ECON 474: Econometrics of Policy Evaluation

Homework 1

John Rosak (rosak2)

02/16/2022

Contents

A Causal Question [30 points]

In the middle of 2014, The City of Fortaleza began an ongoing urban renewal project called "Areninhas." The intervention consists of synthetic football turf, sometimes with a playground and an outdoor gym. Besides, there is a substantial increase in street lighting. Say the Mayor wants to know whether this public policy reduces violence, and the City Hall hires you to evaluate the Areninhas project. Let's start answering a couple of questions:

a. What are the outcome and treatment variables?

This case is looking at the potential reduction in crime in the city, particularly in the areas with low street light areas. The treatment variable is building a synthetic football turf with lights and the outcome would hopefully be the reduction in crime.

b. What are the potential outcomes in this case?

The potential outcomes in this case is finding out that building turfs in low light areas could decrease crime in the area. We could also discover that there is no change to the amount of crime due to turfs.

c. What plausible causal channel runs directly from the treatment to the outcome?

A plausible channel is that there is more light in darker areas so there are safer spaces to be at night, and also football fields give people who might commit crimes something to do and be distracted with.

d. Can you think about possible sources of **selection bias** in the naive comparison of outcomes by treatment status? Which way would you expect the bias to go and why?

There is possible selection bias in the treatment effect in picking places that have better means to afford a new football field. Places like these could experience a reduction in crime or lower crime because they have a proportionately more wealthy population which could lead to incorrect comparisons.

Now, say the City Hall decided to study crime prevention through environmental design. Assume that city blocks were randomized, and 25 got the football fields, while the rest are in the control group.

e. How does randomization solve the selection bias problem you just mentioned?

By randomizing where the fields go, they don't have the option of picking areas that will naturally have less crime or more crime, the areas will hopefully be equally spread out for accurate data.

f. What can you say about the external validity of this study? Can you think about scenarios where the internal validity of this study is violated?

This will depend on the findings, however if this case comes out positively it would be possible to place football fields in the locations that didn't have them to help reduce crime. The internal validity of these cases can be easily violated as with a project this big there are a lot of variables including different levels or types of crimes in areas that could be hard to compare.

Randomized Trials [70 points]

Racial Discrimination in the Labor Market [35 points]

We will use a dataset here from a randomized experiment conducted by Marianne Bertrand and Sendhil Mullainathan for this question. The researchers sent 4,870 fictitious resumes out to employers in response to job adverts in Boston and Chicago in 2001. They varied only the names of job applicants while leaving other relevant candidates' attributes unchanged (i.e., candidates had similar qualifications). Some applicants had distinctly white-sounding names such as Greg Baker and Emily Walsh, whereas other resumes contained stereotypical black-sounding names such as Lakisha Washington or Jamal Jones. Hence, any difference in callback rates can solely be attributed to name manipulation.

a. Illustrate this problem using the Potential Outcomes Framework

Hint: What is the unit of observation? What is the treatment D_i and the observed outcome Y_i? What are the potential outcomes?

We're observing the callback rate for each applicant while the treatment is giving an application a black or white sounding name. The outcome could be equal call back rates or that there is a disparity between the callback rates of balck sounding names and white sounding names.

b. Create a dummy variable named `female` that takes one if `sex=="f"`, and zero otherwise.

```
library(tidyverse)
resume = readRDS("resume.RDS")

resume$female = ifelse(resume$sex == 'f', 1, 0)
```

c. The dataset contains information about candidates' education (`education`), years of experience (`yearsexp`), military experience (`military`), computer and special skills (`computerskills` and `specialskills`), a dummy for gender (`female`), among others. Summarize that information by getting average values by `race` groups.

```
resume %>%
  group_by(race)%>%
  summarise(Education = mean(education),
            Experience = mean(yearsexp),
            Military = mean(military),
            'Computer Skills' = mean(computerskills),
            'Special Skills' = mean(specialskills),
            Female = mean(female))
```

```
## # A tibble: 2 x 7
##   race  Education Experience Military ‘Computer Skills‘ ‘Special Skills‘ Female
##   <chr>     <dbl>      <dbl>    <dbl>             <dbl>            <dbl>  <dbl>
## 1 b          3.62       7.83   0.102              0.832            0.327  0.775
## 2 w          3.62       7.86   0.0924             0.809            0.330  0.764
```

d. Do `education`, `yearsexp`, `military`, `computerskills`, `specialskills` and `female` look balanced between race groups? Use `t.test()` to formally compare resume characteristics and interpret its output. Why do we care about whether those variables are balanced?

```
black = resume[resume$race == 'b', ]
white = resume[resume$race == 'w', ]

t.test(black$education, white$education)
```

```
##
##  Welch Two Sample t-test
##
## data:  black$education and white$education
## t = -0.24048, df = 4855.4, p-value = 0.81
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04510429  0.03524803
## sample estimates:
## mean of x mean of y
##  3.616016  3.620945
```

```
t.test(black$yearsexp, white$yearsexp)
```

```
##
##  Welch Two Sample t-test
##
## data:  black$yearsexp and white$yearsexp
## t = -0.18462, df = 4867.1, p-value = 0.8535
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3101545  0.2567664
## sample estimates:
## mean of x mean of y
##  7.829569  7.856263
```

```
t.test(black$military, white$military)
```

```
##
##  Welch Two Sample t-test
##
## data:  black$military and white$military
## t = 1.1129, df = 4858.8, p-value = 0.2658
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.007193736  0.026084906
## sample estimates:
##  mean of x  mean of y
## 0.10184805 0.09240246
```

```
t.test(black$computerskills, white$computerskills)
```

```
##
##  Welch Two Sample t-test
##
## data:  black$computerskills and white$computerskills
## t = 2.1664, df = 4854.9, p-value = 0.03033
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.002264635 0.045373969
## sample estimates:
## mean of x mean of y
## 0.8324435 0.8086242
```

```
t.test(black$specialskills, white$specialskills)
```

```
##
##  Welch Two Sample t-test
##
## data:  black$specialskills and white$specialskills
## t = -0.21349, df = 4868, p-value = 0.831
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02927347  0.02352398
## sample estimates:
## mean of x mean of y
## 0.3273101 0.3301848
```

```
t.test(black$female, white$female)
```

```
##
##  Welch Two Sample t-test
##
## data:  black$female and white$female
## t = 0.88413, df = 4866.7, p-value = 0.3767
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01299869  0.03435393
## sample estimates:
## mean of x mean of y
## 0.7745380 0.7638604
```

All of the tests have fairly high p-values which indicates that the variables aren't statistically significant from each other, some being extremely close.

e. The output of interest in the data set is `call` - a dummy that takes one if the candidate was called back. Use `t.test()` to compare callbacks between White names and Black names. **Is there a racial gap in the callback?**

```
t.test(black$call, white$call)
```

```
##
##  Welch Two Sample t-test
##
## data:  black$call and white$call
## t = -4.1147, df = 4711.6, p-value = 3.943e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04729503 -0.01677067
## sample estimates:
##  mean of x  mean of y
## 0.06447639 0.09650924
```

From the t test, and having a p-value of .00004, we can tell that their means are statistically different. White sounding names got a much higher call back on average with similar credentials as proven in the previous question.

f. Now, run a regression of `call` on `race`, `education`, `yearsexp`, `military`, `computerskills`, `specialskills`, and `female`. Does the estimate related to race change too much? What is the explanation for that behavior?

```
lm(resume$call ~ resume$race + resume$education + resume$yearsexp + resume$military + resume$
```

```
##
## Call:
## lm(formula = resume$call ~ resume$race + resume$education + resume$yearsexp +
##     resume$military + resume$computerskills + resume$specialskills +
##     resume$female)
##
## Coefficients:
##           (Intercept)            resume$racew       resume$education        resume$yearsexp
##              0.008644                0.031389               0.005050               0.003414
##       resume$military  resume$computerskills   resume$specialskills          resume$female
##              0.006885               -0.019514               0.066585               0.005927
```

From our regression, we can tell that there is an increase in the callback rate if the applicant has a white sounding name.

A/B testing in Practice [30 (+ 10) points]

Let's say you work as a Data Scientist at MeetMeAtTheQuad, a dating app startup company from UIUC alumni. You want to measure the causal effect of the dating app like button size on a couple of important metrics. You decided to run an A/B test to check whether the app developers should increase the like button or not.

a. Define some key metrics you would like to evaluate in this experiment

In this case I'd like to evaluate the number of likes that people give, so we'd use the number of likes, and potentially the number of matches. Both are the main reasons people stay on the app so it'd be in their best interest to increase those numbers.

b. What is the experimental unit? What will the treatment and control group see while using the app? Set up the hypothesis testing you have in mind (i.e., what is your H_0 and H_a)?

The experimental unit will be the current app users, maybe make some of them opt to try/test "new features." The treatment group will see a slightly larger like button and the control group will see the regular size like button. Our null hypothesis is that there is no change in the number of likes (clicking the like button) that a user does, the alternative hypothesis is that users will increase the number of clicks or likes that they give.

c. One essential part of designing an experiment is knowing the sample size needed to test your hypothesis. Say you are running a `t.test()` on the number of likes per user to check differences between the control and treatment groups. Using the `pwr` package, **find the sample size required for this experiment**.

```
library(pwr)
pwr.t.test(d = 0.5, sig.level = 0.05, power = 0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

From the pwr test we find that n is 63.76561

Note: assume a power equal to 0.8, a significance level of .05, two groups, and a minimum effect size of 0.5.

d. What happens with your answer in c) when you try to detect a minimum effect of 0.1?

There is a large increase in n to 1570.733

```
pwr.t.test(d = 0.1, sig.level = 0.05, power = 0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 1570.733
##               d = 0.1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

e. Suppose you saw a statistically significant increase in the **number of likes** in the treatment group, but you did not see any effect on the number of matches. What might be the explanation for this pattern?

Hint: think about the two sides involved. To have a match, two persons need to like each other.

This makes sense because while you can increase the number of likes people could be giving out, it doesn't change the preferences of the user. The matches they were getting previously are going to be the same number of matches they're getting today regardless of the number of likes.

Power Analysis Relationships [+10 points]

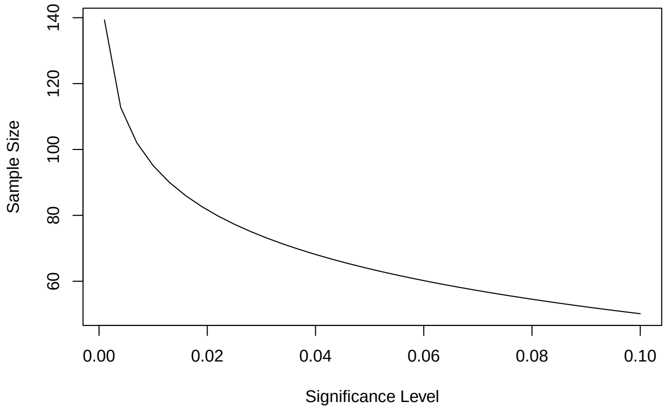
Time for some simulations! Power analysis is a necessary procedure to conduct during the design phase of an experiment. I will establish the first power analysis relationship. Then, you do two more.

1. Holding power and effect size constant, with a higher significance level, you need a smaller sample

Let's simulate the required sample size for different significance levels (ranging from 0.001 to 0.1)

```r
library(pwr)
### alpha is a vector of numbers from 0.001 to 0.1
alpha<-seq(from=0.001, to=0.1, by=0.003)
sample<-matrix(NA, ncol=1, nrow=34)
for(i in 1:length(alpha)){
  sample[i,1]<-pwr.t.test(d = 0.5,
                                    sig.level = alpha[i],
                                    power = 0.8)$n
}

data<-data.frame(alpha, sample)
plot(y=data$sample, x=data$alpha,
     type="l",
     ylab='Sample Size',
     xlab='Significance Level')
```



Now, it is your turn! Show the following:

2. Holding significance level and effect size constant, more power will require more data

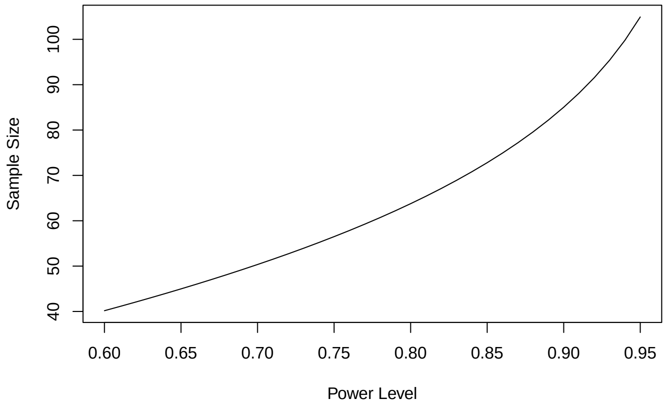
Hint: find the sample size n required for a range of power values (e.g., power from 0.60 to 0.95 by=0.01).

```
library(pwr)
### powerLevel is a vector of numbers from 0.60 to 0.95
powerLevel<-seq(from=0.60, to=0.95, by=0.01)
powerLevel ## There are 36 numbers from 0.60 to 0.95 counting up by 0.01
```

```
##  [1] 0.60 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70 0.71 0.72 0.73 0.74 0.75 0.76
## [20] 0.79 0.80 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89 0.90 0.91 0.92 0.93 0.94 0.95
```

```
sample<-matrix(NA, ncol=1, nrow=36)
for(i in 1:length(powerLevel)){sample[i,1]<-pwr.t.test(d = 0.5,
                                                       sig.level = 0.05,
                                                       power = powerLevel[i])$n
}

data<-data.frame(powerLevel, sample)
plot(y=data$sample, x=data$power, type="l", ylab='Sample Size', xlab='Power Level')
```



3. Holding power and significance level constant, the larger the effect size between groups, the smaller the sample you need to find a statistically significant result (i.e., $p - value < 0.5$)

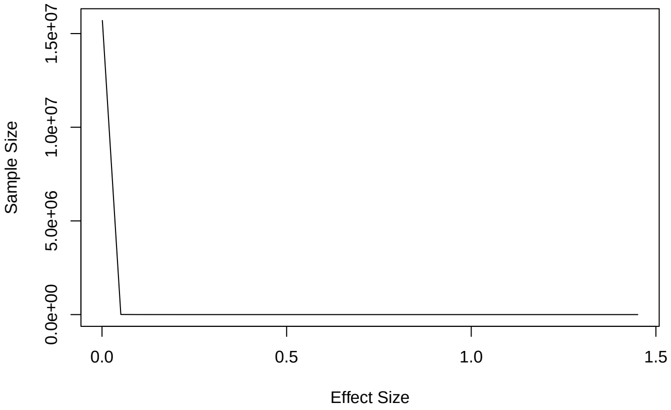
Hint: find the sample size n required for a range of minimum effect values(e.g., d from 0.001 to 1.5 by=0.05).

```
library(pwr)
### effect is a vector of numbers from 0.001 to 1.5 by 0.05
effect<-seq(from=0.001, to=1.5, by=0.05)
effect ## There are 36 numbers from 0.60 to 0.95 counting up by 0.01
```

```
##  [1] 0.001 0.051 0.101 0.151 0.201 0.251 0.301 0.351 0.401 0.451 0.501 0.551 0.601 0.651 0
## [17] 0.801 0.851 0.901 0.951 1.001 1.051 1.101 1.151 1.201 1.251 1.301 1.351 1.401 1.451
```

```
sample<-matrix(NA, ncol=1, nrow=30)
for(i in 1:length(effect)){sample[i,1]<-pwr.t.test(d = effect[i],
                                                        sig.level = 0.05,
                                                        power = 0.8)$n
}

data<-data.frame(effect, sample)
plot(y=data$sample, x=data$effect, type="l", ylab='Sample Size', xlab='Effect Size')
```


Sample Size
1.5e+07
1.0e+07
5.0e+06
0.0e+00
0.0
0.5
1.0
1.5
Effect Size