API210-PS4

Kohei Takata

4/15/2022

```r
## install packages if necessary

#install.packages("ggplot2")
#install.packages("tidyverse")
#install.packages("kableExtra")
#install.packages("psych")
#install.packages("sf")
#install.packages("rnaturalearth")
#install.packages("rnaturalearthdata")
#install.packages("rgeos")

## import packages
library(tidyverse)
library(haven)
library(psych)
library(zoo)
library(gridExtra)
library(kableExtra)
library(sf)
library(rnaturalearth)
library(rnaturalearthdata)
library(rgeos)
```

```r
data_ori <- read_dta("munic.dta")
```

15: Create summary statistics for five variables of your choice.

Variables: r_util94 / r_util98/ r_util02 / voters96 / income

```r
table_15 <- data_ori %>%
            select(r_util94, r_util98, r_util02, gini, income) %>%
            describe() %>%
            select(c("n", "mean", "sd", "median", "min", "max"))

rownames(table_15) <- data_ori %>%
                      select(r_util94, r_util98, r_util02, gini, income) %>%
                      map_dfc(attr, "label")   ##extract labels from attributes

round(table_15 ,2)%>%
  kbl(caption = "Descriptive Statistics") %>%
```

```
  kable_classic(full_width = F, html_font = "Cambria") %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 1: Descriptive Statistics

	n	mean	sd	median	min	max
valid votes/turnout - 1994	4809	0.65	0.10	0.66	0.32	0.91
valid votes/turnout - 1998	5281	0.76	0.09	0.76	0.42	0.97
valid votes/turnout - 2002	5281	0.93	0.03	0.93	0.61	0.99
gini index	5281	0.56	0.06	0.56	0.36	0.82
monthly income	5281	123.13	73.10	106.76	24.98	582.85

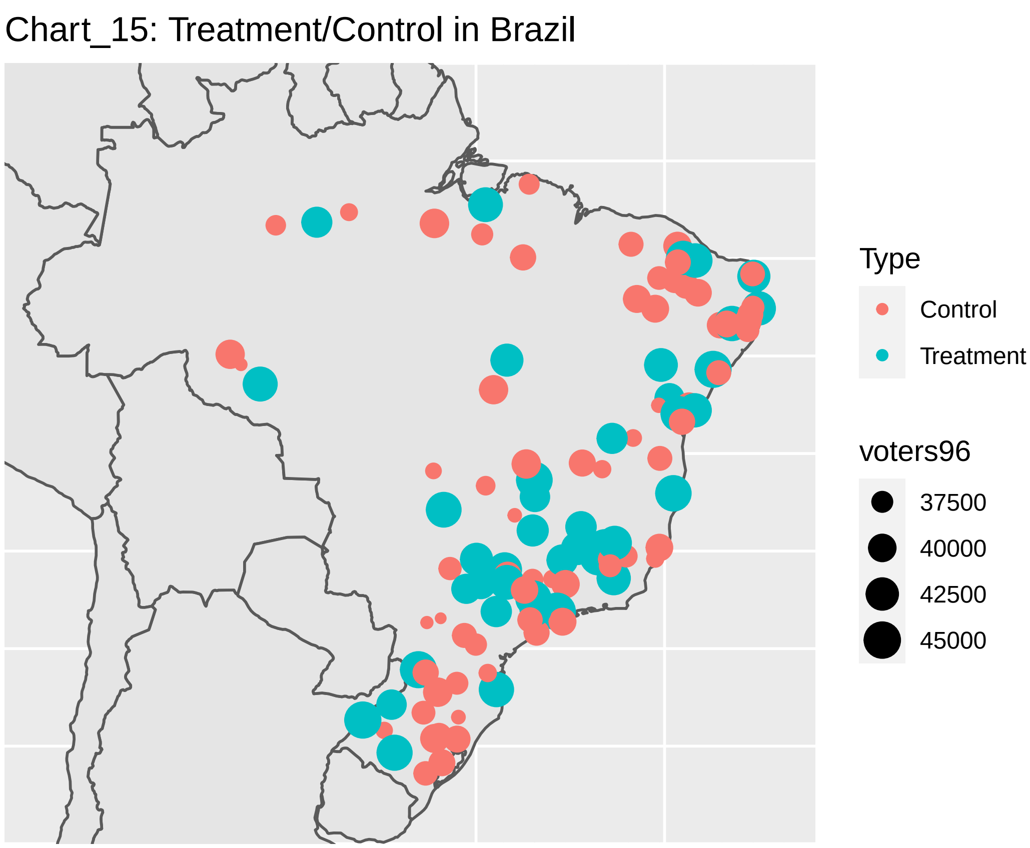
16. Where in Brazil are the treated and control municipalities? Plot a map of Brazil with the following:

a. In one color, the location of the control municipalities, using a bandwidth of 5,000 registered voters.
b. Using another color, the location of the treated municipalities, again using a bandwidth of 5,000 registered voters.
c. The size of each point representing a municipality should be proportional to the number of registered voters in 1996.

```
cut_off = 40500
band    = 5000

data_16b <- data_ori %>%
            mutate(treat = ifelse((voters96 >= cut_off & voters96 <= cut_off + band), 1,
                           ifelse((voters96 <  cut_off & voters96 >= cut_off - band), 0, NA))) %>%  ##
            mutate(Type  = ifelse(treat == 1, "Treatment",
                           ifelse(treat == 0, "Control"  , NA))) %>%                                ##
            mutate(longitude = -1*longitude) %>%                                                     ##
            select(c(voters96, r_util94, r_util98, r_util02, treat, Type, latitude, longitude)) %>% ##
            filter(!is.na(treat))

# Load world's map
world <- ne_countries(scale = "medium", returnclass = "sf")

# Create map of Italy (select coordinates in world map)
Brazil_plot <- ggplot(data = world) +
  geom_sf() +
  coord_sf(ylim = c(5, -35), xlim = c(-75, -32), expand = FALSE)

chart_16 <- Brazil_plot +
            geom_point(data = data_16b, aes(x =  longitude, y = latitude, color = Type, size = voters96
            labs(title = "Chart_15: Treatment/Control in Brazil",
                 x     = NULL,
                 y     = NULL) +
            theme(axis.text.x = element_blank(),
                  axis.text.y = element_blank(),
                  axis.ticks  = element_blank())

ggsave("chart_16.jpeg", chart_16)
```

```
## Saving 6.5 x 4.5 in image
```

```
chart_16
```



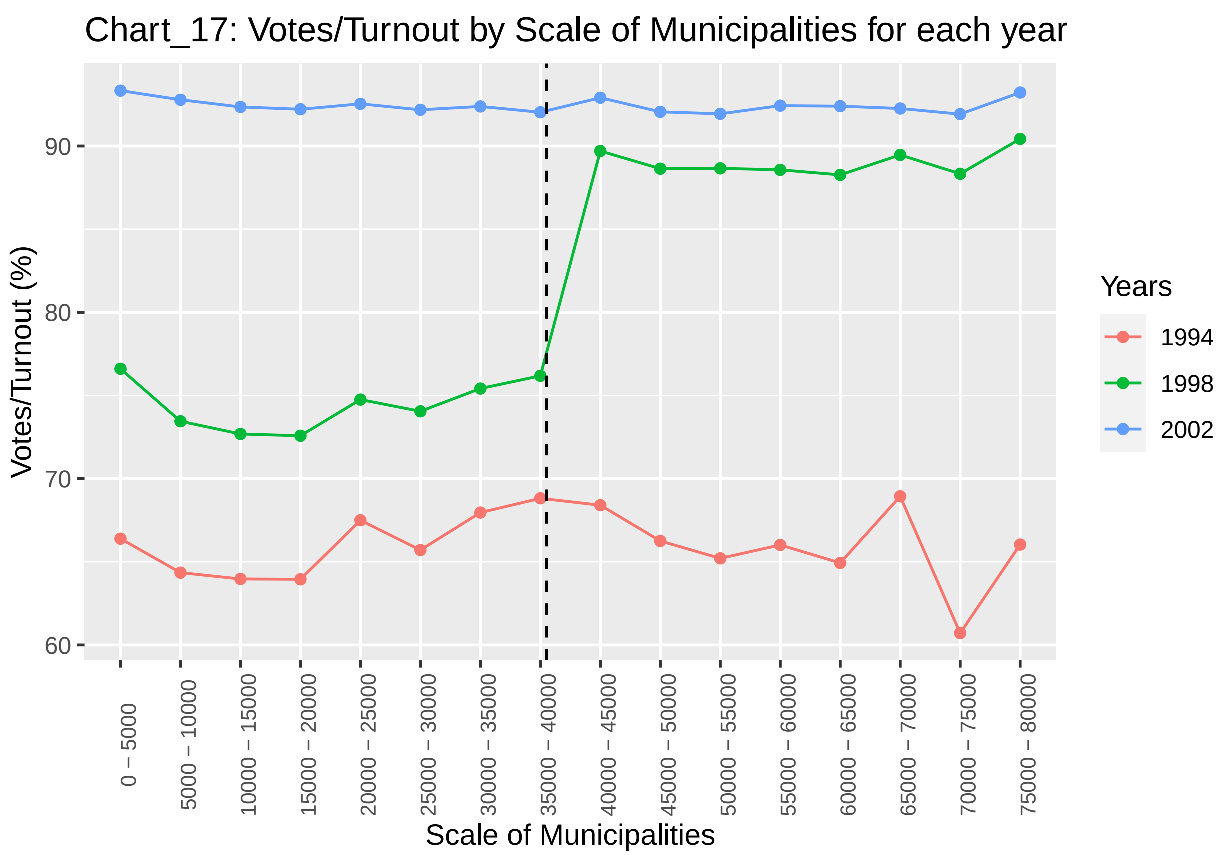
17 Plot the discontinuity in valid votes/turnout in 1994, 1998, and 2002 by plotting the average of valid votes/turnout of bins of municipalities using 5000-voters bins. Make sure to highlight the 40,500 cutoff.

```
## set bins for classification
bins       <- seq(from = 0, to = 80000, by = 5000)

bins_label <- rep(NA, length(bins)-1)  ## create an NA vector

# fill bin labels by "0-5000", "5000-1000" ...
for (i in 1:length(bins) -1){
  bins_label[i] <- paste(bins[i], "-", bins[i + 1])
            }

## create a data set
data_17 <- data_ori %>%
           filter(voters96 <= 80000) %>%  ## filter up to 80,000
           mutate(bins = cut(voters96, breaks = bins,
                                        labels = bins_label)) %>%  ## cut muni? by 5000 each
```

```
        group_by(bins) %>%
        summarize("1994" = mean(r_util94, na.rm = TRUE),        ## mean of each vote/turnout data
                  "1998" = mean(r_util98),
                  "2002" = mean(r_util02)) %>%
        gather(key = "Years", value = "Value", - bins)         ## reshape data to long for plotting

## plot
chart_17 <- ggplot(data = data_17, aes(x = bins, y = Value*100, color = Years, group = Years))+
            geom_line()+
            geom_point()+
            geom_vline(xintercept = 40500/5000, linetype = "dashed") +
            labs(title = "Chart_17: Votes/Turnout by Scale of Municipalities for each year",
                 x     = "Scale of Municipalities",
                 y     = "Votes/Turnout (%)") +
                 theme(axis.text.x = element_text(size = 8, angle = 90))

ggsave("chart_17.jpeg", chart_17)
```

```
## Saving 6.5 x 4.5 in image
```

```
chart_17
```

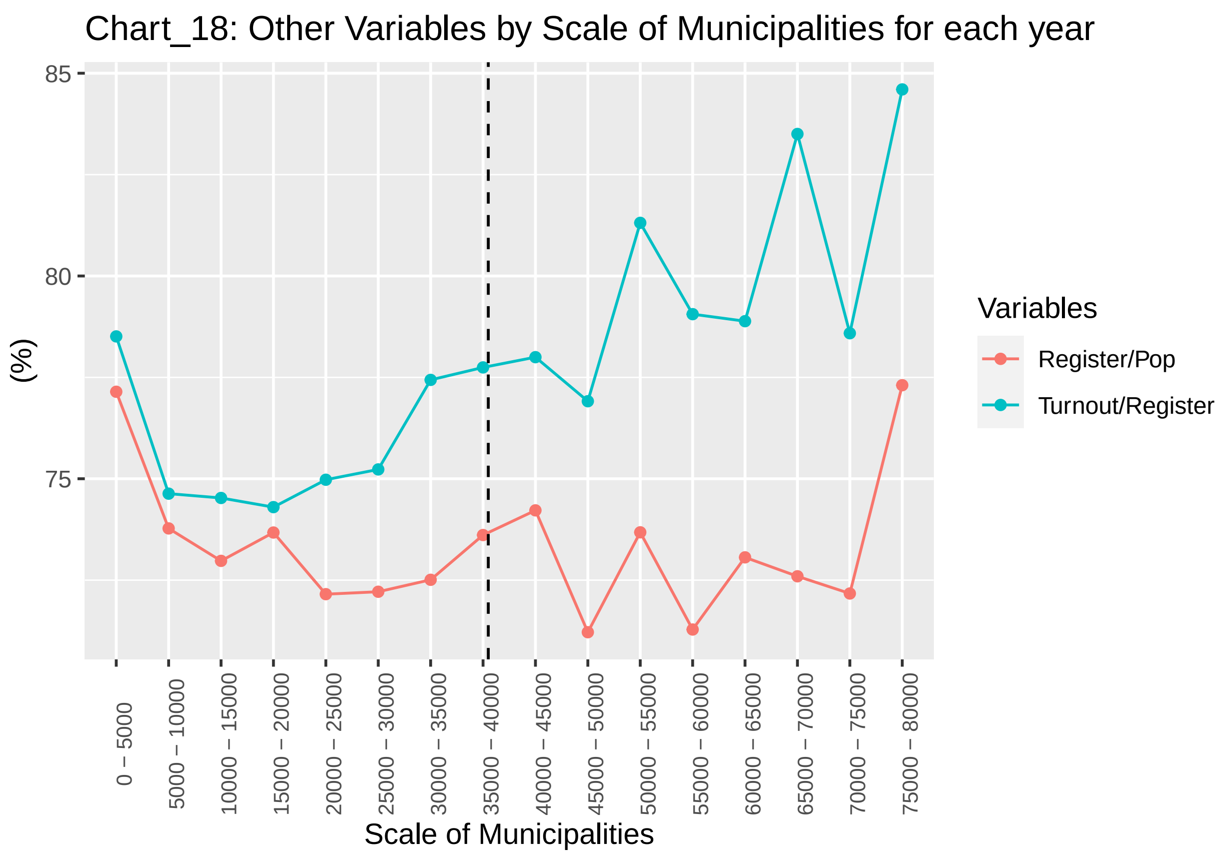


18. [Optional] Make a similar plot (with 5000-voters bins) for the number of registered voters/ population and turnout over registered voters.

```
## create a data set
data_18 <- data_ori %>%
           filter(voters96 <= 80000) %>%  ## filter up to 80,000
           mutate(bins = cut(voters96, breaks = bins,
                                        labels = bins_label)) %>%      ## cut muni? by 5000 each
           group_by(bins) %>%
           summarize("Register/Pop"     = mean(regist, na.rm = TRUE), ## mean of registered voter over p
                     "Turnout/Register" = mean(attend)) %>%           ## mean of turnout over populatio
           gather(key = "Variables", value = "Value", - bins)         ## reshape data to long for plott

## plot
chart_18 <- ggplot(data = data_18, aes(x = bins, y = Value*100, color = Variables, group = Variables))+
            geom_line()+
            geom_point()+
            geom_vline(xintercept = 8 + 500/5000, linetype = "dashed") +
            labs(title = "Chart_18: Other Variables by Scale of Municipalities for each year",
                 x     = "Scale of Municipalities",
                 y     = "(%)") +
                 theme(axis.text.x = element_text(size = 8, angle = 90))

ggsave("chart_18.jpeg", chart_18)
```

```
## Saving 6.5 x 4.5 in image
```

```
chart_18
```



19. Create a function to implement regression (2) in the paper. Your function should take an outcome variable and bandwidth as arguments, and return:

a. The full sample mean of the outcome variable
b. The coefficient from regression (2) on treatment,
c. Its standard error.

$$y_m = \alpha + \beta 1\{v_m > 40,500\} + \gamma * v_m + \delta * v_m 1\{v_m > 40,500\} + \epsilon_m$$

```
func_19 <- function(outcomes, bands){
  ## set results in advance
  results <- list()

  for (j in 1:length(outcomes)){
    ## parameter
    outcome <- outcomes[j]


    ## result
    result <- array(data = NA,
                    dim  = c(length(bands),3),
                    dimnames = list(
                      bands,
```

```r
                      c("Mean", "Treatment Effect", "Standard Error")))

    ## for loop for multiple bands
    for (i in 1:length(bands)){
      ## parameters
      cut_off <-  40500
      band    <-  bands[i]

      ## create a data set
      data_19 <- data_ori %>%
                 mutate(treat = ifelse((voters96 >= cut_off & voters96 <= cut_off + band), 1,
                                ifelse((voters96 <  cut_off & voters96 >= cut_off - band), 0, NA))) %>%
                 filter(!is.na(treat))

      ## select necessary variables
      data_19 <- data_19[, c(outcome, "voters96", "treat")]

      ## rename data set for regression
      colnames(data_19)[1] <- "outcome"

      reg_19  <- lm(outcome*100 ~ treat + voters96 + voters96*treat, data = data_19)


      #a.   The full sample mean of the outcome variable
      result[i, 1] <- round(mean(data_19$outcome, na.rm = TRUE)*100, 2)

      #b.   The coefficient from regression (2) on treatment,
      result[i, 2] <- round(reg_19$coefficients[2], 2)

      #c.   Its standard error.
      result[i, 3] <- round(summary(reg_19)$coefficients[2,2], 2)
    }

    # format result
    results[[j]]<- result %>%
              kbl(caption = attributes(data_19$outcome)$label) %>%  ## extract label data from attribute
              kable_classic(full_width = F, html_font = "Cambria") %>%
              kable_styling(latex_options = "HOLD_position")

  }
  return(results)
}
```

You do not need to add weights. What kernel is this? => rectangular kernel

You do not need to include state fixed effects, but why would you want to include them in this setting? => There may be some unobserved state-specific, time-invariant variables which affects on the voting behaviour. Thus, it would be better to control them by setting state fixed effects.

20. Using the function you wrote in the previous question, report the coefficients you estimate for treatment status for the following outcomes: for the following bandwidths: 15000, 10000, 5000 registered voters.

```
## a.   Valid votes/turnout in 1998
func_19("r_util98", c(5000, 10000, 15000))
```

[[1]]

Table 2: valid votes/turnout - 1998

	Mean	Treatment Effect	Standard Error
5000	81.59	23.45	44.16
10000	80.71	28.92	13.23
15000	79.65	25.19	8.31

```
## b.   Valid votes/turnout in 1994 and 2002
func_19(c("r_util94", "r_util02"), c(5000, 10000, 15000))
```

[[1]]

Table 3: valid votes/turnout - 1994

	Mean	Treatment Effect	Standard Error
5000	68.66	-32.03	55.08
10000	68.13	11.49	17.62
15000	67.31	18.13	11.10

[[2]]

Table 4: valid votes/turnout - 2002

	Mean	Treatment Effect	Standard Error
5000	92.36	-7.34	14.58
10000	92.28	2.69	4.41
15000	92.25	1.99	2.84

```
## c.   Four covariates of your choice to test covariate smoothness.
func_19(c("gini", "income", "regist", "attend"), c(5000, 10000, 15000))
```

[[1]]

Table 5: gini index

	Mean	Treatment Effect	Standard Error
5000	57.64	37.15	29.94
10000	57.46	4.98	9.76
15000	57.43	7.56	5.97

[[2]]

Table 6: monthly income

	Mean	Treatment Effect	Standard Error
5000	17399.03	17662.91	49082.66
10000	16765.52	18723.53	15394.20
15000	16837.69	-3084.61	9909.59

[[3]]

Table 7: registered voters/population

	Mean	Treatment Effect	Standard Error
5000	73.63	-37.76	62.09
10000	73.20	13.49	19.02
15000	72.94	4.08	10.97

[[4]]

Table 8: turnout/registered voters

	Mean	Treatment Effect	Standard Error
5000	77.96	-5.20	46.11
10000	77.82	14.73	15.24
15000	77.41	3.34	9.79

In practice, how would you choose your bandwidth? Comment on your results. => There are some papers proposing optimal choice of bandwiwdth (e.g. Imbens & Kalyanaraman_2009). I would follow them to balance bias and variance. For the results, the larger the bandwidth is, the smaller SE I get. The treatment effect becomes statistically significant. For other variables, I do not see any significant change around the cutoff. This increases the credibility of the treatment effect as causal inference.