Problem set 4

Siwei Dai

October 22, 2021

NOTE1: Start with the file `ps4_2021.Rmd` *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F21/tree/main/assignments). Modify that file to include your answers. Make sure you can "knit" the file (e.g. in RStudio by clicking on the* `Knit` *button). Submit both the Rmd file and the knitted PDF via Canvas*

Question 1

Consider a study that seeks to measure the effect of daily exercise on subjective mental health during finals week in December 2021 (as measured by a survey) among UChicago students.

1a) For a given student, what (in words) are the potential outcomes in this study?

The hypothetical state of subjective mental health the student has during finals week in December 2021 if she exercised daily, and the hypothetical state of subjective mental health the student has during finals week in December 2021 if she didn't exercise daily

1b) With reference to the Fundamental Problem of Causal Inference, explain why it is difficult if not impossible to measure the effect of this treatment on this outcome for any individual student.

For any individual student, she can either exercise daily or not exercise daily, and we can only observe one state of subjective mental health conditional on whether she exercised or not. To measure the effect of the treatment on an individual student, we need to observe both the mental health state given she exercised daily and the mental health state given she had not exercised daily, which, as argued, cannot be observed at the same time.

Even if we want to compare the the mental health state after she exercised daily in year 1 and the mental health state after she didn't exercise daily in year 2, we would have to assume that there are not any confounding factors that would have influenced her subjective mental health. This assumption may hardly stand in reality.

1c) In observational studies, treatment may be related to the potential outcomes. Give one plausible account of how and why treatment may be related to the potential outcomes in this example.

Those who are willing to exercise daily may be better at adjusting to a better mood and a better state of mental health than those who are not willing to exercise daily; in other words, the treatment (whether to exercise daily or not) is related with students' ability to adjust their mental health state, which also affects the potential outcome of subjective mental health we observe.

Question 2

The code below creates a fake dataset with a population of 1000 observations and two variables: `Y0` is $Y(0)$, the potential outcome with treatment set to 0, and `Y1` is $Y(1)$, the potential outcome with treatment set to 1. (Note that observing both potential outcomes is generally not possible; we can do it here because it's a fake dataset.)

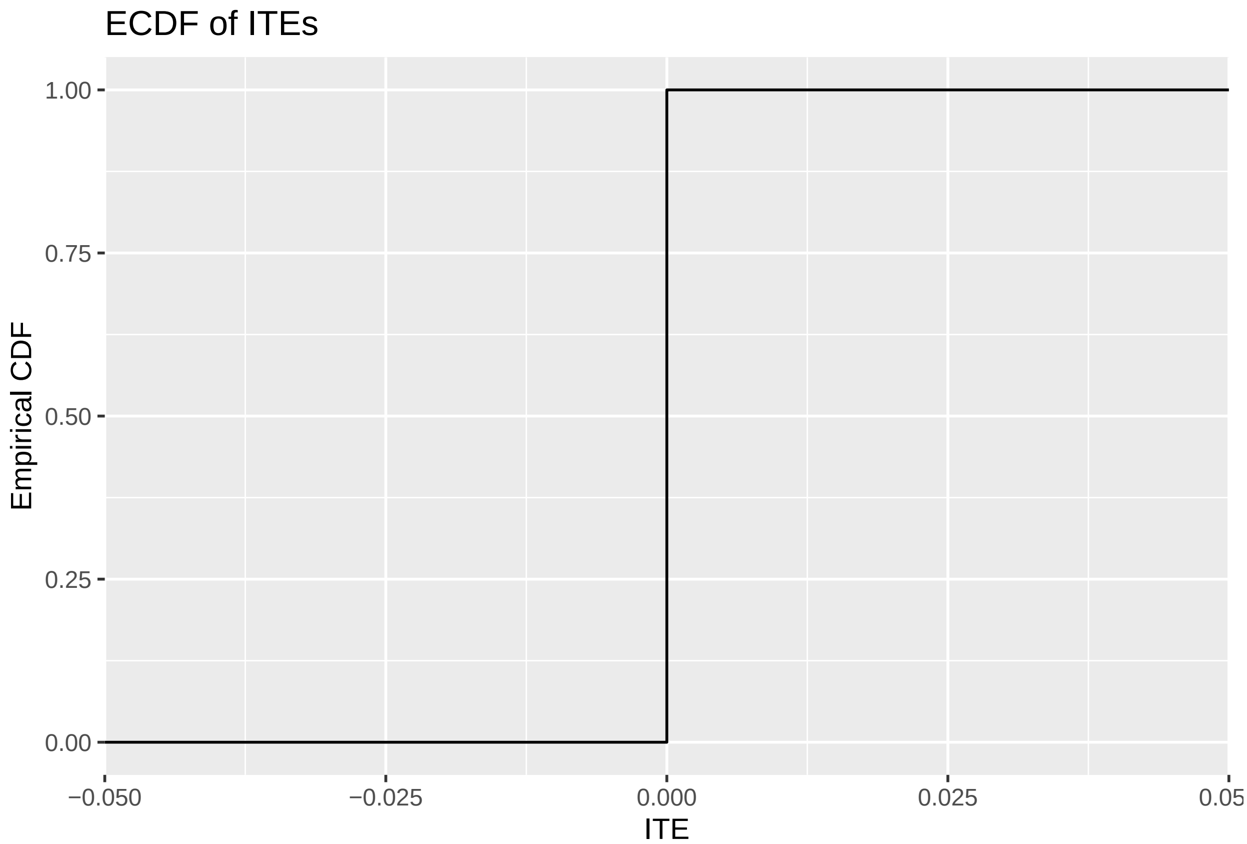
```r
set.seed(30500)
n <- 1000
dat <- tibble(Y0 = runif(n = n, min = 0, max = 1)) %>%
  mutate(Y1 = Y0)
```

2a) Compute the *individual treatment effect (ITE)* for each individual and plot an ECDF of the ITEs. (*Hint*: see ECDF code in lecture 3.2.)

```r
dat <- dat %>%
    mutate('ITE' = Y1 - Y0)
plot2a <- ggplot(data = dat, aes(x = ITE)) +
    stat_ecdf() +
    labs(y = 'Empirical CDF',
         title = 'ECDF of ITEs')
dat
```

```
## # A tibble: 1,000 x 3
##       Y0    Y1   ITE
##    <dbl> <dbl> <dbl>
##  1 0.420 0.420     0
##  2 0.855 0.855     0
##  3 0.842 0.842     0
##  4 0.355 0.355     0
##  5 0.454 0.454     0
##  6 0.681 0.681     0
##  7 0.364 0.364     0
##  8 0.683 0.683     0
##  9 0.539 0.539     0
## 10 0.642 0.642     0
## # ... with 990 more rows
```

```r
plot2a
```



2b) Suppose we choose as our estimand the average treatment effect (ATE). What is the ATE for this population?

```
ATE <- mean(dat$ITE)
ATE
```

```
## [1] 0
```

2c) Add a treatment variable `D` that takes on the value 1 with probability `Y1` and 0 otherwise. (Hint: use the `rbinom()` function.)

```
dat <- dat %>%
    mutate('D' = rbinom(n = n, size = 1, prob = Y1))
dat
```

```
## # A tibble: 1,000 x 4
##       Y0    Y1   ITE     D
##    <dbl> <dbl> <dbl> <int>
##  1 0.420 0.420     0     0
##  2 0.855 0.855     0     1
##  3 0.842 0.842     0     1
##  4 0.355 0.355     0     0
##  5 0.454 0.454     0     0
##  6 0.681 0.681     0     1
##  7 0.364 0.364     0     1
```

```
##   8 0.683 0.683     0     1
##   9 0.539 0.539     0     1
##  10 0.642 0.642     0     1
## # ... with 990 more rows
```

2d) Compute the difference in means using this treatment variable and compare it to the ATE. Why is the difference in means a bad estimator for the ATE in this case?

```
mean_treatment <- dat %>%
    filter(D == 1) %>%
    summarize(mean(Y1, na.rm = TRUE))
mean_control <- dat %>%
    filter(D == 0) %>%
    summarize(mean(Y0, na.rm = TRUE))
diff_in_means <- mean_treatment - mean_control
comparison <- diff_in_means - ATE
comparison
```

```
##   mean(Y1, na.rm = TRUE)
## 1              0.3609576
```

```
# The comparison shows this estimate is 0.3609576 from the ATE.
# It is a bad estimator because he probability of receiving treatment is related
# with the potential outcome Y(1).
```

2e) Create a new treatment variable `D_random` that is assigned at random, as if this were a randomized experiment.

```
dat <- dat %>%
    mutate('D_random' = sample(x = c(0,1), size = 1000, replace = TRUE))
dat
```

```
## # A tibble: 1,000 x 5
##        Y0    Y1   ITE     D D_random
##     <dbl> <dbl> <dbl> <int>    <dbl>
##   1 0.420 0.420     0     0        1
##   2 0.855 0.855     0     1        1
##   3 0.842 0.842     0     1        1
##   4 0.355 0.355     0     0        1
##   5 0.454 0.454     0     0        1
##   6 0.681 0.681     0     1        1
##   7 0.364 0.364     0     1        0
##   8 0.683 0.683     0     1        1
##   9 0.539 0.539     0     1        1
##  10 0.642 0.642     0     1        0
## # ... with 990 more rows
```

2f) Compute the difference in means using this treatment variable and compare it to the ATE.

```
mean_treatment <- dat %>%
    filter(D_random == 1) %>%
```

```r
    summarize(mean(Y1, na.rm = TRUE))
mean_control <- dat %>%
    filter(D_random == 0) %>%
    summarize(mean(Y0, na.rm = TRUE))
diff_in_means <- mean_treatment - mean_control
comparison <- diff_in_means - ATE
comparison
```

```
##   mean(Y1, na.rm = TRUE)
## 1            -0.01801274
```

```r
# This comparison shows that the estimate is -0.01801274 from the ATE.
```

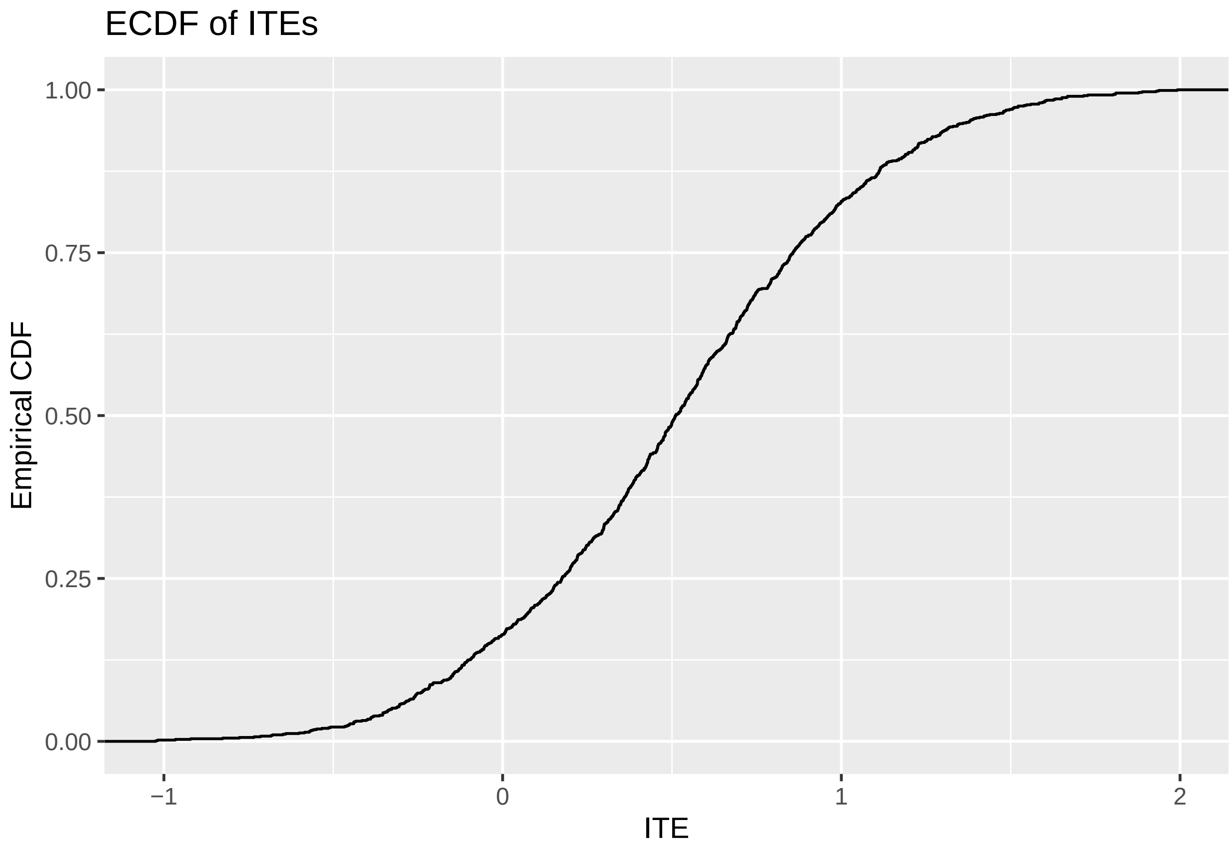
Question 3

The code below creates another fake dataset with a population of 1000 observations and the same two variables, `Y0` and `Y1`.

```r
dat <- tibble(Y0 = rnorm(n = n, mean = 0, sd = 1)) %>%
  mutate(Y1 = Y0 + rnorm(n = n, mean = .5, sd = .5))
```

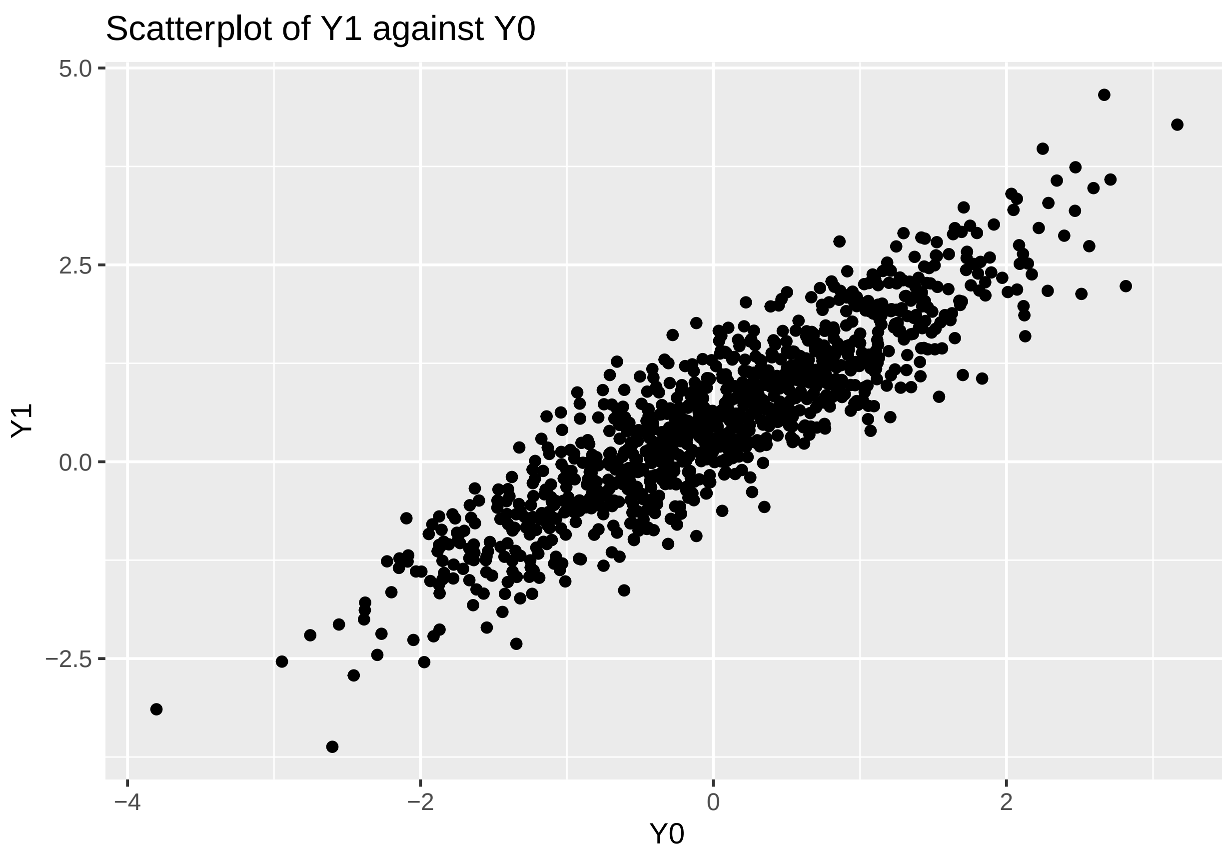
3a) Compute the *individual treatment effect (ITE)* for each individual and plot an ECDF of the ITEs.

```r
dat <- dat %>%
    mutate('ITE' = Y1 - Y0)
plot3a <- dat %>%
    ggplot(aes(x = ITE)) +
    stat_ecdf() +
    labs(y = 'Empirical CDF',
         title = 'ECDF of ITEs')
plot3a
```



3b) Create a scatterplot of `Y1` (vertical axis) against `Y0` (horizontal axis).

```
plot3b <- dat %>%
    ggplot(aes(x = Y0, y = Y1)) +
    geom_point() +
    labs(x = 'Y0',
         y = 'Y1',
         title = 'Scatterplot of Y1 against Y0')
plot3b
```



3c) If this were a study of students, and `Y` were a measure of academic achievement (with `D` a study skills training session), how would you interpret a point at (2,2) on the previous plot? How about a point at (-1, 0)?

(2, 2) would mean that training session has no causal effect on the student's academic achievement; (-1, 0) means the training session caused a positive causal effect of 1 unit on the student's academic achievement

3d) Suppose we choose as our estimand the average treatment effect (ATE). What is the ATE for this population?

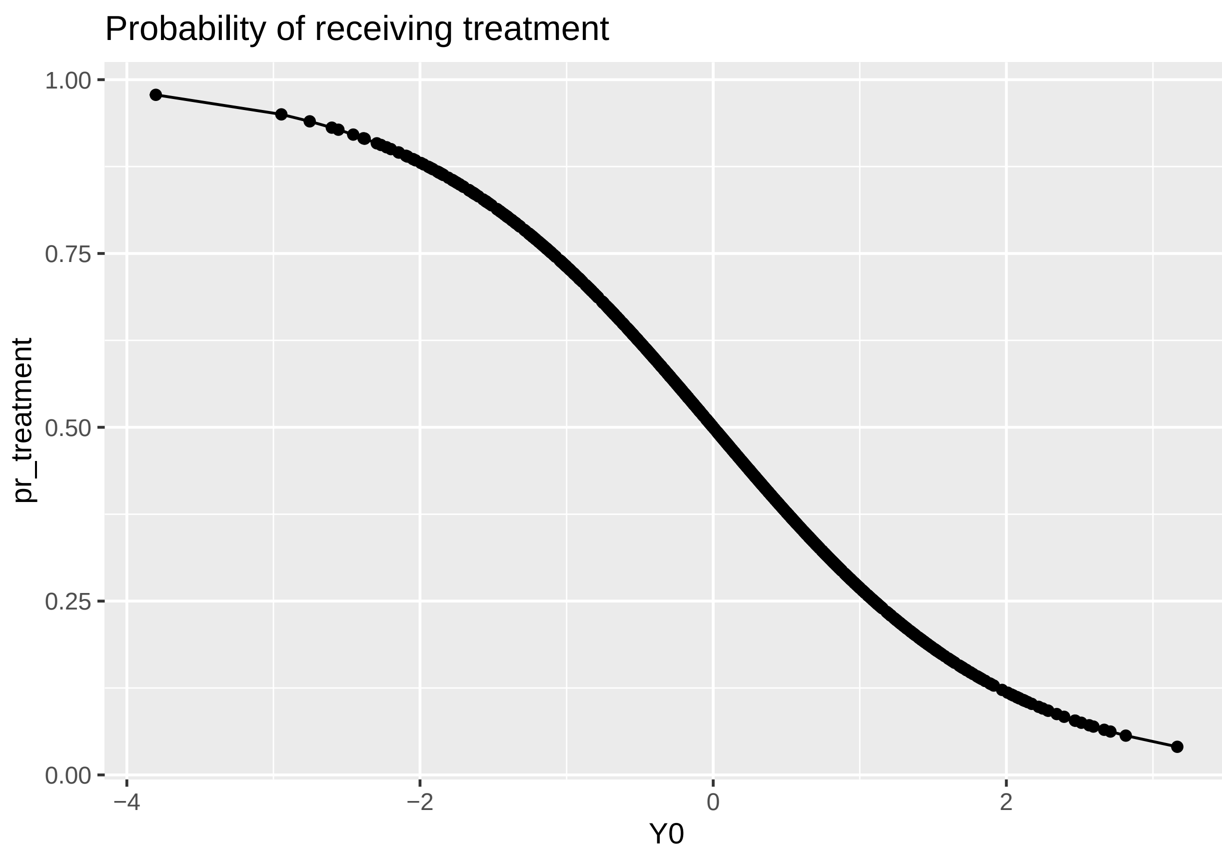
```
ATE <- mean(dat$ITE)
ATE
```

```
## [1] 0.5151057
```

ATE is the average causal effect of the training session on students' academic achievement

3e) Create a new variable `pr_treatment` that is `1 - exp(Y0)/(1 + exp(Y0))`. Plot `pr_treatment` (vertical axis) as a function of `Y0`.

```
dat <- dat %>%
    mutate('pr_treatment' = 1 - exp(Y0)/(1 + exp(Y0)))
plot3e <- dat %>%
    ggplot(aes(x = Y0, y = pr_treatment)) +
    geom_point() +
    geom_line() +
    labs(title = 'Probability of receiving treatment')
plot3e
```



3f) Again supposing `Y` is a measure of academic achievement and `D` a study skill training, why might the probability of treatment be related to `Y0` as in this hypothetical example?

The ones who are more likely to take the treatment (i.e. going to the study skill training), are very likely those who previously are not good at studying or have not mastered study skills, therefore they had a low previous academic achievement (Y0 is low) and wanted to take the training.

3g) Add a treatment variable `D` that takes on the value 1 with probability `pr_treatment` and 0 otherwise. Hint: use the `rbinom()` function.

```
dat <- dat %>%
    mutate('D' = rbinom(n = n, size = 1, prob = pr_treatment))
```

3h) Compute the difference in means using this treatment variable and compare it to the ATE. Why is the difference in means a bad estimator for the ATE in this case?

```
mean_treatment <- dat %>%
    filter(D == 1) %>%
    summarize(mean(Y1, na.rm = TRUE))
mean_control <- dat %>%
    filter(D == 0) %>%
    summarize(mean(Y0, na.rm = TRUE))
diff_in_means <- mean_treatment - mean_control
ATE <- mean(dat$Y1) - mean(dat$Y0)
comparison <- diff_in_means - ATE
comparison
```

```
##   mean(Y1, na.rm = TRUE)
## 1             -0.8530235
```

```
# The comparison means the estimate is -0.8530235 away from the ATE.
# Because the treatment is not randomly assigned and the probability of treatment is correlated with th
```

3i) Create a new treatment variable `D_random` that is assigned at random, as if this were a randomized experiment.

```
dat <- dat %>%
    mutate('D_random' = sample(x = c(0, 1), size = n, replace = TRUE))
```

3j) Compute the difference in means using this treatment variable and compare it to the ATE.

```
mean_treatment <- dat %>%
    filter(D_random == 1) %>%
    summarize(mean(Y1, na.rm = TRUE))
mean_control <- dat %>%
    filter(D_random == 0) %>%
    summarize(mean(Y0, na.rm = TRUE))
diff_in_means <- mean_treatment - mean_control
ATE <- mean(dat$Y1) - mean(dat$Y0)
comparison <- diff_in_means - ATE
comparison
```

```
##   mean(Y1, na.rm = TRUE)
## 1             0.05765998
```

```
# The estimate is 0.05765998 from the ATE.
```