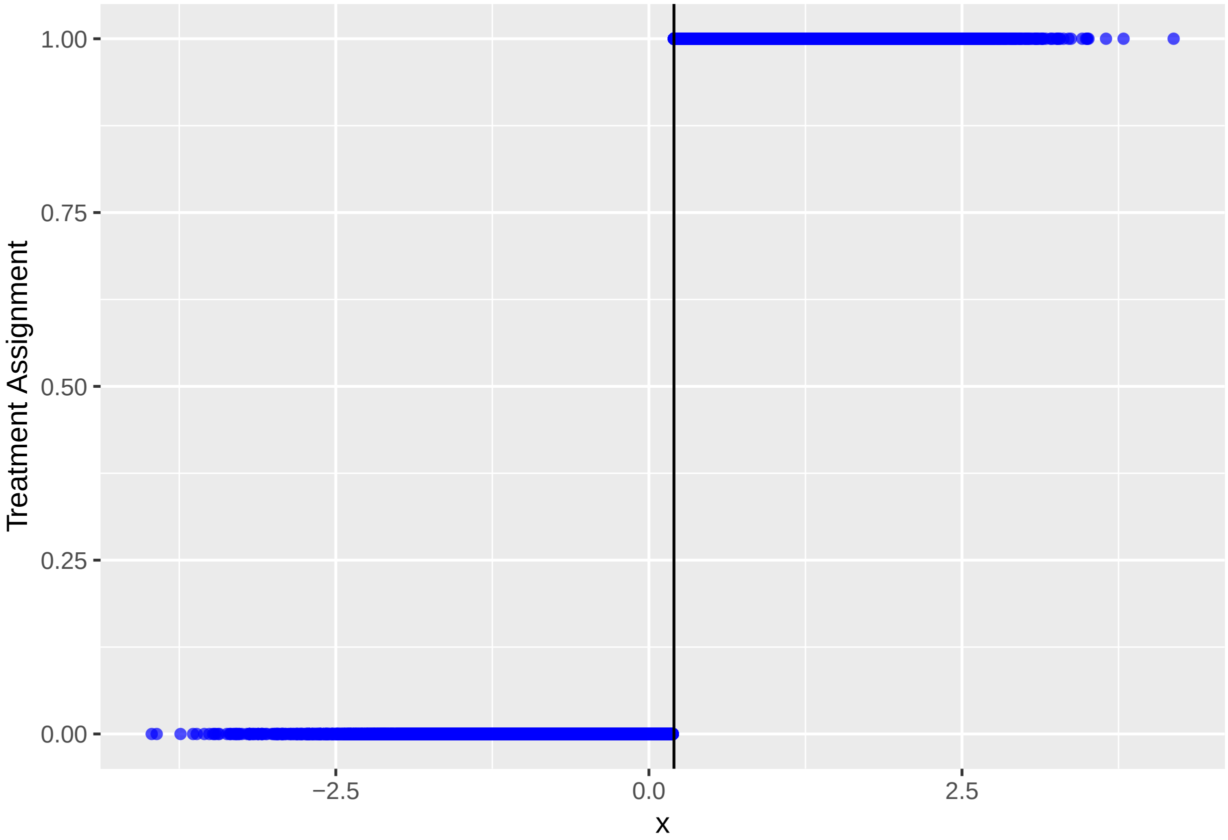
### Question 12 You subtract the mean of the treatment group from the mean of the control group.

This is the intent to treat (**ITT**) because we're simply taking the average of the two groups without regard for compliance-it would also be the negative of the ITT estimate because we're subtracting the wrong way.

---

### Question 13 Is the RD fuzzy or sharp? Illustrate the answer with a single, clear, well-labeled figure.

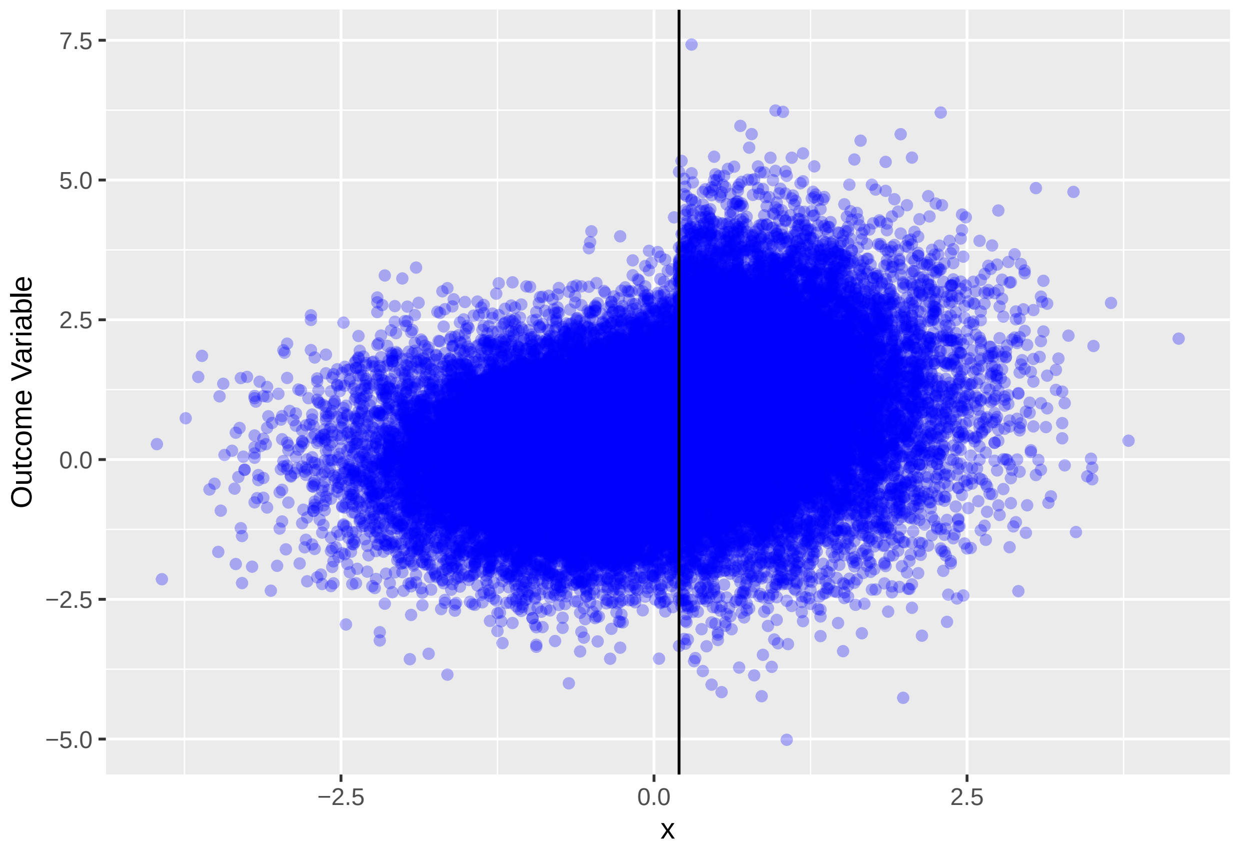
```r
data = read.csv("final-data.csv")

ggplot(data, aes(x = x, y = d)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_vline(xintercept = 0.2) +
  labs(y = "Treatment Assignment",
       x = "x")
```



This demonstrates a clear **sharp regression discontinuity** because the probability of being treated jumps from 0 to 1 at the threshold.

### Question 14 Create a well-labeled scatter plot with the outcome on the y axis and the running variable on the x axis. Does there appear to be a treatment effect? Explain your answer.

```
ggplot(data, aes(x = x, y = y)) +
  geom_point(color = "blue", alpha = 0.3) +
  geom_vline(xintercept = 0.2) +
  labs(y = "Outcome Variable",
       x = "x")
```

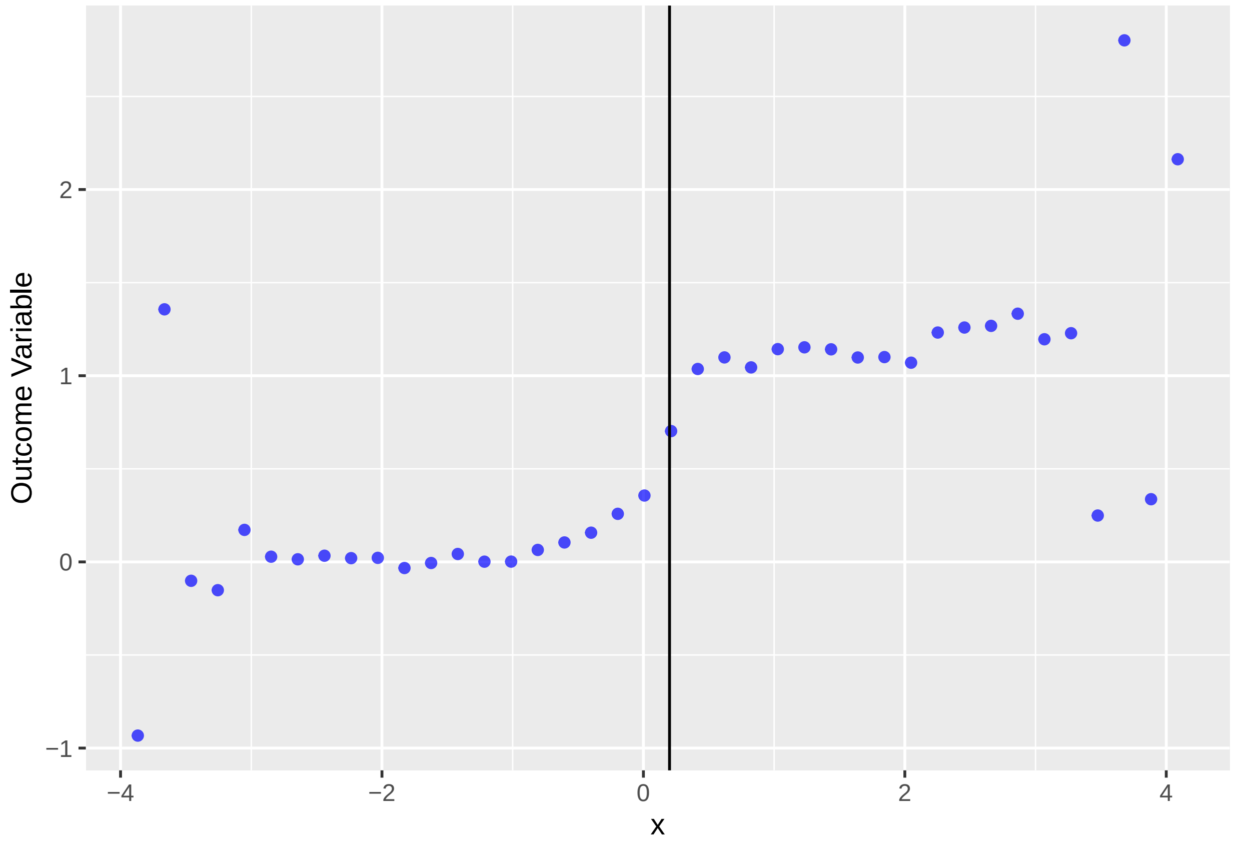


It does not appear as if there is a clear treatment effect. There does appear to be some discontinuity in the top third (or so) of the graph, but it's unclear from the figure if that will translate into an actual treatment effect.

### Question 15 Repeat [14] but now using bin-based summaries (rather than raw data). In other words: summarize the data for each bin, where the bin is defined via the running variable.

```
bins = 40

x_min = min(data$x)
x_max = max(data$x)
x_step = (x_max - x_min)/bins

data2 = data.frame(step_max = seq(from = x_min, to = x_max, by = x_step)) %>%
            mutate(step_min = lag(step_max),
                   group = 0:bins,
                   y = 0) %>%
            filter(group >= 1)

# summarizing the average of y into each bin
for (i in 1:(bins - 1)) {
data2$y[i] = (data %>% filter(x <= data2$step_max[i] & x >= data2$step_min[i]) %>% summarize(mean(y)))[
}
data2$y[bins] = (data %>% filter(data$x > data2$step_max[bins - 1]) %>% summarize(y))[1,1]
```

```
data2 %<>% mutate(x = step_min + x_step/2,
                  d = as.factor(if_else(x > 0.2, 1, 0))) #average of each bin


ggplot(data2, aes(x = x, y = y)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_vline(xintercept = 0.2) +
  labs(y = "Outcome Variable",
       x = "x")
```



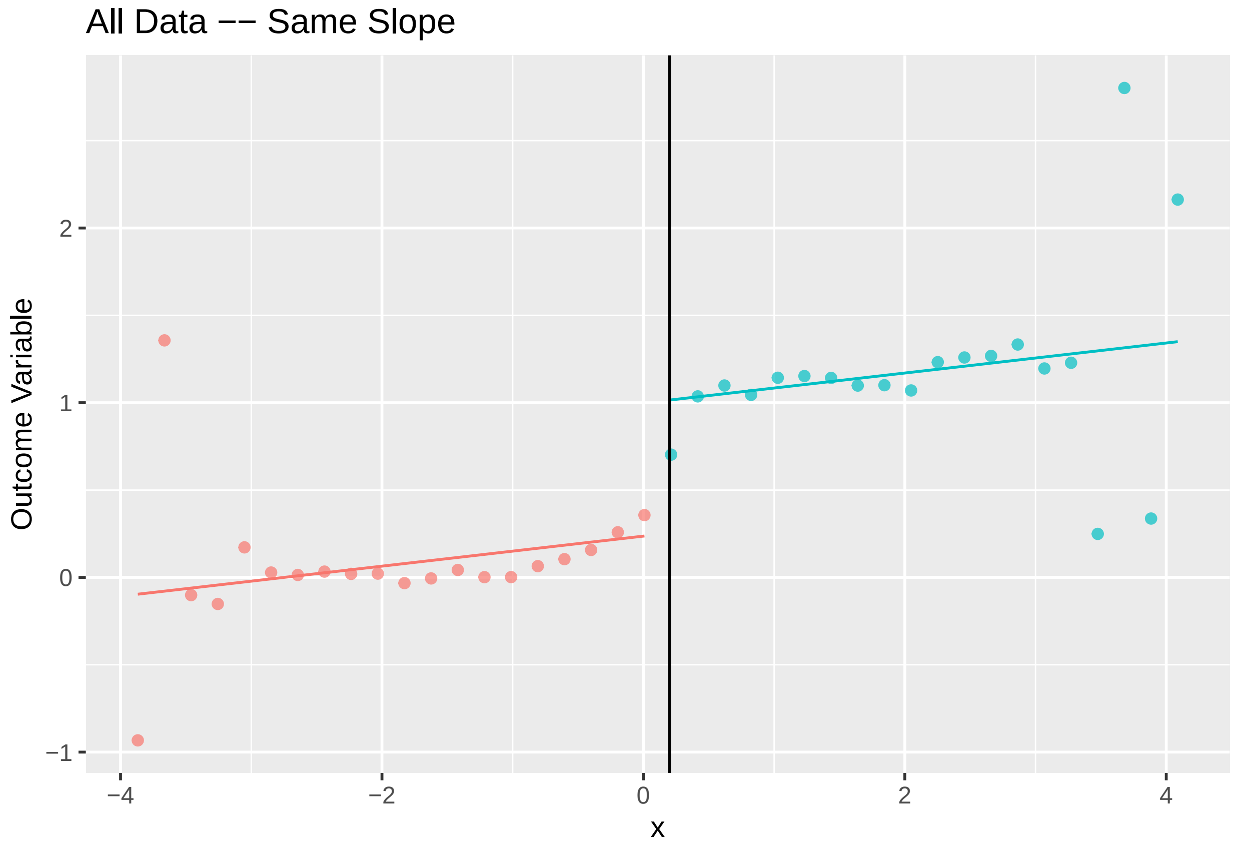
**Question 16 Estimate the treatment effect using an RD design. Each of the following parts gives a different RD specification for you to use in estimating the treatment effect. For each part, provide (1) the point estimate, (2) the standard error, and (3) a clean, well-labeled figure depicting the given estimate and specification.**

```
reg_16a = feols(data = data2 , fml = y ~ x + d)
```

**[16a] Apply a specification that uses constant slope on either side of the discontinuity. Use all of the observations.**

(1) Point estimate: 0.7615*
(2) Standard error: 0.2947
(3) Figure:

```
data2 %<>% mutate(fit_16a = predict(reg_16a))

ggplot(data2, aes(x = x, y = y, color = d)) +
  geom_point(alpha = 0.7) +
  geom_vline(xintercept = 0.2) +
  geom_line(aes(x = x, y = fit_16a)) +
  labs(y = "Outcome Variable",
       x = "x",
       title = "All Data -- Same Slope") +
  theme(legend.position = "none")
```

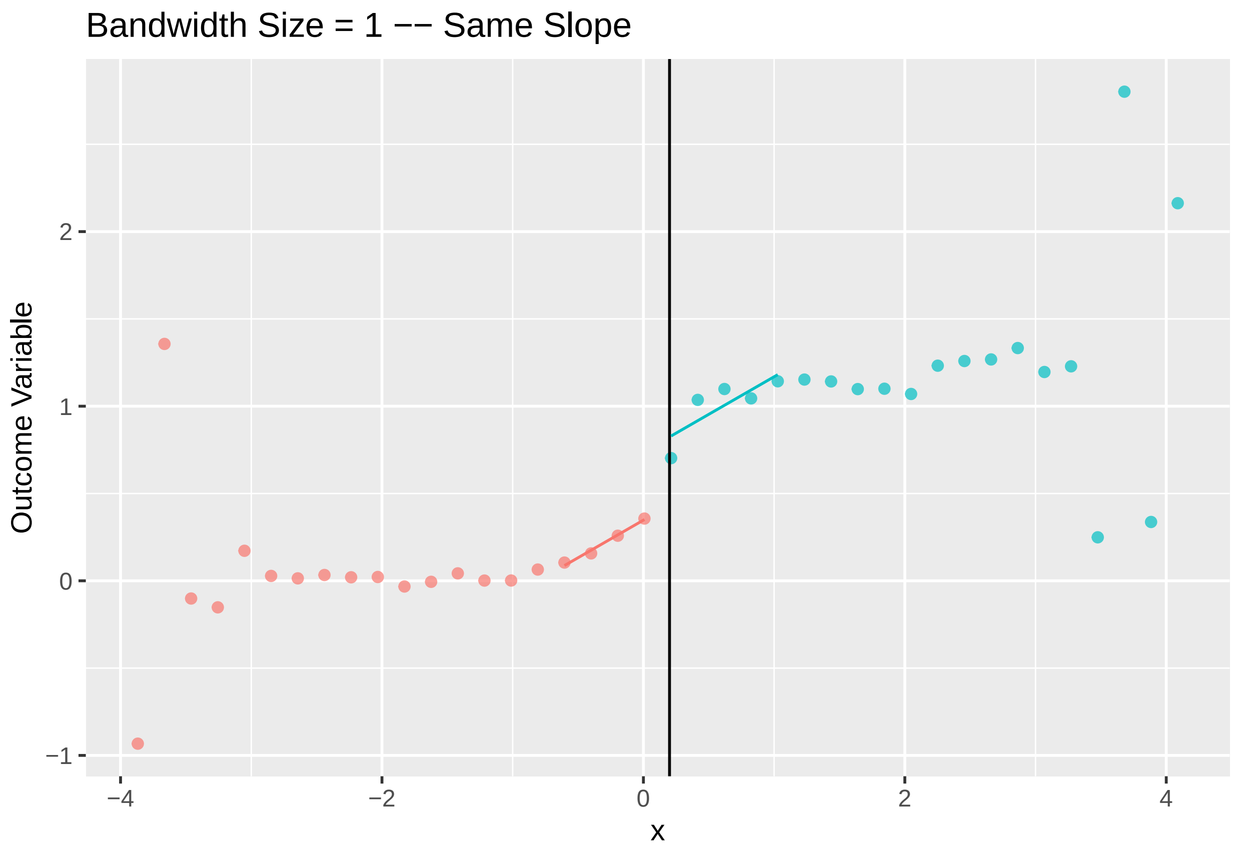


```
# filter so data is within 1 of 0.2
data3 = data2 %>% filter(x >= -0.8, x <= 1.2)

# run regression on filtered data
reg_16b = feols(data = data3, fml = y ~ x + d)

data3 %<>% mutate(fit_16b = predict(reg_16b))
```

**[16b] Repeat, but only use observations within 1 unit of the threshold.**

(1) Point Estimate: 0.3906*
(2) Standard Error: 0.114
(3) Figure:

```
ggplot(data2, aes(x = x, y = y, color = d)) +
  geom_point(alpha = 0.7) +
  geom_vline(xintercept = 0.2) +
  geom_line(data = data3, aes(x = x, y = fit_16b)) +
  labs(y = "Outcome Variable",
       x = "x",
       title = "Bandwidth Size = 1 -- Same Slope") +
  theme(legend.position = "none")
```

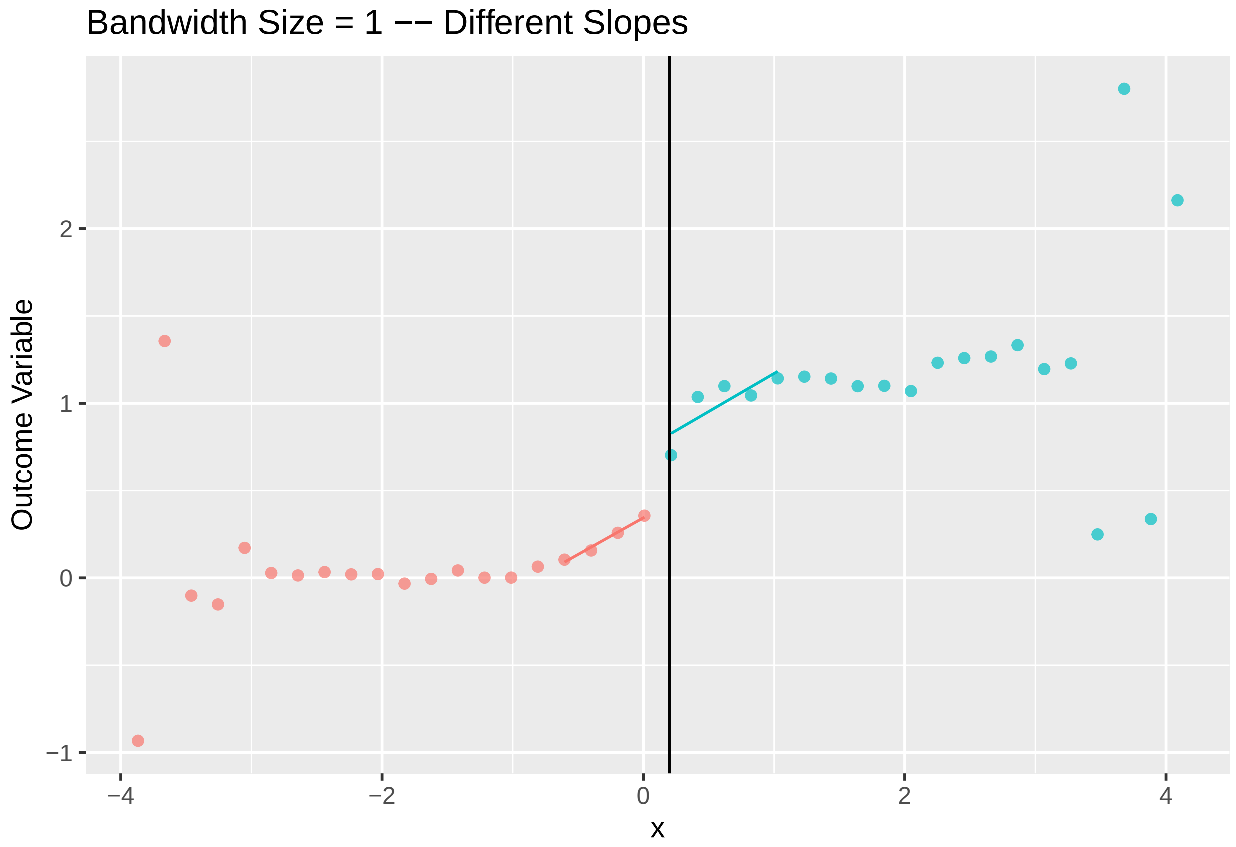


```
reg_16c = feols(data = data3, fml = y ~ x + d + d*x)

data3 %<>% mutate(fit_16c = predict(reg_16c))
```

**[16c] With your tighter bandwidth: Estimate the treatment effect but allow the slope to vary on either side of the discontinuity.**

(1) Point Estimate: 0.3905*
(2) Standard Error: 0.1249
(3) Figure:

```
ggplot(data2, aes(x = x, y = y, color = d)) +
  geom_point(alpha = 0.7) +
  geom_vline(xintercept = 0.2) +
  geom_line(data = data3, aes(x = x, y = fit_16c)) +
  labs(y = "Outcome Variable",
       x = "x",
       title = "Bandwidth Size = 1 -- Different Slopes") +
  theme(legend.position = "none")
```

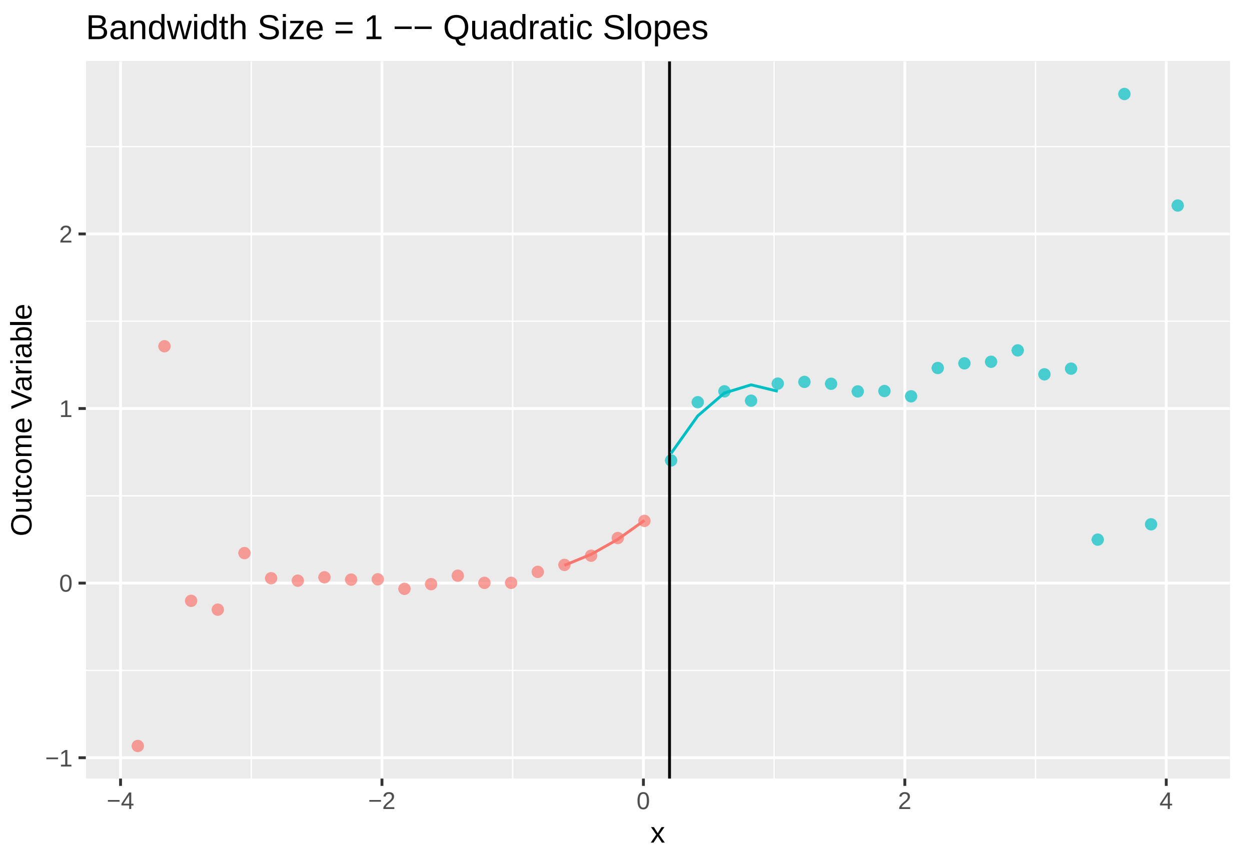


```
reg_16d = feols(data = data3, fml = y ~ d + x + x^2 + d*x + (x^2)*d)

data3 %<>% mutate(fit_16d = predict(reg_16d))
```

**[16d] Keeping your bandwidth tight: Estimate the treatment effect while using quadratic (second-order polynomial) functions on either side of the discontinuity.**

(1) Point Estimate: 0.0769
(2) Standard Error: 0.1865
(3) Figure

```
ggplot(data2, aes(x = x, y = y, color = d)) +
  geom_point(alpha = 0.7) +
  geom_vline(xintercept = 0.2) +
  geom_line(data = data3, aes(x = x, y = fit_16d)) +
  labs(y = "Outcome Variable",
       x = "x",
       title = "Bandwidth Size = 1 -- Quadratic Slopes") +
  theme(legend.position = "none")
```

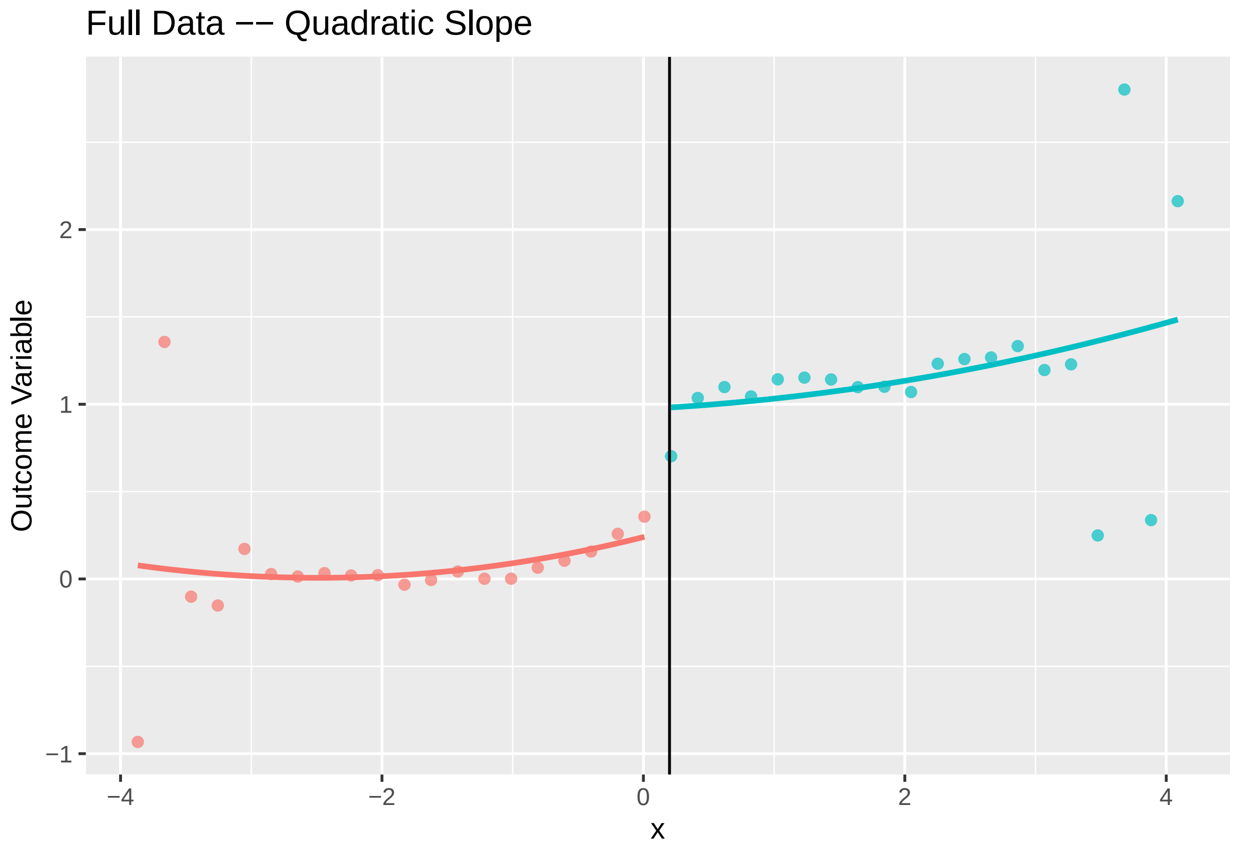


```
reg_16e = feols(data = data2, fml = y ~ d + x + x^2 + d*x + (x^2)*d)
```

**[16e] Repeat the previous quadratic-based estimation on the full dataset.**

(1) Point Estimate: 0.7332
(2) Standard Error: 0.4628
(3) Figure:

```
ggplot(data2, aes(x = x, y = y, color = d)) +
  geom_point(alpha = 0.7) +
  geom_vline(xintercept = 0.2) +
  geom_smooth(method = "lm", formula = y ~ poly(x,2), aes(color = d), se = F) +
  labs(y = "Outcome Variable",
       x = "x",
       title = "Full Data -- Quadratic Slope") +
  theme(legend.position = "none")
```

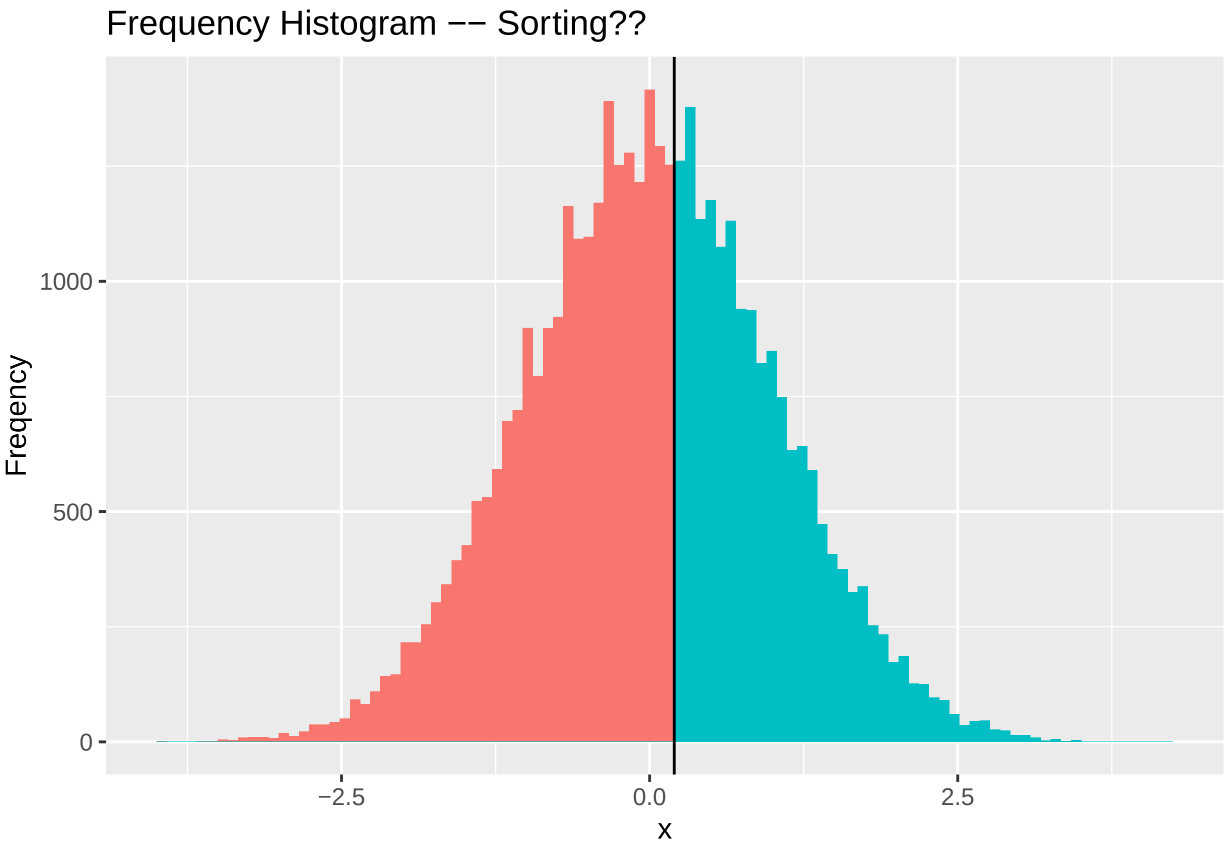


### Question 17 Which of the previous methods worked 'best'? Explain. How much did your treatment-effect estimates vary?

I believe the method in 16d works the best. This is because I believe it balances being flexible (2nd order polynomial) and restricted (by the tight bandwidth) the best out of all of the methods. It should be noted that the estimate from this method is not statistically significant.
The largest estimate is **0.7615** from 16a (using all the data and the same linear slope) and the smallest estimate is **0.0769** from 16d, meaning our estimates vary at most by **0.6846**

### Question 18 Using a single, clear, well-labeled figure: Illustrate whether we should be concerned with sorting. Explain your answer.

```
data %<>% mutate(d = as.factor(if_else(x > 0.2, 1, 0)))

ggplot(data) +
  geom_histogram(aes(x = x, fill = d), bins = 100) +
  geom_vline(xintercept = 0.2) +
  labs(y = "Freqency",
       x = "x",
       title = "Frequency Histogram -- Sorting??") +
  theme(legend.position = "none")
```



It *does not* appear as if any clear discontinuities are present around the cutoff, so I *do not* believe we should be concerned with sorting.

---

### Question 19 Write a simple simulation that demonstrates the M bias. (Hint: DAG lecture) Your simulation should include at least 100 iterations (ideally it would be closer to 10,000, but you've got a short timeline).

We have 5 variables in this simulation: outcome variable $Y$, treatment $D$, and covariates $A, B$, and $C$.
In *M-Bias*, $A$ and $D$ cause $Y$, $B$ causes $D$, and importantly, $A$ and $B$ cause $C$.

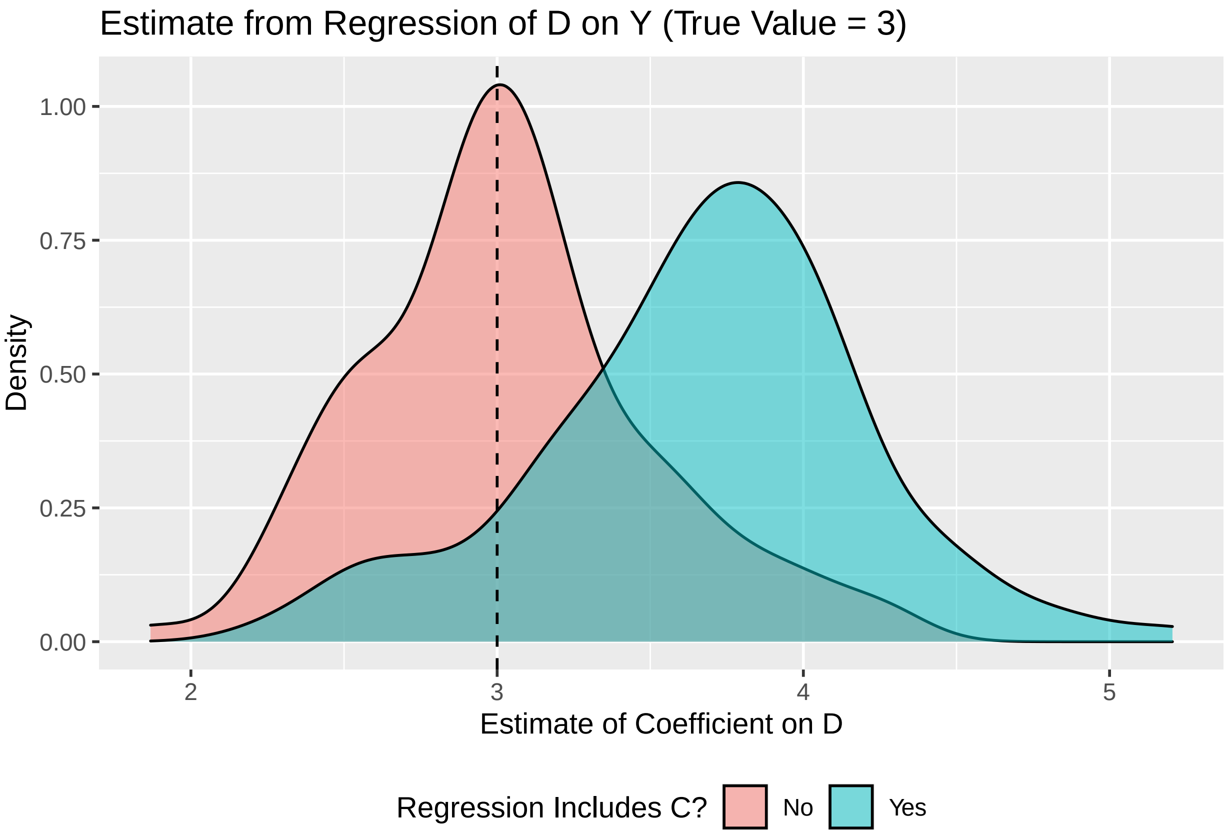
For the *Data Generating Process*, $A$ and $B$ will need to be randomly generated, and every other variable will be some equation containing the appropriate "ancestors." Let these be the equations for the DGP:
- $A \sim U(0,5)$
- $B \sim N(2,4)$
- $C = \frac{1}{2}A - 2B$
- $D = \begin{cases} 0 & ifB<0 \\ 1 & ifB \geq 0 \end{cases}$
- $Y = 1 + \frac{3}{4}A + 3D$

In *each iteration of the simulation* I will generate $n = 50$ observations of the variables above, and run two regressions: 1)$Y \sim \beta_0 + \beta_1 D$ and 2)$Y \sim \beta_0 + \beta_1 D + \beta_2 C$
I will then collect the estimated coefficients for $\beta_1$ and put them into a new data frame for the figure in question 20.

```
# Create function for iterating
func_iter = function(iter, n = 50) {
  # Generate data
  iter_df = tibble(
    A = runif(n, min = 0, max = 5),
    B = rnorm(n, mean = 2, sd = 2),
    C = 0.5*A - 2*B,
    D = if_else(B < 0, 0, 1),
    Y = 1 + (3/4)*A + 3*D
  )

  reg1 = feols(data = iter_df, Y ~ D)
  reg2 = feols(data = iter_df, Y ~ D + C)

  bind_rows(tidy(reg1), tidy(reg2)) %>%
      filter(term == "D") %>%
      select(1:2) %>%
      mutate(reg_type = c("No", "Yes"))
}

# Run simulation (make sure purrr is loaded)
  # set seed
set.seed(31174)

  # run simulation
simulation_list = map(1:100, func_iter)

  # turn into a dataframe
data_sim = bind_rows(simulation_list)
```

**Question 20 Create a nice figure of the simulation's results.**

```
ggplot(data_sim) +
  geom_density(aes(estimate, fill = reg_type),
               position="identity", alpha = 0.5) +
  labs(fill = "Regression Includes C?",
       x = "Estimate of Coefficient on D",
       y = "Density",
       title = "Estimate from Regression of D on Y (True Value = 3)") +
  geom_vline(xintercept = 3, linetype="dashed") +
  theme(legend.position = "bottom")
```


Estimate from Regression of D on Y (True Value = 3)
Density
1.00
0.75
0.50
0.25
0.00
2
3
4
5
Estimate of Coefficient on D
Regression Includes C?
No
Yes