Problem Set 2

Program Evaluation and Causal Inference

Christian Birchler Fenqi Guo Mingrui Zhang Wenjie Tu Zunhan Zhang

Spring Semester 2021

Names are listed in alphabetical order

1. Causal parameters and the selection problem

1(a)

The average treatment effect **ATE** is the expected impact of treatment on the entire population:

$$\begin{aligned}\Delta^{ATE} &= \mathbb{E}(Y_{1i}) - \mathbb{E}(Y_{0i}) \\ &= \mathbb{E}(Y_{1i} - Y_{0i})\end{aligned}$$

- Y_{1i} is the potential outcome of the entire population that gets treated.
- Y_{0i} is the potential outcome of the entire population that gets untreated.
- Only one of the potential outcomes is observable.

The average treatment effect on the treated **ATT** is the expected impact of treatment on those who want to get treated.

$$\begin{aligned}\Delta^{ATT} &= \mathbb{E}(Y_{1i}|D=1) - \mathbb{E}(Y_{0i}|D=1) \\ &= \mathbb{E}(Y_{1i} - Y_{0i}|D=1)\end{aligned}$$

Y_{1i} is observable while Y_{0i} is unobservable.

The average treatment effect on the untreated **ATUT** is the expected impact of treatment on those who do not want to get treated.

$$\begin{aligned}\Delta^{ATUT} &= \mathbb{E}(Y_{1i}|D=0) - \mathbb{E}(Y_{0i}|D=0) \\ &= \mathbb{E}(Y_{1i} - Y_{0i}|D=0)\end{aligned}$$

$Y_{0i}|D=0$ is observable while $Y_{1i}|D=0$ is unobservable.

The naive estimate Δ^N is the difference between the expected value of the observed outcome with treatment and the expected value of the observed outcome without treatment.

$$\Delta^N = \mathbb{E}(Y_{1i}|D=1) - \mathbb{E}(Y_{0i}|D=0)$$

Both $Y_{1i}|D=1$ and $Y_{0i}|D=0$ are observable

1(b)

$$\begin{aligned}
\Delta^{ATE} &= \mathbb{E}(Y_{1i}) - E(Y_{0i}) \\
&= \mathbb{E}(Y_{1i} - Y_{0i}) \\
&= \mathbb{E}_d[E(Y_{1i} - Y_{0i}|D = d)] \\
&= Pr(D = 1) \times \underbrace{\mathbb{E}(Y_{1i} - Y_{0i}|D = 1)}_{\Delta^{ATT}} + (1 - Pr(D = 1)) \times \underbrace{\mathbb{E}(Y_{1i} - Y_{0i}|D = 0)}_{\Delta^{ATUT}} \\
&= Pr(D = 1) \times \Delta^{ATT} + (1 - Pr(D = 1)) \times \Delta^{ATUT}
\end{aligned}$$

Why identification fails?

- $Y_{1i}|D = 1$ and $Y_{0i}|D = 0$ are observable.
- $Y_{1i}|D = 0$ and $Y_{0i}|D = 1$ are unobservable.

Assumptions to identify ATE

- Independence Assumption: the potential outcomes are independent of treatment

$$\mathbb{E}(Y_{1i}|D = 1) = \mathbb{E}(Y_{1i}|D = 0) = \mathbb{E}(Y_{1i})$$

$$\mathbb{E}(Y_{0i}|D = 0) = \mathbb{E}(Y_{0i}|D = 1) = \mathbb{E}(Y_{0i})$$

- Stable Unit Treatment Value Assumption (SUTVA)
  - The potential outcome for any unit do not vary with the treatment assigned to other units.
  - For each unit, there are no different forms of each treatment level, which lead to different potential outcomes.

1(c)

$$\begin{aligned}
\Delta^{ATT} &= \mathbb{E}(Y_{1i} - Y_{0i}|D = 1) \\
&= \underbrace{\mathbb{E}(Y_{1i} - Y_{0i})}_{\Delta^{ATE}} + \underbrace{\mathbb{E}(Y_{1i} - Y_{0i}|D = 1) - \mathbb{E}(Y_{1i} - Y_{0i})}_{\text{bias term}} \\
\Delta^{ATUT} &= \mathbb{E}(Y_{1i} - Y_{0i}|D = 0) \\
&= \underbrace{\mathbb{E}(Y_{1i} - Y_{0i})}_{\Delta^{ATE}} + \underbrace{\mathbb{E}(Y_{1i} - Y_{0i}|D = 0) - \mathbb{E}(Y_{1i} - Y_{0i})}_{\text{bias term}}
\end{aligned}$$

1(d)

$$\begin{aligned}
\Delta^{N} &= \mathbb{E}(Y_{1i}|D = 1) - \mathbb{E}(Y_{0i}|D = 0) \\
&= \mathbb{E}(Y_{1i}|D = 1) - \mathbb{E}(Y_{0i}|D = 0) - \mathbb{E}(Y_{0i}|D = 1) + \mathbb{E}(Y_{0i}|D = 1) \\
&= \underbrace{\mathbb{E}(Y_{1i} - Y_{0i}|D = 1)}_{\Delta^{ATT}} + \underbrace{\mathbb{E}(Y_{0i}|D = 1) - \mathbb{E}(Y_{0i}|D = 0)}_{\text{bias term}}
\end{aligned}$$

$$\mathbb{E}(Y_{0i}|D = 1) = \mathbb{E}(Y_{0i}|D = 0) \implies \Delta^{N} = \Delta^{ATT}$$

Example:

	Master ($D=0$)	PhD ($D=1$)
Master earnings (Y_0)	90000	120000
PhD earnings (Y_1)	70000	150000

In the control sample Y_{0i} (those who got master's degree but did not have a doctor's degree), the earnings between those who did not want to do a PhD ($Y_{0i}|D=0$), and those who did want to do a PhD ($Y_{0i}|D=1$) are different from zero, which leads to a biased estimator for the *Average Treatment Effect on the Treated*. As we can see from the table, those who do not want to do a PhD are more likely to self-select themselves into the control group while those who do want to do a PhD are more likely to self-select themselves into the treatment group. The reason is that both can benefit from this experiment. Therefore, the assumption $\mathbb{E}(Y_{0i}|D=1) = \mathbb{E}(Y_{0i}|D=0)$ does not hold in this example.

2. Self selection

2(a)

$$\begin{aligned}\Delta^{ATE} &= \mathbb{E}(Y_{1i} - Y_{0i}) \\ &= \mathbb{E}[(\beta_0 + \beta_1 + \epsilon_{1i}) - (\beta_0 + \epsilon_{0i})] \\ &= \mathbb{E}(\beta_1 + \epsilon_{1i} - \epsilon_{0i}) \\ &= \mathbb{E}(\beta_1) + \mathbb{E}(\epsilon_{1i}) + \mathbb{E}(\epsilon_{0i}) \\ &= \mathbb{E}(\beta_1)\end{aligned}$$

2(b)

$$\begin{aligned}\Delta^{ATT} &= \mathbb{E}(Y_{1i} - Y_{0i}|D=1) \\ &= \mathbb{E}(\beta_1 + \epsilon_{1i} - \epsilon_{0i}|D=1) \\ &= \mathbb{E}(\beta_1|D=1) > 0\end{aligned}$$

2(c)

If $Y_{1i} < Y_{0i}$, then $D_i = 0$.

$$\begin{cases}\Delta^{ATT} = \mathbb{E}(Y_{1i} - Y_{0i}|D=1) > 0 & Y_{1i} > Y_{0i} \\ \Delta^{ATUT} = \mathbb{E}(Y_{1i} - Y_{0i}|D=0) < 0 & Y_{1i} < Y_{0i}\end{cases}$$

From exercise (1) we know that ATE is a weighted average of the ATT and $ATUT$.

$$\Delta^{ATE} = Pr(D=1) \times \underbrace{\Delta^{ATT}}_{\text{positive}} + (1 - Pr(D=1)) \times \underbrace{\Delta^{ATUT}}_{\text{negative}}$$

Therefore, $\Delta^{ATT} > \Delta^{ATE}$.

Intuition: Since there exists a self-selection bias, the average treatment effect on the treated would be more pronounced than it should have to be while the average treatment effect on the untreated would be negative.

3. Randomization at work

```r
# load libraries
library(tidyverse)
library(haven)
library(randomizr)
library(stargazer)

# set seed for reproducibility
set.seed(123)

Data <- read_dta('randomization_2021.dta')

head(Data)
```

```
## # A tibble: 6 x 5
##   black birthday  year    y0    y1
##   <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1     1       16  2005  24.4  25.0
## 2     1       10  2005  24.3  25.0
## 3     0       28  2006  25.1  25.8
## 4     0       29  2005  27.1  27.9
## 5     1        5  2006  24.3  25.0
## 6     1        4  2005  24.3  25.0
```

3(a)

From previous question, we derived

$$\Delta^N = \Delta^{ATT} + \underbrace{\mathbb{E}(Y_{0i}|D=1) - \mathbb{E}(Y_{0i}|D=0)}_{\text{selection bias}}$$

The naive estimator uses the observable $\mathbb{E}(Y_{0i}|D=0$ in place of the missing counterfactual $\mathbb{E}(Y_{0i}|D=1)$. Randomization is a method that minimizes the selection bias in the above formula. In other words, we want to make sure that in the realized outcome where people who get untreated, those who want to get treated and those who do not want to get treated are distributed in a way such that their conditional expected outcomes are almost the same.

3(b)

If I implement randomization, I will take the following points into account:

- Subjects in groups should not be systematically different. In other words, participants in treatment should not be different from:
  - ⋆ all of the **eligible** population
  - ⋆ those who did **not** get treated
  - ⋆ themselves **prior** to the start of the experiment
- If possible, I would use a "double-blind" design to avoid **Experimenter Bias**. In other words, both experimenters and subjects should have no prior knowledge of group assignment.
- I will try to conduct the experiment in a way such that the subjects even do not feel that they are being observed. To some extent, this can reduce the **Hawthorne Effects**.

- I will also try to hide as much information as possible so that subjects cannot self-select themselves into the experiment in their own self-interest.

In practice, the situation gets much more complicated and we can never remove every single problem in our randomization. All we can do is to minimize those "bad'' effects in order to get a plausible outcome based on the randomization.

3(c)

```
# randomize treatment
Data$d <- simple_ra(N=dim(Data)[1], prob=0.5)
```

3(d)

```
# before randomization
x <- mean(Data$y0[Data$black==1])
y <- mean(Data$y0[Data$black==0])
bias <- x - y

# after randomization
x.ra <- mean(Data$y0[Data$d==1])
y.ra <- mean(Data$y0[Data$d==0])
bias.ra <- x.ra - y.ra

cat(sprintf('Before the randomization, the selection bias is %.3f
            \nAfter the randomization, the selection bias is %.3f',
            bias, bias.ra))
```

Before the randomization, the selection bias is -2.195

After the randomization, the selection bias is 0.040

The randomization worked as we see the selection bias gets closer to zero.

3(e)

```
# create a variable y containing realized outcome
Data$y <- ifelse(Data$d==1, Data$y1, Data$y0)

# estimate ATE
ATE <- mean(Data$y1-Data$y0)

# estimate ATT
ATT <- mean(Data$y1[Data$d==1]-Data$y0[Data$d==1])

cat(sprintf('The average treatment effect is %.3f
            \nThe average treatment effect on treated is %.3f',
            ATE, ATT))
```

The average treatment effect is 0.696

The average treatment effect on treated is 0.682

3(f)

```
# regress the realized outcome on the randomized treatment variable
model1 <- lm(y ~ d, data=Data)
stargazer(model1, keep.stat=c('n', 'rsq'), font.size='small', header=F)
```

Table 2:

	Dependent variable:
	y
d	0.722**
	(0.295)
Constant	24.549***
	(0.202)
Observations	96
R^2	0.060
Note:	*p<0.1; **p<0.05; ***p<0.01

$$y_i = \alpha + \beta d_i + u_i$$

$$\begin{aligned}\mathbb{E}(y_i|d_i) &= \mathbb{E}(\alpha + \beta d_i + u_i|d_i) \\ &= \alpha + \beta d_i + \underbrace{\mathbb{E}(u_i|d_i)}_{\text{zero by assumption}} \\ &= \alpha + \beta d_i\end{aligned}$$

$$\begin{aligned}\mathbb{E}(y_i) &= \mathbb{E}(\mathbb{E}(y_i|d_i)) \\ &= \mathbb{E}(\alpha + \beta d_i) \\ &= \alpha + \beta\mathbb{E}(d_i)\end{aligned}$$

$$\begin{aligned}\mathbb{E}(y_0) &= \alpha + \beta\mathbb{E}(d_0) \\ &\Downarrow \\ \hat{\alpha} &= \overline{y}_0\end{aligned}$$

$$\begin{aligned}\mathbb{E}(y_1) &= \alpha + \beta\mathbb{E}(d_1) \\ &\Downarrow \\ \hat{\beta} &= \overline{y}_1 - \hat{\alpha} \\ &= \overline{y}_1 - \overline{y}_0\end{aligned}$$

3(g)

```
# include birthday as a regressor
model2 <- lm(y ~ d + birthday, data=Data)
```

```
# include birthday and year as regressors
model3 <- lm(y ~ d + birthday + year, data=Data)
stargazer(model1, model2, model3, keep.stat=c('n','rsq'),
          font.size='small', header=F, omit='Constant')
```

Table 3:

	Dependent variable:		
	y		
	(1)	(2)	(3)
d	0.722**	0.688**	0.584**
	(0.295)	(0.297)	(0.292)
birthday		0.008	0.005
		(0.017)	(0.016)
year			−0.732**
			(0.293)
Observations	96	95	95
R^2	0.060	0.057	0.118
Note:		*p<0.1; **p<0.05; ***p<0.01	

From column(1) to column(2) to column(3), we see a decrease in the causal parameter as we include more regressors. We might have omitted variable bias in our model and the sign of the bias is positive.

3(h)

It is useful to include additional variables in the regression model since adding irrelevant variables will never affect the consistency of the estimator (it only affects the variance) while omitting relevant variables will render the estimator inconsistent, not to mention the variance. However, we should never add post-treatment controls as regressors in our model.

There are two effects in our case:

- Adding additional regressors reduces the variance of the error term, lowering the variance of the estimator.
- Adding additional regressors reduces the usable variation in the explanatory variable, increasing the variance of the estimator.

From column(1) to column(2), we see an increase in the standard error. The second effect dominates.

From column(2) to column(3), we see an decrease in the standard error. The first effect dominates.

Overall, the first effect dominates as we can see from column(1) to column(3).