HW4_ANS

Zihao Li

2023-11-11

Load the packages we want.

```
### Quietly load all packages needed
### insert your code here
library(readxl) # for loading the excel-formatted data
library(knitr)
library(dplyr)
library(tidyverse)
```

Part 1

Load the data here

```
### do not change anything in {}
### r work for answering question goes here
cluster = read.csv("cluster.csv")
summary(cluster)
```

```
##    treatment      store_id          sale         customer_id
##  Min.   :0.0   Min.   :  1.00   Min.   :  0.00   Min.   :   1.0
##  1st Qu.:0.0   1st Qu.: 50.75   1st Qu.: 62.71   1st Qu.: 250.8
##  Median :0.5   Median :100.50   Median : 77.21   Median : 500.5
##  Mean   :0.5   Mean   :100.50   Mean   : 76.95   Mean   : 500.5
##  3rd Qu.:1.0   3rd Qu.:150.25   3rd Qu.: 91.34   3rd Qu.: 750.2
##  Max.   :1.0   Max.   :200.00   Max.   :163.84   Max.   :1000.0
```

1. ATE and CI, original data

Calculate the average treatment effect on sales and the 95% confidence interval on that average under the assumption that you can treat each of the 200,000 observations in the data as independent. (That is, just analyze this the way you have analyzed other data so far in this course).

```
### do not change anything in {}
### r work for answering question goes here

# ATE
ate <- mean(cluster$sale[cluster$treatment == 1]) - mean(cluster$sale[cluster$treatme
nt == 0])

# SE
se <- sqrt(var(cluster$sale[cluster$treatment == 1]) / sum(cluster$treatment == 1) +
var(cluster$sale[cluster$treatment == 0]) / sum(cluster$treatment == 0))

# CI
lb = ate - 1.96*se
ub = ate + 1.96*se

print(ate)
```

```
## [1] 3.600389
```

```
print(lb)
```

```
## [1] 3.416286
```

```
print(ub)
```

```
## [1] 3.784491
```

Q1 ANSWER HERE

Treatment does not have a direct impact on the likelihood of people coming into the store which means treatment and control groups should ensure that any difference in the number of people entering the store is not influenced by the treatment itself.

2. ATE and CI, store-level data

```
### do not change anything in {}
### r work for answering question goes here
### I had using CHATGPT for helping me to correct error, other wise, it keeping showi
ng "summarise()` has grouped output by 'treatment'. You can override using the `.grou
ps` argument"

store_avg_sales <- cluster %>%
  group_by(treatment, store_id) %>%
  summarise(avg_sales = mean(sale), .groups = 'drop')

# ATE
ate_store_level <- mean(store_avg_sales$avg_sales[store_avg_sales$treatment == 1]) -
mean(store_avg_sales$avg_sales[store_avg_sales$treatment == 0])

# SE
se_store_level <- sqrt(var(store_avg_sales$avg_sales[store_avg_sales$treatment == 1])
/ sum(store_avg_sales$treatment == 1) + var(store_avg_sales$avg_sales[store_avg_sales
$treatment == 0]) / sum(store_avg_sales$treatment == 0))

# CI
lb_store_level <- ate_store_level - 1.96 * se_store_level
ub_store_level <- ate_store_level + 1.96 * se_store_level

print(ate_store_level)
```

```
## [1] 3.600389
```

```
print(lb_store_level)
```

```
## [1] -0.4908496
```

```
print(ub_store_level)
```

```
## [1] 7.691627
```

Q2 ANSWER HERE

store level ate: 3.6, lb:-0.5, ub:7.7

3. Discussion

3. Discuss the changes (if any) you see in the ATE and CI range going from the individual-level to the store-aggregated data.

Q3 ANSWER HERE

It does have changing

Individual ate: 3.6, lb:3.4, ub:3.9

store level ate: 3.6, lb:-0.5, ub:7.7

Which means store level having same ate to the individual but it is more unstable compare to individual

Part 2: Flu shot RCT

Load the data here

```
### do not change anything in {}
### r work for answering question goes here
flu = read.csv("flu_rct_2023.csv")
summary(flu)
```

```
##        id          treatment         quiz          flu_shot
##  Min.   :  1.0   Min.   :0.00   Min.   :0.000   Min.   :0.000
##  1st Qu.:125.8   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.000
##  Median :250.5   Median :1.00   Median :0.000   Median :1.000
##  Mean   :250.5   Mean   :1.03   Mean   :0.346   Mean   :0.592
##  3rd Qu.:375.2   3rd Qu.:2.00   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :500.0   Max.   :2.00   Max.   :1.000   Max.   :1.000
```

Q4. What fraction of people in each treatment arm actually took the quiz?

```
### do not change anything in {}
### r work for answering question goes here
fraction <- flu %>%
  group_by(treatment) %>%
  summarise(sum_quiz = sum(quiz))

F_1 = fraction$sum_quiz[2]/ sum(flu$treatment == 1)
F_2 = fraction$sum_quiz[2]/ sum(flu$treatment == 2)

print(F_1)
```

```
## [1] 0.3167702
```

```
print(F_2)
```

```
## [1] 0.2881356
```

Q4 ANSWER HERE

treatment fraction_quiz

1 0.3167702

2 0.2881356

Q5. Respond to pat

Your boss, Pat, after seeing these statistics says to you: “Don’t we want to look at what share of people got the shot among the ones who took the quiz in the treatment groups?” What do you say to Pat?

Q5 ANSWER HERE

Data we had showes fraction for people who take the quiz for both treatment is almsot samee, So we do want to look at what share of people got the shot among the ones who took the quiz in the treatment groups which can bring some more meaningful and valuable data to us.

Q6. ITT for each treatment arm

What is the Intention to Treat estimate of the treatment effect for each treatment relative to control?

```
### do not change anything in {}
### r work for answering question goes here
fraction <- flu %>%
  group_by(treatment) %>%
  summarise(sum_flu = sum(flu_shot))
ITT_0 = fraction$sum_flu[1]/ sum(flu$treatment == 0)
ITT_1 = fraction$sum_flu[2]/ sum(flu$treatment == 1)
ITT_2 = fraction$sum_flu[3]/ sum(flu$treatment == 2)
print(ITT_0)
```

```
## [1] 0.5
```

```
print(ITT_1)
```

```
## [1] 0.5776398
```

```
print(ITT_2)
```

```
## [1] 0.6892655
```

Q6 ANSWER HERE

ITT_0 0.5

ITT_1 0.5776398

ITT_2 0.6892655

Q7. Excludabiltiy

In order for a Treatment on the Treated estimate to be valid, we have to believe that the only way that offering a treatment could affect behavior is through the person actually taking up the treatment (not just learning about it). Briefly discuss why that assumption might not be right in this setting.

Q7 ANSWER HERE

Becuase it is in the same company which people might share what they get based on their different email, which people in the control group also might being affect by the people in the treatment group which might cause this assumption might not be right in this setting.

Q8. TOT estimates

Setting aside any concern from question 7, find the TOT for effect of survey on flu shot, comparing Treatment 1 to control? What about comparing Treatment 2 to control? What about comparing Treatment 2 to Treatment 1? [Note: this third question is not exactly something we have covered in class. Asking you to extend the logic of what we have done.]

```
### do not change anything in {}
### r work for answering question goes here
fraction <- flu %>%
  group_by(treatment) %>%
  summarise(sum_quiz = sum(quiz), sum_F_and_Q = sum(quiz ==1 & flu_shot == 1) )
tot_1 = fraction$sum_F_and_Q[2]/fraction$sum_quiz[2]
tot_2 = fraction$sum_F_and_Q[3]/fraction$sum_quiz[3]
TOT2_to_1 = tot_2/tot_1
print(tot_1)
```

```
## [1] 0.6862745
```

```
print(tot_2)
```

```
## [1] 0.7295082
```

```
print(TOT2_to_1)
```

```
## [1] 1.062998
```

Q8 ANSWER HERE

tot_1 0.6862745

tot_2 0.7295082

TOT2_to_1 1.062998

Q9 What's going on?

Briefly give a conjecture (i.e., a guess/hypothesis) about what could be driving the difference in the Treatment on the Treated effects between Treatment 1 and Treatment 2.

Q9 ANSWER HERE

It looks like that treatment 2 is more useful comparing to the treatment 1 which might caused by HR visiting which will let employee to be worry about if they dont do the flu and they will be in the trouble which push them to finish quiz and get the flu shot.

10 Generative AI

Did you use generative AI? If so, please explain how.

Q10 ANSWER HERE

I did use chatgpt for question 2 for helping me to correct error, other wise, it keeping showing "summarise() `has grouped output by 'treatment'. You can override using the` .groups` argument" and I just type in the code I had and ask it to help me correct it based on error was showing.