Problem Set 2

Applied Stats/Quant Methods 1

Due: October 16, 2022

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Sunday October 16, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1 (40 points): Political Science

The following table was created using the data from a study run in a major Latin American city.[1] As part of the experimental treatment in the study, one employee of the research team was chosen to make illegal left turns across traffic to draw the attention of the police officers on shift. Two employee drivers were upper class, two were lower class drivers, and the identity of the driver was randomly assigned per encounter. The researchers were interested in whether officers were more or less likely to solicit a bribe from drivers depending on their class (officers use phrases like, "We can solve this the easy way" to draw a bribe). The table below shows the resulting data.

[1] Fried, Lagunes, and Venkataramani (2010). "Corruption and Inequality at the Crossroad: A Multimethod Study of Bribery and Discrimination in Latin America. *Latin American Research Review.* 45 (1): 76-97.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	14	6	7
Lower class	7	7	1

(a) Calculate the χ^2 test statistic by hand/manually (even better if you can do "by hand" in R).

```
df <- as.data.frame(PS02_data)

# Firstly, I subset the row and column totals (these # will be used to
    find expected frequencies and #later to find proportions.
Total.sample.size <- df[3,4]
Total.not.stopped <- df[3,1]
Total.bribe.req <- df[3,2]
Total.stopped <- df[3,3]
Total.Upper <- df[1,4]
Total.Lower <- df[2,4]

# I then subset frequency observed for each cell and # find the frequency
     expected.
fo1 <- df[1,1]
fe1 <- ((Total.not.stopped)*(Total.Upper))/(Total.sample.size)
fo2 <- df[2,1]
fe2 <- ((Total.not.stopped)*(Total.Lower))/(Total.sample.size)
fo3 <- df[1,2]
fe3 <- ((Total.bribe.req) * (Total.Upper))/(Total.sample.size)
fo4 <- df[2,2]
fe4 <- ((Total.bribe.req) * (Total.Lower))/(Total.sample.size)
fo5 <- df[1,3]
fe5 <- ((Total.stopped)*(Total.Upper))/(Total.sample.size)
fo6 <- df[2,3]
fe6 <- ((Total.stopped)*(Total.Lower))/(Total.sample.size)

# Then using the formula, I find the chi statistic (3.79912)
#  by adding the results for each cell
chi.stat1 <- ((fo1 - fe1) ^ 2 / fe1) + ((fo2 - fe2) ^ 2 / fe2) +
  ((fo3 - fe3) ^ 2 / fe3) + ((fo4 - fe4) ^ 2 / fe4) + ((fo5-fe5) ^ 2 /
    fe5) + ((fo6 - fe6) ^ 2 / fe6)
print(chi.stat1)
```

(b) Now calculate the p-value from the test statistic you just created (in R).[2] What do you conclude if $\alpha = 0.1$?

[2]Remember frequency should be > 5 for all cells, but let's calculate the p-value here anyway.

```
1
2 # To find the p-value, calculate the degrees of freedom (rows -1) x (cols
      -1)
3 # Using pchsq() function with the chi-stat and degrees of freedom, I can
     get p-value 0.29508
4 deg_free <- (nrow(df) - 1) * (ncol(df) - 1)
5 p.value <- pchisq(chi.stat1, deg_free)
6 p.value
7
8 # As the p-value > 0.1 we cannot reject the null hypothesis that the two
     variables are statistically #independent
```

(c) Calculate the standardized residuals for each cell and put them in the table below.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.523

```
1 z1 <- (fo1 - fe1) / sqrt(fe1 * ((1 - (Total.Upper/Total.sample.size))*(1
     - (Total.not.stopped/Total.sample.size))))
2 z2 <- (fo2 - fe2) / sqrt(fe2 * ((1 - (Total.Lower/Total.sample.size))*(1
     - (Total.not.stopped/Total.sample.size))))
3 z3 <- (fo3 - fe3) / sqrt(fe3 * ((1 - (Total.Upper/Total.sample.size))*(1
     - (Total.bribe.req/Total.sample.size))))
4 z4 <- (fo4 - fe4) / sqrt(fe4 * ((1 - (Total.Lower/Total.sample.size))*(1
     - (Total.bribe.req/Total.sample.size))))
5 z5 <- (fo5 - fe5) / sqrt(fe5 * ((1 - (Total.Upper/Total.sample.size))*(1
     - (Total.stopped/Total.sample.size))))
6 z6 <- (fo6 - fe6) / sqrt(fe6 * ((1 - (Total.Lower/Total.sample.size))*(1
     - (Total.stopped/Total.sample.size))))
```

(d) How might the standardized residuals help you interpret the results?

```
1 # The positive residual of 1.5 suggests there were more upper class
     drivers stopped with a warning than our hypothesis of independence
     predicts. Conversely, there were less lower class drivers that were
     just stopped with a warning than we would expect with the hypothesis
     of statistical independence. Similarly, the negative residual -1.642
     suggests there were less upper class drivers asked for a bribe than
     lower class drivers than our hypothesis of independence predicts.
```

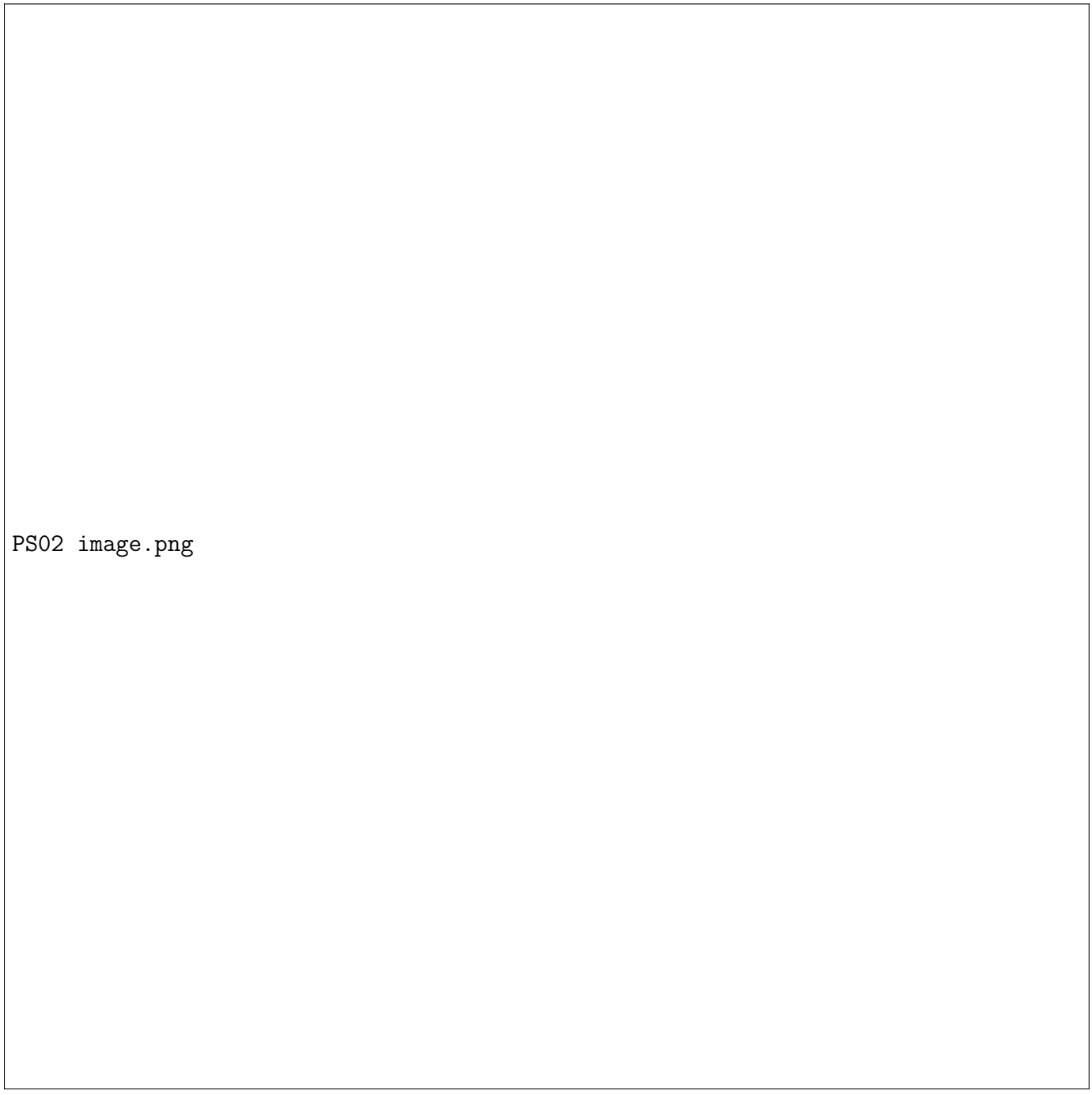
Question 2 (40 points): Economics

Chattopadhyay and Duflo were interested in whether women promote different policies than men.[3] Answering this question with observational data is pretty difficult due to potential confounding problems (e.g. the districts that choose female politicians are likely to systematically differ in other aspects too). Hence, they exploit a randomized policy experiment in India, where since the mid-1990s, $\frac{1}{3}$ of village council heads have been randomly reserved for women. A subset of the data from West Bengal can be found at the following link: `https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv`

Each observation in the data set represents a village and there are two villages associated with one GP (i.e. a level of government is called "GP"). Figure 1 below shows the names and descriptions of the variables in the dataset. The authors hypothesize that female politicians are more likely to support policies female voters want. Researchers found that more women complain about the quality of drinking water than men. You need to estimate the effect of the reservation policy on the number of new or repaired drinking water facilities in the villages.

[3]Chattopadhyay and Duflo. (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica*. 72 (5), 1409-1443.

Figure 1: Names and description of variables from Chattopadhyay and Duflo (2004).



(a) State a null and alternative (two-tailed) hypothesis.

Null Hypothesis: The reservation policy, in which 1/3 of village council heads have

randomly reserved for women, has no impact on the number of new/repaired drinking facilities in villages. (The slope of the regression line would equal zero)

Alternative Hypothesis: The reservation policy does affect the number of new/repaired drinking facilities in villages. (the slope of the regression line would not equal zero)

1. Null: Beta = 0
2. Alternative: Beta $\neq$ 0

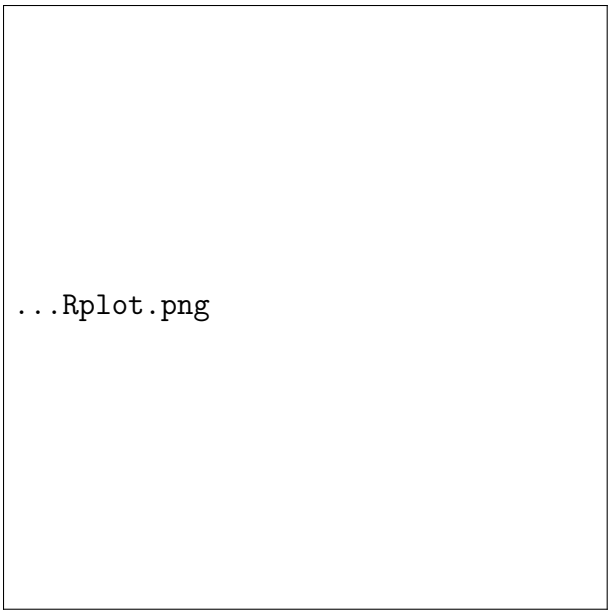
(b) Run a bivariate regression to test this hypothesis in R (include your code!).

```
plot(df$reserved, df$water)
lin <- lm(df$water ~ df$reserved)
summary(lin)
Call:
lm(formula = df$water ~ df$reserved)

Residuals:
    Min      1Q  Median      3Q     Max
-23.991 -14.738  -7.865   2.262 316.009

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    14.738      2.286   6.446 4.22e-10 ***
df$reserved     9.252      3.948   2.344   0.0197 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
            1

Residual standard error: 33.45 on 320 degrees of freedom
Multiple R-squared:  0.01688, Adjusted R-squared:  0.0138
F-statistic: 5.493 on 1 and 320 DF,  p-value: 0.0197
```



(c) Interpret the coefficient estimate for reservation policy.

The intercept coefficient (14.738) is the average increase in drinking water facilities for an increase in the reserved policy variable.

The slope coefficient suggests that the number of drinking water facilities increases by about 9.3 with the introduction of the reservation policy