EC325 Problem Set 1

Othar Zaldastani II

February 27, 2026

Question 1: The Oregon Health Insurance Experiment

Part A Finkelstein et al. describe previous similar studies as having been hampered by the unobserved differences that exist in the insured and uninsured population. This study randomized participants to control for these differences, allowing them to determine the true causal effect of health insurance on a population.

Part B In 2008, Oregon had a waiting list for low-income adults to join its Medicaid program. They decided to determine who would be allowed into the program by the use of a lottery system. "Winning" the lottery meant the winner would be given the opportunity to apply for Medicaid, not that they would automatically be placed on Medicaid. First, an individual would be selected by the lottery, then they would have to apply for Medicaid. Not all winners would end up actually obtain coverage through Medicaid.

Part C First, Medicaid is less attractive to doctors and hospitals because they pay less. As a result, not all hospitals and doctors accept Medicaid, decreasing the effectiveness of it when compared to private insurance. Second, they mention the effect of Medicaid may not be as strong in places with robust public health clinic systems that provide their services for very low or no cost.

Question 2: Loading and Exploring the Data

Part A Setup and Data Observation

```r
library(tidyverse)
library(dplyr)

# ohie <- read.csv("~/RStudio Stuff/EC325/EC325 Problem Sets/EC325 Problem Set 1/Ohie_sample.csv") # Ma
ohie<- read_csv("~/EC325 Windows Git/data/ohie/ohie_sample.csv") # Windows
```

`ohie` has 23777 observations of 11 variables.

Part B `ohie` Data Representation

```r
head(ohie)
```

```
## # A tibble: 6 x 11
##   person_id treatment female race   hisp educ   hhinc ever_medi any_hosp any_doc
##       <dbl>     <dbl>  <dbl> <chr> <dbl> <chr>  <dbl>     <dbl>    <dbl>   <dbl>
## 1         1         1      0 white     0 hsdeg      1         0        0       0
## 2         2         1      1 white     0 hsdeg      5         0        0       0
## 3         5         1      1 <NA>      1 hsdeg     11         0        0       0
## 4         6         1      0 white     0 colld~     4         0        0       1
## 5         8         0      0 white     0 lths       1         0        0       0
## 6         9         0      1 white     0 hsdeg      5         0        0       1
## # i 1 more variable: good_health <dbl>
```

Each row in `ohie` represents an observation of a unique participant in the study.

Part C Glimpse of the Data

```
glimpse(ohie)
```

```
## Rows: 23,777
## Columns: 11
## $ person_id   <dbl> 1, 2, 5, 6, 8, 9, 10, 13, 14, 23, 25, 26, 27, 28, 29, 30, ~
## $ treatment   <dbl> 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0~
## $ female      <dbl> 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1~
## $ race        <chr> "white", "white", NA, "white", "white", "white", "white", ~
## $ hisp        <dbl> 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0~
## $ educ        <chr> "hsdeg", "hsdeg", "hsdeg", "colldeg", "lths", "hsdeg", "mt~
## $ hhinc       <dbl> 1, 5, 11, 4, 1, 5, 5, 2, 10, 9, 4, 2, 4, 11, 5, 1, 14, 6, ~
## $ ever_medi   <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ any_hosp    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ any_doc     <dbl> 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0~
## $ good_health <dbl> 1, 1, 1, 1, NA, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, ~
```

R assigned the character data type to the `educ` variable and a numeric variable for `treatment`.

Part D Summary of `ohie`

```
summary(ohie)
```

```
##    person_id       treatment          female           race          
##  Min.   :    1   Min.   :0.0000   Min.   :0.000   Length:23777      
##  1st Qu.:18532   1st Qu.:0.0000   1st Qu.:0.000   Class :character  
##  Median :37243   Median :0.0000   Median :1.000   Mode  :character  
##  Mean   :37303   Mean   :0.4967   Mean   :0.595                     
##  3rd Qu.:56007   3rd Qu.:1.0000   3rd Qu.:1.000                     
##  Max.   :74921   Max.   :1.0000   Max.   :1.000                     
##                                   NA’s   :81                        
##       hisp            educ               hhinc          ever_medi     
##  Min.   :0.0000   Length:23777       Min.   : 1.000   Min.   :0.0000  
##  1st Qu.:0.0000   Class :character   1st Qu.: 3.000   1st Qu.:0.0000  
##  Median :0.0000   Mode  :character   Median : 6.000   Median :0.0000  
##  Mean   :0.1171                      Mean   : 6.696   Mean   :0.2004  
##  3rd Qu.:0.0000                      3rd Qu.: 9.000   3rd Qu.:0.0000  
##  Max.   :1.0000                      Max.   :22.000   Max.   :1.0000  
##  NA’s   :248                         NA’s   :1284                     
##     any_hosp          any_doc        good_health    
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.00000   Median :1.0000   Median :1.0000  
##  Mean   :0.07133   Mean   :0.6042   Mean   :0.5696  
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000  
##  NA's   :168       NA's   :249      NA's   :380     
```

i) The variables with the missing values are `hhinc`, `goodhealth`, `any_doc`, `hisp`, `any_hosp`, and `female`.\
ii) `hhinc` has the most missing values with 1284.\
iii) The range of household incomes seems low. $1000 is a very low annual household income for anyone in the United States, and a maximum income of $22,000 seems low, even for people on Medicaid. The number that is most surprising to me is the median income of $6,000 which implies many of the people on Medicaid work part time jobs with not many hours.

Question 3: Descriptive Statistics

Part A Data Means

```
vars <- c("female", "any_doc", "any_hosp", "good_health")
for (var in vars) {
    cat(var, ":", round(mean(ohie[[var]], na.rm = TRUE), 2), "\n")
  }
```

```
## female : 0.59 
## any_doc : 0.6 
## any_hosp : 0.07 
## good_health : 0.57 
```

`female` = 0.59 means 59% of participants are Female.\ `any_doc` = 0.6 means 60% of participants reported a visit to the doctor in the previous six months.\ `any_hosp` = 0.07 means 7% of participants reported being hospitalized in the previous six months.\ `good_health` = 0.57 means 57% of participants reporting being in good health.

Part B Count of `educ`

```
count(ohie, educ)
```

```
## # A tibble: 5 x 2
##   educ        n
##   <chr>   <int>
## 1 colldeg  2601
## 2 hsdeg   11434
## 3 lths     3907
## 4 mths     5072
## 5 <NA>      763
```

The most common education level present in the sample is a high school degree, making up nearly half of all observations in the sample. Given that education has a positive correlation with income, the distribution seems reasonable. 41% of participants do not have a high school degree and only 11% have a college degree.

Part C Creation of a College Binary Variable

```
ohie <- ohie |>
  mutate(college = ifelse(educ == "colldeg", 1, 0))

mean(ohie$college, na.rm = TRUE)
```

```
## [1] 0.1130182
```

≈ 11.3% of participants received a college degree.

```
ohie |>
  group_by(female) |>
  summarize(mean_good_health = mean(good_health, na.rm = TRUE))
```

Part D

```
## # A tibble: 3 x 2
##   female mean_good_health
##    <dbl>            <dbl>
## 1      0            0.563
## 2      1            0.574
## 3     NA            0.562
```

There is a slight difference, with females reporting being in good health at a rate 1% higher than males. This could be due to males working in more physical jobs or possibly being on disability due to an injury sustained during their job. We could run a difference in means test to determine if this difference is statistically significant.

Question 4: Selection Bias and Causal Inference

Part A Y_{1i} is the outcome for i if treated and Y_{0i} is the outcome for i if not treated. The fundamental problem of causal inference is that we can only observe one or the other, so the true effect cannot be entirely determined. You can never see both the treatment and control results on one person/observation.

Part B The difference in means can be decomposed into two parts. The causal effect term which represents the true impact of having Medicaid on the participants who enrolled and the selection bias term which represents differences that existed without Medicaid. People who enrolled in Medicaid may have seen a greater need for it, whether due to worse health, more tenuous financial situations or other reasons. This causes a confounding effect in the results of the experiment due to the fundamental problem of causal inference.

Part C The Oregon lottery solves the selection bias problem by randomizing the members of the population who get access to Medicaid enrollment. This helps to solve the selection bias problem by randomly assigning people into the treatment group, regardless of personal motivation or other potentially confounding characteristics. The Law of Large Numbers is relevant here because it means as the number of participants increase, the average characteristics of lottery winners and losers shrinks.

Part D I would not expect the treatment and control groups to be balanced. Selection into the groups would no longer be randomized with many potential selection bias effects present. People in worse health have a stronger incentive to enroll immediately, individuals who are not well-informed or who have less free time would likely hear about the program later. First-come, first-served would add in a very meaningful selection bias effect.

Question 5: Checking Balance

Female

```
# Code is the same format for all tables, with only one variable changed
ohie |>
  group_by(treatment) |>
  summarize(mean_female = mean(female, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   treatment mean_female
##       <dbl>       <dbl>
## 1         0       0.602
## 2         1       0.588
```

Hispanic

```
## # A tibble: 2 x 2
##   treatment mean_hisp
##       <dbl>     <dbl>
## 1         0     0.114
## 2         1     0.120
```

Household Income

```
## # A tibble: 2 x 2
##   treatment mean_household_income
##       <dbl>                 <dbl>
## 1         0                  6.54
## 2         1                  6.85
```

College Degree

```
## # A tibble: 2 x 2
##   treatment mean_college
##       <dbl>        <dbl>
## 1         0        0.116
## 2         1        0.110
```

Part B The means are relatively similar across the treatment and control groups. The average difference in means is approximately 4%. This is an indication of successful randomization.

Part C If the observable characteristics are effectively randomized, it likely means that all variables are randomized, regardless of their observability. The more people in a randomized trial, the more likely the trial is to be truly and effectively randomized because random differences between people cancels out more the larger the size of the pool.

Question 6: Estimating the Causal Effect of the Lottery

```
means_table <- ohie |>
  group_by(treatment) |>
  summarize(
    ever_medi   = mean(ever_medi,   na.rm = TRUE),
    any_doc     = mean(any_doc,     na.rm = TRUE),
    any_hosp    = mean(any_hosp,    na.rm = TRUE),
    good_health = mean(good_health, na.rm = TRUE)
  )

round(means_table, 4)
```

Part A

```
## # A tibble: 2 x 5
##   treatment ever_medi any_doc any_hosp good_health
##       <dbl>     <dbl>   <dbl>    <dbl>       <dbl>
## 1         0     0.102   0.576   0.0704       0.543
## 2         1     0.300   0.633   0.0722       0.597
```

Part B All outcomes show increases with treatment, with `any_hosp` showing an increase of 2.5%, `good_health` and `any_doc` both showing approximately 10% increases, and `ever_medi` showing an increase of nearly 300%. The direction of all changes make sense. I would expect Medicare enrollment to slightly increase the number of people who go to a hospital (potentially swaying people on the fence due to financial constraints), the number of people in good health as they now have improved access to healthcare, the number of people who visit a doctor for the same reason as hospitalization, and of course, the number of people who have ever enrolled in Medicaid, as insurance coverage was a primary focal point of the study. The outcome that shows the largest treatment effect is `ever_medi`, which is the variable showing whether a participant enrolled in Medicaid during the course of the study.

```
summary(lm(good_health ~ treatment, data = ohie))
```

Part C

```
##
## Call:
## lm(formula = good_health ~ treatment, data = ohie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5967 -0.5427  0.4033  0.4573  0.4573
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.542719   0.004559  119.05   <2e-16 ***
## treatment   0.053981   0.006465    8.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4944 on 23395 degrees of freedom
##   (380 observations deleted due to missingness)
## Multiple R-squared:  0.002971,   Adjusted R-squared:  0.002929
## F-statistic: 69.72 on 1 and 23395 DF,  p-value: < 2.2e-16
```

The intercept is 0.543 and the slope coefficient for `treatment` is 0.054. The intercept represents the model's expected percentage of people in good health without winning the lottery, and the slope coefficient represents the percentage increase in good health of people in the population that won the lottery. It can also be thought of as the probability increase a person will be in good health depending on their lottery status.

Part D The differences are exactly the same. The linear model gives the exact same result as the difference in means because for a linear model with a binary (dummy) variable, the regression line is the line between the two group means. Since the variable has no possible middle values, there is no way for the linear model to be different from a difference in means test.

Part E Based on these results, it appears that winning the Oregon health insurance lottery causes increases in self-reported health. Participants that won the lottery experienced an increase in self-reported good health of $\approx$ 10%. That is a large increase in practical terms. With something as complex and multifaceted as personal health, an increase of 10% is an impressive result.

Question 7: Reflection

Part A The effects of winning the lottery versus having Medicaid may be different between these two groups because not all who won the lottery end up with Medicaid. Some non-winners may still end up with insurance through their employer or programs such as Medicare. Also, the lottery only gives winners access to the application for Medicaid which does not give them access to healthcare directly, even if they win. Some Medicaid enrollees may still have limited access to hospitals which would mute its effect.

Part B Some outcomes from this study worth considering are the long term effects of the study. The study only tracks participants for around a year, which is not long enough to see how chronic conditions improve and other potential diseases are caught or treated early, which would improve health outcomes. This program also decreases healthcare inequality and could potentially help treat some disability causing chronic conditions over time which could be a massive benefit for the state due to decreased disability spending and increased labor force participation.

Part C While it is possible these results could generalize to other populations or settings, there are many aspects of this study and population that might make this less likely. First, a new population could have a higher insurance rate or better access to care which would decrease the effect because baseline healthcare statistics are already high relative to their maximum. A location with lower rates of Medicaid acceptance would see a smaller positive effect. Also, if the program is implemented in states with more untreated conditions and lower healthcare utilization numbers, the effect could be larger.