Final Exam

Vinish Shrestha

12/8/2021

```
user = 2
if(user == 1) {
  datapath <- "/home/vinish/Dropbox/Demand for women health products/dta"
} else {
  datapath <- "/home/user1/Dropbox/health economics teaching/Problem Set 2021"
}
```

Section 1. Data Creation

Refer to the Difference-in-Differences Lecture; you can find the slides https://github.com/vinishshrest/Health-Economics-Course/tree/main/Lecture%20Slides%202020. Here, we are going to analyze a simulated version of Snow's study.

In summary, Snow hypothesized that cholera is transmitted through water. He utilized a natural experiment – among two major water companies (SV and Lambeth), Lambeth changed its water source in 1852 to a cleaner location that was not contaminated by the sewage. Then he implemented difference-in-differences approach by comparing Lambeth (before and after 1852) vs. SV (before and after 1852).

I have simulated Snow's data below to help understand the difference-in-differences setup.

Simulate SV 1849 and 1854

```
set.seed(1)
SV_func <- function(mean_deaths, sd_deaths, n, year, mean_pop, sd_pop) {
SVonly <- round(rnorm(mean = mean_deaths, sd = sd_deaths, n = n), digits = 0)
SV_district <- paste("SV", seq(1:n), sep = "_")
year <- rep(year, length(SVonly))
population <- round(rnorm(mean = mean_pop, sd = sd_pop, n = n), digits = 0) + SVonly*10
SV <- data.frame(cbind(deaths = SVonly, district_code = SV_district, year = year, population = populati
SV <- SV %>% mutate(deaths = as.numeric(as.character(deaths)),
                            population = as.numeric(as.character(population)))
return(SV)

}
# SV1849
SV1849 <- SV_func(mean_deaths = 120, sd_deaths = 50, n = 20, year = 1849, mean_pop = 8045, sd_pop = 100
head(SV1849)
```

```
##   deaths district_code year population
## 1     89          SV_1 1849       9854
## 2    129          SV_2 1849      10117
```

```
## 3     78          SV_3 1849       8900
## 4    200          SV_4 1849       8056
## 5    136          SV_5 1849      10025
## 6     79          SV_6 1849       8779
```

```
(sum(SV1849$deaths)/sum(SV1849$population))*10000
```

```
## [1] 138.746
```

```
write.csv(SV1849, paste(datapath, "SV1849.csv", sep = "/"), row.names = FALSE)

# SV1854
SV1854 <- SV_func(mean_deaths = 60, sd_deaths = 50, n = 20, year = 1849, mean_pop = 8045, sd_pop = 1000
head(SV1854)
```

```
##   deaths district_code year population
## 1     52          SV_1 1849      10967
## 2     47          SV_2 1849       8476
## 3     95          SV_3 1849       9685
## 4     88          SV_4 1849       8953
## 5     26          SV_5 1849       7562
## 6     25          SV_6 1849       8484
```

```
write.csv(SV1854, paste(datapath, "SV1854.csv", sep = "/"), row.names = FALSE)
(sum(SV1854$deaths)/sum(SV1854$population))*10000
```

```
## [1] 75.93889
```

Simulate Lambeth 1849 and 1854

```
set.seed(1)
Lam_func <- function(mean_deaths, sd_deaths, n, year, mean_pop, sd_pop) {
Lamonly <- round(rnorm(mean = mean_deaths, sd = sd_deaths, n = n), digits = 0)
Lam_district <- paste("Lambeth", seq(1:n), sep = "_")
year <- rep(year, length(Lamonly))
population <- round(rnorm(mean = mean_pop, sd = sd_pop, n = n), digits = 0) + Lamonly*15
Lam <- data.frame(cbind(deaths = Lamonly, district_code = Lam_district, year = year, population = popula
Lam <- Lam %>% mutate(deaths = as.numeric(as.character(deaths)),
                          population = as.numeric(as.character(population)))
return(Lam)

}
# Lam1849
Lam1849 <- Lam_func(mean_deaths = 200, sd_deaths = 50, n = 20, year = 1849, mean_pop = 9045, sd_pop = 1
head(Lam1849)
```

```
##   deaths district_code year population
## 1    169     Lambeth_1 1849      12683
## 2    209     Lambeth_2 1849      13119
## 3    158     Lambeth_3 1849      11504
## 4    280     Lambeth_4 1849      10858
## 5    216     Lambeth_5 1849      13029
## 6    159     Lambeth_6 1849      11363
```

```
(sum(Lam1849$deaths)/sum(Lam1849$population))*10000
```

```
## [1] 172.0061
```

```
write.csv(Lam1849, paste(datapath, "Lambeth1849.csv", sep = "/"), row.names = FALSE)

# Lam1854
Lam1854 <- Lam_func(mean_deaths = 40, sd_deaths = 50, n = 20, year = 1849, mean_pop = 9045, sd_pop = 120
head(Lam1854)
```

```
##   deaths district_code year population
## 1     32     Lambeth_1 1849      12407
## 2     27     Lambeth_2 1849       9403
## 3     75     Lambeth_3 1849      10998
## 4     68     Lambeth_4 1849      10099
## 5      6     Lambeth_5 1849       8243
## 6      5     Lambeth_6 1849       9347
```

```
write.csv(Lam1854, paste(datapath, "Lambeth1854.csv", sep = "/"), row.names = FALSE)
(sum(Lam1854$deaths)/sum(Lam1854$population))*10000
```

```
## [1] 47.5614
```

Section 2. Data work

You have 4 data files on blackboard for this exercise: “SV1849.csv”, “SV1854.csv”, “Lambeth1849.csv”, “Lambeth1854.csv”. Open each of these files and store them in R. For example:

```
#Lam1849 <- read.csv(paste(datapath, "Lambeth1849.csv", sep = "/"), row.names = FALSE)
#head(Lam1849)
```

File description:

- X = row name
- deaths = number of cholera related deaths
- district_code = disctrict code. Note that district code begins with the name of the water company (e.g., Lambeth_1).
- year = year (1849 vs 1854)
- population = popuation of the district

Questions

a. Calculate the total number of deaths (due to cholera for each data file) and store them as objects. Be clear with the choice of names that you give. These need to be intuitive.
b. Calcuate the number of deaths per 10,000 people using the population values. Do this for each data file. Store them as objects.
c. Estimate the first (Lambeth before and after) and second (SV before and after) differences, respectively.
d. Then estimate the difference-in-differences estimates (per 10,000 people). Store the estimate as "DD."
e. Interpret your findings from the first difference.
f. Next, bring all 4 data files together. You can do:

```
bigdata <- rbind(SV1849, SV1854, Lam1849, Lam1854)
head(bigdata)
```

```
##   deaths district_code year population
## 1     89          SV_1 1849       9854
## 2    129          SV_2 1849      10117
## 3     78          SV_3 1849       8900
## 4    200          SV_4 1849       8056
## 5    136          SV_5 1849      10025
## 6     79          SV_6 1849       8779
```

g. Create an indicator for after (year = 1954).
h. Create an indicator for Lambeth districts.
i. Create DD variable as $dd = After * Lambeth$.
j. Run the following regression: $deaths = \alpha + \beta DD + Lambeth + After$.
k. Next control for population. Run the following regression: $deaths = \alpha + \beta DD + \gamma Lambeth + \eta After + \sigma log(population) + \epsilon$.
l. Talk about the DD estimate in k.
m. Thoughts: DD assumption