Problem Set 1: Solutions

Your Name - Net ID - Section Number

Due Oct 6th, 2023

This homework must be turned in on Brightspace by Oct 6th 2023. It must be your own work, and your own work only – you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. **You may consult with others, but when you write up, you must do so alone.**

Your homework submission must be written and submitted using Rmarkdown. No handwritten solutions will be accepted. You should submit:

1. A compiled PDF file named yourNetID_solutions.pdf containing your solutions to the problems.
2. A .Rmd file containing the code and text used to produce your compiled pdf named yourNetID_solutions.Rmd.

Note that math can be typeset in Rmarkdown in the same way as Latex. Please make sure your answers are clearly structured in the Rmarkdown file:

1. Label each question part
2. Do not include written answers as code comments. Write out answers and explanations separately.
3. The code used to obtain the answer for each question part should accompany the written answer. Comment your code!

Question 1. Definitions and Examples (20 points)

Answer the following questions. Be as specific and detailed as possible. Give examples.

1. What is the fundamental problem of causal inference? (5 points)

The fundamental problem of causal inference is that it is (usually) impossible to observe both potential outcomes of treatment and control; if the unit receives the treatment it is impossible to observe the outcome of the counterfactual (receiving no treatment) since the unit has in fact received treatment and vice versa. For example, when testing the efficacy of a drug for every subject you can only see the outcome after receiving or not receiving the treatment as the subject can't both take and not take the drug.

2. Why are experiments important? (5 points)

Experiments, especially randomized controlled trials (RCTs), are the fundamental "solution" to the fundamental problem of causal inference. Experiments are studies where the researcher controls the probability of being assigned the treatment. One can engineer the experiment to meet two of the requirements for causal inference, ignorability and positivity. Through randomization both of these assumptions can be met; by randomizing, stratifying, etc. we can hopefully eliminate any confounding effects and also ensure every unit has a greater than 0 probability of being assigned to the treatment.

3. What does ignorability mean? (5 points)

Ignorability, or unconfoundedness refers to the treatment assignment being independent of the potential outcomes given some observed covariates. In other words, whether or not a case is assigned to treatment will not affect the case's potential outcome and any covariates which affect both treatment and outcome have been identified and controlled for. Here is an example of an ignorability violation: suppose you were carrying out a study to examine the effects of a exercise regimen on body weight. However, instead of randomly assigning participants to treatment and control groups, you allow the participants to choose themselves. One of the obvious issues with this is that people who voluntarily choose to participate in this weight-loss regimen are much more likely to already care about their body weight and actively try (even outside of the study) to lower their body weights. As a result their outcome is not independent from treatment assignment.

4. What is SUTVA? (5 points)

SUTVA stands for Stable Unit Treatment Value Assumption. It is one of the requirements for causal inference. SUTVA calls for no spillover (or interference) and single value of treatment. No spillover means one unit's treatment status will not affect the potential outcomes of other units. An example of a spillover violation is this: let's say you had a study on the effect of the presence of a scarecrow on pest-related crop damage. You put a scarecrow in one field as the treatment, and leave a nearby nearly identical field untouched as the control. However, the scarecrow manages to scare away some pestilent animals from the treatment to the scarecrow-free control field increasing the pest damage in the control field. This is an example of spillover since the treatment of a treatment case affected the outcome of a case in the control group. Single value of treatment refers to a single version of treatment for all treatments of the same level; in other words the process by which the treatment is given does not affect the treatment potential outcome. Here's an example of a single treatment value violation: imagine you conducted a study on the effect of psychological therapy on subjects' sense of fulfillment (only on mentally healthy people for obvious ethical reasons). You randomly assign subjects to treatment (therapy sessions) and control (no therapy), however you fail to standardize the actual therapy experience. Instead, subjects assigned to different therapists effectively receive different therapy treatments. As a result, those who were lucky enough to be assigned to more skilled or experienced therapists had "better" outcomes (higher sense of fulfillment) than those assigned to inexperienced or unenthusiastic therapists, who might end up worse off. This is a violation of single treatment value as differences in how exactly the treatment was received affected the individual treatment outcomes.

Question 2. Bed Nets and Malaria (20 points)

Article: Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment by Jessica Cohen and Pascaline Dupas

Some economists have argued that “cost-sharing” makes it more likely that a product will be used (versus giving it away for free). Cohen and Dupas partnered with 20 Kenyan prenatal clinics to distribute subsidized anti-malarial bed nets. For each clinic, they varied the extent of the subsidy: either full (free bed-nets, $D_i = 1$) or partial (90% cheaper bed-nets, $D_i = 0$). They measure (among other things) whether women who received bed nets used them (Y_i).

1. What is $\mathbb{E}[Y_i|D_i = 0]$? (5 points)

The expected outcome for subjects who don’t receive the treatment, or how likely it is that a subject will use the bed net given that the subject received a partial subsidy.

2. What is $\mathbb{E}[Y_i(1)]$? (5 points)

The expected potential outcome if everyone received the treatment, or how likely it is that a subject will use the bed net if it was given out for free to everyone.

3. What is $\mathbb{E}[Y_i(1)|D_i = 0]$? (5 points)

The expected outcome for subjects who are in clinics which only give out the partial subsidy, or the process to assign the treatment to the subject doesn’t occur, if they received the treatment (the full subsidy).

4. Cohen and Dupas randomized treatment at the level of the clinic, but the outcomes of interest are at the individual level. Is there a violation of consistency/SUTVA? Why or why not? Argue your case. (5 points)

It can’t be concluded with certainty whether or not a consistency/SUTVA violation exists based on the given information, but most likely not. There would be a violation if there were some spillover effect or different values of treatment. A single version of treatment violation is very unlikely, as it seems unreasonable that these clinics could somehow carry out different versions of treatment since the treatment is a simple cost subsidy. A spillover effect would be if a clinic giving out the treatment to one patient would affect other patients’ (from the same clinic or from different clinics) treatment outcome. This is unlikely, but could still occur in theory. For example, if one woman receiving a free net would cause her to discuss with and convince other participating women about the benefits of using a net, this could affect the probability of using the nets for the other women in the study. Randomizing on the clinic level as opposed to the individual level isn’t inherently a violation either, as this is essentially a randomized clustered experiment.

Question 3. Application (Coding) (30 points)

The STAR (Student-Teacher Achievement Ratio) Project is a four year *longitudinal study* examining the effect of class size in early grade levels on educational performance and personal development.

This exercise is in part based on[1]:

Mosteller, Frederick. 1997. "The Tennessee Study of Class Size in the Early School Grades." *Bulletin of the American Academy of Arts and Sciences* 50(7): 14-25.

A longitudinal study is one in which the same participants are followed over time. This particular study lasted from 1985 to 1989 involved 11,601 students. During the four years of the study, students were randomly assigned to small classes, regular-sized classes, or regular-sized classes with an aid. In all, the experiment cost around $12 million. Even though the program stopped in 1989 after the first kindergarten class in the program finished third grade, collection of various measurements (e.g., performance on tests in eighth grade, overall high school GPA) continued through the end of participants' high school attendance.

We will analyze just a portion of this data to investigate whether the small class sizes improved performance or not. The data file name is `STAR.csv`, which is a CSV data file. The names and descriptions of variables in this data set are:

Name	Description
`race`	Student's race (White = 1, Black = 2, Asian = 3, Hispanic = 4, Native American = 5, Others = 6)
`classtype`	Type of kindergarten class (small = 1, regular = 2, regular with aid = 3)
`g4math`	Total scaled score for math portion of fourth grade standardized test
`g4reading`	Total scaled score for reading portion of fourth grade standardized test
`yearssmall`	Number of years in small classes
`hsgrad`	High school graduation (did graduate = 1, did not graduate = 0)

Note that there are a fair amount of missing values in this data set. For example, missing values arise because some students left a STAR school before third grade or did not enter a STAR school until first grade.

1. Create a new factor variable called `kinder` in the data frame. This variable should recode `classtype` by changing integer values to their corresponding informative labels (e.g., change 1 to `small` etc.). Similarly, recode the `race` variable into a factor variable with four levels (`white`, `black`, `hispanic`, `others`) by combining Asians and Native Americans as the `others` category. For the `race` variable, overwrite the original variable in the data frame rather than creating a new one. Recall that `na.rm = TRUE` can be added to functions in order to remove missing data. (5 points)

```r
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
STAR <- read.csv("STAR2.csv")
```

[1] I have provided you with a sample of their larger dataset. Empirical conclusion drawn from this sample may differ from their article.

```r
STAR <- STAR %>%
  mutate(kinder = factor(case_when(classtype == 1 ~ "small",
                                   classtype == 2 ~ "regular",
                                   .default = "regular with aid")),
         race = factor(case_when(race == 1 ~ 'white',
                                 race == 2 ~ 'black',
                                 race == 4 ~ 'hispanic',
                                 .default = 'others')))

head(STAR)
```

```
##    race classtype yearssmall hsgrad g4math g4reading           kinder
## 1 white         2          0     NA     NA        NA          regular
## 2 black         3          0      1     NA        NA regular with aid
## 3 white         1          4      1    707       719            small
## 4 white         1          2      1     NA        NA            small
## 5 black         2          0      0     NA        NA          regular
## 6 black         2          0     NA    674       571          regular
```

2. How does performance on fourth grade reading and math tests for those students assigned to a small class in kindergarten compare with those assigned to a regular-sized class? Do students in the smaller classes perform better? Use means to make this comparison while removing missing values. Give a brief substantive interpretation of the results. To understand the size of the estimated effects, compare them with the standard deviation of the test scores. (10 points)

```r
meandiffReading <- filter(STAR, kinder == 'small') %>% pull(g4reading) %>% mean(na.rm = T) - filter(STAR
meandiffMath <- filter(STAR, kinder == 'small') %>% pull(g4math) %>% mean(na.rm = T) - filter(STAR, kin
sdReading <- pull(STAR, g4reading) %>% sd(na.rm = T)
sdMath <- pull(STAR, g4math) %>% sd(na.rm = T)

cat('difference in reading test means:', meandiffReading, '\n')
```

```
## difference in reading test means: 4.084932
```

```r
cat('standard deviation of reading scores:', sdReading, '\n')
```

```
## standard deviation of reading scores: 52.33846
```

```r
cat('difference in math test means:', meandiffMath, '\n')
```

```
## difference in math test means: -0.6222809
```

```r
cat('standard deviation of math scores:', sdMath, '\n')
```

```
## standard deviation of math scores: 43.8088
```

Students assigned to small classes in kindergarten performed on average 4.085 points higher on the reading test and 0.622 points lower on the math test in comparison to students assigned to regular-sized classes. These differences are very small compared to the standard deviations of both test scores, which suggests that being assigned to a small vs regular sized class had no effect on reading and math test performance, or at least there is nowhere near enough evidence to reasonably conclude an effect exists.

3. Instead of comparing just average scores of reading and math tests between those students assigned to small classes and those assigned to regular-sized classes, look at the entire range of possible scores. To do so, compare a high score, defined as the 66th percentile, and a low score (the 33rd percentile) for small classes with the corresponding score for regular classes. These are examples of *quantile treatment effects*. Does this analysis add anything to the analysis based on mean in the previous question? (Hint: You will use the quantile() function in r.) (5 points)

```
readingpctSmall <- filter(STAR, kinder == 'small') %>% pull(g4reading) %>% quantile(na.rm = T, probs = c
readingpctRegular <- filter(STAR, kinder == 'regular') %>% pull(g4reading) %>% quantile(na.rm = T, probs
mathpctSmall <- filter(STAR, kinder == 'small') %>% pull(g4math) %>% quantile(na.rm = T, probs = c(0.33
mathpctRegular <- filter(STAR, kinder == 'regular') %>% pull(g4math) %>% quantile(na.rm = T, probs = c(0

cat('reading test score percentile differences:\n')
```

```
## reading test score percentile differences:
```

```
print(readingpctSmall - readingpctRegular)
```

```
## 33% 66%
##   3   1
```

```
cat('math test score percentile differences:\n')
```

```
## math test score percentile differences:
```

```
print(mathpctSmall - mathpctRegular)
```

```
##   33%   66%
## -1.00  2.36
```

The differences in scores for both tests at both the 33rd and 66th percentiles was very small which supports the previous conclusion that small class size had no effect on fourth grade test performance.

4. We examine whether the STAR program reduced the achievement gaps across different racial groups. Begin by comparing the average reading and math test scores between white and minority students (i.e., Blacks and Hispanics) among those students who were assigned to regular classes with no aid. Conduct the same comparison among those students who were assigned to small classes. Give a brief substantive interpretation of the results of your analysis. (5 points)

```
white <- STAR %>% filter(race == 'white')
minority <- STAR %>% filter(race %in% c('black', 'hispanic'))

readingdiffSmall <- filter(white, kinder == 'small') %>% pull(g4reading) %>% mean(na.rm = T) - filter(mi
readingdiffRegular <- filter(white, kinder == 'regular') %>% pull(g4reading) %>% mean(na.rm = T) - filte
mathdiffSmall <- filter(white, kinder == 'small') %>% pull(g4math) %>% mean(na.rm = T) - filter(minority
mathdiffRegular <- filter(white, kinder == 'regular') %>% pull(g4math) %>% mean(na.rm = T) - filter(mino

cat('difference in reading test means for small size:', readingdiffSmall, '\n')
```

```
## difference in reading test means for small size: 29.53008
```

```
cat('difference in reading test means for regular size:', readingdiffRegular, '\n')
```

```
## difference in reading test means for regular size: 33.05486
```

```
cat('difference in math test means for small size:', mathdiffSmall, '\n')
```

```
## difference in math test means for small size: 14.7655
```

```
cat('difference in math test means for regular size:', mathdiffRegular, '\n')
```

```
## difference in math test means for regular size: 12.34691
```

On average, white students scored higher than minority students on both reading and math tests for every class size. However, for the reading test, the gap between average white student score and average minority student score was smaller for small class sizes than for regular class sizes. The opposite is seen for math;

the difference in means was larger at the small class size than at the regular class size. Looking only at the means, it is unclear if these differences are statistically significant.

5. We consider the long term effects of kindergarten class size. Compare high school graduation rates across students assigned to different class types. Also, examine whether graduation rates differ by the number of years spent in small classes. Finally, as done in the previous question, investigate whether the STAR program has reduced the racial gap between white and minority students' graduation rates. Briefly discuss the results. (5 points)

```
STAR %>%
  group_by(kinder) %>%
  summarise(`graduation rate` = mean(hsgrad, na.rm = T))
```

```
## # A tibble: 3 x 2
##   kinder           `graduation rate`
##   <fct>                        <dbl>
## 1 regular                      0.823
## 2 regular with aid             0.836
## 3 small                        0.848
```

```
STAR %>%
  group_by(yearssmall) %>%
  summarise(`graduation rate` = mean(hsgrad, na.rm = T))
```

```
## # A tibble: 5 x 2
##   yearssmall `graduation rate`
##        <int>             <dbl>
## 1          0             0.825
## 2          1             0.798
## 3          2             0.852
## 4          3             0.827
## 5          4             0.883
```

```
cat('high school graduation rate differences:\n', mean(white$hsgrad, na.rm = T) - mean(minority$hsgrad,
```

```
## high school graduation rate differences:
##  0.1221848
```

```
cat('\nhigh school graduation rate differences based on class size:\n')
```

```
##
## high school graduation rate differences based on class size:
```

```
print(tapply(white$hsgrad, white$kinder, mean, na.rm = T) - tapply(minority$hsgrad, minority$kinder, mea
```

```
##          regular regular with aid            small
##        0.1214189        0.1387456        0.1005110
```

```
cat('\nhigh school graduation rate differences based on years in small classes:\n')
```

```
##
## high school graduation rate differences based on years in small classes:
```

```
print(tapply(white$hsgrad, white$yearssmall, mean, na.rm = T) - tapply(minority$hsgrad, minority$yearssm
```

```
##          0          1          2          3          4
## 0.12973203 0.07014829 0.11618590 0.12500000 0.10214067
```

Aid and small class sizes did have higher graduation rates than regular class sizes but the difference is very small. Spending 4 years in small classes had a noticeably higher graduation rate than the other years, but extra years did not necessarily equate to higher graduation rates as the graduation rate for 0 years in small classes was larger than for 1 year and similarly the graduation rate for 3 years was less than 2. There seems to generally be a positive correlation between years spent in small classes and graduation rates but the dips during the odd year counts are strange. The difference in graduation rate between white and minority students is similarly high regardless of class size or years spent in small classes, which suggests that the STAR program had little to no effect on the racial gap for high school graduation rates.

Question 4. Design Your Experiment (30 points)

Design your own experiment from start to finish. Choose an *interesting* question. Explain why observational data may give you the wrong answer. Detail the potential outcomes and a well-defined treatment. Explain the type of experiment (completely random, cluster-design, block/stratified). Will your design ensure a causal treatment effect? (Remember: Be as specific as possible and give examples.)

Question: Do product availability urgency indicators ("there are n items left") increase e-commerce sales during holiday (peak shopping) season? To answer this question we will use a block/stratified design. Since consumers' urgency to buy a product probably varies between different types of products, we will stratify by product type, e.g. electronics, clothing, and beauty products. Within these blocks we will randomly assign divide shoppers (half and half) into treatment and control groups. The treatment will be a "n items left in stock" label at the top of the product page. The potential outcome $Y_i(1)$ will be how much the customer spends on that product with the label present while the potential outcome $Y_i(0)$ will be how much the customer spends with no label present. This design will ensure a causal treatment effect by satisfying the three requirements of ignorability, SUTVA, and positivity. By randomizing assignment into the treatment and control groups, we will control for any confounders and selection effects to ensure ignorability. Since it's e-commerce, there is a risk of spillover since customers may share product pages with their friends. To mitigate this we can limit the experiment to customers who are browsing from the website's home page as most people share using URLs. To eliminate the risk of single treatment violation, we can make sure the limited availability indicators are identical in design and positioning. By randomizing the treatment assignment we ensure that all subjects of the experiment have a chance of being assigned the treatment so the positivity assumption is met. Observational data can be quite misleading due to the inability to control for confounding factors. If data on limited availability labels exists, it probably is not randomized so as to control for factors such as concurrent promotions and market campaigns, or individual product popularity. Both of these have an effect on sales, so randomization is important to control for them.