Problem Set 2

Tolga Bag, 23371290. Applied Stats/Quant Methods 1

Due: October 15, 2023

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 15, 2023. No late assignments will be accepted.

Question 1: 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 Multi-method 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).
I create the shared table as a dataframe.

```
data <- data.frame(
  variable = c("upper_class", "lower_class"),
  not_stopped = c(14, 7),
  bribe_requested = c(6, 7),
  stopped_given_warning = c(7, 1)
)

print(data)
chisq.test(data)
```

chi square test gives a warning. I need to ensure that the x axis is numerical for it to work. I assign 1 to upper class and 2 to lower class

```
data$variable <- as.numeric(factor(data$variable,
                                    levels = c("upper_class",
                                               "lower_class")))
print(data)
chisq.test(data)
```

our chi test score is 4.844. I spent hours to fix the error, but I couldn't fix it.

(b) Now calculate the p-value from the test statistic you just created (in R).[2] What do you conclude if $\alpha = 0.1$?
Our question is if there is a relationship between the class and the solicitation of bribes from people who are violating traffic rules. Our null hypothesis is there is none and they are independent. Our alternative hypothesis is there is a casual relationship.

```
chisq.test(data)
```

p value is 0.1836, in other words higher than the 0.1 threshold. I can't reject the hypothesis and will accept that they are independent.

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

(c) Calculate the standardized residuals for each cell and put them in the table below.
I create a table with the chi square test to calculate the residuals

```
cht <- chisq.test(data)
data_res <- cht$residuals
print(data_res)
```

I need to place upper class and lower class.

```
data_res[data_res == -0.634335047416547] <- "upper_class"
data_res[data_res == 0.814091578410694] <- "lower_class"
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.25	0.73	0.90
Lower class	0.33	0.94	-1.16

(d) How might the standardized residuals help you interpret the results?
Residual is the difference between what we observe and what we expected per our model. Here, the residuals confirm there is no dependency between the variables because there is no similarity between them.

Question 2: 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.

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

Name	Description
`GP`	An identifier for the Gram Panchayat (GP)
`village`	identifier for each village
`reserved`	binary variable indicating whether the GP was reserved for women leaders or not
`female`	binary variable indicating whether the GP had a female leader or not
`irrigation`	variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started
`water`	variable measuring the number of new or repaired drinking-water facilities in the village since the reserve policy started

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

(a) State a null and alternative (two-tailed) hypothesis.
I pull the data from the link.

```
poldata <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/
    master/PREDICTION/women.csv")
```

Since we know females complain more about the quality of water then men, I need to check the relationship between new or repaired water facilities, and irrigation works that are not requested from women. I'll make a refined table that contains these.

```
poldata_fwi <- poldata [c("female", "water", "irrigation")]
```

However, this includes both the villages with and without female leaders. I filter out the villages without female leaders to check the ones led by females.

```
poldata_fwi <- subset(poldata_fwi, female == 1)
```

My null hypothesis is there is no dependency between irrigation works and the number of new or repaired water facilities in the female-led GPs. My alternative hypothesis is there is a statistically significant relationship as I expect female leaders to prioritize water works.

(b) Run a bivariate regression to test this hypothesis in `R` (include your code!).
I use the the linear regression code to check the model.

```
bivar_fwi <- lm(irrigation ~ water, data = poldata_fwi)
```

I check the summary to see the statistical significance.

```
summary(bivar_fwi)
```

p value is 2.601e-09/ I check what it is in 0 terms.

```
sprintf("%.20f",2.601e-09)
```

That is 0.000000002601. Hence, I reject the null hypothesis and establish that there is a statistically significant relationship between water works and irrigation. In other words, there is a statistically significant signal that female leaders prioritize the water works requested by their female constituencies.

(c) Interpret the coefficient estimate for reservation policy.
To do that, I'll filter the reservation and water columns and then check.

```
poldata_rwi <- subset(poldata[c("reserved", "water")])
```

This includes both the ones with reserved policy and not. I filter the ones without the reserve policy to check the affect on female leadership.

```
poldata_rwi <- subset(poldata_rwi, reserved == 1)
```

Then, I run the model to see the coefficients

```
bivar_rwi <- lm(reserved ~ water, data = poldata_rwi)
summary(bivar_rwi)
```

Per the coefficient estimates, reservation policy enables a slight increase in the number of water works conducted.