HW 1

Erik Andersen

2022-04-27

Contents

Setup

```
# Load packages

pacman::p_load(tidyverse, broom, haven, data.table, here, magrittr, stargazer, DT)

# Load data

ohp_df = read_dta(here('data', 'ohp.dta'))

# Set as a data.table for quick manipulations

setDT(ohp_df)
```

Question 1

There is a subtle difference between the two variables. The treatment reports whether the person was a winner of the OHP lottery, while the survey variable reports if the person enrolled in the medicaid program. This may sound the same since the OHP lottery gave each winner the chance to enroll for the closed medicaid program. The difference is the compliance rate. Not everyone who wins the lottery will choose to join the medicaid program, so the two variables will have different values.

Treatment is the treatment variable instead of the survey variable because it is a random selection while who joins medicaid is not. It is reasonable to assume that there is some difference between the people who choose to join medicaid after winning the lottery and those who do not, so there is a selection bias between the two groups. If we used the survey variable as the treatment variable, then our results would be biased by the selection bias. But if we use the treatment variable, then the two groups have been selected randomly, so there should not be any selection bias between them.

Question 2

The variables I chose to test are gender, age, education, the percentage of patients diagnosed with hypertension before the lottery, the number of people diagnosed with diabetes before the lottery, and the number of people diagnosed with depression pre lottery. It was important to choose variables that are determined before the treatment so we don't introduce bias.

```r
# Get the means of the relevant variables

ohp_means = ohp_df[treatment == 0,
       lapply(.SD, mean, na.rm = T),
       .SDcols = c("gender_inp",
                   'age_inp',
                   'edu_inp',
                   'hbp_dx_pre_lottery',
                   'dia_dx_pre_lottery',
                   'dep_dx_pre_lottery')]



# Table of means

stargazer(ohp_means, type = 'text', summary.stat = 'mean', title = 'Control Group Means', covariate.labe
```

```
##
## Control Group Means
## ==============================================
## Statistic                              Mean
## --------------------------------------------
## Gender                                0.569
## Age                                   40.606
## Education                             2.238
## High Blood Pressure pre Randomization 0.183
## Diabetes pre Randomization            0.072
## Depression pre Randomzation           0.350
## --------------------------------------------
```

Question 3

```
# Run all the balance regressions

balance_reg1 = ohp_df %>% lm(gender_inp ~ treatment,.) |> tidy()
balance_reg2 = ohp_df %>% lm(age_inp ~ treatment,.) |> tidy()
balance_reg3 = ohp_df %>% lm(edu_inp ~ treatment,.) |> tidy()
balance_reg4 = ohp_df %>% lm(hbp_dx_pre_lottery ~ treatment,.) |> tidy()
balance_reg5 = ohp_df %>% lm(dia_dx_pre_lottery ~ treatment,.) |> tidy()
balance_reg6 = ohp_df %>% lm(dep_dx_pre_lottery ~ treatment,.) |> tidy()

# Vector of differences from regression

diffs = c(balance_reg1$estimate[2], balance_reg2$estimate[2], balance_reg3$estimate[2], balance_reg4$es

# Vector of standard errors

se = c(balance_reg1$std.error[2], balance_reg2$std.error[2], balance_reg3$std.error[2], balance_reg4$st

# Vector of means

means = c(ohp_means$gender_inp[1], ohp_means$age_inp[1], ohp_means$edu_inp[1], ohp_means$hbp_dx_pre_lot
# Data frame of characteristics, means, and differences

balance_table = data.table(Characteristics = c("gender_inp", "age_inp", "edu_inp", "hbp_dx_pre_lottery"

# Print table for pdf

balance_table
```

```
##       Characteristics Control_Mean Treatment-Control_Difference Standard_Errors
##                <char>        <num>                        <num>           <num>
## 1:         gender_inp   0.56881205                 -0.006106555     0.008977078
## 2:            age_inp  40.60606061                  0.380317975     0.211772551
## 3:            edu_inp   2.23839699                  0.021675126     0.016445278
## 4: hbp_dx_pre_lottery   0.18264293                 -0.001337153     0.006984929
## 5: dia_dx_pre_lottery   0.07172201                 -0.000796696     0.004659082
## 6: dep_dx_pre_lottery   0.35022253                 -0.018298305     0.008579026
```

```
# Printed table. This works for html, but not pdf output. Its a better table, so I'm emailing you the h

# datatable(balance_table, class = 'cell-border stripe') |> formatRound(c("Control_Mean", "Treatment-Co
```

Question 4

The balance table is consistent with random assignment into treatment and control groups. What we are looking for in the balance table to confirm this is that there is no statistical difference between the control mean and the treated mean for any of the characteristics. Another consideration (that we don't run into here) is that a difference could be statistically insignificant, but still of a large magnitude. This would be a concerning result despite the lack of precision estimating it.

In our case, we can see from the table that none of the differences between the control and treatment means (right column) are significant. For gender, we see that in the control group, 57% of participants are female, and that is only -0.0062 different than the treated group gender composition. The standard error is 0.0091 which we can easily see makes the estimate non-significant. Also important, the -0.0062 estimate is two orders of magnitude smaller than the control mean, so even if the difference is estimated imprecisely, we are not concerned that there is a meaningful difference between the groups. The same analysis holds true for each variable I chose, so we can conclude that there is strong evidence the random assignment was successful.

Question 5

```
compliance_reg = ohp_df %>% lm(ohp_all_ever_survey ~ treatment,.) |> tidy()


compliance_reg
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    0.158   0.00571      27.7 2.12e-164
## 2 treatment      0.254   0.00790      32.1 2.27e-217
```

The above regression output shows us that if a person is enrolled in the treatment group, they are approximately 25 percentage points more likely to enroll in medicaid. The result is highly significant, so we can be confident that there was an effect of the treatment on enrolling in medicaid.

Question 6

The variables I chose for this question are if the patient was diagnosed with depression, diabetes, or hypertension after the lottery, the number of doctors visits, and blood pressure. For the last two, I am not entirely sure that they are outcome variables. The descriptions are not explicit whether they are blood pressure readings, and number of doctors visits before the treatment or after it. I am assuming that they are after the treatment for two reasons. First (a meta reason), the question asks for 4 to 6 health outcome variables, and only three variables explicitly say they are post lottery, so there must be 1-3 more health outcome variables. Second, it seems reasonable that they would only have for after the lottery. The blood pressure variable for instance says its an average of three consecutive readings. It seems unlikely that they would be able to find data on that for all 12,000 people in the study before the lottery, so I am assuming it is after. For the same reasoning, I am also assuming the number of doctors visits is also recorded after the lottery.

```r
# Run all the outcome regressions

outcome_reg1 = ohp_df %>% lm(dep_dx_post_lottery ~ treatment,.) #|> tidy()
outcome_reg2 = ohp_df %>% lm(dia_dx_post_lottery ~ treatment,.) #|> tidy()
outcome_reg3 = ohp_df %>% lm(hbp_dx_post_lottery ~ treatment,.) #|> tidy()
outcome_reg4 = ohp_df %>% lm(bp_sar_inp ~ treatment,.) #|> tidy()
outcome_reg5 = ohp_df %>% lm(doc_num_mod_inp ~ treatment,.) #|> tidy()

# Printed table

stargazer(outcome_reg1, outcome_reg2, outcome_reg3, outcome_reg4, outcome_reg5, type = 'text', keep.sta
```

```
##
## Intent to Treat Effects
## =====================================================================================================
##                                                  Dependent variable:
##               ---------------------------------------------------------------------------------------
##               dep_dx_post_lottery dia_dx_post_lottery hbp_dx_post_lottery   bp_sar_inp   doc_num_mod_
##                   Depression           Diabetes          Hypertension     Blood Pressure # Doctor Vis
## -----------------------------------------------------------------------------------------------------
## Treatment            0.005             0.009***             0.002            -0.058         0.396*
##                     (0.004)             (0.002)            (0.004)          (0.300)        (0.216)
##
## Control Mean       0.049***            0.012***           0.057***         119.130***      5.746***
##                     (0.003)             (0.002)            (0.003)          (0.217)        (0.156)
##
## -----------------------------------------------------------------------------------------------------
## Observations        12,095              12,186             11,945            12,188         12,158
## =====================================================================================================
## Note:                                                                          *p<0.1; **p<0.05; ***p<0
```

Question 7

```
# First I'll make a vector of all the effects

effects = c(outcome_reg1$coefficients[2], outcome_reg2$coefficients[2], outcome_reg3$coefficients[2],ou

# Now we divide each effect by the compliance rate to get the treatment on the treated effect

atet = effects / compliance_reg$estimate[2]

# And give each value its name so we can tell which value is for which

names(atet) = c("Depression", "Diabetes", "Hypertension", "Blood Pressure", "# Doctor Visits"); atet
```

```
##      Depression        Diabetes    Hypertension  Blood Pressure # Doctor Visits
##     0.018130398     0.033947483     0.009484628    -0.229734246     1.560334292
```

The above printout shows the average treatment effect on the treated for each of the chosen variables. We can see that each one went up except for blood pressure. This may seem at first glance to be wrong, but it makes sense with more thought. The first three variables are depression, diabetes, and hypertension *diagnoses*. If someone goes to a doctor more, as we would expect, and see when they have medicaid, it is more likely that such a disease will be noticed and diagnosed. If someone isn't going to the doctor they cannot have any disease diagnosed, so it is reasonable that these went up. The blood pressure measure goes down, which is also what we expect to see. Since the blood pressure variable is an objective measurement, there isn't the diagnosis effect we saw in the other variables. We expect that doctors visits should make someone healthier, and lowering blood pressure is one sign of that.

To calculate these values, we took the intent to treat effects from the previous problem for each variable, and divided them by the compliance rate we found a few problems ago. I will explain the intuition behind this. The compliance rate is the portion of people who won the lottery who actually got on medicaid. In our case it was about 25%. The intent to treat effect is the raw difference between the treated group and the control group for each variable. Since only 25% of people in the control group complied, we are not picking up the whole effect of the treatment with that because 75% of that estimate is full of people who didn't comply with the treatment, and so were the same as the control group. We want to know what the effect on a group of people who get medicaid is, so we take the raw estimate, and multiply it by 4 (divide by 25%). This attempts to get at the true effect we would see if everyone in the treatment group complied with the treatment. This assumes that the people who complied, and those who didn't are comparable. It is possible that the people who complied with the treatment are those on whom medicaid has the biggest effect, in which case the ATET would over estimate the true effect.

Question 8

Attrition bias occurs when individuals in a study exit non-randomly before its completion because of death, unwillingness to participate, or any other factors. I think there is potential for attrition bias in this study. As I said, one cause of attrition is death of participants. For this to bias the estimate, people in the control group (or treatment group) would have to die non-randomly at a different rate than those in the treatment group (or control group). I think there is a strong possibility that is the case. The study provided health care to the treatment group, and the Ur-purpose of health care is to stop people from dying. Since the treated group had more access to health care, they probably died less often than the control group who did not have access to medicaid. The control group would then non-randomly die faster than the treated group, and create attrition bias.

Question 9

Abstract A 2008 program in Oregon allowed a unique opportunity for a natural experiment studying the effects of medicaid on a variety of health factors. Oregon decided to expand their medicaid coverage by lottery, randomly giving low-income, uninsured people a chance to apply for medicaid coverage. Because of the randomization, I could study the causal effect medicaid insurance on various health outcomes such as blood pressure, and diagnoses of depression, diabetes, and hypertension. The sample contains about 25,000 participants from the Portland area, and covers their health outcomes over 18 months. I find people on medicaid who won the lottery increase their doctor visits by about 1.5, while seeing a drop in blood pressure of 0.23. I also find that diagnoses of depression, diabetes, and hypertension increase due to the increased frequency of doctor visits. We conclude that being on medicaid increased overall health of participants in the study.