# Empirical Project 1

Devin Bunch

Coauthored by Joey and Tillman

4/29/2021

## Contents



### 0.1 Abstract

**ABSTRACT** There exists extreme variety of views on Medicaid: it's controversial. While it lowers out of pocket costs when you go to the doctor. we expect demand will increase. vyoure not facing the true market price so yu could overuse it too. The FDA relies on randomized control trials. To disperce the program, they randomly assign public health insurance program, medicaid coverage to some people and not others. In 2008, Oregon has a medicare programs for low income, uninsured adults who are not eligible for medicaid through other means, but because of budget shortfalls, there was no room for openings. Until they got approved to let in 10,000 additional people into the program. Since the demand for these spots were high, policy makers decided the fairest thing to do was to run a huge public relations campaign to apply for the lottery for medicaid. Tracked the outcomes of both those who won the free medicaid coverage and those who lost. How do we allocate limited resources? It's a perennial problem that governments, nonprofits, and everyone faces. Get rid of privilege and follow federal law to run a lottery for fairness. In this empirical study of rich data, the Oregon Health randomly assigned 30,000 out of 75,000 who showed up the ability to apply for medicaid.

## 0.2 Question 1

> Explain the difference between the variables treatment and ohp_all_ever_survey. Explain why treatment is the treatment variable (Di), rather than ohp_all_ever_survey.

The difference between the treatment and ohp_all_ever_survey is that treatment is individuals that were entered into medicaid specifically through the lottery, whereas ohp_all_ever_survey are those that have been enrolled in medicaid independent of the lottery.

```
#Load packages and libraries

library(pacman)

p_load(readr,dplyr, tidyverse, ggplot2, skimr, haven, stargazer, tidymodels, skimr, janitor, magrittr, c
```

## 0.3 Question 2

> Provide evidence that the OHP lottery really did randomly assign individuals to treatment and control groups. Similar to Table 1 in Taubman et al. (2014), please create a nicely formatted table that reports means of 4 to 6 relevant characteristics for individuals in the control group. Note: Part of this question is to get you to think about which variables should be balanced in a randomized experiment. You need to read carefully through all the variables in the dataset (documentation attached at the end of this file) and decide which 4 to 6 you will summarize.

```
#load data from downloads since Rconsole cannot download ".dta" files
ohp_data <- read_dta("./data/ohp.dta")
```

```
#Skim the data to gain further information
#glimpse(ohp_df)
#ohp_df %>% skim()

#create our regression variables
gender = lm(gender_inp ~ treatment, data = ohp_data)
age = lm(age_inp ~ treatment, data = ohp_data)
not_white = lm(race_nwother_inp ~ treatment, data = ohp_data)
education = lm(edu_inp ~ treatment, data = ohp_data)
medicine = lm(rx_num_mod_inp ~ treatment, data = ohp_data)
cholesterol = lm(chl_inp ~ treatment, data = ohp_data)

#interpreting the table: the constant is the mean of the variable (like, age),
#and the treatment yields the difference in means.
#now make a table with stargazer
```

```
#load data from downloads since Rconsole cannot download ".dta" files
ohp_data <- read_dta("./data/ohp.dta")

#Skim the data to gain further information
#glimpse(ohp_df)
#ohp_df %>% skim()
#names(ohp_df)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:huxtable':
##
##     add_footnote

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```r
#create an object that is just a summary
RecurveSum1 <- ohp_data %>%
    group_by(treatment) %>%
    summarize(
      Gender = mean(gender_inp, na.rm = TRUE),
      Age = mean(age_inp, na.rm = TRUE),
      Nonwhite = mean(race_nwother_inp, na.rm = TRUE),
      Education = mean(edu_inp, na.rm = TRUE),
      Number_of_Medication = mean(rx_num_mod_inp, na.rm = TRUE),
      Cholesterol = mean(chl_inp, na.rm = TRUE))


kable(RecurveSum1, format = "latex", booktabs = TRUE,
        caption = "Weather Normalized (TMY3) Predicted Annual Therms",
        format.args = list(big.mark= ",")) %>%
    kable_styling(latex_options = c("HOLD_position","scale_down"))
```

Table 1: Weather Normalized (TMY3) Predicted Annual Therms

| treatment | Gender | Age | Nonwhite | Education | Number_of_Medication | Cholesterol |
|---|---|---|---|---|---|---|
| 0 | 0.5688121 | 40.60606 | 0.1424648 | 2.238397 | 1.838114 | 205.7692 |
| 1 | 0.5627055 | 40.98638 | 0.1458202 | 2.260072 | 1.966447 | 205.1270 |

This table is called a balance table. A balanced table checks observable characteristics of the treatment and control group to check if they differ. Why would we want to do this? Well if treatment assignment is truly random, then the observable characteristics of the treatment and the control group should be reasonably similar. So they should have similar proportions of men, women, blacks, whites, similar levels of education.

## 0.4 Questions 3 & 4

We want to check to make sure the randomization actually worked. The basics of reporting results from randomized experiments. First we will do a balancing table that shows that treatment and control groups do not differ by demographic characteristics or pre-treatment outcomes. Then we want to find the impact of the treatment on the outcome called the effect estimation. Thus, to do this we want to look at some variables, like cholesterol, and make sure their means are not biased before the treatment begins.

```
stargazer(gender, age, not_white, education, cholesterol,
          title = "Overall Regression Results for Explanatory Variables",
          type = 'latex', robust = TRUE, font.size = "small",
          omit.stat = c("f", "ser"),
          column.sep.width = "6pt")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Apr 30, 2021 - 20:20:02

Table 2: Overall Regression Results for Explanatory Variables

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | gender_inp | age_inp | race_nwother_inp | edu_inp | chl_inp |
| | (1) | (2) | (3) | (4) | (5) |
| treatment | −0.006 | 0.380* | 0.003 | 0.022 | −0.642 |
| | (0.009) | (0.212) | (0.006) | (0.016) | (0.613) |
| Constant | 0.569*** | 40.606*** | 0.142*** | 2.238*** | 205.769*** |
| | (0.006) | (0.153) | (0.005) | (0.012) | (0.443) |
| Observations | 12,229 | 12,228 | 12,190 | 12,218 | 12,174 |
| $R^2$ | 0.00004 | 0.0003 | 0.00002 | 0.0001 | 0.0001 |
| Adjusted $R^2$ | −0.00004 | 0.0002 | −0.0001 | 0.0001 | 0.00001 |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Apr 30, 2021 - 20:20:02

Table 3: Overall Regression Results for Explanatory Variables

| TRUE |
|---|

> Is the balance table consistent with individuals having been randomly assigned to treatment group and control groups? Why or why not? For each of the variables you summarized above, calculate: (i) the difference between the mean in the treatment group and the mean in the control group. (ii) the standard error for the difference in means. Add these as columns two and three to the table you started in question 2. **Already shown above.**

Essentially it is taking a difference of means (of ages) between the treatment and control group and it includes a standard error. So if my regression returns an estimate of the difference in means of 0.380 and a standard error of 0.212 then if I take the ratio of the two (0.380/0.212 = 1.80ish) gives me a t-stat. The t-stat will be statistically significant at the 5% level if $> 1.96$ (10% if $> 1.645$). So using the estimate (difference in means) and the standard error, I can make an inference statement about whether the treatment and control group have statistically significant difference in means. This is important to test how well balanced the random treatment assignment was across ages. A sufficiently randomized treatment assignment would hopefully show no difference between the two groups. However even if treatment is randomized well, slight differences may come up. If a lot of significant differences come up it would suggest we did not randomize our groups well, but it seems we have.

This is useful for designing and implementing randomized evaluations for policy making. From our table, we can gather some of the bigger findings: age is statistically significant at the 90 % confidence interval level.

We are just checking the difference between the control group and the treatment group so as to see if there were predispositions in bias selection.

Note: Difference in Group Means = Treatment Effect +Selection Bias Note: Selection Bias = 0 with if units are randomly assigned to treatment and control groups

## 0.5 Question 5

> 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? Hint: For this question and question 7, you can use the same regression as in question 3, just changing the dependent variable.

```
#to find the compliance rate, regress those enrolled in medicade on the treatment
c_rate <- lm(data= ohp_data, ohp_all_ever_survey ~ treatment)
c_rate
```

```
##
## Call:
## lm(formula = ohp_all_ever_survey ~ treatment, data = ohp_data)
##
## Coefficients:
## (Intercept)    treatment
##      0.1583       0.2536
```

The compliance rate for the OHP experiment is 0.2536. The effect of being assigned to the treatment group on the probability of being enrolled in Medicaid is 25.36%.

## 0.6 Question 6

> What is the intent-to-treat (ITT) effect of the OHP experiment on health outcomes? Please create a nicely formatted table that reports ITT estimates on 4 to 6 relevant health outcomes. Again, part of this question is to get you to think about which 4 to 6 variables could be used as health outcome variables.

This question is asking for a simple regression estimate. Choosing some health factored variables now to look for correlations, we want to regress treatment on: Diagnosed with diabetes after the lottery, Number of doctor's visits, Diagnosed with diabetes after the lottery, Number of prescription medications currently taking, and total cholesterol status.

```
#This reports ITT estimates on the 6 relevant health outcomes we chose
dia <- lm(data=ohp_data, dia_dx_post_lottery~treatment)

doc <- lm(data = ohp_data, doc_num_mod_inp~treatment)

hdp <- lm(data=ohp_data, hbp_dx_post_lottery~treatment)

meds <- lm(data=ohp_data,rx_num_mod_inp~treatment)

chl <- lm(data=ohp_data,chl_inp~treatment)
```

```
stargazer(dia, doc, hdp, meds, chl, title = "Overall Regression Results for Outcome Variables", align =
          omit.stat = c("f", "ser"),
          column.sep.width = "6pt")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Apr 30, 2021 - 20:20:03 % Requires LaTeX packages: dcolumn

Table 4: Overall Regression Results for Outcome Variables

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | dia_dx_post_lottery | doc_num_mod_inp | hbp_dx_post_lottery | rx_num_mod_inp | chl_inp |
| | (1) | (2) | (3) | (4) | (5) |
| treatment | 0.009*** | 0.396* | 0.002 | 0.128** | -0.642 |
| | (0.002) | (0.216) | (0.004) | (0.053) | (0.613) |
| Constant | 0.012*** | 5.746*** | 0.057*** | 1.838*** | 205.769*** |
| | (0.002) | (0.156) | (0.003) | (0.038) | (0.443) |
| Observations | 12,186 | 12,158 | 11,945 | 11,912 | 12,174 |
| $R^2$ | 0.001 | 0.0003 | 0.00003 | 0.0005 | 0.0001 |
| Adjusted $R^2$ | 0.001 | 0.0002 | -0.0001 | 0.0004 | 0.00001 |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Apr 30, 2021 - 20:20:03 % Requires LaTeX packages: dcolumn

Table 5: Overall Regression Results for Outcome Variables

| TRUE |
|---|

The ITT effect is equal to the treatment's impact on each of the variables we chose. For the number of medications a person has, the ITT effect is seen in the table as 0.128, significant at the 95% level. For the number of doctor's visits, we have an ITT effect of 0.396 with a lesser significance level. Horizontally alongside "treatment" we can see our variable's corresponding ITT effects.The ITT is the coefficient is the difference in mean of the variable between the control and the treatment. The intercept is the mean of the control, and the treatment coefficient is the ITT estimate. you get one ITT estimate per variable of interest regressed on treatment.

#Question 7 > 7. What is the "treatment on the treated" effect (ATET) of the OHP experiment, i.e. the effect among those who applied for Medicaid? Estimate it for every health outcome you chose in question 6 and provide some intuition for the calculation of this estimate.

For this question, we are taking the beta coefficients from question 6, and then dividing each of them by the compliance rate found in question 5. This yields the average treatment effect on the treated. The intend to treat effect divided by the compliance rate.

```
ATET = list(c(0.009, 0.396, 0.002, 0.128, -0.642))

#List outcome
Outcome = list(c("Diabetes Post", "Medical Visits", "Hypertension Post", "Medication","Cholesterol" ))

ATET = as.data.frame(ATET)
```

Table 6: ATET

| ATET | Outcome |
|---|---|
| 0.035488959 | Diabetes Post |
| 1.561514196 | Medical Visits |
| 0.007886435 | Hypertension Post |
| 0.504731861 | Medication |
| -2.531545741 | Cholesterol |

```
#Create data frame
Outcome = as.data.frame(Outcome)
Outcome = Outcome %>% mutate(ATET = ATET/.2536)%>% mutate(Outcome = Outcome)
Outcome = Outcome %>% select(ATET, Outcome)
#Output table
kable(Outcome, format = "latex", booktabs = TRUE,
        caption = "ATET",
        format.args = list(big.mark= ","))
```

**Reporting on the ATET's:** The effect for the people who wanted it do not see significant change in their own health outcomes, but we see significant change in "useage" variables related to medicaid. From the table, we can see the treatment's coefficients are equal to the ATET of that column's variable. We see this trend in a surge of medicaid usage, because the use of medication decreases significantly alongside a decrease in the actual usage among the people who got medicaid anyways. The other people who didn't have access to medicaid, and wanted it, were seen to use it more. There was a greater change in medicaid coverage services.

## 0.7 Question 8

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

There's only two years of data since the medicaid coverage expanded afterwards to include everyone and that got rid of the control group. Everyone is now offered medicaid coverage due to the crazy demand. 10th of March 2008 to September 30 2009. In the lectures by Amy Finkelstein, she addresses the controversy surrounding medicaid's expansion and the necessities for Randomized Experiment Designs to be implemented into policy making. We should always worry about a little bit of attrition bias just because of unforeseeable attributes like survey fatigue, when people get treatment, leave and don't come back to complete it, since it is such a short time period causes the attrition rate to be lower but it is still there. hi