Project One

Isaac Wilfong

2025-04-11

```
#loading in the required libraries
library(tidyverse)
library(sandwich)
library(lmtest)
library(stargazer)
library(systemfit)
library(modelsummary)
library(coefplot)


# loading in the data
OHP <- read.csv("ohp.csv")
```

```
#Understanding the data
columns <- colnames(OHP)
#columns
#head(OHP)
```

Number One

Explain the key difference between the variables *treatment* and *ohp_all_ever_survey*. Explain why *treatment* should be the experimental treatment variable we talked about in class (Di), rather than *ohp_all_ever_survey*.

We want the intent to treat estimate to find the average treatment effect. With ohp_all_ever_survey, this is people who has ever had medicaid. This isn't random selection. There is something that makes these people different for the simple fact that they enrolled in medicaid. They are all linked through that. This introduces sample bias and we are not finding the average treatment effect but the average person who would ever get medicaid. We use the treatment because the treatment is randomly selected. This eliminated any type of sample bias pending that the treatment and control group are not different in any statistically significant manner. With the treatment variable, this is the people that received the medicaid lottery. Here we can use the intent to treat method to find the average treatment effect. This is the randomized control experiment.

Number Two

Provide evidence that the OHP lottery really did randomly assign individuals to treatment and control groups. Think about which variables should be balanced in a randomized experiment, then look through the variables in the dataset (documentation attached at the end of this file) and decide which 5 you will

summarize in the table described below. I recommend using regressions to calculate all the information you need for this table, and then manually putting the table together.

For this we can create a balance table to see if there is any statistical difference between the treatment group and the control group.

```
# this is for a comprehensive balance table of all the variable
# excluding ids, and post variables

library(tableone)

# Define variables of interest
columns <- colnames(OHP) # getting a variable name of all the columns
columns <- columns[-(1:4)] # eliminating the ID variables
columns <- columns[-c(4,6,11)] # eliminating all the post variables

# Define grouping variable
grouping_var <- "treatment"

# Generate the balance table
balance_table <- CreateTableOne(vars = columns, strata = grouping_var, data = OHP)

# Print the table
print(balance_table)
```

```
##                                 Stratified by treatment
##                                  0              1              p      test
##   n                                5842           6387
##   age_inp (mean (SD))             40.61 (11.69)  40.99 (11.71)  0.073
##   bp_sar_inp (mean (SD))         119.13 (16.73) 119.07 (16.39)  0.846
##   chl_inp (mean (SD))            205.77 (34.17) 205.13 (33.45)  0.295
##   dep_dx_pre_lottery (mean (SD))   0.35 (0.48)    0.33 (0.47)   0.033
##   dia_dx_pre_lottery (mean (SD))   0.07 (0.26)    0.07 (0.26)   0.864
##   doc_num_mod_inp (mean (SD))      5.75 (11.83)   6.14 (11.96)  0.067
##   edu_inp (mean (SD))              2.24 (0.90)    2.26 (0.91)   0.188
##   gender_inp (mean (SD))           0.57 (0.50)    0.56 (0.50)   0.496
##   hbp_dx_pre_lottery (mean (SD))   0.18 (0.39)    0.18 (0.39)   0.848
##   hispanic_inp (mean (SD))         0.18 (0.38)    0.18 (0.39)   0.502
##   itvw_english_inp (mean (SD))     0.88 (0.32)    0.87 (0.34)   0.036
##   numhh_list (mean (SD))           1.20 (0.40)    1.29 (0.46)  <0.001
##   ohp_all_ever_survey (mean (SD))  0.16 (0.37)    0.41 (0.49)  <0.001
##   race_black_inp (mean (SD))       0.11 (0.31)    0.10 (0.30)   0.165
##   race_nwother_inp (mean (SD))     0.14 (0.35)    0.15 (0.35)   0.598
##   race_white_inp (mean (SD))       0.69 (0.46)    0.69 (0.46)   0.720
##   rx_num_mod_inp (mean (SD))       1.84 (2.82)    1.97 (2.96)   0.016
```

```
# Regressions for the 5 variables

#setting up the models with respective robust standard errors
model_age <- lm(age_inp ~ treatment, data = OHP)
rob_se_age <- sqrt(diag(vcovHC(model_age, type = "HC1")))

model_SBP <- lm(bp_sar_inp ~ treatment, data = OHP)
rob_se_SBP <- sqrt(diag(vcovHC(model_SBP, type = "HC1")))
```

```r
model_gender <- lm(gender_inp ~ treatment, data = OHP)
rob_se_gender <- sqrt(diag(vcovHC(model_gender, type = "HC1")))

model_edu <- lm(edu_inp ~ treatment, data = OHP)
rob_se_edu <- sqrt(diag(vcovHC(model_edu, type = "HC1")))

model_white <- lm(race_white_inp ~ treatment, data = OHP)
rob_se_white <- sqrt(diag(vcovHC(model_white, type = "HC1")))

#stargazer table with all 5 of the models and the respective RSE

stargazer(
  model_age, model_SBP, model_gender, model_edu, model_white,
  se = list(rob_se_age,rob_se_SBP,  rob_se_gender, rob_se_edu, rob_se_white),
  type = "latex",
  header = FALSE,
  title = "Balance Table with Robust Standard Errors",
  column.labels = c("Age" , "Blood Pressure", "Gender", "Education", "Doctor Visits"),
  omit.stat = c("f", "ser"),
  digits = 3
)
```

Table 1: Balance Table with Robust Standard Errors

	Dependent variable:				
	age_inp Age	bp_sar_inp Blood Pressure	gender_inp Gender	edu_inp Education	race_white_inp Doctor Visits
	(1)	(2)	(3)	(4)	(5)
treatment	0.380*	−0.058	−0.006	0.022	−0.003
	(0.212)	(0.300)	(0.009)	(0.016)	(0.008)
Constant	40.606***	119.130***	0.569***	2.238***	0.690***
	(0.153)	(0.219)	(0.006)	(0.012)	(0.006)
Observations	12,228	12,188	12,229	12,218	12,190
R^2	0.0003	0.00000	0.00004	0.0001	0.00001
Adjusted R^2	0.0002	−0.0001	−0.00004	0.0001	−0.0001
Note:					*p<0.1; **p<0.05; ***p<0.01

I did this is two ways. I found a really neat balance table package in R and I thought it looked like ones I seen in development economics. The second table includes the regression estimates while the first one does not.

Number 3

Is your balance table consistent with individuals having been randomly assigned to treatment group and control groups? Why or why not?

The balance table seems to be consistent with individuals having been randomly assigned to the treatment and control. The majority of our p-values in the balance table are above the .10 meaning that there is

no statistical difference between the control group and the treatment group. There are a few variables of concern that showed p-values lower than 0.05. I believe that is important to note, however, overall this is similar to most balance tables I have seen in other academic papers.

Number 4

Estimate the compliance rate for the OHP experiment. That is, what is the effect of being assigned to the treatment group on the probability of being enrolled in Medicaid? For this question and question 7, you can use the same regression model as in question 3, just changing the dependent variable.

```
#OHP %>% filter(treatment == 1) %>% summarize(mean(ohp_all_ever_survey))

Comp_model <- lm(ohp_all_ever_survey ~ treatment, data = OHP)
rob_se_Comp <- sqrt(diag(vcovHC(Comp_model, type = "HC1")))
model_summary <- summary(Comp_model) # this is to extract the p values later

#creating a data frame to then create a knitr table
Compliance_DF <- data.frame(variable = c("Treatment"),
                            compliance = c(coef(Comp_model)),
                            Robust_SE = c(rob_se_Comp),
                            P_Value = c(model_summary$coefficients[,4])) #extracting p values

#note - the pvalue is so small that to R is it basically a zero

# getting rid of the intercept column on my data frame
Compliance_DF <- Compliance_DF[-1, ]

# now on to the knitr version to make it look fancy

library(knitr)

# Generate a formatted table
kable(
  Compliance_DF,
  caption = "Compliance Rate Table",
  digits = 4,
  format = "markdown"
)
```

Table 2: Compliance Rate Table

	variable	compliance	Robust_SE	P_Value
treatment	Treatment	0.2536	0.0078	0

Number 5

What is the intent-to-treat (ITT) effect of the OHP experiment on health outcomes? Please create a clearly formatted table that reports ITT regression estimates on 5 relevant health outcomes. The table should include intercept terms (control group means) and their standard errors, treatment effect coefficients and their standard errors, and the number of observations per model. Make sure your standard errors are robust.

For this question, I am going to use these 5 health outcomes [dep_dx_post_lottery , dia_dx_post_lottery, doc_num_mod_inp , bp_sar_inp , rx_num_mod_inp]. After running these, I decided to add two extra variables.

```
outcome_model_1 <- lm(dep_dx_post_lottery ~ treatment , data = OHP)
rob_se_outcome_1 <- sqrt(diag(vcovHC(outcome_model_1, type = "HC1")))

outcome_model_2 <- lm(dia_dx_post_lottery ~ treatment, data = OHP)
rob_se_outcome_2 <- sqrt(diag(vcovHC(outcome_model_2, type = "HC1")))

outcome_model_3 <- lm(doc_num_mod_inp  ~ treatment, data = OHP)
rob_se_outcome_3 <- sqrt(diag(vcovHC(outcome_model_3, type = "HC1")))

outcome_model_4 <- lm(bp_sar_inp ~ treatment, data = OHP)
rob_se_outcome_4 <- sqrt(diag(vcovHC(outcome_model_4, type = "HC1")))

outcome_model_5 <- lm(rx_num_mod_inp   ~ treatment, data = OHP)
rob_se_outcome_5 <- sqrt(diag(vcovHC(outcome_model_5, type = "HC1")))

outcome_model_6 <- lm(chl_inp    ~ treatment, data = OHP)
rob_se_outcome_6 <- sqrt(diag(vcovHC(outcome_model_6, type = "HC1")))

outcome_model_7 <- lm(hbp_dx_post_lottery   ~ treatment, data = OHP)
rob_se_outcome_7 <- sqrt(diag(vcovHC(outcome_model_7, type = "HC1")))

stargazer(
  outcome_model_1 , outcome_model_2, outcome_model_3, outcome_model_4,
  se = list(rob_se_outcome_1 , rob_se_outcome_2, rob_se_outcome_3, rob_se_outcome_4),
  type = "latex",
  header = FALSE,
  title = "Regression Results with Robust Standard Errors",
  column.labels = c("Depression" , "Diabetes", "Doctor Visits", "Blood Pressure"),
  omit.stat = c("f", "ser"),
  digits = 3
  )
```

```
stargazer(
  outcome_model_5, outcome_model_6, outcome_model_7,
  se = list(rob_se_outcome_5, rob_se_outcome_6, rob_se_outcome_7),
  type = "latex",
  header = FALSE,
  title = "Regression Results with Robust Standard Errors",
  column.labels = c("Medications", "Chloresteral","Hypertension"),
  omit.stat = c("f", "ser"),
  digits = 3
)
```

I had to have two separate tables. For some reason when I added that 5th model the stargazer table became too big to print.

Table 3: Regression Results with Robust Standard Errors

	Dependent variable:			
	dep_dx_post_lottery Depression	dia_dx_post_lottery Diabetes	doc_num_mod_inp Doctor Visits	bp_sar_inp Blood Pressure
	(1)	(2)	(3)	(4)
treatment	0.005	0.009***	0.396*	−0.058
	(0.004)	(0.002)	(0.216)	(0.300)
Constant	0.049***	0.012***	5.746***	119.130***
	(0.003)	(0.001)	(0.155)	(0.219)
Observations	12,095	12,186	12,158	12,188
R^2	0.0001	0.001	0.0003	0.00000
Adjusted R^2	0.00003	0.001	0.0002	−0.0001
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: Regression Results with Robust Standard Errors

	Dependent variable:		
	rx_num_mod_inp Medications	chl_inp Chloresteral	hbp_dx_post_lottery Hypertension
	(1)	(2)	(3)
treatment	0.128**	−0.642	0.002
	(0.053)	(0.614)	(0.004)
Constant	1.838***	205.769***	0.057***
	(0.037)	(0.448)	(0.003)
Observations	11,912	12,174	11,945
R^2	0.0005	0.0001	0.00003
Adjusted R^2	0.0004	0.00001	−0.0001
Note:			*p<0.1; **p<0.05; ***p<0.01

Number 6

Can any of your health outcome treatment effect coefficient estimates be directly compared to one another without adjustment? If so, test for whether these coefficients are the same.
In R, use the systemfit procedure we practiced in the Boston Housing data (note that method = "SUR" must be declared in the systemfit command for this to work properly). Don't worry about robust standard errors here. Discuss your findings, thinking carefully about what "statistical significance" of an estimate really means. If not, standardize your outcome variables to make them comparable and do the above.

The three health outcomes that are "post-lottery can be compared to one another because they are all three binary variables in our data. The two outcomes I will test will be the Depression post lottery variable and the diabetes post lottery variable.

```
eq_Depress <-  dep_dx_post_lottery ~ treatment
eq_Diabetes <- dia_dx_post_lottery ~ treatment
eq_system <- list(eq_Depress, eq_Diabetes)
lm_system <- systemfit(eq_system,
                       data = OHP,
                       method = "SUR")
linearHypothesis(lm_system, c("eq1_treatment = eq2_treatment"))
```

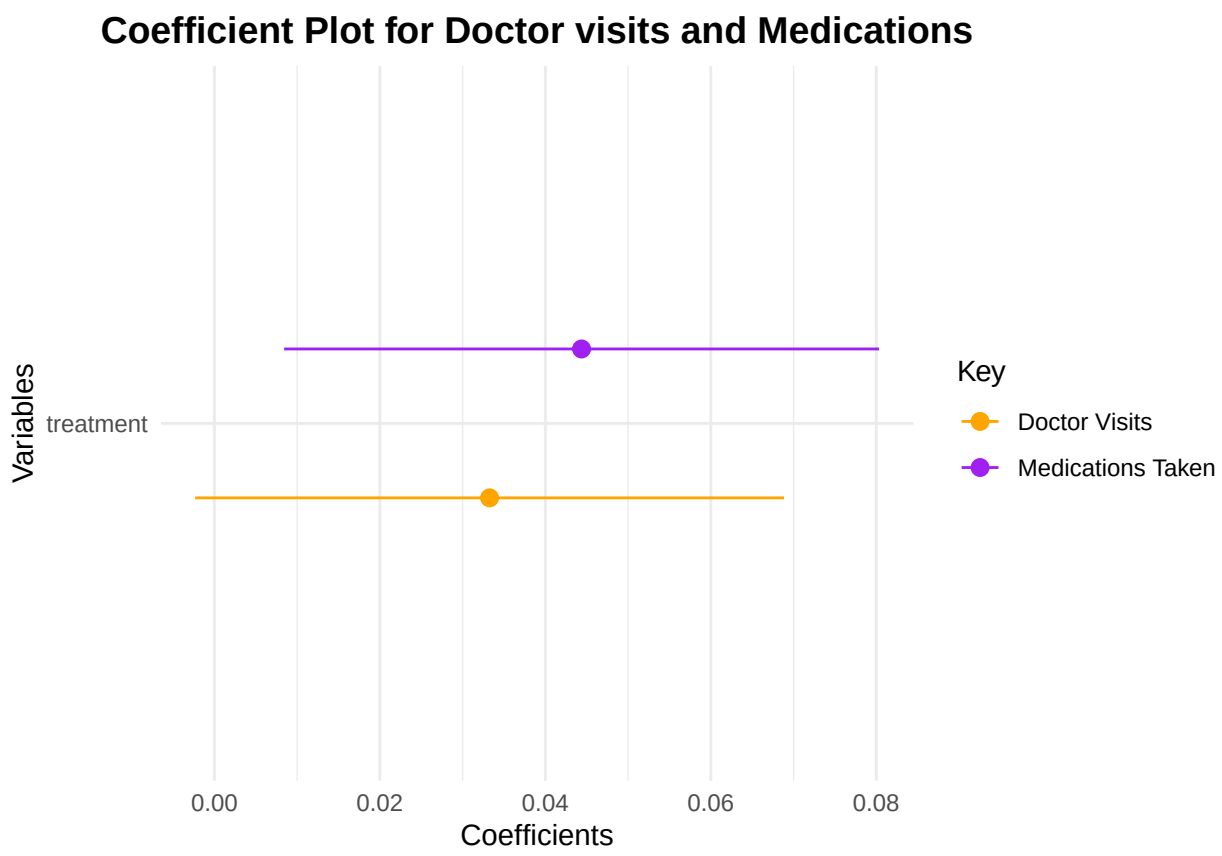
```
## Linear hypothesis test (Theil's F test)
##
## Hypothesis:
## eq1_treatment - eq2_treatment = 0
##
## Model 1: restricted model
## Model 2: lm_system
##
##   Res.Df Df      F Pr(>F)
## 1  24278
## 2  24277  1 0.7633 0.3823
```

In this test, our null hypothesis is that the intent to treat coefficient for depression and the intent to treat coefficient for diabetes are statistically different from one another. What our test shows is that we do not have enough information to reject the null hypothesis and can not say with certainty that the two coefficients are in any way statistically different from each other. We can come to this conclusion because the p-value of the test is 0.38 which is above the threshold we would need to reject the null hypothesis.

Number 7

Create a plot of standardized outcome treatment effect coefficient estimates. If you had to standardize some outcomes to do part 6, then use these. If not, then standardize the outcome variables you used in part 5 but not part 6, and use these. I recommend the "modelplot" command within the "modelsummary" package in R, or you can take take the approach I showed in the Boston Housing Data example, or you can try using the "ggstats" package and "ggcoef_compare" command within it (I think this is a little less clean, but some people prefer it. See here for a description. You should produce something that looks like just the top panel ("Agriculture") of the first example under the "Comparing several models" section, except having the dot fill represent the p-value is silly: skip this).
You should produce something that looks just like the second example under the "Model names as coefficient names" section (except add a dashed line at zero to assist people visualizing the significance of the coefficients).

```r
#standardizing my variables
OHP$medications_std <- scale(OHP$rx_num_mod_inp)
OHP$doctors_std <- scale(OHP$doc_num_mod_inp)

#getting my models to plot
model_doctor <- lm(doctors_std ~ treatment, data=OHP)
model_meds <- lm(medications_std ~ treatment, data=OHP)

#saving my model plot to a variable to be able to use with ggplot

plot <- modelplot(list(model_doctor,model_meds), coef_omit = "(Intercept)")

#using modelplot with ggplot to edit key and colors


plot + theme_minimal() +
    labs(title = "Coefficient Plot for Doctor visits and Medications",
         x = "Coefficients", y = "Variables") +
    scale_color_manual(values = c("orange", "purple"), name = "Key",
                       labels = c("Doctor Visits", "Medications Taken")) +
    theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
```



Number 8

What is the "treatment on the treated" effect (ATET) of the OHP experiment, i.e. the effect among those who applied for Medicaid? Calculate it for every health outcome you chose in question 6. Describe how you calculated this estimate and what it means. You don't need standard errors here for full credit, but if you want to try, see the coding guide below about using the delta method in R or use the "suest" postestimation commands in Stata.

```
# ATET = ITT / Compliance
compliance <- Compliance_DF$compliance #taking the compliance rate I found earlier

#Saving the ITT effects to a variables
Depression_ITT <- outcome_model_1$coefficients # For depression
Diabetes <- outcome_model_2$coefficients # diabetes
Visits <- outcome_model_3$coefficients # doctor visits
BloodPressure <- outcome_model_4$coefficients #blood pressure
Medications <- outcome_model_5$coefficients # medications

#calculating the ATET
ATET_depression <- Depression_ITT/compliance
ATET_Diabetes <- Diabetes/compliance
ATET_Visits <- Visits / compliance
ATET_BloodPressure <- BloodPressure / compliance
ATET_Medications <- Medications / compliance

# setting up a data frame to make a table out of the ATET

ATET_df <- data.frame(DepressionATET = c(ATET_depression),
                      DiabetesATET = c(ATET_Diabetes),
                      VisitsATET = c(ATET_Visits),
                      BloodPressureATET = c(ATET_BloodPressure),
                      MedicationsATET = c(ATET_Medications))

# removing the intercept term in my data frame
ATET_df <- ATET_df[-1, ]

#creating a table using knitr
kable(
  ATET_df,
  caption = "Average Treatment Effect on Treated Table",
  digits = 4,
  format = "markdown"
)
```

Table 5: Average Treatment Effect on Treated Table

	DepressionATET	DiabetesATET	VisitsATET	BloodPressureATET	MedicationsATET
treatment	0.0181	0.0339	1.5603	-0.2297	0.5061

I found the Average Treatment Effect on the Treated by taking the compliance rate that was calculated earlier in the assignment and then taking each beta coefficient found in the first 5 health outcome regressions I did and divided that by the compliance rate.

The estimates above are the effect the treatment had specifically on the people that received the treatment. For the depression variable, people that received the treatment and complied were diagnosed with depression 1.8 percentage points higher than those that did not receive the treatment. We can compare this to the Intent to Treat effect which was only .5 percentage points higher. This means that the effect of the treatment was higher for people that actually enrolled in medicaid versus the people who won the lottery.

Number 9

Do you have to worry about attrition bias in analyzing this data? Explain why or why not.

This is the type of data you would want to analyze attrition bias. The one example I can think of is what is someone becomes disqualified from medicaid because they received it through the lottery and then was able to boost their income because of medicaid. Another way of saying that is that financial impacts of negative health shocks were alleviated through receiving medicaid. That would be a real issue that would change our results. This is just one example but we would need to analyze why people dropped out of each group and if it was non-random or we had high attrition bias, then that would cause some errors with our analysis. This non-random attrition was one problem that the RAND health experiment had.

Number 10

Suppose that you are submitting these results to a general interest journal such as Science for publication. Write an abstract of 200 or fewer words describing what you have found in your analysis of the OHP data, similar to the abstract in Taubman et al. (2014).

In 2008, Oregon expanded its Medicaid program through a lottery system. In this system, everyone that was eligible for the limited expansion was put into a drawing and names were drawn at random. This led to the ability to study the effects of health insurance, specifically medicaid through a randomized controlled experiment. The tools used to determine the effect of winning the lottery was linear regression. We found that health outcomes were not changed within the year time frame of the treatment. However, we did find that there were positive effects when it came to medications prescribed and doctor office visits. There was a 12.8 percentage point increase in medications prescribed and 39.6 percentage point increase in doctor visits within the treatment group. The results show us that there is a positive intent to treat effects as it concerns preventative care. That preventative care could have lasting impacts on health that were not seen in a year's time frame.