Assignment 6

Vincent Zhang

Due 2/21 @ 11:59pm

Overview

You will work with data from Butler & Broockman's study (2011), "Do Politicians Racially Discriminate Against Constituents?".

In addition to the dataset, you will need the authors' replication files. The complete replication files can be downloaded at https://isps.yale.edu/research/data/d037. The relevant replication file is file `D037F04`.

Setup

Load in the dataset. If you are downloading the dataset from the website above, you will note that there are `.csv` and `.dta` versions of the dataset. You may use either. Alternatively, the `.dta` version is available at Canvas.

(1) Table 1

Replicate Table 1 on p. 469 of the study:

- You must replicate Table 1 Column 1 manually, along the lines of part 1 of Tuesday's in-class exercise (`6.1_butler-broockman_2011.Rmd`). If you completed the in-class exercise, you can simply copy-paste your code from the exercise below.
- You may replicate remaining columns using `t.test()`.
- Be sure to print your output in your final PDF.
- It is *not* necessary to insert your output into a publishable table, but you may do so if you wish.

<YOUR REPLICATION OF TABLE 1 HERE>

```
no_part <- dat$treat_noprimary
rep_part <- dat$treat_repprimary
dem_part <- dat$treat_demprimary

dj_rep_p <- 0
dj_dem_p <- 0
jm_rep_p <- 0
jm_dem_p <- 0
```

```r
calculate_stat <- function(partisanship) {
  p1 <- mean(dat$reply_atall[which(partisanship == 1 & dat$treat_deshawn == 1)])
  p2 <- mean(dat$reply_atall[which(partisanship == 1 & dat$treat_jake == 1)])

  n1 <- length(dat$reply_atall[which(partisanship == 1 & dat$treat_deshawn == 1)])
  n2 <- length(dat$reply_atall[which(partisanship == 1 & dat$treat_jake == 1)])

  race_diff <- p1 - p2

  se <- sqrt((p1*(1-p1)/n1) + (p2*(1-p2)/n2))
  t <- (p1-p2)/se
  dof <- n1 + n2 - 2
  p <- pt(t, dof) * 2

  if (all(partisanship == rep_part)) {
    dj_rep_p <<- p1
    jm_rep_p <<- p2
  } else if (partisanship == dem_part) {
    dj_dem_p <<- p1
    jm_dem_p <<- p2
  }

  cat("\nDeshawn Jackson:", "percentage = ", p1, ", N = ", n1)
  cat("\nJake Mueller: ", "percentage = ", p2, ", N = ", n2)
  cat("\nRace Differential", race_diff, " p = ", p)
}

#-------------------------------------------
# First column (No Partisanship Signal)
#-------------------------------------------
calculate_stat(no_part)
```

```
##
## Deshawn Jackson: percentage =  0.5533499 , N =  806
## Jake Mueller:  percentage =  0.6046798 , N =  812
## Race Differential -0.05132993  p =  0.03643861
```

```r
#-------------------------------------------
# Second column (Republican Signal)
#-------------------------------------------
calculate_stat(rep_part)
```

```
##
## Deshawn Jackson: percentage =  0.5432099 , N =  810
## Jake Mueller:  percentage =  0.5646341 , N =  820
## Race Differential -0.02142427  p =  0.3843112
```

```r
#-------------------------------------------
# Third column (Democratic Signal)
#-------------------------------------------
calculate_stat(dem_part)
```

```
##
## Deshawn Jackson: percentage =  0.5726601 , N =  812
## Jake Mueller:  percentage =  0.5531915 , N =  799
## Race Differential 0.01946861  p =  1.569064
```

```
#-------------------------------------------
# Fourth column (Party Differential)
#-------------------------------------------
# i did not know how to get the p values for these, but here are the percentages:"

cat("\nDeshawn Jackson:", "percentage = ", (dj_rep_p - dj_dem_p))
```

```
##
## Deshawn Jackson: percentage =  -0.02945022
```

```
cat("\nJake Mueller: ", "percentage = ", (jm_rep_p - jm_dem_p))
```

```
##
## Jake Mueller:  percentage =  0.01144266
```

(2) Replication File

Open the authors' replication file (`Butler_Broockman_AJPS_2011_public_R.R`). Run the file from start to finish. Note that you will need to modify the authors' function for loading in the data.

(2.1)

Compare your code for Table 1 against the authors'. You'll note that the authors constructed Table 1 using OLS regression (`lm()`). Examine the output of their regression models and see if you can visualize their data and the OLS regression line that runs through them. Then, write a brief paragraph describing how / where key pieces of information from Table 1 are expressed in the OLS regression output. (It is not necessary to explain why; you may simply describe.)

The authors' analysis involves using an ordinary least squares regression model. I commented out all the code except the code relating to table 1 so that I could better look at what the regression is doing (from reading the code, I believe this to be the out.1." -out.2.* codes). If we look at out.1.1 (which the authors' label as no primary and deshawn), as well as out. it was sort of hard for me to see where the similarity between this and my data. Some of the numbers I noticed were similar were in the out.1.1 to out.1.3, where I noticed the (intercept) estimate would be the same as the percentages found in table 1. However, the part I was confused about is that out.1.1 - out.1.3 is doing analysis on DeShawn Jackson, evident in how in the R file it says, "reply_atall ~ treat_deshawn", however, these percentages are matching the percentages in the table found for Jake Mueller. For instance in out.1.1 (no_partisanship signal), we see under coefficients, that the estimate for the intercept is 0.60468, and in the table 1, we see that the percentage is exactly the same for Jake Mueller: 60.5%. However, the weird part is that out.1.1 is doing the regression on Deshawn, or at least that's what it seems like when I look at the R code, where it is applying lm() to reply_atall ~ treat_deshawn. However, this was still a similarity between table 1 and the results I see from running the R file, and I can't figure out why. Another area of similarity is the F-statistic: where the p-value we find in the F-statistic after running the R file is the same as the p-value we see in Table 1. Similarly to how we see these results in out.1.1 - out.1.3, we also see more similarities in more of the out.1 outputs, but again, it seems unintuitive to me. For example, in out.1.6, we see that it is running lm() on data[data$treat_jake], so for treating Jake Mueller, and also that it is running it on reply_atall ~ treat_repprimary, which suggests

that it is comparing it checking the republican signal, but when we check the intercept value, we see that it is 0.55319, which is about 55.3% which again is found in table 1, but instead of under the republican signal column, its under democratic signal column, and in the Jake Mueller row (which at least this is correct), so again I see that some data is found in the R file ran by the authors and the table 1, but I do not understand why the key pieces of information of jumbled.

(2.2)

After you have worked through the entire replication file, think about the file itself. Was it well written and self explanatory? If not, what might make the file more user friendly? Write a few sentences up to a short paragraph here.

The file was wack and not self-explanatory to be honest. I think that comments explaing what each of the code lines does, and what the output means would be useful. Also, its honestly just hard to find the information too; when we have an R markdown file, we have the data integrated into the code lines, so its very easy to see where things are coming from. However, when we run this large R file, all the code just pops out all at once, and so its a little hard to see where exactly each output came from which lead to a lot of squinting of the eyes. I think that obviously having a stronger understanding of what each of the statistical variables in these models mean whould help, but I still do think that the presentation of all these numbers are hard to find, and thus not too well written, and that a lot of self digging into where the code came from and what it actually means and does needs to be done. More comments mainly would improve understandability and readability