Problem set 4

Your name here

Due 10/22/2021 at 5pm

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?

One potential outcomes for a given student is the student's survey response if she undertakes daily exercise (call this $Y(1)$). The other is the student's survey response if she does not undertake daily exercise (call this $Y(0)$).

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.

Either a student undertakes daily exercise during finals week in a given year or she does not. Thus we only see one potential outcome for that student and cannot observe the difference between them. This is the FPOCI.

One could compare a student's mental health during a quarter when she does undertake daily exercise with her mental health during a quarter when she does not undertake daily exercise, but other things will certainly be different between those two quarters, and some of those may affect her mental health, making it difficult to distinguish the effect of the treatment from the effect of these other factors.

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.

Poor mental health may make it harder to exercise. If so, then students who undertake daily exercise likely would have had better mental health even in the absence of exercise than students who do not undertake daily exercise.

Question 2

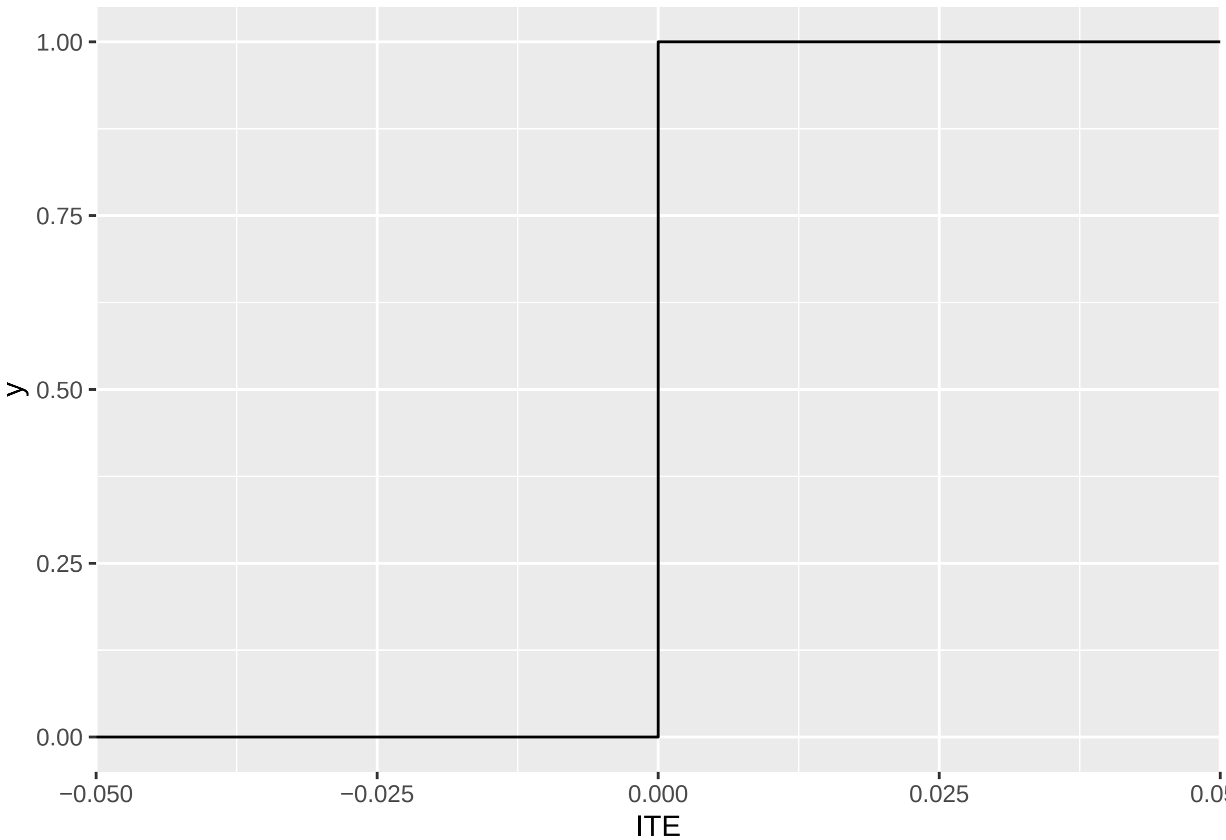
The code below creates a fake dataset with a population of 1000 observations and two variables: `Y0` is $Y(0)$, the potential outcome with treatment set to 0, and `Y1` is $Y(1)$, the potential outcome with treatment set to 1. (Note that observing both potential outcomes is generally not possible; we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

```
set.seed(30500)
n <- 1000
```

```r
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 %>%
  mutate(ITE = Y1 - Y0) %>%
  ggplot(aes(x = ITE)) +
  stat_ecdf(geom = "step")
```



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

It's obvious that the ATE is zero from the setup, but we can compute it thus:

```r
dat %>%
  summarize(ate = mean(Y1) - mean(Y0))
```

```
## # A tibble: 1 x 1
##     ate
##   <dbl>
## 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.)

```r
dat2 <- dat %>%
  mutate(D = rbinom(n, size = 1, prob = Y1))
```

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?

```
dat2 %>%
  summarize(mean(Y1[D == 1]) - mean(Y0[D==0]))
```

```
## # A tibble: 1 x 1
##   `mean(Y1[D == 1]) - mean(Y0[D == 0])`
##                                   <dbl>
## 1                                 0.361
```

The treament is related to the potential outcomes: here, subjects with higher $Y(1) = Y(0)$ have a higher probability of being treated. This means that the treated units would have a higher value of the outcome variable than the control units even if (as is true here) the treatment has no effect.

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

```
dat3 <- dat %>%
  mutate(D_random = sample(x = rep(c(0,1), n/2)))
```

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

```
dat3 %>%
  summarize(mean(Y1[D_random == 1]) - mean(Y0[D_random==0]))
```

```
## # A tibble: 1 x 1
##   `mean(Y1[D_random == 1]) - mean(Y0[D_random == 0])`
##                                                 <dbl>
## 1                                            -0.00811
```

It should be close. It differs from the ATE only due to random variation in the treatment variable `D`.

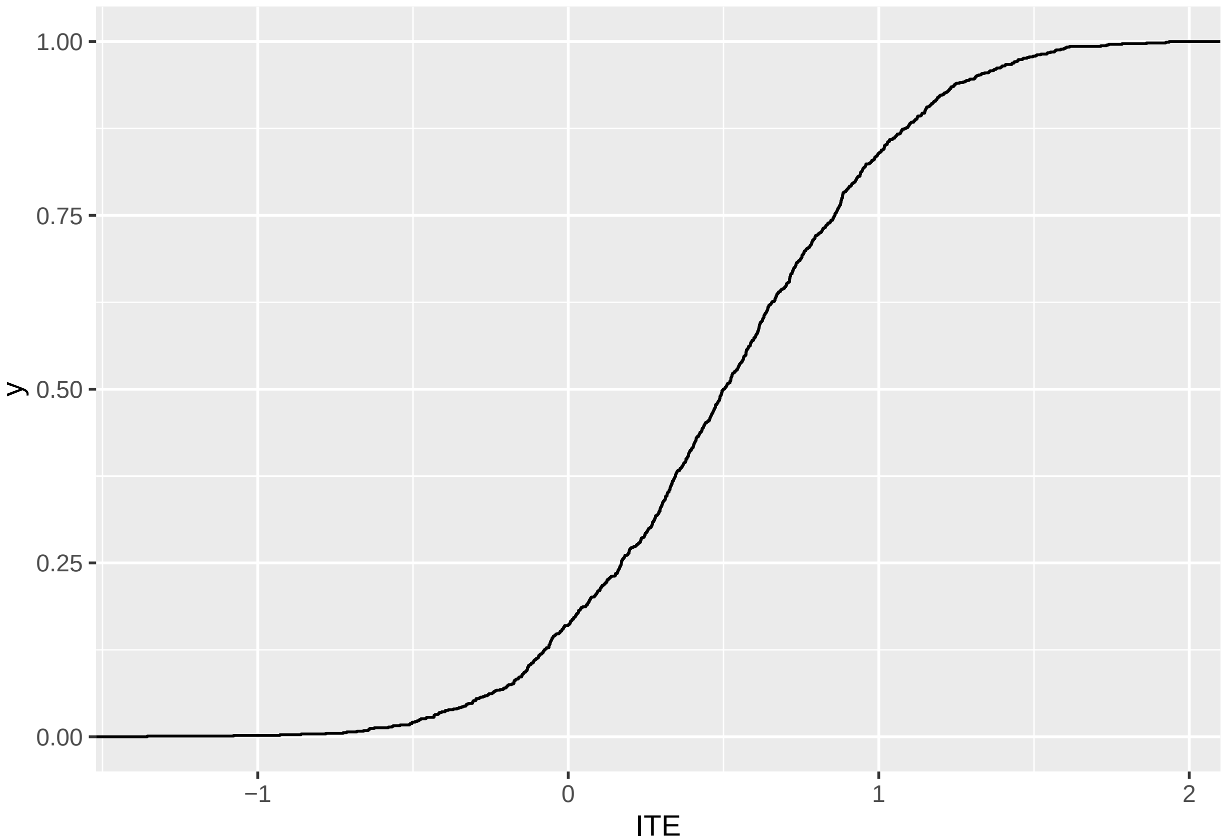
Question 3

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

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

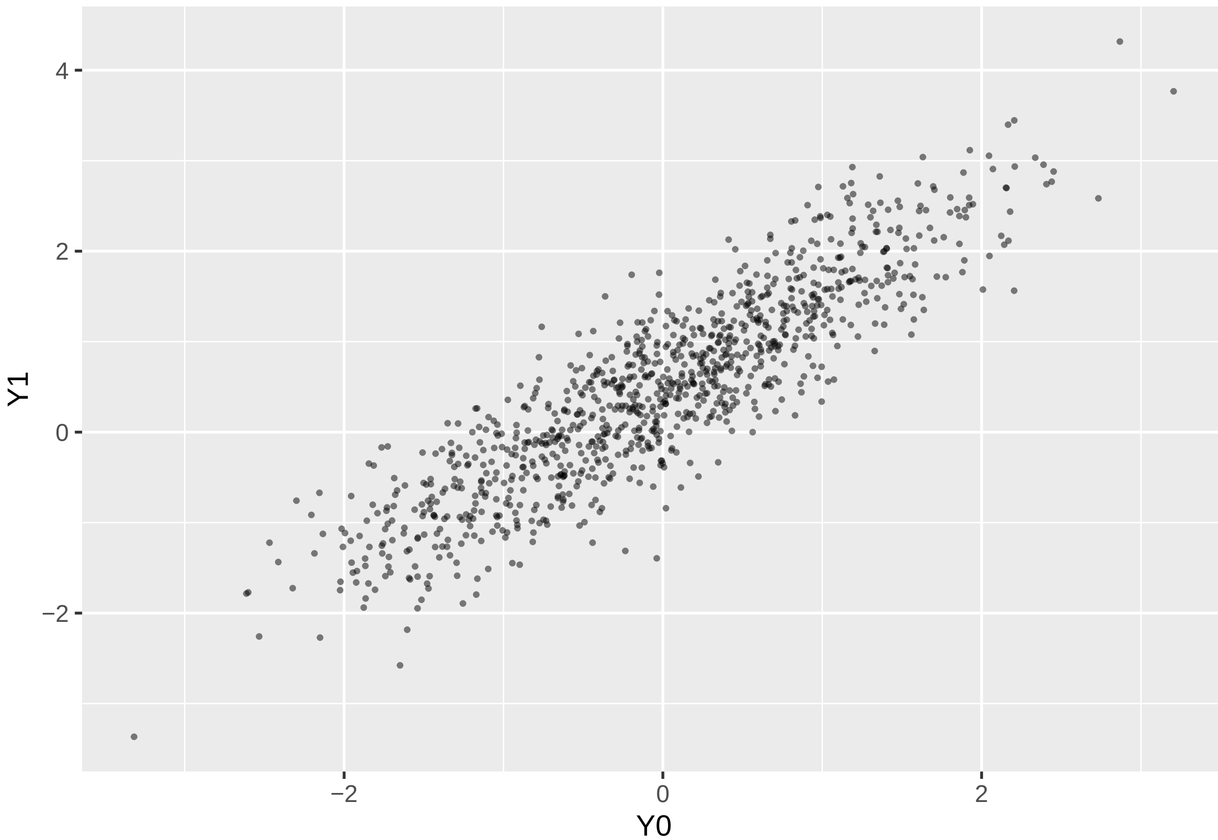
3a) Compute the *individual treatment effect (ITE)* for each individual and plot an ECDF of the ITEs.

```
dat %>%
  mutate(ITE = Y1 - Y0) %>%
  ggplot(aes(x = ITE)) +
  stat_ecdf(geom = "step")
```



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

```
dat %>%
  ggplot(aes(x = Y0, y = Y1)) +
  geom_point(alpha = .5, size = .5)
```



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)?

A point at (2,2) corresponds to a student who is totally unaffected by the treatment and would do well whether or not she attends the study skills session. A point at (-1,0) corresponds to a weaker student who does 1 point better when she attends the session than when she does not.

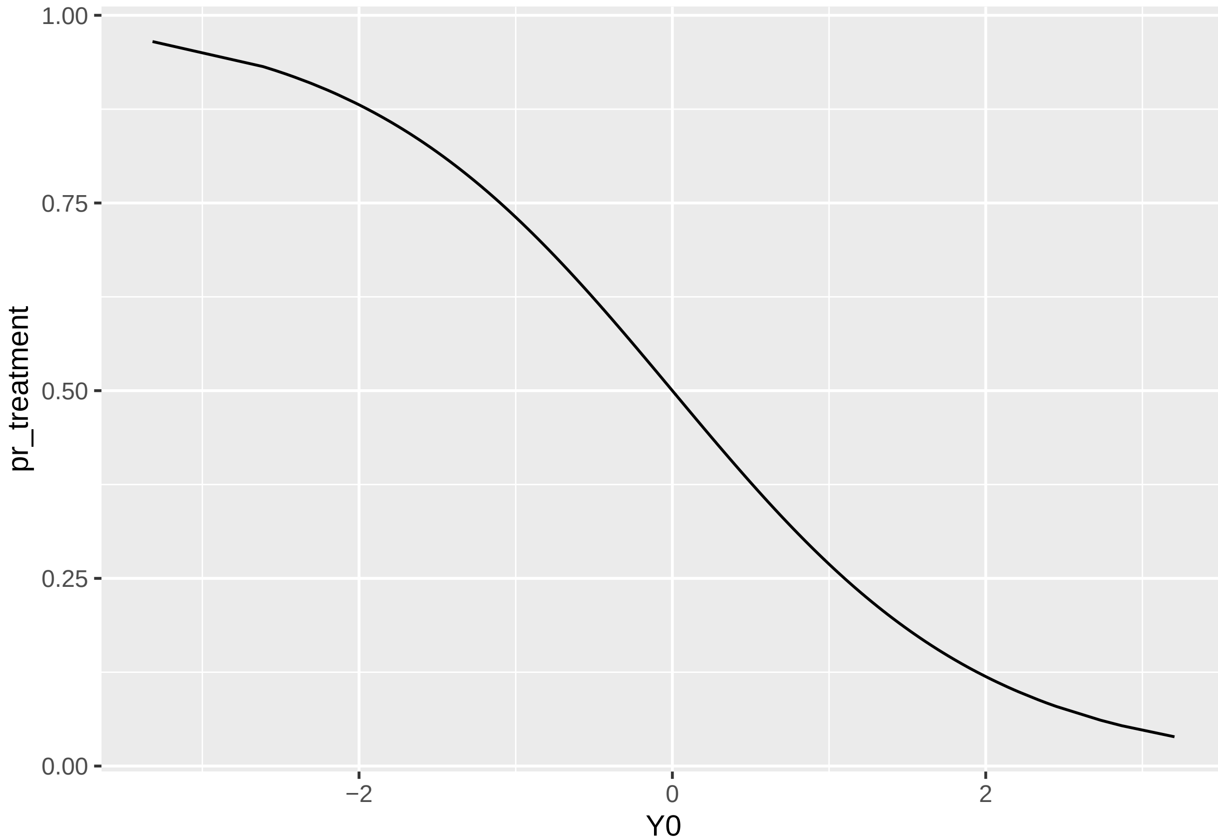
3d) Suppose we choose as our estimand the average treatment effect (ATE). What is the ATE for this population?

```
dat %>%
  summarize(mean(Y1) - mean(Y0))
```

```
## # A tibble: 1 x 1
##   `mean(Y1) - mean(Y0)`
##                   <dbl>
## 1                 0.505
```

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`.

```
dat2 <- dat %>%
  mutate(pr_treatment = 1 - exp(Y0)/(1 + exp(Y0)))
dat2 %>%
  ggplot(aes(x = Y0, y = pr_treatment)) +
  geom_line()
```



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?

It could be that the weaker students are more likely to seek out support. Or perhaps they are more likely to be asked or required to attend.

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

```
dat3 <- dat2 %>%
  mutate(D = rbinom(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?

```
dat3 %>%
  summarize(mean(Y1[D == 1]) - mean(Y0[D==0]))
```

```
## # A tibble: 1 x 1
##   `mean(Y1[D == 1]) - mean(Y0[D == 0])`
##                                    <dbl>
## 1                                 -0.282
```

Again, the treatment is related to the potential outcomes. Here the students who undertake the training would have done worse than those who didn't even if none of them had undertaken the training. In this case it makes it seem like the training makes students do worse even though it actually makes them do .5 better on average.

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

```r
dat3 <- dat2 %>%
  mutate(D_random = sample(x = rep(c(0,1), n/2)))
```

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

```r
dat3 %>%
  summarize(mean(Y1[D_random == 1]) - mean(Y0[D_random==0]))
```

```
## # A tibble: 1 x 1
##   `mean(Y1[D_random == 1]) - mean(Y0[D_random == 0])`
##                                                 <dbl>
## 1                                               0.500
```

This is very close to the target of .5; any difference you get reflect random variation in the treatment.