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`).

Step 1: Load and Process The Data:

```
Bribe <- read.csv(”C:/Users/Caesar/Documents/GitHub/StatsI_Fall2022/problemSets/PS02/My_Answers/Bribe.csv”)

names(Bribe_num) <- Bribe[, 1]
Bribe_num <- Bribe[, -1]
head(Bribe_num)
# First Column: 1 for Upper Class, 2 for Lower Class
```

Step 2: Calculate the expected frequency of each cell

```
exp_Frq_Upper_NS <- sum(Bribe_num[1, ])*sum(Bribe_num[, 1])/sum(Bribe_num)
exp_Frq_Upper_NS

exp_Frq_Upper_BR <- sum(Bribe_num[1, ])*sum(Bribe_num[, 2])/sum(Bribe_num)
exp_Frq_Upper_BR

exp_Frq_Upper_SGW <- sum(Bribe_num[1, ])*sum(Bribe_num[, 3])/sum(Bribe_num)
exp_Frq_Upper_SGW

```

```
exp_Frq_Lower_NS <- sum(Bribe_num[2, ])*sum(Bribe_num[, 1])/sum(Bribe_num)
exp_Frq_Lower_NS

exp_Frq_Lower_BR <- sum(Bribe_num[2, ])*sum(Bribe_num[, 2])/sum(Bribe_num)
exp_Frq_Lower_BR

exp_Frq_Lower_SGW <- sum(Bribe_num[2, ])*sum(Bribe_num[, 3])/sum(Bribe_num)
exp_Frq_Lower_SGW
```

Expected Frequency Upper Class is not stopped 13.5
Expected Frequency Upper Class is bribe requested: 8.36
Expected Frequency Upper Class Stopped/Given warning: 5.14
Expected Frequency Lower Class Not Stopped: 7.5
Expected Frequency Lower Class bribe Requested: 4.64
Expected Frequency Lower Class Stopped/Given warning: 2.86
Step 3: Calculate the sum of X-square/Test Statistics

```
TS <-(13.5-14)^2/13.5+(8.36-6)^2/8.36+(5.14-7)^2/5.14+(7.5-7)^2/7.5+(4.64-7)^2/4.64+(2.86-1)^2/2.86
TS
```

The sum of X-square/Test Statistics is 3.80.

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

```
df <- (3-1)*(2-1)
df
```

The degree of freedom is 2. The Chi square value with degree of freedom 2, alpha = 0.1 is 4.61. The test statistics 3.8 < 4.61. The obtained chi-square value did not exceed the critical value of 4.61. Therefore, we cannot reject the null hypothesis that soliciting a bribe by police or not is independent from the class of driver.

or

```
pchisq(3.80, df =2, lower.tail = FALSE)
```

p-value is 0.15, which is larger than 0.1. Therefore, we cannot reject the null hypothesis that soliciting a bribe by police or not is independent from the class of driver.

[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.

```r
Res_Upper_NS <- (14-13.5)/sqrt(13.5*(1-sum(Bribe_num[1, ])/sum(Bribe_num))*(1-sum(Bribe_num[ ,1])/sum(Bribe_num)))
Res_Upper_NS

Res_Upper_BR <- (6-8.36)/sqrt(8.36*(1-sum(Bribe_num[1, ])/sum(Bribe_num))*(1-sum(Bribe_num[ ,2])/sum(Bribe_num)))
Res_Upper_BR

Res_Upper_SGW <- (7-5.14)/sqrt(5.14*(1-sum(Bribe_num[1, ])/sum(Bribe_num))*(1-sum(Bribe_num[ ,3])/sum(Bribe_num)))
Res_Upper_SGW

Res_Lower_NS <- (7-7.5)/sqrt(7.5*(1-sum(Bribe_num[2, ])/sum(Bribe_num))*(1-sum(Bribe_num[ ,1])/sum(Bribe_num)))
Res_Lower_NS

Res_Lower_BR <- (7-4.64)/sqrt(4.64*(1-sum(Bribe_num[2, ])/sum(Bribe_num))*(1-sum(Bribe_num[ ,2])/sum(Bribe_num)))
Res_Lower_BR

Res_Lower_SGW <- (1-2.86)/sqrt(2.86*(1-sum(Bribe_num[2, ])/sum(Bribe_num))*(1-sum(Bribe_num[ ,3])/sum(Bribe_num)))
Res_Lower_SGW
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.3220	-1.6437	1.5258
Lower class	-0.3220	1.6445	-1.5246

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

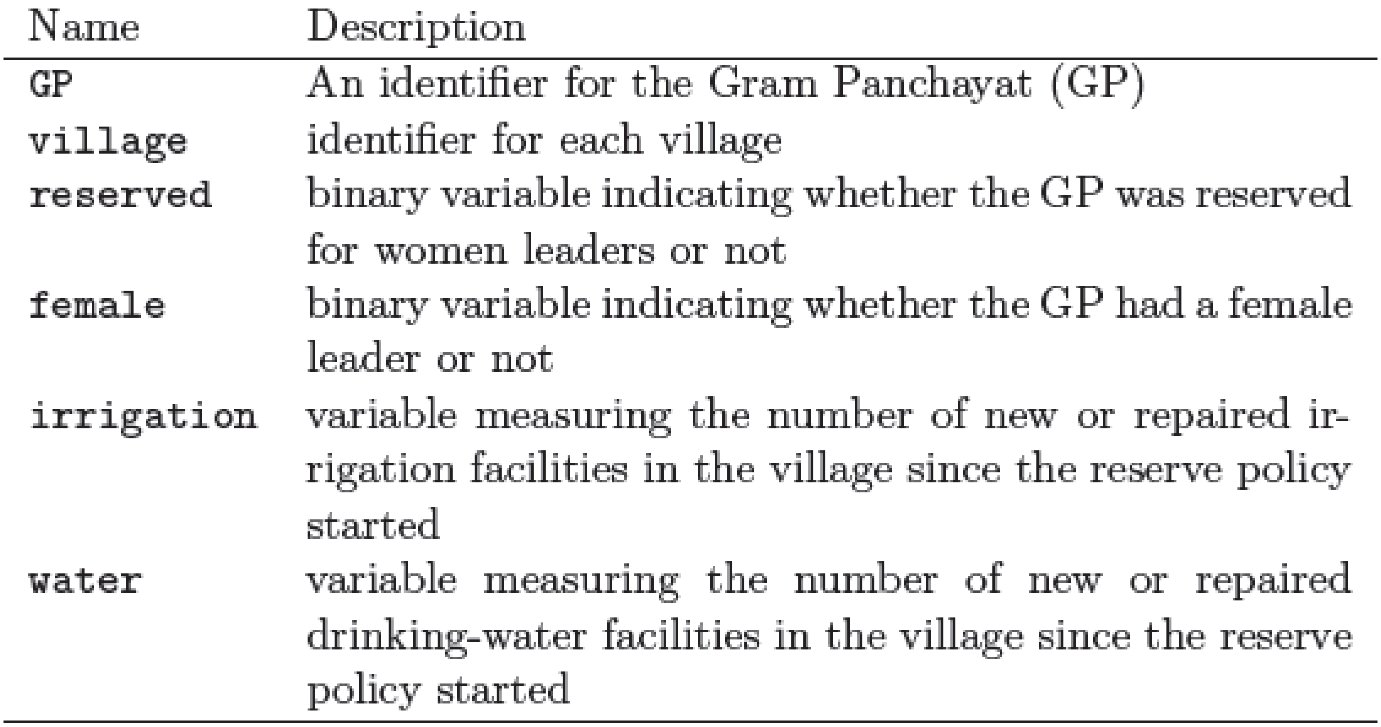
The absolute values of standard residuals in all cells are smaller than 2, we can therefore informally conclude that we cannot reject the null hypothesis that the two variables, drivers' class and police's bribing behaviour, are independent. In the cells of upper class not stopped and lower class not stopped, the absolute values of standard residuals are closer to 0, which indicate there is a stronger evidence that the drivers' class and they are not stopped by police are independent from each other.

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.

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.
Null Hypothesis: There is no correlation between the existence of reservation policy and the number of new or repaired drinking water facilities in the villages.

(b) Run a bivariate regression to test this hypothesis in `R` (include your code!). Step 1: load data

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

Step 2: run a bivariate regression

```
Reg_Gen_Wat <- lm(water ~ reserved, data = WESTB)
summary(Reg_Gen_Wat)

library(stargazer)
stargazer(Reg_Gen_Wat, type = "html", out = "Reg_Gen_Wat.html")
```

(c) Interpret the coefficient estimate for reservation policy.
The p-value of the coefficient of female is 0.0197, which is lower than 0.05. Therefore, we can reject the null hypothesis that there is no correlation between the existence of reservation policy and the number of new or repaired drinking water facilities in the villages at the 95 per cent level.
The coefficient of reserved is 9.252, which indicates that when there is a reservation policy in place, we expect to see an average difference by 9.252 higher in the number of new or repaired drinking water facilities in the villages.