Final Exam

Fan Bu

11/23/2020

```
library(tidyverse)
library(ggplot2)
library(broom)
```

Question 1

(a)

Suppose we want to study the effect of sunlight exposure to bone health among middle-aged women.

Outcome: whether or not a woman receives a diagnosis of osteoporosis by age 50.

Treatment: natural sunlight exposure of at least 30 minutes per day.

Confounders: place of living, lifestyle, general health habits, etc.

Conditions that would violate SUTVA: People who live in different regions/countries and/or go out at different times during the day or different seasons of the year may get very different levels of ultra-violet light exposure for 30 minutes of sunlight. Therefore, "30 minutes of sunlight exposure" can indicate a wide range of actual treatments received that vary by the intensity of sunlight.

(b)

Suppose we want to study the effect of a campaign ad launched on social media (e.g., Facebook) on voter turnout.

Outcome: whether or not someone votes on the Election Day.

Treatment: exposure to the campaign ad (e.g., the ad appears on the user's Facebook homepage as a "sponsored" post)

Confounders: the user's frequency of using Facebook, personal characteristics, political inclinations, etc.

Conditions that would violate SUTVA: There may be a spillover effect between users who are friends. A person who doesn't get selected for the campaign may still get exposed to it if his friends re-post or "like" the campaign ad, or talk to him/her about it offline.

Question 2

(b) and **(d)** are correct.

(a) is wrong because it indicates an individual treatment effect, but ATE measures the average effect for the study population instead.

(c) is wrong because it mistakes ATE as a 2-group difference estimand (and doesn't assume randomization), whereas ATE is a comparison between counterfactuals.

Question 3

(a)

Unconfoundedness:

$$(Y_i(0), Y_i(1)) \perp Z_i \mid X_i, \text{ for each unit } i.$$

In other words, we need to verify that $Pr(Z \mid Y(0), Y(1), X) = Pr(Z \mid X)$.

For $X = 1$, we have

$$\begin{aligned}
&Pr(Z = 1 \mid X = 1) = (1/6 + 1/9 + 1/18)/(1/2) = 2/3;\\
&Pr(Z = 1 \mid Y(1) = 1, Y(0) = 1, X = 1) = (1/6)/(1/6 + 1/12) = 2/3,\\
&Pr(Z = 1 \mid Y(1) = 1, Y(0) = 0, X = 1) = (1/9)/(1/9 + 1/18) = 2/3,\\
&Pr(Z = 1 \mid Y(1) = 0, Y(0) = 0, X = 1) = (1/18)/(1/18 + 1/36) = 2/3.
\end{aligned}$$

That is, $Pr(Z \mid Y(0), Y(1), X = 1) = Pr(Z \mid X = 1) = 2/3$.

And for $X = 0$, we have

$$\begin{aligned}
&Pr(Z = 1 \mid X = 0) = (1/36 + 1/18 + 1/12)/(1/2) = 1/3;\\
&Pr(Z = 1 \mid Y(1) = 1, Y(0) = 1, X = 0) = (1/36)/(1/36 + 1/18) = 1/3,\\
&Pr(Z = 1 \mid Y(1) = 1, Y(0) = 0, X = 0) = (1/18)/(1/18 + 1/9) = 1/3,\\
&Pr(Z = 1 \mid Y(1) = 0, Y(0) = 0, X = 1) = (1/12)/(1/12 + 1/6) = 1/3.
\end{aligned}$$

That is, $Pr(Z \mid Y(0), Y(1), X = 0) = Pr(Z \mid X = 0) = 1/3$.

Therefore, we indeed have $Pr(Z \mid Y(0), Y(1), X) = Pr(Z \mid X)$, and the unconfoundedness asumption holds.

(b)

We only need to show that $Pr(Z \mid Y(0), Y(1)) \neq Pr(Z)$.

Since $Pr(Z = 1) = 1/6 + 1/9 + 1/18 + 1/36 + 1/18 + 1/12 = 1/2$, but

$$Pr(Z = 1 \mid Y(1) = 1, Y(0) = 1) = \frac{1/6 + 1/36}{1/6 + 1/12 + 1/36 + 1/18} = 7/12,$$

we can easily see that these two quantities are not equal, and thus marginal unconfoundedness doesn't hold.

(c)

We need (1) unconfoundedness (this already holds), and (2) SUTVA.

The joint distribution of (Y, X, Z) is summarized in the table below.

Y	X	Z	probability
1	1	1	5/18
0	1	1	1/18
1	0	1	1/12
0	0	1	1/12
1	1	0	1/12
0	1	0	1/12
1	0	0	1/18
0	0	0	5/18

(d)

$$\mathbb{E}(Y \mid Z = 1) = (5/18 + 1/12)/(1/2) = 13/18,$$

while

$$\mathbb{E}(Y(1)) = 1 - (1/18 + 1/36 + 1/12 + 1/6) = 2/3.$$

So $\mathbb{E}(Y \mid Z = 1) \neq \mathbb{E}(Y(1))$.

Moreover,

$$\mathbb{E}(Y \mid Z = 0) = (1/12 + 1/18)/(1/2) = 5/18,$$

while

$$\mathbb{E}(Y(0)) = 1/18 + 1/36 + 1/12 + 1/6 = 1/3.$$

So obviously $\mathbb{E}(Y \mid Z = 0) \neq \mathbb{E}(Y(0))$.

This is because the treated group and control group are not really comparable (the two sub-populations are not the same).

(e)

Since

$$\begin{aligned}
&\mathbb{E}_X\{\mathbb{E}(Y \mid Z = 1, X)\} \\
=&\mathbb{E}(Y \mid Z = 1, X = 1)Pr(X = 1) + \mathbb{E}(Y \mid Z = 1, X = 0)Pr(X = 0) \\
=&5/6 \times 1/2 + 1/2 \times 1/2 \\
=&2/3,
\end{aligned}$$

we do have $\mathbb{E}(Y(1)) = \mathbb{E}_X\{\mathbb{E}(Y \mid Z = 1, X)\}$.

Moreover, since

$$\begin{aligned}
&\mathbb{E}_X\{\mathbb{E}(Y \mid Z = 0, X)\} \\
=&\mathbb{E}(Y \mid Z = 0, X = 1)Pr(X = 1) + \mathbb{E}(Y \mid Z = 0, X = 0)Pr(X = 0) \\
=&1/2 \times 1/2 + 1/6 \times 1/2 \\
=&1/3,
\end{aligned}$$

we also have $\mathbb{E}(Y(0)) = \mathbb{E}_X\{\mathbb{E}(Y \mid Z = 0, X)\}$.

For $z = 1$,

$$\begin{aligned}
&\mathbb{E}[I(Z = 1)Y/P(Z = 1 \mid X)] \\
=&\mathbb{E}[I(Z = 1)Y/P(Z = 1 \mid X) \mid X = 1]Pr(X = 1) \\
&+ \mathbb{E}[I(Z = 1)Y/P(Z = 1 \mid X) \mid X = 0]Pr(X = 0) \\
=&\mathbb{E}[I(Z = 1)Y \mid X = 1]/(2/3) \times 1/2 + \mathbb{E}[I(Z = 1)Y \mid X = 0]/(1/3) \times 1/2 \\
=&\frac{5/9}{2/3} \times 1/2 + \frac{1/6}{1/3} \times 1/2 \\
=&\frac{15}{36} + \frac{1}{4} \\
=&2/3 \\
=&\mathbb{E}(Y(1)).
\end{aligned}$$

And for $z = 0$

$$\begin{aligned}
&\mathbb{E}[I(Z = 0)Y/P(Z = 0 \mid X)] \\
=&\mathbb{E}[I(Z = 0)Y/P(Z = 0 \mid X) \mid X = 1]Pr(X = 1) \\
&+ \mathbb{E}[I(Z = 0)Y/P(Z = 0 \mid X) \mid X = 0]Pr(X = 0) \\
=&\mathbb{E}[I(Z = 0)Y \mid X = 1]/(1/3) \times 1/2 + \mathbb{E}[I(Z = 0)Y \mid X = 0]/(2/3) \times 1/2 \\
=&\frac{1/6}{1/3} \times 1/2 + \frac{1/9}{2/3} \times 1/2 \\
=&\frac{1}{4} + \frac{1}{12} \\
=&1/3 \\
=&\mathbb{E}(Y(0)).
\end{aligned}$$

(f)

Causal risk difference: $\mathbb{E}(Y(1)) - \mathbb{E}(Y(0)) = 1/3$.

Causal risk ratio: $\frac{\mathbb{E}(Y(1))}{\mathbb{E}(Y(0))} = 2$.

Causal odds ratio: $\frac{Pr(Y(1)=1)/(1-Pr(Y(1)=1))}{Pr(Y(0)=1)/(1-Pr(Y(0)=1))} = 4$.

I think all these estimands represent the causal effect of the treatment for the target population, so there shouldn't be a fundamental difference between them.

That being said, causal risk ratio and odds ratio can be better choices when we want to measure the causal effect in a **relative** sense for a binary outcome. For example, the absolute difference between 90% and 80% risks may be the same as the difference between 20% and 10%, but the relative difference between the latter risks may be more significant in some application scenarios, and using causal risk ratio and/or odds ratio can reflect that.

To some extent, the "advantage"/"disadvantage" of an estimand depends on our interpretation needs, i.e., how we want to understand and explain the causal effects.

(g)

$$\mathbb{E}(Y \mid Z = 1, X = 1) = 5/6;$$
$$\mathbb{E}(Y \mid Z = 0, X = 1) = 1/2;$$
$$\mathbb{E}(Y \mid Z = 1, X = 0) = 1/2;$$
$$\mathbb{E}(Y \mid Z = 0, X = 0) = 1/6.$$

For $X = 1$:

- conditional causal risk difference: $5/6 - 1/2 = 1/3$.
- conditional causal risk ratio: $\frac{5/6}{1/2} = 5/3$.
- conditional causal odds ratio: $\frac{(5/6)/(1/6)}{(1/2)/(1/2)} = 5$.

For $X = 0$:

- conditional causal risk difference: $1/2 - 1/6 = 1/3$.
- conditional causal risk ratio: $\frac{1/2}{1/6} = 3$.
- conditional causal odds ratio: $\frac{(1/2)/(1/2)}{(1/6)/(5/6)} = 5$.

There is an interaction between Z and X since the (conditional) causal effect seems to differ between the $X = 0$ stratum and the $X = 1$ stratum.

Question 4

(a)

Define the principal strata S_i for person i based on the potential outcomes of intermediate variables D_i and W_i:

$$S_i = (D_i(0), D_i(1), W_i(0), W_i(1)).$$

Then all partipants can be classified into 16 different groups based on the 16 different binary combinations for the entries in S_i.

(b)

According to the 3 facts:

(i) –> $W_i(z) = 0$ if $D_i(z) = 0$, or $Pr(W_i(z) = 1 \mid D_i(z) = 0) = 0$ for $z = 0, 1$. (Treatment receipt is "one-sided"; only those diagnosed can receive treatment.)

(ii) –> $Pr(W_i(z) = 1 \mid D_i(z) = 1) < 1$, or $Pr(W_i(z) = 0 \mid D_i(z) = 1) > 0$ for $z = 0, 1$. (Similar to the positivity assumption.)

(iii) –> $W_i(0) \leq W_i(1)$, and there exists i with $W_i(0) = 0$ and $W_i(1) = 1$ (Pr(W_i(0)=0, W_i(1)=1) > 0). ((Strict) Monotonicity assumption.)

We can reduce the total number of possible strata to 7 given these assumptions. These principal strata are summarized and explained in the table below.

$D_i(0)$	$D_i(1)$	$W_i(0)$	$W_i(1)$	Meaning in our context
0	0	0	0	those who wouldn't get diagnosed or treated in either arm
0	1	0	0	those who wouldn't get diagnosed in control arm but would get diagnosed in treated arm, and wouldn't get treated in either arm
0	1	0	1	those who wouldn't get diagnosed or treated in control arm, but would get diagnosed and treated in treated arm
1	0	0	0	those who would get diagnosed in control arm but wouldn't get diagnosed in treated arm, and wouldn't get treated in either arm
1	1	0	0	those who would get diagnosed in both arms, but wouldn't get treated in either arm
1	1	0	1	those who would get diagnosed in both arms, and wouldn't get treated in control arm but would be treated in treated arm
1	1	1	1	those who would get diagnosed and treated in both arms

Question 5

```
# load the data
BHdat = read_csv("BHdata.csv")
```

(a)

The DiD estimate is

```
Diffs = BHdat %>% group_by(G) %>%
  summarise(mean_diff = mean(Yafter) - mean(Ybefore))
Diffs$mean_diff[Diffs$G == 1] - Diffs$mean_diff[Diffs$G == 0]
```

```
## [1] 7.144014
```

Here the variance is estimated through the robust variance formula (**not** assuming homoscedasticity):

$$\mathbb{V}(\hat{\tau}_{DID}) = \frac{\hat{\sigma}^2_{00}}{N_{00}} + \frac{\hat{\sigma}^2_{01}}{N_{01}} + \frac{\sigma^2_{10}}{N_{10}} + \frac{\sigma^2_{11}}{N_{11}}.$$

Thus the estimated standard error (square root of variance) is

```
BHdat %>% group_by(G) %>%
  summarise(N = n(), B = var(Ybefore), A = var(Yafter)) %>%
  mutate(B = B/N, A = A/N) %>%
  select(B, A) %>%
  sum() %>%
  sqrt()
```

```
## [1] 1.72659
```

Therefore the DiD estimate of the causal effect is `7.144`, with a standard error of `1.727`.

The necessary causal assumptions include:

1. Randomization. That is, there is no confounder.
2. Parallel trends. That is, the treatment and control groups would experience the same trend without the treatment.
3. SUTVA.

(b)

Fit a linear model of Y_{t+1} on Y_t and G and check out the coefficient estimates.

```
LDV.mod = lm(Yafter ~ Ybefore + as.factor(G), data=BHdat)
#summary(LDV.mod)
tidy(LDV.mod)
```

```
## # A tibble: 3 x 5
##   term          estimate std.error statistic   p.value
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     -2.77     1.17       -2.37 1.86e-  2
## 2 Ybefore          0.997    0.0276     36.2  9.82e-111
## 3 as.factor(G)1    7.12     0.749       9.51 6.84e- 19
```

We can see that the estimated causal effect is (the coefficient for `G`) `7.121`, with a standard error of `0.749`.

Necessary assumptions include:

1. Ignorability conditioned on the LDV; that is,

$$Y_{i,t+1}(0) \perp G \mid Y_{i,t}, \quad \text{for each unit } i.$$

2. SUTVA

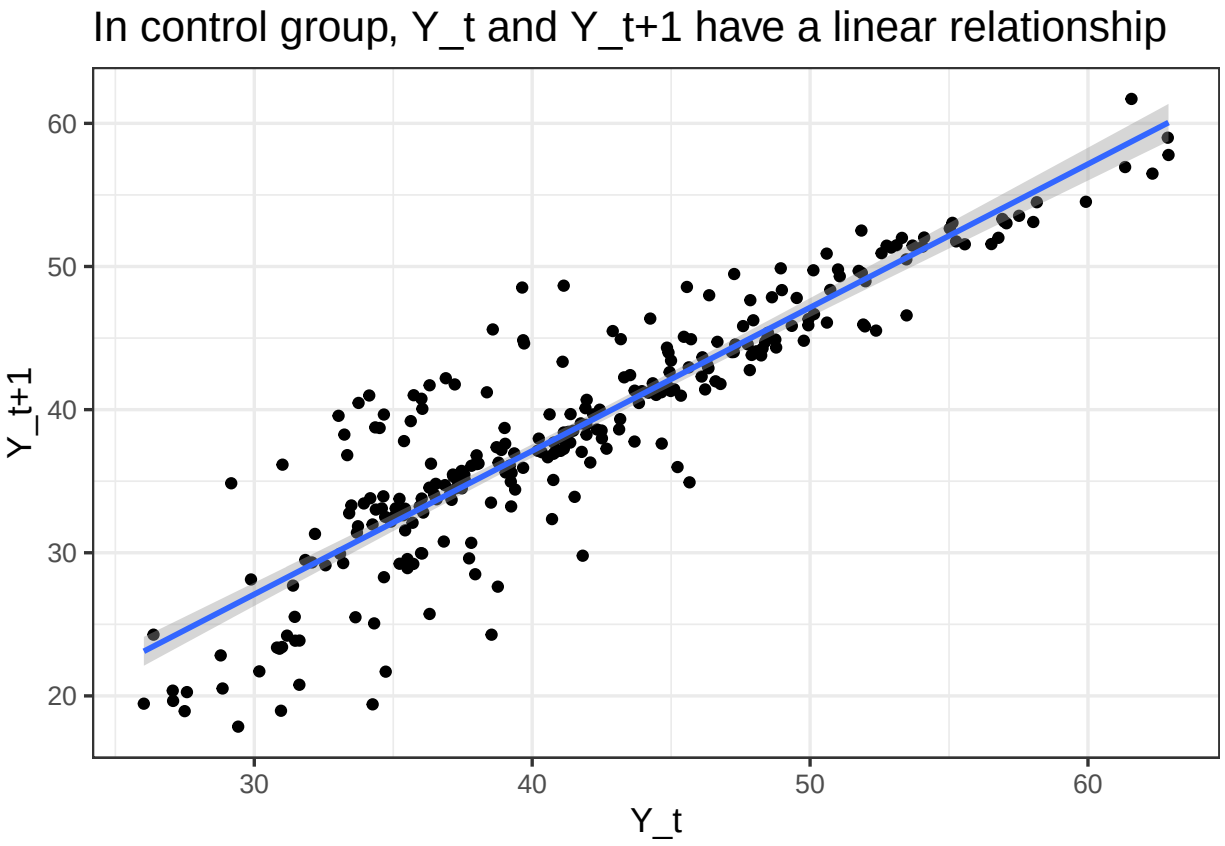
(c)

The results from these two methods are similar.

DID and LDV estimates would be identical if the linear coefficient β for Y_t in the LDV model is 1. And in fact, in the estimated linear model in (b), $\hat{\beta} = 0.997$, which is very close to 1. So it's not surprising at all that these two methods above produce very similar results in this case.

We can also see from the graph below (Y_{t+1} vs Y_t in the control group) that Y_{t+1} and Y_t clearly have a linear relationship (and the slope is also close to 1).

```
ggplot(data=(BHdat %>% filter(G==0)), aes(x=Ybefore, y=Yafter)) +
  geom_point() +
  geom_smooth(method='lm') +
  labs(x='Y_t', y='Y_t+1',
title = 'In control group, Y_t and Y_t+1 have a linear relationship') +
  theme_bw(base_size = 13)
```



Question 6

```
FDR = read_csv('FRD.csv')
```

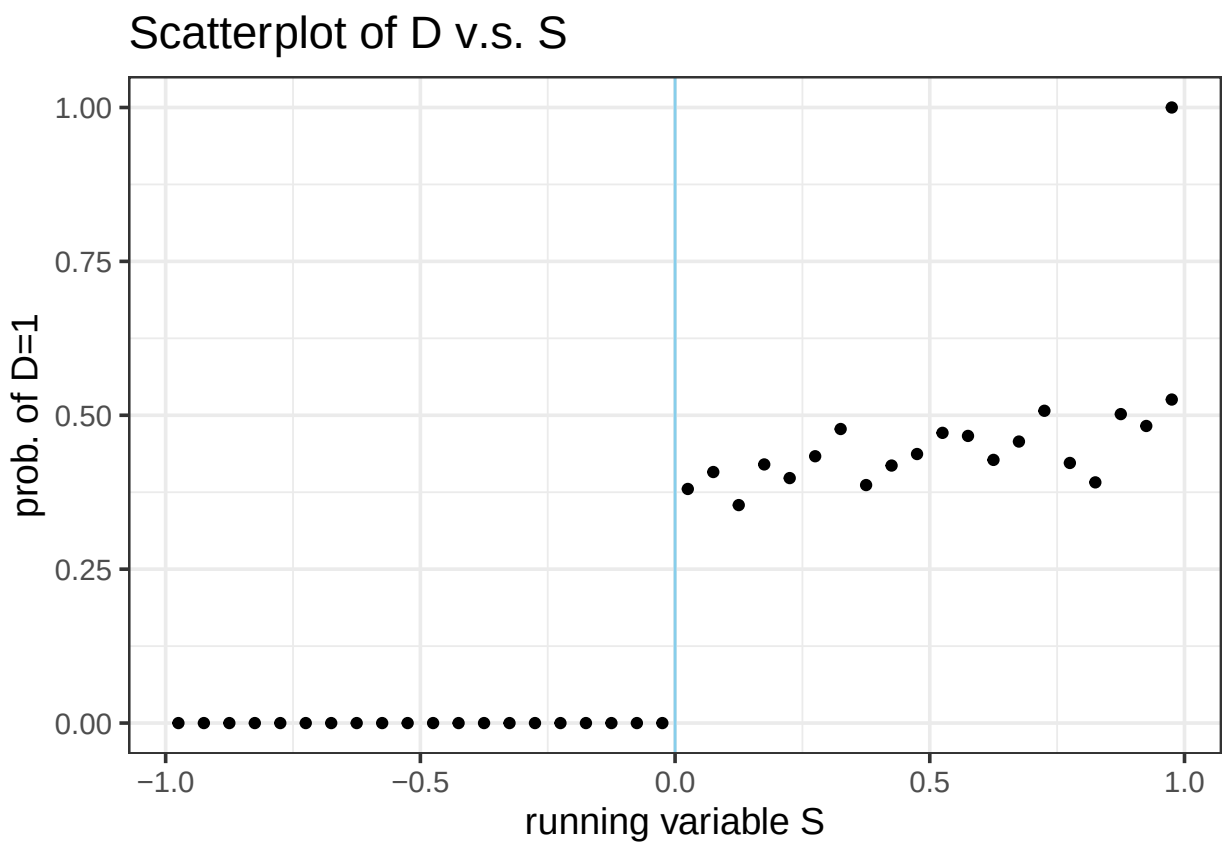
(a)

Following the hint, divide the range of `rv0` into 0.05-wide bins and calcuate the average of `D` (and `outcome`) in each bin. And then draw scatterplots.

```
# width of bin
w = 0.05
# map values into bins
FDR_new = FDR %>% mutate(S_bin = rv0 %/% w) %>%
  mutate(S = ifelse(S_bin * w + w/2 < 1, S_bin * w + w/2, 1 - w/2)) %>%
  group_by(S) %>%
  summarise(Y_bin = mean(outcome), D_bin = mean(D))
```

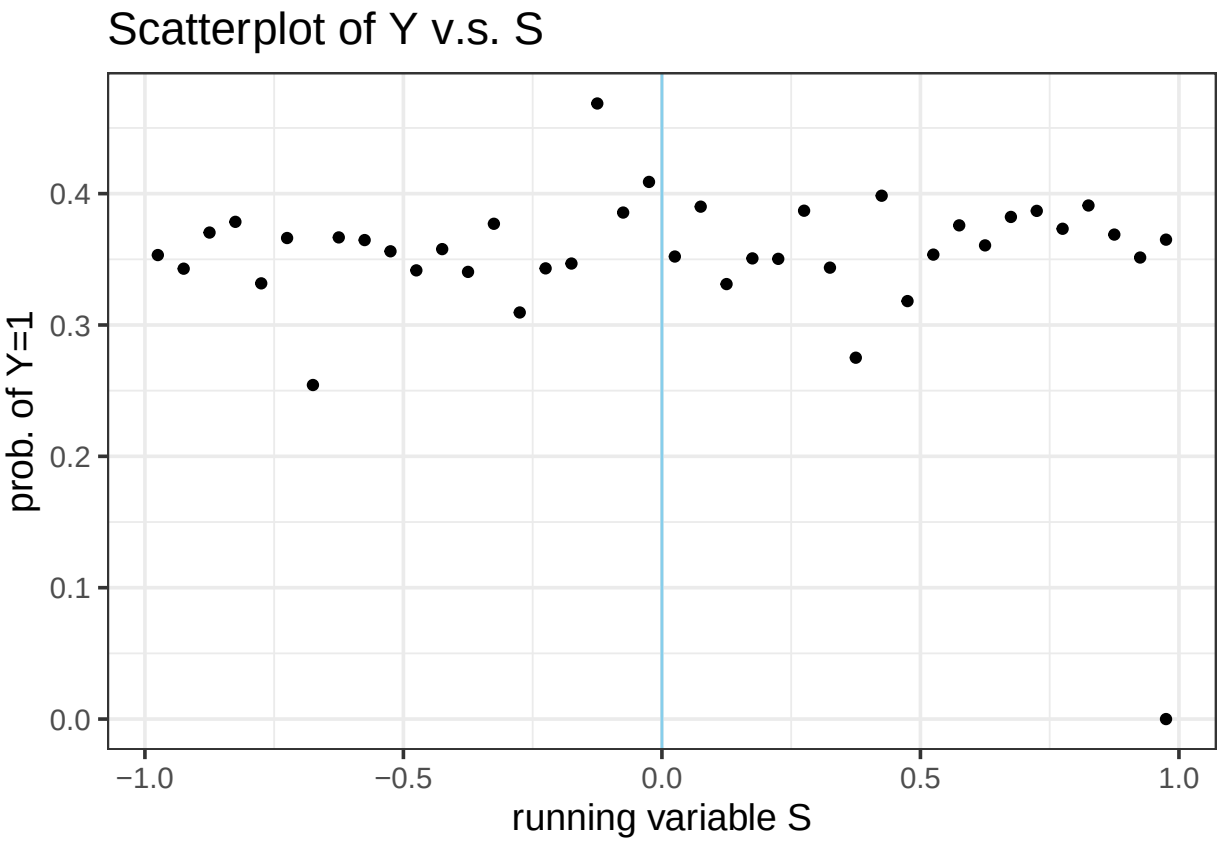
The scatterplot of S v.s. D:

```
ggplot(data = FDR_new) +
  geom_point(aes(x=S, y=D_bin)) +
  geom_vline(xintercept = 0, color='skyblue') +
  labs(x='running variable S', y='prob. of D=1',
       title = 'Scatterplot of D v.s. S') +
  theme_bw(base_size = 14)
```



The scatterplot of S v.s. Y:

```
ggplot(data = FDR_new) +
  geom_point(aes(x=S, y=Y_bin)) +
  geom_vline(xintercept = 0, color='skyblue') +
  labs(x='running variable S', y='prob. of Y=1',
       title = 'Scatterplot of Y v.s. S') +
  theme_bw(base_size = 14)
```



We can see that:

1. A unit would only be treated if he is encouraged; i.e., $Pr(D_i = 1 \mid Z_i = 0) = 0$.
2. Not all units who are encouraged would be actually treated; i.e., $Pr(D_i = 0 \mid Z_i = 1) > 0$.
3. Around the threshold $S = 0$ (within a very narrow local neighborhood), $Pr(Y = 1)$ seems to decrease as S moves from the left side to the right side of threshold 0.

(b)

Create variables `L` and `R` first for all i.

```
FDR = FDR %>%
  mutate(L = ifelse(Z==0, rv0, 0), R = ifelse(Z==1, rv0, 0))
```

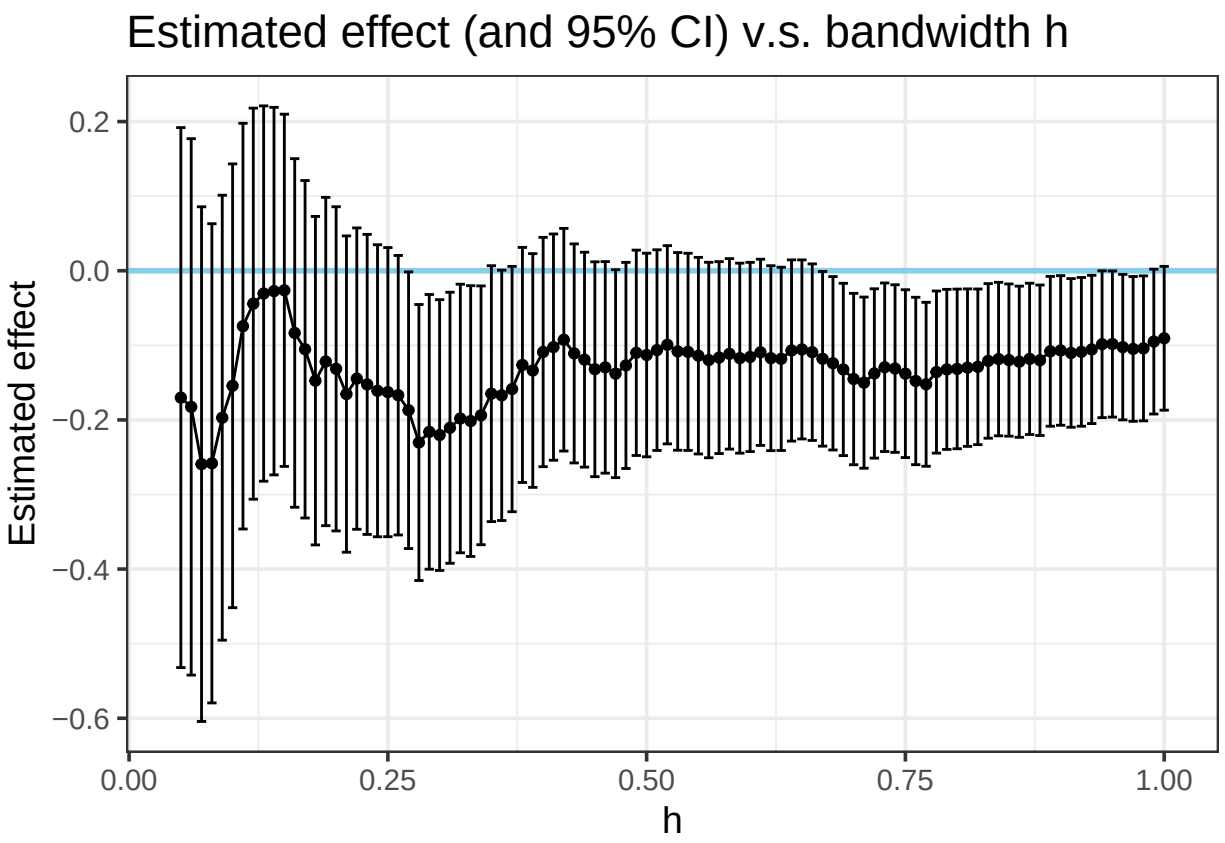
Then for every `h in seq(0.05, 1, 0.01)`, do steps 2-3.

```
Hs = seq(from = 0.05, to = 1, by = 0.01)
res = NULL
for(h in Hs){
  FDR_sub = FDR %>% filter(rv0 < h, rv0 > -h)
  # 1st LS
  D.mod = lm(D ~ Z + L + R, data=FDR_sub)
  FDR_sub$Dfit = predict(D.mod)
  # 2nd LS
  Y.mod = lm(outcome ~ Dfit + L + R, data=FDR_sub)
  # get estimate and 95% CI
  res = rbind(res, c(coef(Y.mod)['Dfit'], confint(Y.mod)['Dfit',]))
}
res.dat = as.data.frame(res)
```

```
names(res.dat) = c('estimate', 'lb', 'ub')
res.dat$h = Hs
```

Produce the plot of `h` v.s. estimates (and confidence bands).

```
ggplot(data=res.dat, aes(x=h, y=estimate)) +
  geom_hline(yintercept = 0, color='skyblue', size=1) +
  geom_point()+
  geom_line()+
  geom_errorbar(aes(min = lb, max=ub))+
  labs(title='Estimated effect (and 95% CI) v.s. bandwidth h',
       y = 'Estimated effect') +
  theme_bw(base_size = 14)
```



From the plot above, we can see that

- when we increase h, variance gets reduced (since the confidence bands get narrower),
- when h is small, the estimates are not very stable (another evidence of high variance with small bandwidth), but the estimates stabilize as h gets larger,
- overall, with all bandwidth h between 0.05 and 1, the estimate is negative; this suggests that the FDR estimand is very likely negative, and this conclusion is robust to the choice of bandwidth h.

(c)

Note that L_i is non-zero only if $S_i \leq 0$ and R_i is non-zero only if $S_i > 0$. Also, $Z_i = 0$ if $S_i \leq 0$ and $Z_i = 1$ if $S_i > 0$.

Then the fitted model from Step 2 is

$$\begin{aligned}\hat{D}_i =& \hat{\alpha}_0 + \hat{\alpha}_Z Z_i + \hat{\alpha}_L L_i + \hat{\alpha}_R R_i \\ =& \begin{cases} \hat{\alpha}_0 + \hat{\alpha}_L L_i & \text{if } S_i \leq 0 \\ (\hat{\alpha}_0 + \hat{\alpha}_Z) + \hat{\alpha}_R R_i & \text{if } S_i > 0 \end{cases} \\ =& \begin{cases} \hat{\alpha}_0 + \hat{\alpha}_L S_i & \text{if } S_i \leq 0 \\ (\hat{\alpha}_0 + \hat{\alpha}_Z) + \hat{\alpha}_R S_i & \text{if } S_i > 0 \end{cases},\end{aligned}$$

which is equivalent to fitting a linear regression model separately for units on each side of the threshold.

Similarly, the fitted model from Step 3 is

$$\begin{aligned}\hat{Y}_i =& \hat{\beta}_0 + \hat{\beta}_D \hat{D}_i + \hat{\beta}_L L_i + \hat{\beta}_R R_i \\ =& \begin{cases} \hat{\beta}_0 + \hat{\beta}_D \hat{D}_i + \hat{\beta}_L L_i & \text{if } S_i \leq 0 \\ \hat{\beta}_0 + \hat{\beta}_D \hat{D}_i + \hat{\beta}_R R_i & \text{if } S_i > 0 \end{cases} \\ =& \begin{cases} \hat{\beta}_0 + \hat{\beta}_D \hat{D}_i + \hat{\beta}_L S_i & \text{if } S_i \leq 0 \\ \hat{\beta}_0 + \hat{\beta}_D \hat{D}_i + \hat{\beta}_R S_i & \text{if } S_i > 0 \end{cases} \\ =& \begin{cases} \hat{\beta}_0 + \hat{\beta}_D(\hat{\alpha}_0 + \hat{\alpha}_L S_i) + \hat{\beta}_L S_i & \text{if } S_i \leq 0 \\ \hat{\beta}_0 + \hat{\beta}_D((\hat{\alpha}_0 + \hat{\alpha}_Z) + \hat{\alpha}_R S_i) + \hat{\beta}_R S_i & \text{if } S_i > 0 \end{cases} \\ =& \begin{cases} (\hat{\beta}_0 + \hat{\beta}_D \hat{\alpha}_0) + (\hat{\beta}_D \hat{\alpha}_L + \hat{\beta}_L) S_i & \text{if } S_i \leq 0 \\ [\hat{\beta}_0 + \hat{\beta}_D(\hat{\alpha}_0 + \hat{\alpha}_Z)] + (\hat{\beta}_D \hat{\alpha}_R + \hat{\beta}_R) S_i & \text{if } S_i > 0 \end{cases},\end{aligned}$$

which is also equivalent to fitting a separate linear regression model for each side of the threshold.

(d)

Here the "encouragement" Z is an IV: it directly affects D but has no direct effect on outcome Y. The estimation procedure in (b) is essentially the procedure to obtain the 2-stage-least-squares (2SLS) estimator.

Question 7

The lecture(s) on instrument variables (as well as noncompliance and principal stratification) are great! I never thought of econometric approaches (like 2SLS) from a causal inference perspective, so it was quite enlightening to see some of the connections there.

(This may sound rebellious but...) It would be great to have a survey lecture on Pearl's causal framework, particularly since we are a Bayesian department and people do have some basic understanding of graphical models; just a brief introduction like the one on machine learning methods would do - it seems that these days a lot of CS people are using Pearl's framework.