HW2_Markdown

2024-02-07

Manay Divatia

Problem 1

a)

```r
turnout_df = read.csv("~/Downloads/turnout.csv")
summary(turnout_df)
```

```
##       year           VEP              VAP              total       
##  Min.   :1980   Min.   :159635   Min.   :164445   Min.   : 64991  
##  1st Qu.:1986   1st Qu.:171192   1st Qu.:178930   1st Qu.: 73179  
##  Median :1993   Median :181140   Median :193018   Median : 89055  
##  Mean   :1993   Mean   :182640   Mean   :194226   Mean   : 89778  
##  3rd Qu.:2000   3rd Qu.:193353   3rd Qu.:209296   3rd Qu.:102370  
##  Max.   :2008   Max.   :213314   Max.   :230872   Max.   :131304  
##                                                                   
##       ANES           felons          noncit         overseas        osvoters  
##  Min.   :47.00   Min.   : 802   Min.   : 5756   Min.   :1803   Min.   :263  
##  1st Qu.:57.00   1st Qu.:1424   1st Qu.: 8592   1st Qu.:2236   1st Qu.:263  
##  Median :70.50   Median :2312   Median :11972   Median :2458   Median :263  
##  Mean   :65.79   Mean   :2177   Mean   :12229   Mean   :2746   Mean   :263  
##  3rd Qu.:73.75   3rd Qu.:3042   3rd Qu.:15910   3rd Qu.:2937   3rd Qu.:263  
##  Max.   :78.00   Max.   :3168   Max.   :19392   Max.   :4972   Max.   :263  
##                                                                NA's   :13   
```

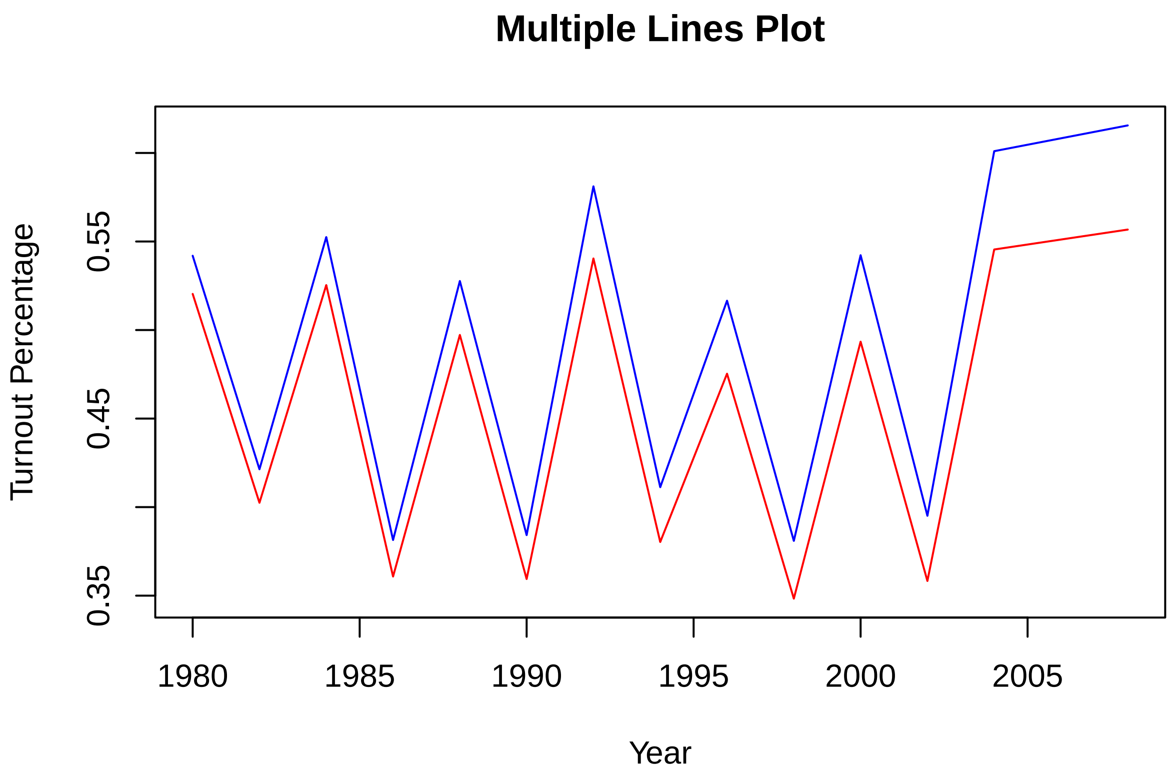
```r
dim(turnout_df)
```

```
## [1] 14  9
```

The year range goes from 1980 to 2008 in 2 year increments since that is the election cycle for presidential and midterm elections. There are 14 observations (rows) and 9 values (columns). From the summary, it looks like the osvoters columns is the only one with NA's (13). Another interesting thing is that the total votes range from 64,991 to 131,304 (in thousands).

b)

```r
y1 = turnout_df$total / (turnout_df$VAP + turnout_df$overseas)
y2 = turnout_df$total / turnout_df$VEP
matplot(turnout_df$year, cbind(y1, y2), type = "l", lty = 1,
        col = c("red", "blue"), xlab = "Year",
        ylab = "Turnout Percentage", main = "Multiple Lines Plot")
```



The turnout rate based on VEP was usually about 2% higher than the turnout rate using VAP and overseas voters. I also made a little plot where the blue line is turnout rate using VEP and the red line is turnout rate using VAP and overseas voters. The plot can show how the two turnouts compare and show how turnout rate based on VEP was consistently slightly higher.

c)

```
vap_anes_diff = ((100* y1) -  turnout_df$ANES)
mean(vap_anes_diff)
```

```
## [1] -20.32914
```

```
range(vap_anes_diff)
```

```
## [1] -26.17150 -11.06116
```

```
vep_anes_diff = ((100* y2) -  turnout_df$ANES)
mean(vep_anes_diff)
```

```
## [1] -16.83634
```

```
range(vep_anes_diff)
```

```
## [1] -22.489359  -8.581054
```

We can see through the data that, on average, ANES overestimates voter turnout by 20.3% when compared to VAP which is very high. Also, ANES overestimates voter turnout by 16.8% when compared to VEP. The range for VAP is from 26.1% overestimate to 11% overestimate. The range for VEP is slightly lower with a range of 8.6% overestimate to 22.5% overestimate. This could be due to that bias that was mentioned earlier.

d)

```r
vep_diff <- function(df) {
  vep_turnout <- df$total / df$VEP
  vep_anes_diff <- ((100 * vep_turnout) -  df$ANES)
  print(mean(vep_anes_diff))
  print(range(vep_anes_diff))
}

turnout_df_midterm <- turnout_df[turnout_df$year %% 4 != 0, ]
vep_diff(turnout_df_midterm)
```

```
## [1] -15.4288
## [1] -22.489359  -8.581054
```

```r
turnout_df_presidential <- turnout_df[turnout_df$year %% 4 == 0, ]
vep_diff(turnout_df_presidential)
```

```
## [1] -17.892
## [1] -21.34207 -16.44567
```

The ANES estimates of voter turnout get worse in presidential elections and better in midterms. Still in all cases, ANES is overestimating how many people are actually voting which goes back to the idea that bias is present.

e)

```r
turnout_df_before <- turnout_df[turnout_df$year <= 1992, ]
turnout_df_after <- turnout_df[turnout_df$year > 1992, ]
vep_diff(turnout_df_before)
```

```
## [1] -15.85378
## [1] -18.751404  -8.581054
```

```r
vep_diff(turnout_df_after)
```

```
## [1] -17.81891
## [1] -22.48936 -13.90684
```

From the information, in the first half of election years, ANES overestimated voter turnout by about 15.8% with a range of 8.6% - 18.8% of overestimation. In the second half of election years, ANES overestimated voter turnout by about 17.8% with a range of 13.9% - 22.5% of overestimation. So from the first to second half, the mean and range increased which means that the bias of the ANES increased over time.

f

```r
turnout_df_2008 = turnout_df[turnout_df$year == 2008, ]
adj_vap = turnout_df_2008$VAP - (turnout_df_2008$felons + turnout_df_2008$overseas)
vap_turnout_adj = (turnout_df_2008$total - turnout_df_2008$overseas) / adj_vap
vap_turnout_raw = turnout_df_2008$total / turnout_df_2008$VAP
vep_turnout_raw = turnout_df_2008$total / turnout_df_2008$VEP
anes_turnout_raw = turnout_df_2008$ANES / 100
```

The adjusted VAP turnout rate was around 56.7%. The raw VAP turnout rate was 56.8% so very close to the adjusted rate. The raw VEP turnout rate was 61.5% which was higher than the adjusted. Finally, the

ANES raw turnout rate was 78% which was greatly higher than the adjusted VAP turnout rate. I think this was kind of expected since the ANES was overestimating voter turnout in all the analysis so far.

Problem 2

a)

```
star_df = read.csv("~/Downloads/STAR.csv")
star_df$kinder <- ifelse(star_df$classtype == 1, "small", ifelse(star_df$classtype == 2, "regular", "reg
star_df$race <- ifelse(star_df$race == 1, "White", ifelse(star_df$race == 2, "Black", ifelse(star_df$rac
```

b)

```
star_df_small = star_df[star_df$classtype == 1, ]
star_df_reg = star_df[star_df$classtype == 2, ]
mean(star_df_small$g4math, na.rm = TRUE)
```

```
## [1] 709.1851
```

```
mean(star_df_small$g4reading, na.rm = TRUE)
```

```
## [1] 723.3912
```

```
mean(star_df_reg$g4math, na.rm = TRUE)
```

```
## [1] 709.5214
```

```
mean(star_df_reg$g4reading, na.rm = TRUE)
```

```
## [1] 719.89
```

```
sd(star_df_small$g4math, na.rm = TRUE)
```

```
## [1] 43.57318
```

```
sd(star_df_small$g4reading, na.rm = TRUE)
```

```
## [1] 51.54494
```

```
sd(star_df_reg$g4math, na.rm = TRUE)
```

```
## [1] 41.02063
```

```
sd(star_df_reg$g4reading, na.rm = TRUE)
```

```
## [1] 53.16788
```

The mean math score for students in small classes was 709.2 and the mean math score for students in regular classes was 709.5. We could say that students in the regular class performed better but the difference is so small, I feel like we can't say if there is a difference or not. For reading, the mean score for students in the small class was 723.4 and the mean score for students in regular classes was 719.9. For both, it seems that the students performed better on reading than on math. However for reading, the small class had the better mean. The standard deviation for math scores for students in small classes was 43.6 while for regular classes it was 41. This seems like a lot compared to the scores and could show that both scores overlap and we don't know if one group performed better. This is the same case for reading where the standard deviation for students in the small classes was 51.5 while for students in regular classes, it was 53.2.

c)

```
quantile(star_df_small$g4math, c(.33, .66), na.rm = TRUE)
```

```
## 33% 66%
## 694 726
```

```
quantile(star_df_small$g4reading, c(.33, .66), na.rm = TRUE)
```

```
## 33% 66%
## 705 741
```

```
quantile(star_df_reg$g4math, c(.33, .66), na.rm = TRUE)
```

```
## 33% 66%
## 696 724
```

```
quantile(star_df_reg$g4reading, c(.33, .66), na.rm = TRUE)
```

```
## 33% 66%
## 705 740
```

This analysis tells us some information about the math scores but really nothing about the reading scores. For math, the small class had a 33rd percentile of 694 and a 66th percentile of 726. For math, the regular class had a 33rd percentile of 696 and a 66th percentile of 724. It shows that the small class had a slightly wider range between the 33rd and 66th percentile. For reading however, both classes had a 33rd percentile of 705 and for the 66th percentile, the small class had a score of 741 while the regular class had a score of 740 which shows a very little difference overall. It does help us to know that maybe the class size didn't have very much of an impact on score.

d)

```
nrow(star_df[star_df$yearssmall == 4, ])
```

```
## [1] 857
```

```
nrow(star_df[star_df$yearssmall == 3, ])
```

```
## [1] 353
```

```
nrow(star_df[star_df$yearssmall == 2, ])
```

```
## [1] 390
```

```
nrow(star_df[star_df$yearssmall == 1, ])
```

```
## [1] 768
```

```
nrow(star_df[star_df$yearssmall == 0, ])
```

```
## [1] 3957
```

```
mean(star_df[star_df$yearssmall == 4, ]$g4math, na.rm = TRUE)
```

```
## [1] 710.0519
```

```
mean(star_df[star_df$yearssmall == 4, ]$g4reading, na.rm = TRUE)
```

```
## [1] 724.6651
```

```
mean(star_df[star_df$yearssmall == 3, ]$g4math, na.rm = TRUE)
```

```
## [1] 709.617
```

```r
mean(star_df[star_df$yearssmall == 3, ]$g4reading, na.rm = TRUE)
```

```
## [1] 719.8986
```

```r
mean(star_df[star_df$yearssmall == 2, ]$g4math, na.rm = TRUE)
```

```
## [1] 711.914
```

```r
mean(star_df[star_df$yearssmall == 2, ]$g4reading, na.rm = TRUE)
```

```
## [1] 717.8681
```

```r
mean(star_df[star_df$yearssmall == 1, ]$g4math, na.rm = TRUE)
```

```
## [1] 707.5524
```

```r
mean(star_df[star_df$yearssmall == 1, ]$g4reading, na.rm = TRUE)
```

```
## [1] 723.1471
```

```r
mean(star_df[star_df$yearssmall == 0, ]$g4math, na.rm = TRUE)
```

```
## [1] 707.9793
```

```r
mean(star_df[star_df$yearssmall == 0, ]$g4reading, na.rm = TRUE)
```

```
## [1] 719.8754
```

Students spent 0-4 years in a small class. The majority was by far students spending 0 years in a small class as 3957 students fit that category. 857 students spent 4 years, 353 students spent 3 years, 390 students spent 2 years, and 768 students spent 1 year in a small class. In terms of math scores, time spent in a small class didn't seem to affect scores too much. For 4 years in a small class, the mean was 710. For 3 years in a small class, the mean was 709.6. For 2 years in a small class, the mean was 711.9. For 1 years in a small class, the mean was 707.5. And for 0 years in a small class, the mean was 707.9. This shows not much change overall. For reading scores, it was the same story. For 4 years in a small class, the mean was 724.7. For 3 years in a small class, the mean was 719.9. For 2 years in a small class, the mean was 717.9. For 1 years in a small class, the mean was 723.1. And for 0 years in a small class, the mean was 719.9. Overall, the math and reading scores didn't seem to be correlated with how long a student spent in the small class.

e)

```r
mean(star_df_small[star_df_small$race == "White", ]$g4math, na.rm = TRUE)
```

```
## [1] 711.19
```

```r
mean(star_df_small[star_df_small$race == "White", ]$g4reading, na.rm = TRUE)
```

```
## [1] 727.8388
```

```r
mean(star_df_small[star_df_small$race == "Black", ]$g4math, na.rm = TRUE)
```

```
## [1] 697.5043
```

```r
mean(star_df_small[star_df_small$race == "Black", ]$g4reading, na.rm = TRUE)
```

```
## [1] 698.614
```

```r
mean(star_df_small[star_df_small$race == "Hispanic", ]$g4math, na.rm = TRUE)
```

```
## [1] NaN
```

```r
mean(star_df_small[star_df_small$race == "Hispanic", ]$g4reading, na.rm = TRUE)
```

```
## [1] NaN
```

```r
mean(star_df_reg[star_df_reg$race == "White", ]$g4math, na.rm = TRUE)
```

```
## [1] 711.4104
```

```r
mean(star_df_reg[star_df_reg$race == "White", ]$g4reading, na.rm = TRUE)
```

```
## [1] 725.1158
```

```r
mean(star_df_reg[star_df_reg$race == "Black", ]$g4math, na.rm = TRUE)
```

```
## [1] 698.5323
```

```r
mean(star_df_reg[star_df_reg$race == "Black", ]$g4reading, na.rm = TRUE)
```

```
## [1] 689.3548
```

```r
mean(star_df_reg[star_df_reg$race == "Hispanic", ]$g4math, na.rm = TRUE)
```

```
## [1] NaN
```

```r
mean(star_df_reg[star_df_reg$race == "Hispanic", ]$g4reading, na.rm = TRUE)
```

```
## [1] NaN
```

It appears that in both reading and math in both small and regular classes, the white students got better scores than the black students. However, in the small classes, the Hispanic kids, on average, performed the best in reading and math.

f)

```r
sum(star_df_small$hsgrad, na.rm = TRUE)
```

```
## [1] 754
```

```r
sum(star_df_reg$hsgrad, na.rm = TRUE)
```

```
## [1] 892
```

```r
mean(star_df[star_df$yearssmall == 4, ]$hsgrad, na.rm = TRUE)
```

```
## [1] 0.877551
```

```r
mean(star_df[star_df$yearssmall == 3, ]$hsgrad, na.rm = TRUE)
```

```
## [1] 0.8324607
```

```r
mean(star_df[star_df$yearssmall == 2, ]$hsgrad, na.rm = TRUE)
```

```
## [1] 0.8131868
```

```r
mean(star_df[star_df$yearssmall == 1, ]$hsgrad, na.rm = TRUE)
```

```
## [1] 0.7910448
```

```r
mean(star_df[star_df$yearssmall == 0, ]$hsgrad, na.rm = TRUE)
```

```
## [1] 0.828602
```

Overall, it looks like students in regular classes had more graduates than students in small classes. It is important to note that there are a lot of NAs for this data which could skew it. There is some correlation with graduation and how many years they were in a small class. Students in a small class for 4 years had the highest graduation rate with a steady decline for every year less in a small class with the exception of students who spent no years in a small class.

Problem 3

The whole point of randomization or random assignment is to make the groups as comparable as possible. By randomizing, we can say that the 2 groups are comparable and because of this, we can make one group a treatment group and one a control group and compare the difference of whatever outcome we are looking at. We cannot estimate causal effect of COVID lockdowns between lockdown and non-lockdown states because there are other factors involved. We cannot say that the lockdown and non-lockdown states are the same. Because of this, we cannot be sure that other factors are not involved like the economy, state politics, population, average age, etc.