Lab 6 Solutions

2024-05-10

```r
# Grab 2 packages
library(pacman)
p_load(tidyverse, ggplot2, dplyr, fixest, broom)
```

Group 1: Difference in Differences and Parallel Trends

The **parallel trends** assumption in the difference-in-differences model says that absent treatment, the treatment group *would have* followed the same trend that the control group did.

Consider the following scenario: three states (A, B, and C) are considering implenting near identical investment policies. It is passed in states A and B, not in C. However, in state B, the day before the policy was to be implemented, it was halted by a Federal Judge. In state C, since all hope is lost, residents move away (to state D, suppose) and state C begins a recession.

Suppose GDP (Y) in each state i at time t is determined by the following data generating process:

$$Y_{i,t} = \alpha + \gamma_i + \beta \cdot t + \delta \cdot \mathbb{I}(policy_{i,t}) + \eta \cdot t \cdot \mathbb{I}(i = C, t \geq 0) + \varepsilon_{i,t}$$

Where:

- $\alpha = 10$
- Fixed effects: $\gamma_A = 2, \gamma_B = -1, \gamma_C = 0.5$
- Time trend: $\beta = 0.5$
- $\delta = 1$, and $\mathbb{I}(policy_i) = 1$ if the policy is implemented, 0 otherwise
- $\eta = -0.7$
- $\varepsilon_{i,t} \sim N(0, 0.3)$

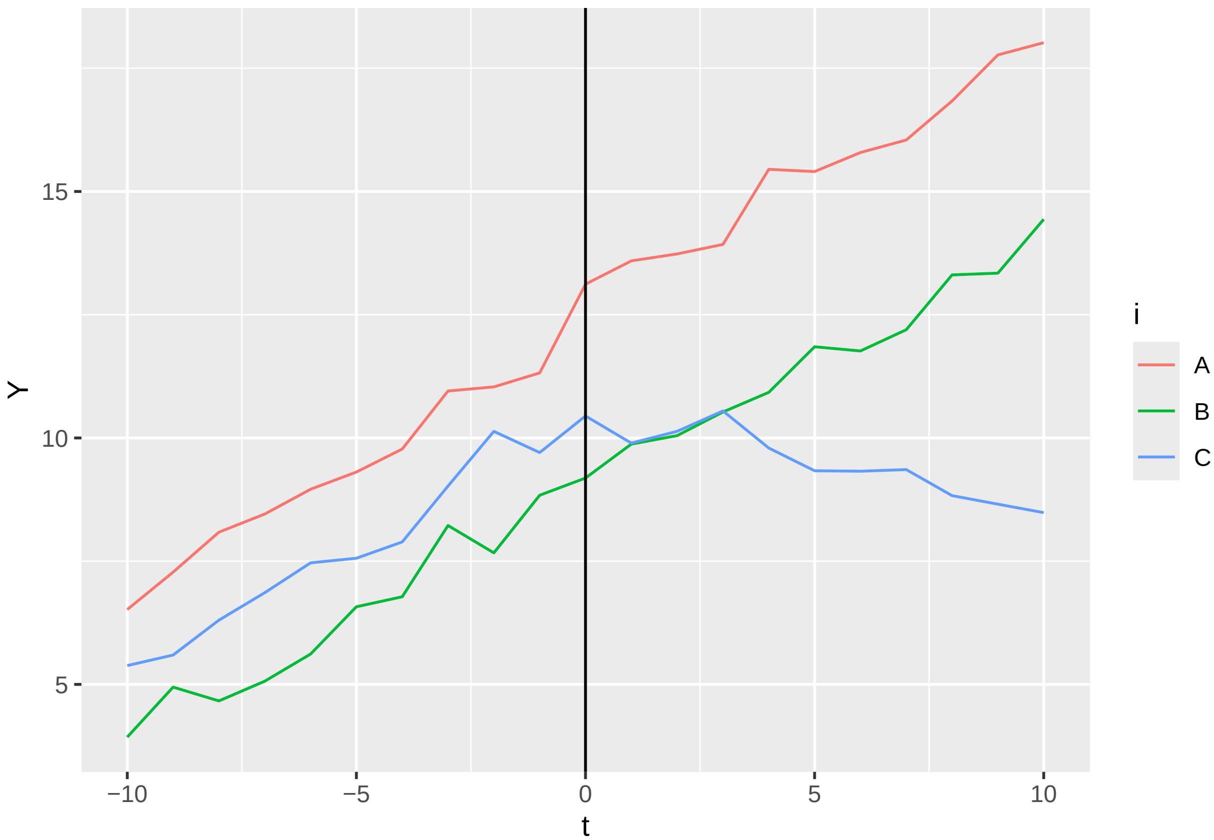
You, as an economist interested in the causal effect of the policy's implementation.

1.1 Create data Write a function that generates data for states A, B, and C over periods -10 to 10, where the policy is enacted at time $t = 0$.

Graph an example dataset, showing each state's output (Y) overtime on the same graph, depicting when the policy was implemented.

```r
data_1 = function(alpha = 10, beta = 0.5, delta = 1, eta = -0.7){

  data = tibble(i = rep(c("A", "B", "C"), 21)) %>%
    group_by(i) %>%
    mutate(t = -10:10) %>% ungroup() %>%
    mutate(D = if_else(t >= 0, 1, 0),
           policy = if_else(i == "A" & D == 1, 1, 0),
           C_ind = if_else(i == "C" & D == 1, 1, 0),
           gamma = if_else(i == "A", 2,
                           if_else(i == "B", -1, 0.5)),
```

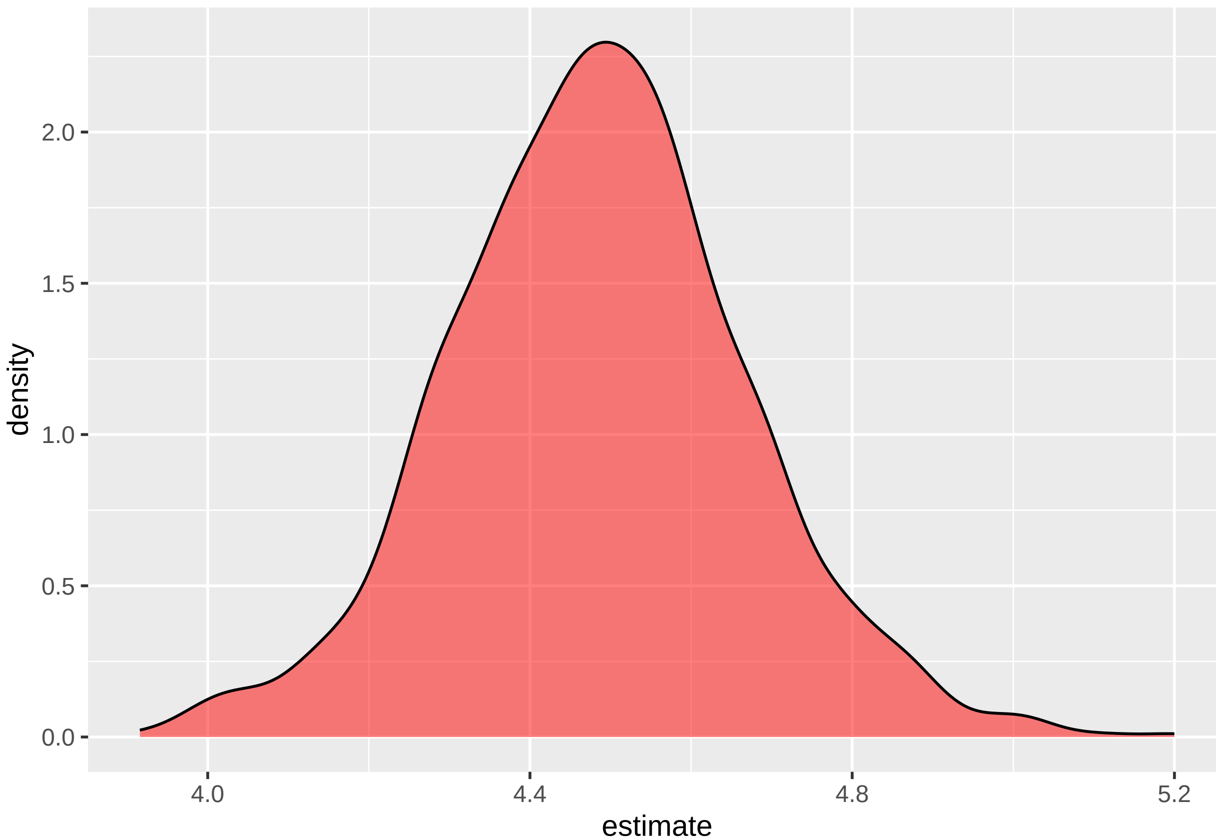
```
          e = rnorm(63, 0,0.3),
          Y = alpha + gamma + beta*t + delta*policy + eta*t*C_ind + e)


  return(data)
}

test_1 = data_1()

ggplot(test_1, aes(x = t, y = Y, color = i)) +
  geom_line() +
  geom_vline(xintercept = 0)
```



1.2 Simulate with C as Control Run 1000 iterations of a simulation where you run a difference in differences regression with state A as the treatment group and C as the control. Then graph the estimates of δ in a density plot.

```
# Function for simulating
sim_1.2 = function(iter){

  data_i = data_1() %>% filter(i != "B") # data except from state B

  reg_i = feols(data_i, Y ~ factor(i) + factor(t) + policy)

  tidy(reg_i) %>%
      # only want the estimate of b
        filter(term == "policy") %>%
```

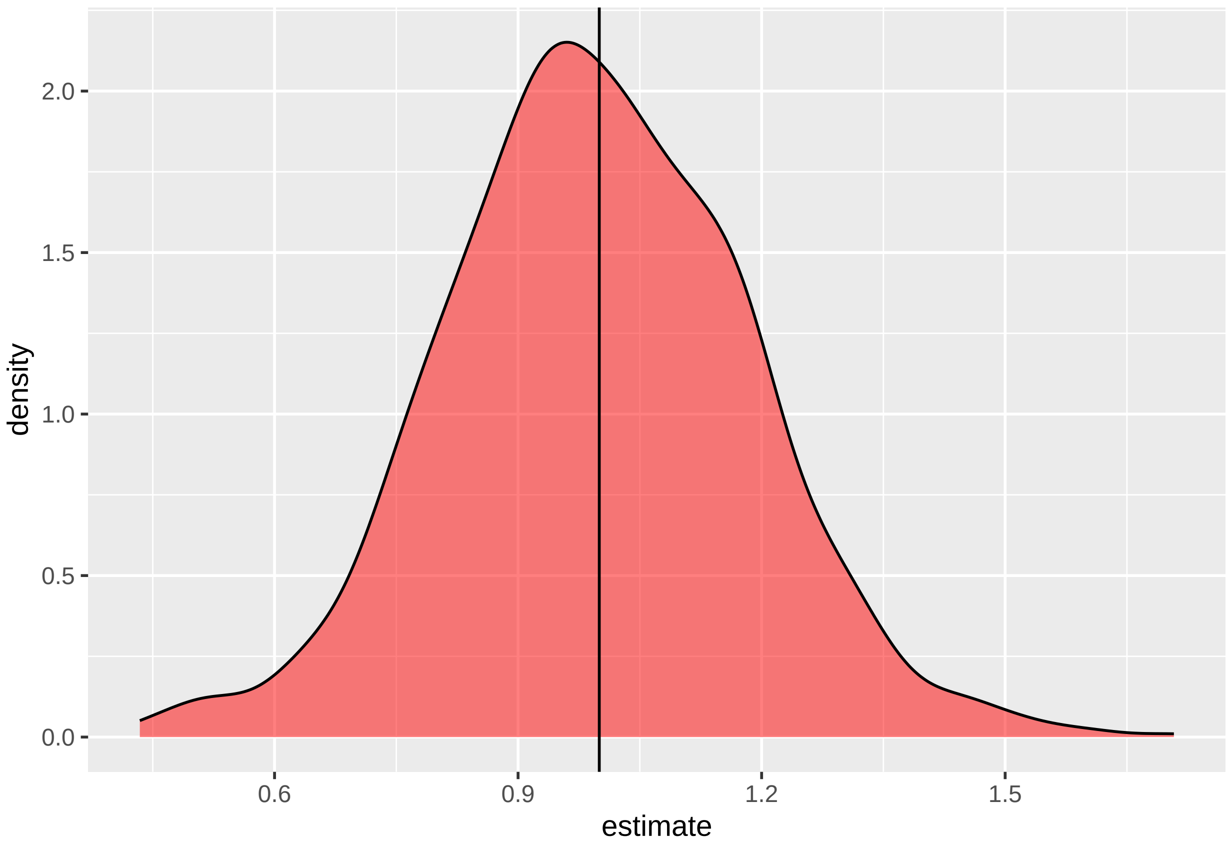
```
      # grab the estimate
        select(2)
}

# Simulate!
iter = 1000

results_1.2 = bind_rows(map(1:iter, sim_1.2))

# And graph:
ggplot(results_1.2, aes(estimate)) +
  geom_density(fill = "red", alpha = 0.5)
```



1.3 Simulate with B as Control Repeat **1.2**, this time with B as the control.

```
sim_1.3 = function(iter){

  data_i = data_1() %>% filter(i != "C") # data except from state C

  reg_i = feols(data_i, Y ~ factor(i) + factor(t) + policy)

  tidy(reg_i) %>%
      # only want the estimate of b
        filter(term == "policy") %>%
      # grab the estimate
        select(2)
}
```

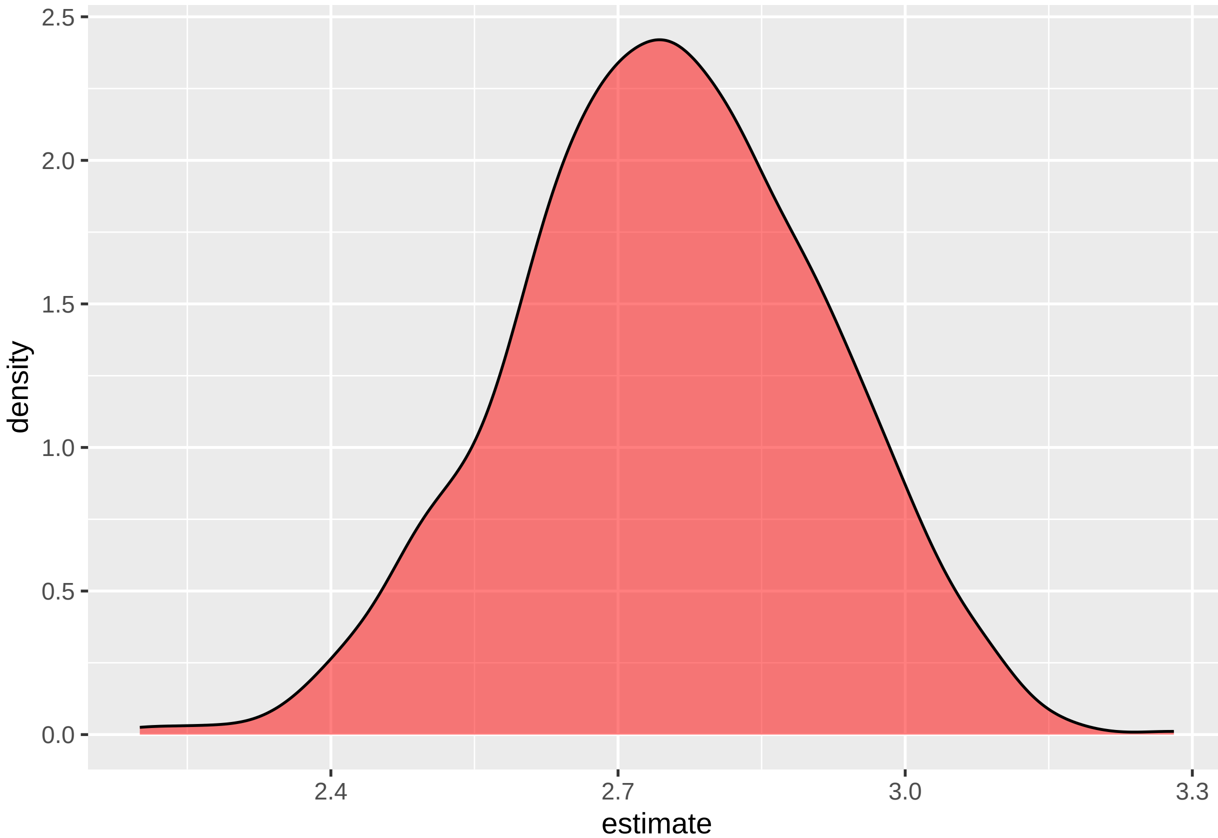
```r
# Simulate!
results_1.3 = bind_rows(map(1:iter, sim_1.3))

# And graph:
ggplot(results_1.3, aes(estimate)) +
  geom_density(fill = "red", alpha = 0.5) +
  geom_vline(xintercept = 1)
```



1.4 BONUS! Run 1.2/1.3 again, this time using both B and C as control groups.

```r
sim_1.4 = function(iter){

  data_i = data_1() # ALL DATA

  reg_i = feols(data_i, Y ~ factor(i) + factor(t) + policy)

  tidy(reg_i) %>%
      # only want the estimate of b
        filter(term == "policy") %>%
      # grab the estimate
        select(2)
}

# Simulate!
results_1.4 = bind_rows(map(1:iter, sim_1.4))
```

```
# And graph:
ggplot(results_1.4, aes(estimate)) +
  geom_density(fill = "red", alpha = 0.5)
```



Notice: still biased, but less so than in **1.3**

Group 2: Instrumental Variable

Consider an agriculture market where equilibrium is determined by the two following equations for Supply and Demand:

$$\begin{aligned} \text{Supply: } q_t &= \gamma \cdot p_t + \eta \cdot w_t + \nu_t \\ \text{Demand: } q_t &= \delta \cdot p_t + \varepsilon_t \end{aligned}$$

Where $\gamma, \eta > 0$, and $\delta < 0$. p_t and q_t are de-meaned measures of price and quantity of the good, w_t represents a de-meaned measure of weather, where higher levels of w_t increase crop yields.

2.1 Create Data To generate the data, first solve for the market clearing price p_t. The exogenous variables are drawn i.i.d. (each period) from the following distributions:

- $w_t \sim U(-3, 3)$
- $\nu_t \sim N(0, 1)$
- $\varepsilon \sim N(0, 2)$

Generate the data for 100 periods $t = 1$ to $t = 100$ where:

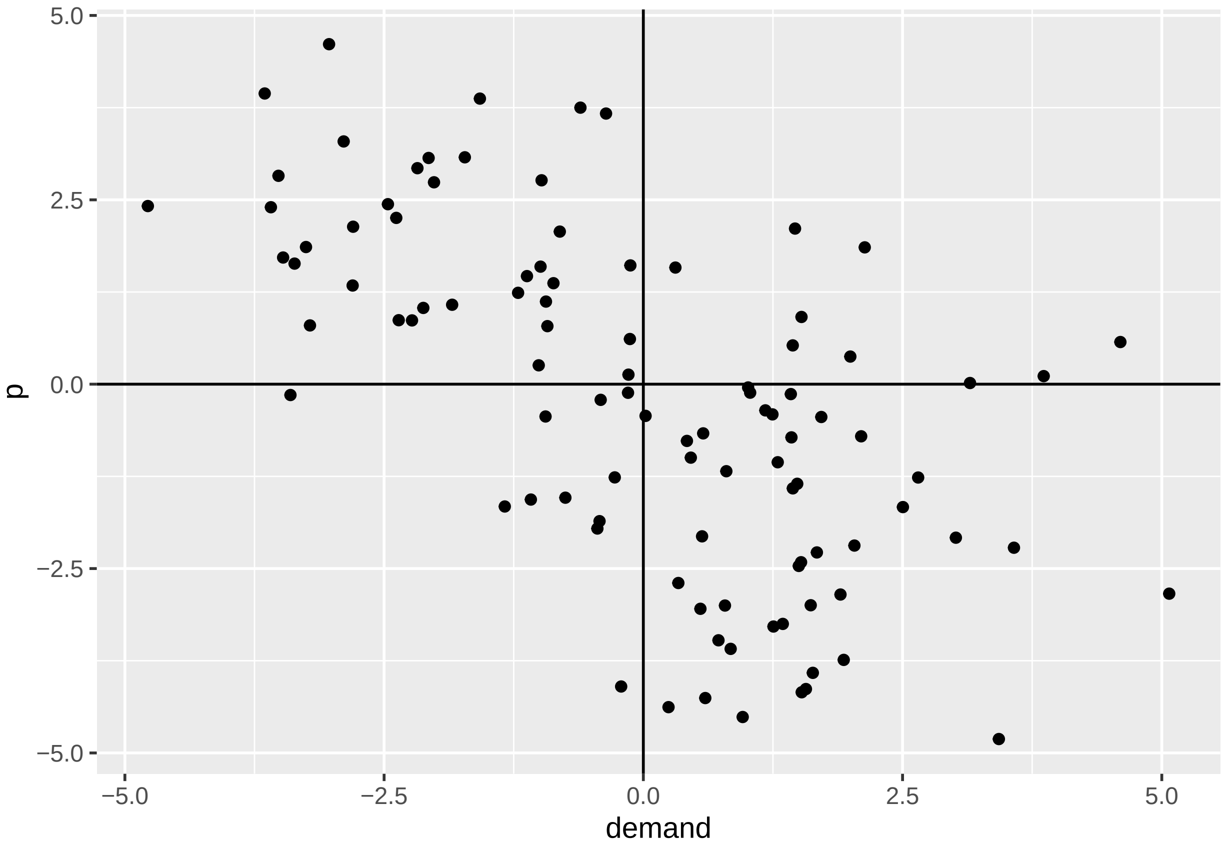
$$\gamma = 0.5, \quad \eta = 1.5, \quad \delta = -1$$

Graph and example data set with p_t against quantity demanded (q_t).

Market Clearing Price

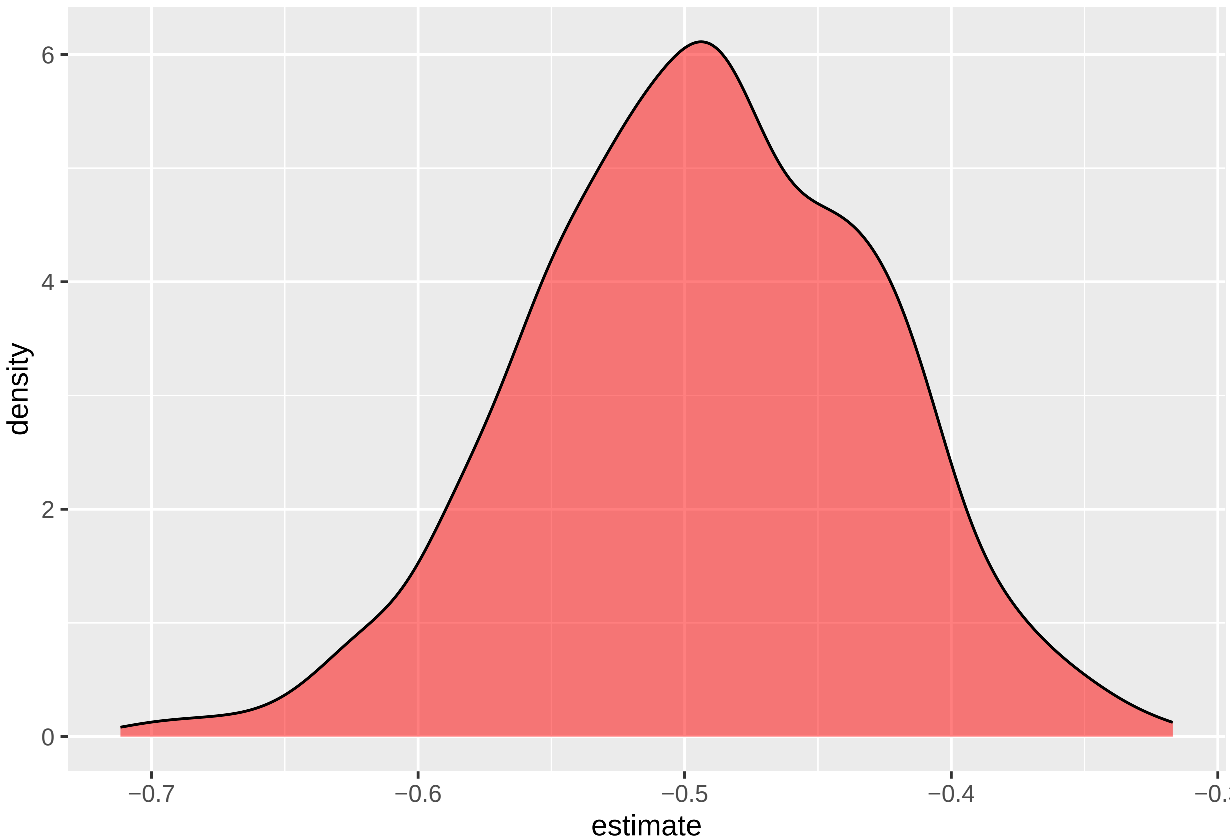
$$p_t = \frac{1}{\delta - \gamma} (\eta \cdot w_t + \nu_t - \varepsilon_t)$$

```r
data_2 = function(delta = -1, gamma = 0.5, eta = 1.5){

  data = tibble(
    t = 1:100,
    w = runif(100, -3, 3),
    v = rnorm(100, 0, 1),
    e = rnorm(100, 0, 2),
    p = (1/(delta - gamma))*(eta*w + v - e),
    demand = delta*p + e,
    supply = gamma*p + eta*w + v
  )

  return(data)
}


# graph quantity and price
test_2 = data_2()

ggplot(test_2, aes(y = p)) +
  geom_point(aes(x = demand)) +
  geom_vline(xintercept = 0) +
  geom_hline(yintercept = 0)
```



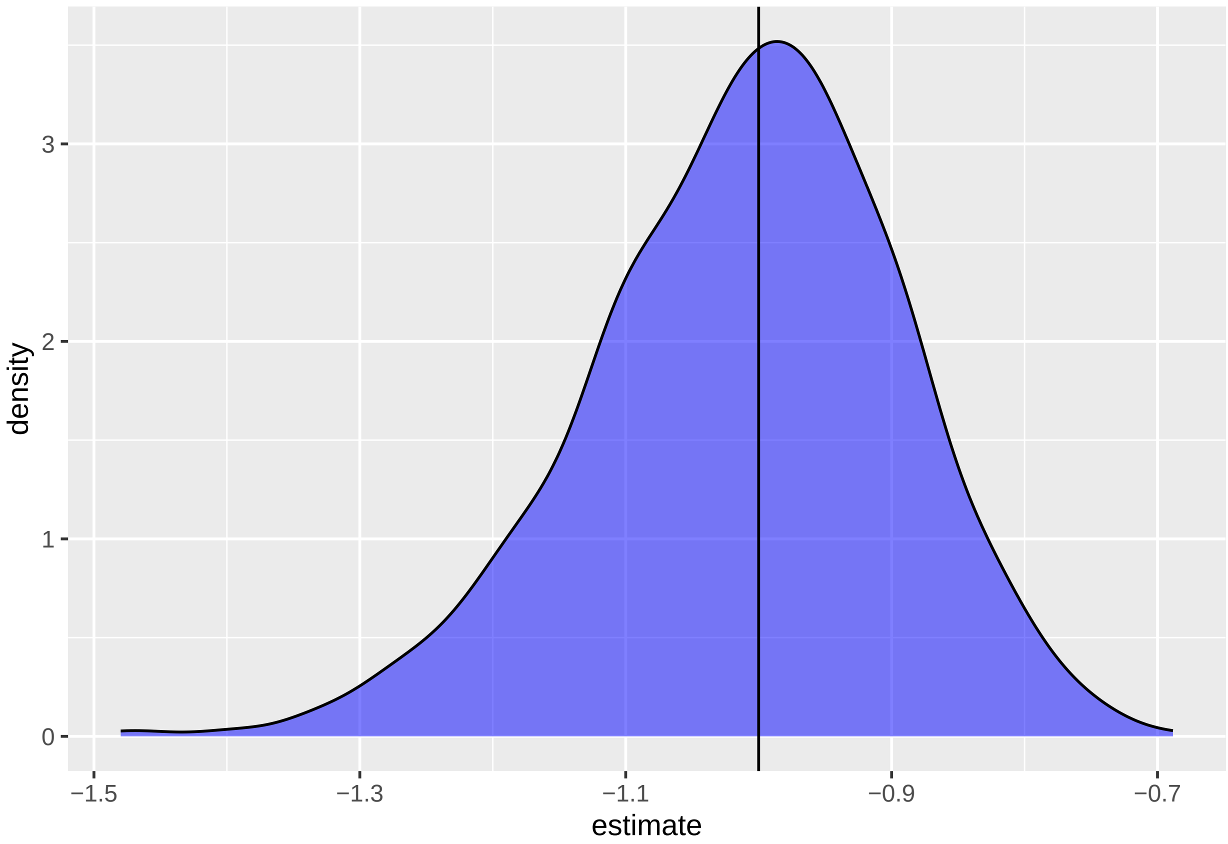
2.2 Estimating Demand Write a function that simulates 1000 iterations of this data set, running a regression estimating δ. Collect the estimates and graph on a density plot.

```
sim_2.2 = function(iter){

  data_i = data_2()

  reg_i = feols(data_i, demand ~ p)

  tidy(reg_i) %>%
      # only want the estimate of b
        filter(term == "p") %>%
      # grab the estimate
        select(2)
}

# Simulate!
results_2.2 = bind_rows(map(1:iter, sim_2.2))

# And graph:
ggplot(results_2.2, aes(estimate)) +
  geom_density(fill = "red", alpha = 0.5)
```



2.3 IV Now repeat **2.2** using w_t as an instrument for price.

```
p_load(ivreg)


sim_2.3 = function(iter){

  data_i = data_2()

  iv_reg = ivreg(demand ~ p | w, data = data_i)

  tidy(iv_reg) %>%
      filter(term == "p") %>%
      select(2)
}

# Simulate!
results_2.3 = bind_rows(map(1:iter, sim_2.3))

# And graph:
ggplot(results_2.3, aes(estimate)) +
  geom_density(fill = "blue", alpha = 0.5) +
  geom_vline(xintercept = -1)
```



2.4 Invalid Instrument? Generate the data again, this time splitting observations into two groups of equal size, and make the following correction:

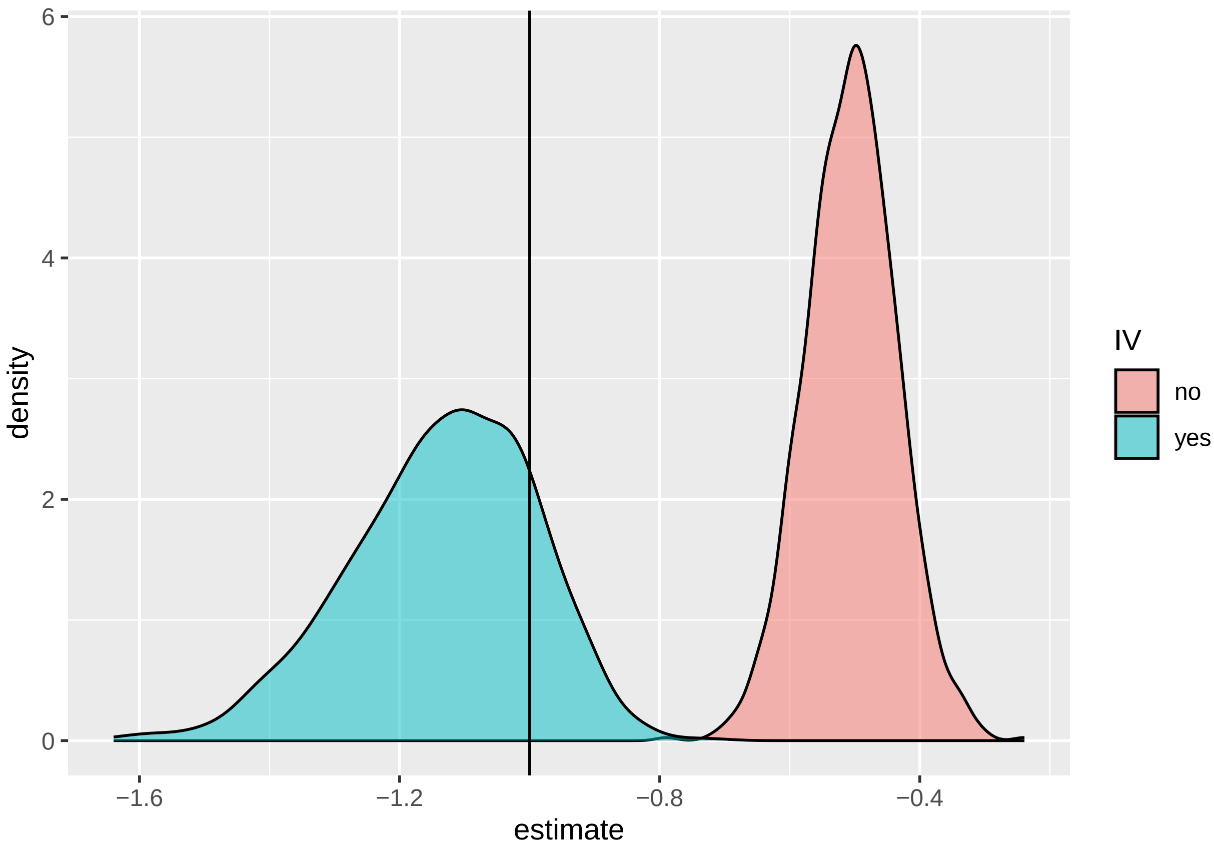
- In odd periods: $\varepsilon_t \sim N(0, 2)$
- In even periods: $\varepsilon_t \sim N(0, 2) + 0.2 \cdot w_t$

That is, there is correlation between the demand disturbances and the weather (bad weather makes consumers discouraged).

Repeat **2.2** and **2.3** again with this data generating process

```
data_2.4 = function(delta = -1, gamma = 0.5, eta = 1.5){

  data = tibble(
    t = 1:100,
    w = runif(100, -3, 3),
    v = rnorm(100, 0, 1),
    D = rep(0:1, 50),
    e = rnorm(100, 0, 2) + D*0.2*w,
    p = (1/(delta - gamma))*(eta*w + v - e),
    demand = delta*p + e,
    supply = gamma*p + eta*w + v
  )

  return(data)
}

sim_2.4 = function(iter){
```

```
  data_i = data_2.4()

  reg_i = feols(data_i, demand ~ p)

  iv_reg = ivreg(demand ~ p | w, data = data_i)

  bind_rows(tidy(reg_i), tidy(iv_reg)) %>%
      filter(term == "p") %>%
      select(2) %>%
      mutate(IV = c("no", "yes"))
}



# Simulate!
results_2.4 = bind_rows(map(1:iter, sim_2.4))

# mean of iv estimates
mean_iv = results_2.4 %>% filter(IV == "yes") %>% summarize(mean(estimate))

# And graph:
ggplot(results_2.4, aes(estimate)) +
  geom_density(aes(fill = IV), alpha = 0.5) +
  geom_vline(xintercept = -1)
```



The IV is no longer valid :/

Group 3: Regression Discontinuity

Consider the following scenario:

You want to estimate the effect of college on earnings. A state college only accepts SAT math scores above 400, you have access to a high school's record of 1000 students' SAT scores (SAT_i) and annual income many years later (Y_i). (Assume everyone who scored over 400 went to college).

The data generating process for income is determined by the following equation:

$$Y_i = \alpha + \delta \cdot D_i + \beta_1 \cdot \ SAT_i \cdot \ (1 - D_i) + \beta_2 \cdot \ SAT_i \cdot D_i + \varepsilon_i$$

$$\text{Where} : D_i = \begin{cases} 1 & \text{if} \quad SAT_i \geq 400 \\ 0 & \text{if} \quad SAT_i < 400 \end{cases}$$

Assume $\beta_1 < \beta_2, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon)$.

Additionally, suppose SAT scores are distributed according to $SAT_i \sim N(500, 120)$ (pretty close to reality)

Which variable determines the causal effect of going to college on income?

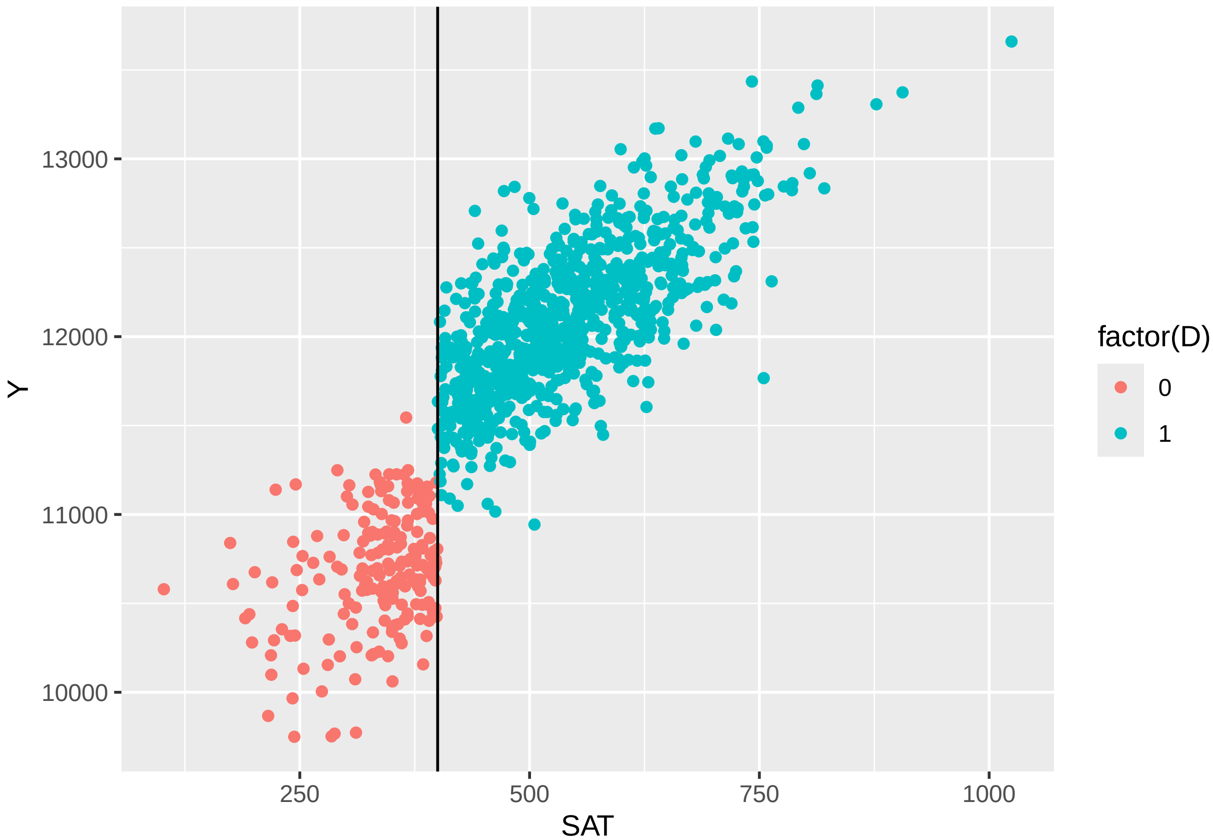
3.1 Create Data: Let:

- $\alpha = 10,000$
- $\delta = 500$
- $\beta_1 = 2$
- $\beta_2 = 3$
- $\sigma_\varepsilon = 300$

Create a function that generates a dataset of 1,000 students according to this scenario.

Plot a sample generation, labeling the cut off and whether an observation went to college or not.

```r
# Data Generating Function

data_3 = function(N = 1000, alpha = 10000, delta = 500, beta1 = 2, beta2 = 3){

  data = tibble(
    SAT = rnorm(N, mean = 500, sd = 120),
    e = rnorm(N, 0, 300),
    D = if_else(SAT >= 400, 1, 0),
    Y = alpha + delta*D + beta1*SAT*(1 - D) + beta2*SAT*D + e
  )

  return(data)
}

test_3 = data_3()

ggplot(test_3, aes(x = SAT, y = Y, color = factor(D))) +
  geom_point() +
  geom_vline(xintercept = 400)
```

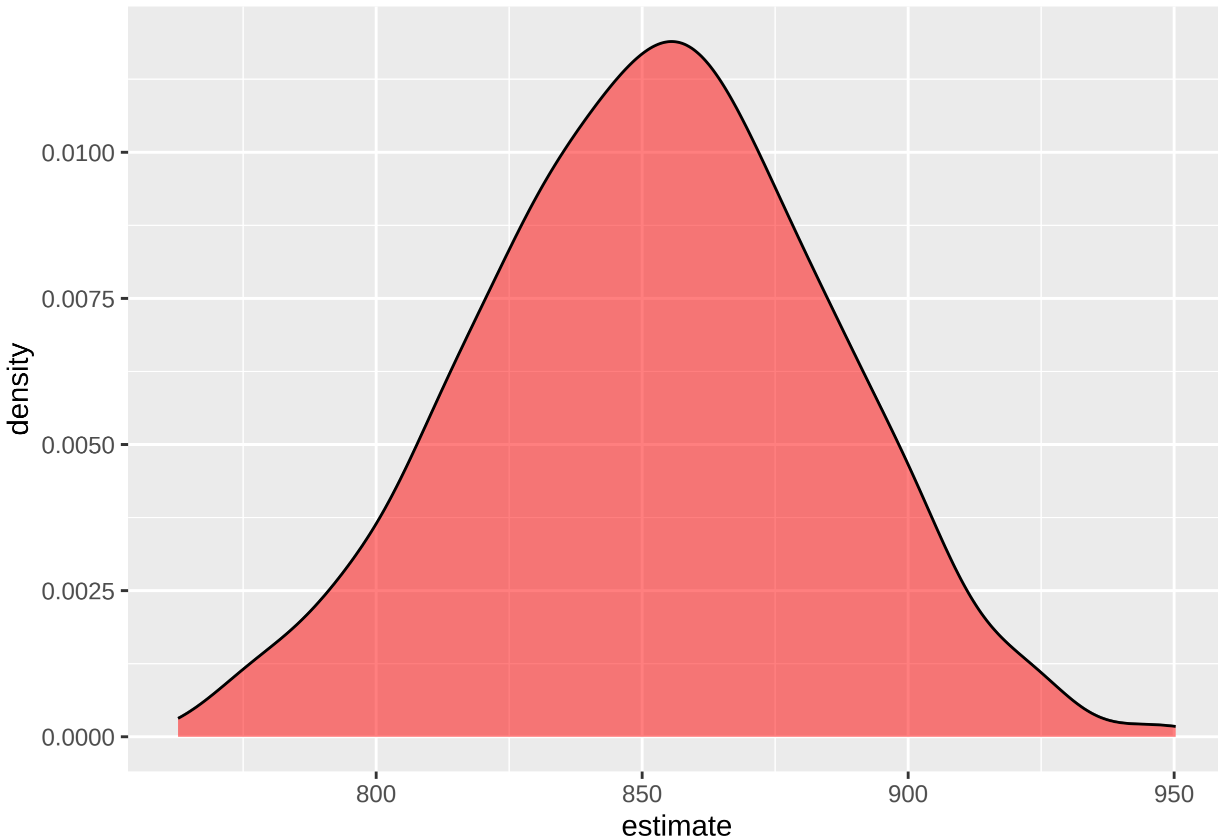


3.2 Regression with Same Slope Write a function that simulates generating the data and running a *Regression Discontinuity Design* linear regression that assumes the slope of the line is the same on either side of the cut off.

Run this simulation 1,000 times, collecting the estimates of δ, and plot the density of these estimates.

Are these estimates biased, unbiased, or ambiguous?

```
# Function for simulating
sim_3.2 = function(iter){

  data_i = data_3()

  reg_i = feols(data_i, Y ~ D + SAT)

  tidy(reg_i) %>%
      # only want the estimate of b
        filter(term == "D") %>%
      # grab the estimate
        select(2)
}

# Simulate!
results_3.2 = bind_rows(map(1:iter, sim_3.2))

# And graph:
ggplot(results_3.2, aes(estimate)) +
  geom_density(fill = "red", alpha = 0.5)
```



Clearly biased

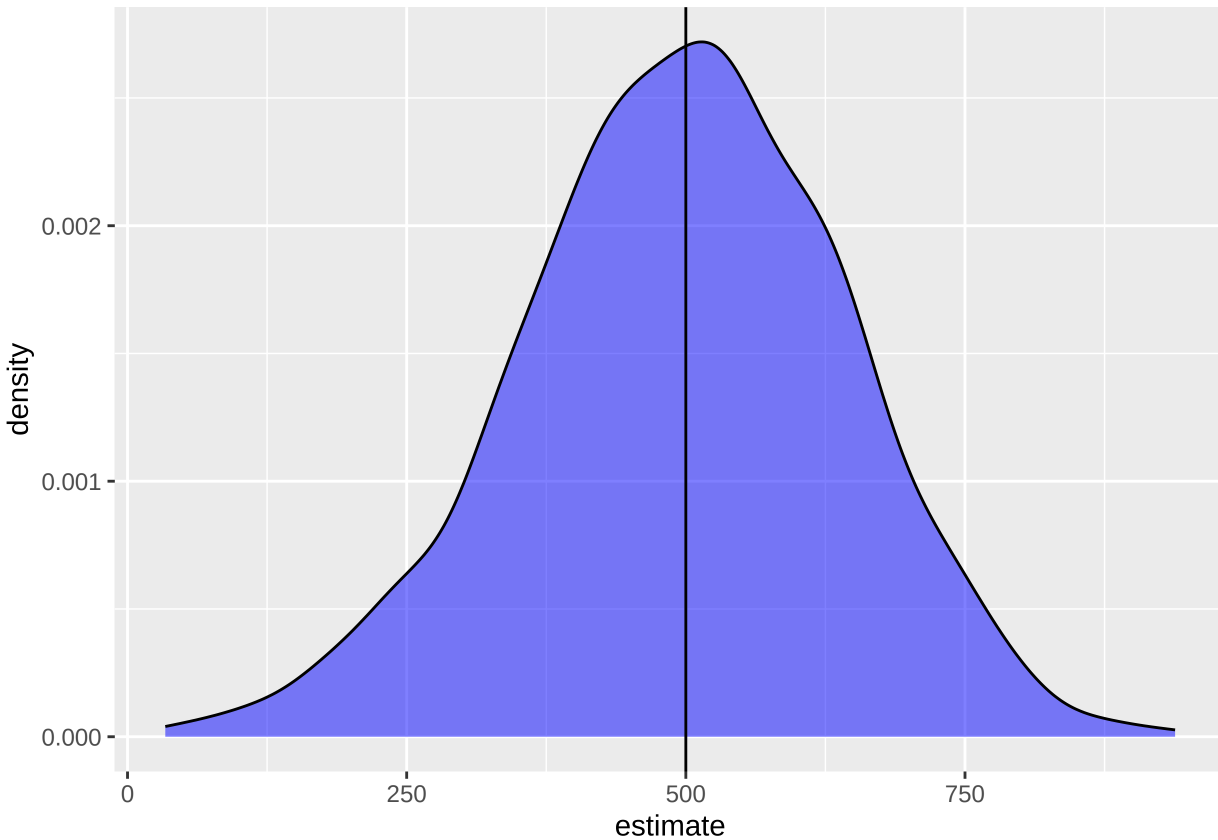
3.3 Regression with Different Slopes Repeat **4.2**, this time allowing for differences in the slope of the regression line on either side of the cut off.

Run this simulation 1,000 times, collecting the estimates of δ, and plot the density of these estimates.

Are these estimates biased, unbiased, or ambiguous?

```
# Function for simulating
sim_3.3 = function(iter){

  data_i = data_3()

  reg_i = feols(data_i, Y ~ D + SAT*(1-D) + SAT*D)

  tidy(reg_i) %>%
      # only want the estimate of b
        filter(term == "D") %>%
      # grab the estimate
        select(2)
}

# Simulate!
results_3.3 = bind_rows(map(1:iter, sim_3.3))

# And graph:
ggplot(results_3.3, aes(estimate)) +
```

```r
  geom_density(fill = "blue", alpha = 0.5)+
  geom_vline(xintercept = 500)
```



UNBIASED!