R Notebook

Contents

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(ivreg)
```

Questions 1-2

1

> Explain briefly how and why simple models, optimization, equilibrium (particularly supply and demand), and an understanding of property rights together form an important part of the core of microeconomic analysis.

Simple models allow us to take the complex world around us and reduce it to something that is easier to digest and work with. We can then take that model and optimize it so that we can get the best result possible. We can then look at how we can influence the world around us to induce that optimal result. The optimal result is not always the equilibrium result which is the point in supply and demand or in another system that is naturally reached without outside influence. Property rights are deeply intertwined with microeconomics because they define how people perceive ownership of goods. Wikipedia notes that property rights have three attributes and that people have "the right to use the good," "the right to earn income from the good," and "the right to transfer the good to others, alter it, abandon it, or destroy it (the right to ownership cessation)."

2

> Suppose riding coaster A provides exactly the same total satisfaction as riding coaster B, and that each seats 50 riders per cycle, but that coaster A takes only 2/3 as much time as coaster B to complete a cycle. Use the idea of equilibrium to estimate the length of the line for B relative to the line for A. Explain your approach.

$$50a = \frac{2}{3}50b50a = 33.33b$$

Let's say that there are 50 people in line for only ride B, the line will move fully in time T. If there are 50 people in line for ride A, the line will take time $\frac{2}{3}T$. People will want the lines to take as little time as possible, and thus gravitate towards the line that takes the least amount of time. If there are 50 people or less, this is relatively simple. They will choose ride A. However, if there are more than 50 people in line for ride A, then they would want to know the proportion. From the math above, we learn that the ride wait times will be equal if the number of people in line B is two thirds that of line A, with the equation being $a = \frac{2}{3}b$ where a and b are the number of people in each line. If the b side of the equation is larger, then people will shift towards ride A. If a is larger, people will go to ride B.

Questions 3-6

> **One-hundred people fish in two lakes, East Lake and West Lake, on any given day. Each may choose to fish in either lake. Let n be the number that fish in East Lake and $100 - n$ be the number that fish in West Lake. West Lake is very large and relatively plain with dispersed fish populations. The value someone receives fishing in W is 20. East Lake is spectacular with concentrated fish populations. The value someone gets from fishing in East Lake is $50 - 0.5n$.**

3

> If people individually choose which lake to fish in, how many will fish in East Lake and how many in West Lake, and what is the total value to everyone fishing? (Total value is just the number fishing East multiplied by the individual value of dishing in East, plus the number fishing in West multiplied by the individual value of fishing in West.)

$$20 = 50 - .5n - 30 = -.5nn = 60$$

The number of people fishing in East Lake is 60 which leaves 40 people in West Lake.

$$20 * 40 = 80060 * (50 - .5(60)) = 1200800 + 1200 = 2000$$

The total value of everyone fishing in the lake is 2000 where the people in West Lake have a combined total of 800 and East Lake is 1200.

4

> If a benevolent dictator could limit the number fishing in East, how many people fishing in East and how many in West would maximize total value, and what would total value be?

```
fishW <- 0
fishE <- 0
maxP <- 0
wMax <- 0
eMax <- 0
max <- 0

for(i in 0:100) {
  fishE <- i
  fishW <- 100 - fishE
  max <- (20*fishW)+(fishE*(50-(.5*fishE)))
  if(max > maxP) {
    eMax <- fishE
    wMax <- fishW
    maxP <- max
  }
}
paste(maxP, eMax, wMax)
```

```
## [1] "2450 30 70"
```

The maximum profit is 2450. This happens when the number of people fishing in East Lake is 30 and the number of people fishing in West Lake is 70.

5

> Relate the difference between outcomes in #3 and #4 to property rights and the tragedy of the commons.

The tragedy of the commons is that every person is seeking to maximize their own profit, rather than the profit of the community, or the commons. In question three, people seek to maximize their own profit, and thus the total profit is only 2000. In question four, the benevolent Dr Dewey decides to maximize collective profit and the total profit becomes 2450.

6

> Suppose the government can charge a fee to fish in East Lake. How large would the fee need to be to induce the optimum number of people to fish there? How much revenue would the government collect?

$V = v(x)v(w) = v(e) - f20 = 50 - .5(n) - f20 = 50 - .5n - f - 30 = -.5n - f30 = .5n + f0 = .5n + f - 30$

```
n <- 30

for(f in 0:1000) {
  if (0 == (.5 * n) + f - 30) {
    break
  }
}

print(paste0("n: ", n, ", f: ", f))
```

```
## [1] "n: 30, f: 15"
```

```
print(paste0("Total profit: ", 20*(100-n) + (n*(50-(.5*n))-(n*f))))
```

```
## [1] "Total profit: 2000"
```

```
n <- 30

for(f in 0:(50-(.5*n))) {
  if (20*(100-n) == (n*(50-(.5*n)-f))) {
    print(paste(n, f))
    break
  }
}

20*(100-n) - (n*(50-(.5*n)-f))
```

```
## [1] 1400
```

```
print(paste0("n: ", n, ", f: ", f))
```

```
## [1] "n: 30, f: 35"
```

```
print(paste0("Total profit: ", 20*(100-n) - (n*(50-(.5*n)-f))))
```

```
## [1] "Total profit: 1400"
```

I could be totally wrong, but I don't want to think I am. From above, we have the equation $v(w) = v(e) - f$ where $v(w)$ and $v(e)$ are the profit functions for West and East Lakes respectively, and f is the fee applied to fishing in East Lake. Written out, that is $20(100 - n) = n(50 - .5n) - fn$ where n is the number of people fishing in East Lake. Since we know that the optimal result is achieved at $n = 30$, we can constrain the loop as such and iterate over a value for the fee, f, and we get $f = 15$ which is a total government revenue of $n * f = 30 * 15 = 450$.

Questions 7-10

7

> Discuss the limitations of randomized control trials covered in the article by Deaton and Cartwright.

The single biggest issue with randomized controlled trials is that you cannot make predictions for which you have no data. So even though your RCT may cover an extremely wide swathe of the population, there are people outside that group who you cannot make predictions for. As you increase the size of the study, the costs and complexity increase and so researchers must find an appropriate balance between the size and cost.

8

> Discuss what gave rise to the credibility revolution.

Everyone was publishing papers about whatever hot topic they could find, but if more than one person studied the same thing, they were likely to make different assumptions about the topic, and thus influence the outcome of the study. This gave rise to the credibility revolution where researchers looked to hold each other accountable for their research assumptions and to design better studies.

9

> List and discuss several factors that combined to increase the credibility of empirical economics over the course of the credibility revolution.

- Sensitivity Analysis: Adding sensitivity analysis allows research authors to test their assumptions and see how much their results would change if everything was just a little bit different.
- Randomized Controlled Trials: As discussed above, RCTs can be a powerful tool to conduct studies which might otherwise not be possible. While not common, they have been used in cases such as Andrew Yang's Universal Basic Income pilot program.
- Better and More Data: More and better data seems like a given. As the quality and quantity increase, your studies will have better results that are more representative of the world at large.
- Better Research Design: Again, this seems like a given. As the quality of the research design improves, then the quality of the results should improve as well.

10

Discuss a potential downside of the credibility revolution.

One criticism, called `External Validity` is that study based research is too specific and not suited for the world at large. In essence, the need for Better Research Design has created a problem where the studies are so narrow in scope in order to control for outside influences, the study may not be useful in general.

All Remaining Questions

You are interested in whether a tutoring program improves performance. You have a measure of initial ability, S_1. You know who signed up for and received voluntary tutoring, T. You later observe a second measure of performance, S_2. You are worried unobservables, U, influence the initial score, the probability of signing up for tutoring for a given initial score, and the second score.

11

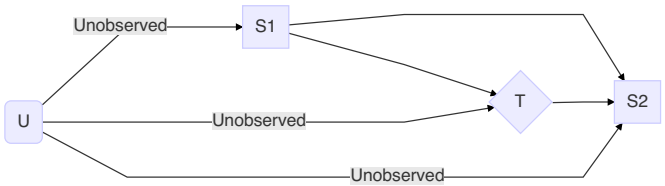
Give some examples of unobservables relevant to this case.

- Some students may or may not drink coffee before the exam
- Some students may have overbearing parents that insist on their little munchkin being given tutoring, regardless of S_1
- Some students may have had a panic attack during the exam due to the lack of a clock in the exam room. Ultimately, a clock was installed before S_2

12

Draw the DAG.

```
library(webshot)
DiagrammeR::mermaid("
graph LR
A(U) -->|Unobserved|B[S1]
    B --> D
    A -->|Unobserved|C{T}
    A -->|Unobserved|D[S2]
    B --> C
    C --> D
")
```



13

What is a backdoor path? Why do you have to block them? How do you block them? Based on the DAG, what backdoor path or paths, if any, can't be blocked with the data you have?

Questions 14-17

Suppose there are only four students and the data is as shown in the table ~~to the right~~ below.

Subject	Y^0	Y^1	T
1	84	76	1
2	76	82	0
3	74	64	1
4	80	78	0

14

What are the observed sample difference, ATT, ATU, ATE, and selection bias?

```
dt <- data.table(
  subject = c(1, 2, 3, 4),
  y0 = c(84, 76, 74, 80),
  y1 = c(76, 82, 64, 78),
  t = c(1, 0, 1, 0)
```

```
)

dt[, TE := y1 - y0][t == TRUE, TT := TE][t == FALSE, TU := TE]
data.table(
  ATE = c(mean(dt$TE)),
  ATT = c(mean(dt$TT, na.rm = T)),
  ATU = c(mean(dt$TU, na.rm = T))
)
```

```
##     ATE ATT ATU
## 1: -3.5  -9   2
```

15

Why, in reality, can't you actually calculate ATT, ATU, ATE, or selection bias? That is, what is the fundamental problem with causal inference?

16

Can big data or AI, on their own, possibly resolve this fundamental problem of causal inference? Why or why not?

17

How does random assignment help break selection bias?

Questions 18-21

Assume treatment was randomly assigned.

18

What exactly is Fischer's sharp null in this context? Why is Fisher's sharp null particularly useful for randomization inference?

19

Show there are 6 possible ways to assign treatment to 2 of 4 students.

```
combn(4, 2)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    1    1    2    2    3
## [2,]    2    3    4    3    4    4
```

I admit that I cheated a little with the code above, but it works. There are six combinations of choosing two students from a group of four. The general formula is $C(n, r) = \frac{n!}{r!(n-r!)}$. We can plug in and get:

$$C(4, 2) = \frac{4!}{2!(4-2)!} = \frac{24}{2!(2)!} = \frac{24}{4} = 6$$

20

> With the small sample you can construct all possible assignments manually. What are the six sample differences under Fischer's sharp null?

```
for(r1 in 1:4) {
  for(r2 in r1:4) {
    print(dt[r1:r2, .(y0, y1)])
  }
}
```

```
##    y0 y1
## 1: 84 76
##    y0 y1
## 1: 84 76
## 2: 76 82
##    y0 y1
## 1: 84 76
## 2: 76 82
## 3: 74 64
##    y0 y1
## 1: 84 76
## 2: 76 82
## 3: 74 64
## 4: 80 78
##    y0 y1
## 1: 76 82
##    y0 y1
## 1: 76 82
## 2: 74 64
##    y0 y1
## 1: 76 82
## 2: 74 64
## 3: 80 78
##    y0 y1
## 1: 74 64
##    y0 y1
## 1: 74 64
## 2: 80 78
##    y0 y1
## 1: 80 78
```

21

> Discuss how strong the evidence is for an actual effect based on this (small) sample. Basically, what is the relevant p-value and what do you make of it?

Questions 22-25

Suppose you have a sample of 400 with 200 randomly treated.

22

Show the expression for the number of ways to randomly choose 200 from 400. Don't bother trying to calculate how many that is. Why not?

$$C(400, 200) = \frac{400!}{200!(400 - 200)!}$$

For one, *400!==64034522846623895262347970319503005850702583026002959458684445942802397169186831436278478647*
and *200!==7886578673647905035523632139321850622951359776871732632947425332443594499634033429203042840119846*
and I'm a pretty smart guy, but I'm not smart enough for this shit. The end result is a really big number that is so big it has lost all meaning.

23

Practically, how can we approximate the Fisher exact test p-value with large samples?

24

Now suppose treatment is not randomly assigned. Instead, everyone who scored under a 60 got extra tutoring and no one else did. Explain how a regression discontinuity design can get at the causal effect of tutoring. Draw a picture to illustrate the LATE.

25

What is a LATE? Why can we only estimate the LATE, not the ATE? Why do you often need a lot of data to make an RDD work?

Questions 26-27

Now suppose everyone with an initial score under 60 was assigned to tutoring, but some did not take it, and that a couple of students who scored a bit over 60 were able to get tutoring anyway.

26

How would you go about employing regression discontinuity to get at the causal effect in this case? What do you use as an instrument for treatment? Why do you need an instru- ment? What is an instrumental variable, anyway?

27

Draw a figure clearly illustrating the difference between sharp and fuzzy RDD. Hint: This is not about the outcome measure.

Questions 28-32

Data on 400 (hypothetical) observations for these questions is in the ac- companying file "tutoring.csv".

```
tut <- fread("tutoring.csv")
head(tut)
```

```
##    s1 t s2
## 1: 62 1 77
## 2: 65 0 73
## 3: 66 0 71
## 4: 43 1 63
## 5: 67 1 78
## 6: 54 1 71
```

28

Calculate the simple sample difference.

```
tut$diff <- tut$s2 - tut$s1
```

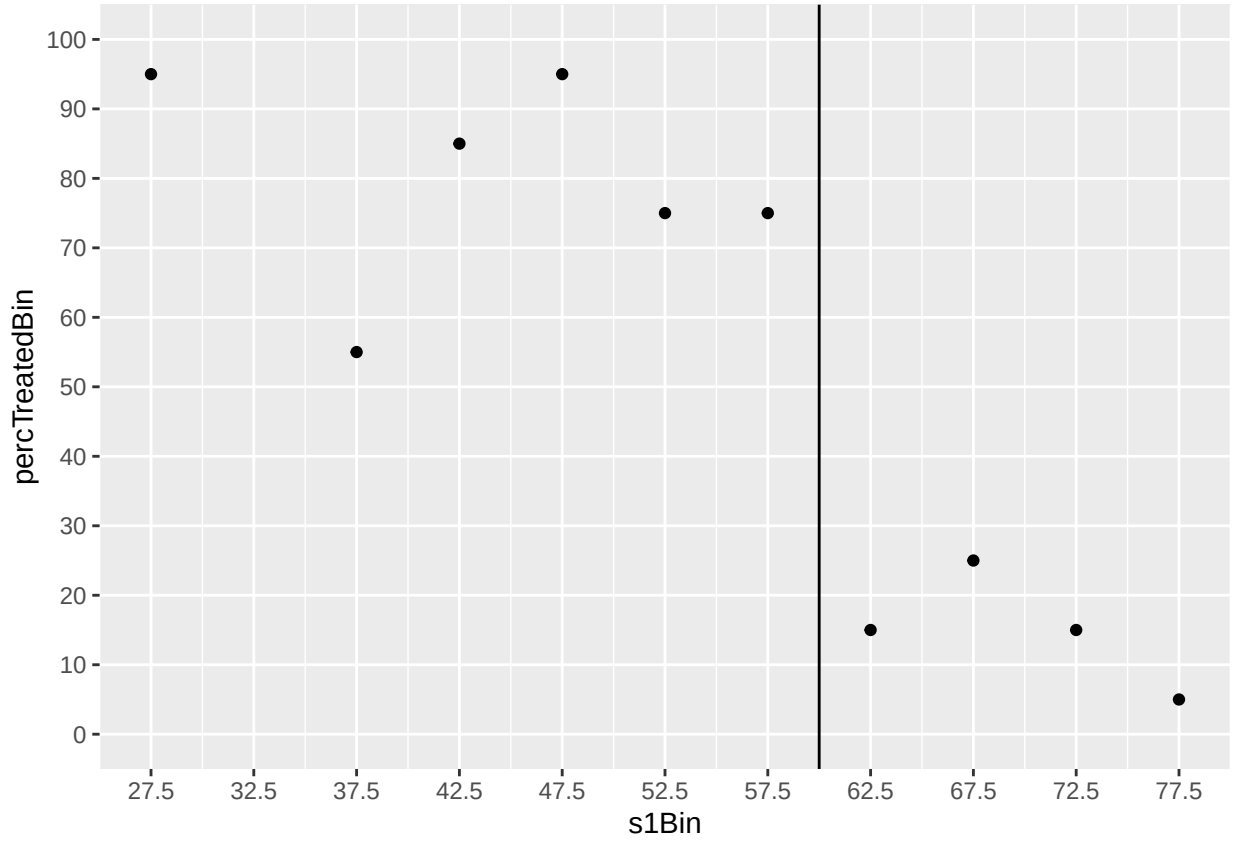
29

Bin the data in 5-point increments, letting each bin be represented by its midpoint. For each bin, calculate the percent that were treated. Plot this against the bin midpoint. Plot a vertical line at the cut point. Interpret the figure.

```
cutPoint <- 60
bins <- seq(2.5,100,5)

# tut$s1Bin <- cut(tut$s1, breaks = seq(0,100,5), labels = bins)
# tut$s2Bin <- cut(tut$s2, breaks = seq(0,100,5), labels = bins)
#
# q29 <- tut %>%
#   group_by(s1Bin) %>%
#   count(t) %>%
#   mutate(percTreatedBin = round(n / sum(n) * 100, 2)) %>%
#   filter(t == 1) %>%
#   select(-t)

q29 <- tut[, c("s1Bin",
               "s2Bin") := .(cut(tut$s1, breaks = seq(0,100,5), labels = bins),
                             cut(tut$s2, breaks = seq(0,100,5), labels = bins))][
            order(s1Bin, t), .N, by = .(s1Bin, t)][
            , percTreatedBin := round(N / sum(N) * 100, 2), by = s1Bin][
            t == 1, !c("t")]

q29$s1Bin <- as.numeric(levels(q29$s1Bin)[q29$s1Bin])
head(q29)
```

```
##    s1Bin  N percTreatedBin
## 1:  27.5  1         100.00
## 2:  37.5  3          60.00
## 3:  42.5  5          83.33
## 4:  47.5 15          93.75
## 5:  52.5 26          76.47
## 6:  57.5 48          71.64
```

```
q29 %>%
  ggplot() +
  geom_point(aes(x = s1Bin, y = percTreatedBin)) +
  geom_vline(xintercept = cutPoint, color = "black") +
  scale_x_continuous(breaks = bins) +
  scale_y_binned(n.breaks = 10, limits = c(0,100))
```

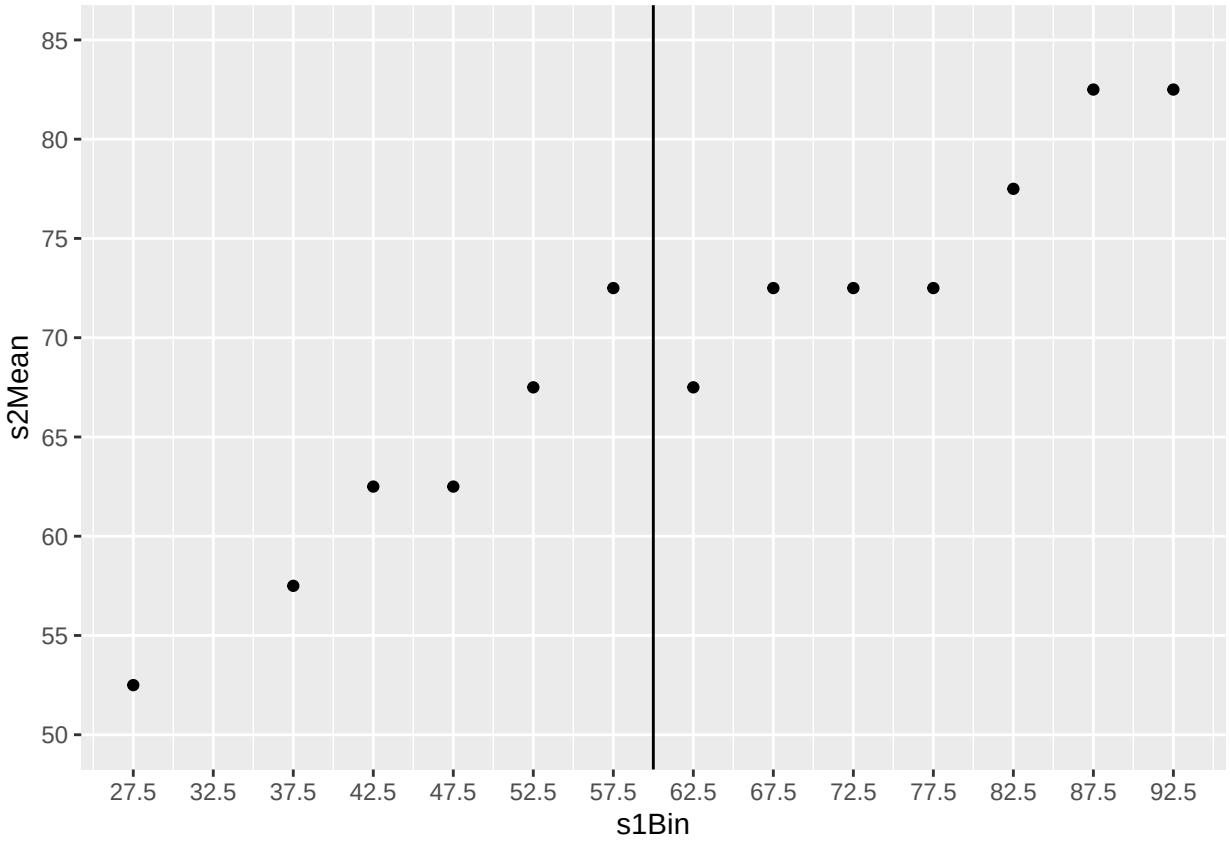


30

For each bin from the previous question, calculate the average score on the second exam. Plot this against the bin midpoint. Plot a vertical line at the cut point. Interpret the figure.

```
# q30 <- tut %>%
#   group_by(s1Bin) %>%
#   summarise(s2Mean = round(mean(s2), 2))
q30 <- tut[, .(s2Mean = round(mean(s2)), 2), by = "s1Bin"][order(s1Bin), -c("V2")]
```

```
q30$s1Bin <- as.numeric(levels(q30$s1Bin)[(q30$s1Bin)])

q30 %>%
  ggplot() +
  geom_point(aes(x = s1Bin, y = s2Mean)) +
  geom_vline(xintercept = cutPoint) +
  scale_x_continuous(breaks = bins) +
  scale_y_binned(n.breaks = 10, limits = c(50, 85))
```



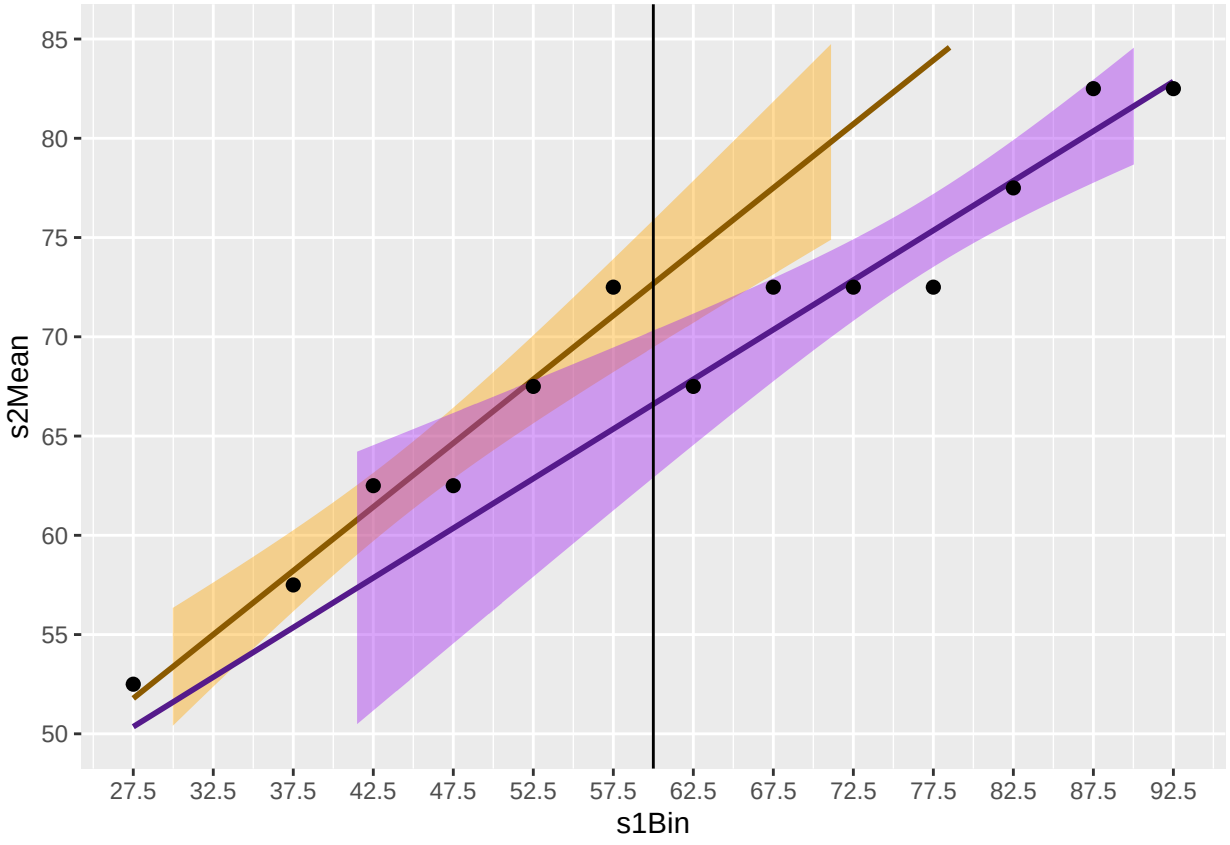
31

Calculate the local effect of intention to treat. You will need to do a regression to accomplish this.

```
# ggplot() +
#   stat_smooth(data = q30[q30$s1Bin <= cutPoint,], aes(x = s1Bin, y = s2Mean),
#               method = "lm", formula = "y~x", color = "orange4", fill = "orange", fullrange = TRUE) +
#   stat_smooth(data = q30[q30$s1Bin > cutPoint,], aes(x = s1Bin, y = s2Mean),
#               method = "lm", formula = "y~x", color = "purple4", fill = "purple", fullrange = TRUE) +
#   geom_point(data = q30, aes(x = s1Bin, y = s2Mean), size = 2) +
#   geom_vline(xintercept = cutPoint) +
#   scale_x_continuous(breaks = bins) +
#   scale_y_binned(n.breaks = 10, limits = c(50, 85))
#
```

```
# lm(formula = s2Mean ~ s1Bin, data = q30[q30$s1Bin <= cutPoint])
# lm(formula = s2Mean ~ s1Bin, data = q30[q30$s1Bin > cutPoint])

ggplot() +
  stat_smooth(data = q30[s1Bin <= cutPoint], aes(x = s1Bin, y = s2Mean),
              method = "lm", formula = "y~x", color = "orange4", fill = "orange", fullrange = TRUE) +
  stat_smooth(data = q30[s1Bin > cutPoint,], aes(x = s1Bin, y = s2Mean),
              method = "lm", formula = "y~x", color = "purple4", fill = "purple", fullrange = TRUE) +
  geom_point(data = q30, aes(x = s1Bin, y = s2Mean), size = 2) +
  geom_vline(xintercept = cutPoint) +
  scale_x_continuous(breaks = bins) +
  scale_y_binned(n.breaks = 10, limits = c(50, 85))
```

```
## Warning: Removed 17 rows containing missing values (geom_smooth).
```




```
lm(formula = s2Mean ~ s1Bin, data = q30[s1Bin <= cutPoint])
```

```
##
## Call:
## lm(formula = s2Mean ~ s1Bin, data = q30[s1Bin <= cutPoint])
##
## Coefficients:
## (Intercept)        s1Bin
##       38.05         0.58
```

```
lm(formula = s2Mean ~ s1Bin, data = q30[s1Bin > cutPoint])
```

```
##
## Call:
## lm(formula = s2Mean ~ s1Bin, data = q30[s1Bin > cutPoint])
##
## Coefficients:
## (Intercept)        s1Bin
##     30.7143       0.5714
```

32

Calculate an estimate of the LATE. You will need to use an instrumental variables estimator.