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).

```r
#First create your variables including row and column totals
upper_Class <- c(14, 6, 7)
lower_Class <- c(7, 7, 1)
tot_ucrow <- sum(upper_Class)
tot_lcrow <- sum(lower_Class)
col_tot1 <- sum(14, 7)
col_tot2 <- sum(6, 7)
col_tot3 <- sum(7,1)

# Calculate the expected frequencies by dividing the row total
# by the grand total of the rows and multiplying by the column total

fe_1<- tot_ucrow/sum(tot_lcrow, tot_ucrow)*col_tot1
fe_2 <- tot_ucrow/sum(tot_lcrow, tot_ucrow)*col_tot2
fe_3 <- tot_ucrow/sum(tot_lcrow, tot_ucrow)*col_tot3

fe_4 <- tot_lcrow/sum(tot_lcrow, tot_ucrow)*col_tot1
fe_5 <- tot_lcrow/sum(tot_lcrow, tot_ucrow)*col_tot2
fe_6 <- tot_lcrow/sum(tot_lcrow, tot_ucrow)*col_tot3

#complete your test statistic by subtracting the expected frequency
# from the observed frequency. Square this number and divide by the expexted frequency.

a <- ((14-fe_1)*(14-fe_1)/fe_1)
b <- ((6-fe_2)*(6-fe_2)/fe_2)
c <- ((7-fe_3)*(7-fe_3)/fe_3)
d <- ((7-fe_4)*(7-fe_4)/fe_4)
e <- ((7-fe_5)*(7-fe_5)/fe_5)
f <- ((1-fe_6)*(1-fe_6)/fe_6)

#sum these values

x2 <- sum(a,b,c,d,e,f)

##Therefore my X2 test statistic is 3.79
```

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

```
# First we work out the degrees of freedom
#df = (rows - 1)(columns - 1)

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

#Our df = 2

#Now we can work out our p value

p_Value = pchisq(x2, df=2, lower.tail=F)

p_Value

#> p_Value
[1] 0.1502306
```

The p value is greater than alpha.

Therefore we fail to reject the null hypothesis that officers were neither more or less likely to solicit a bribe from a driver depending on their class.

[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.

```
# Now we need our standardized residuals = fobserved - fexpected / se

14 - fe_1 / sqrt(fe_1(1-(tot_lcrow/42)(1-col_tot1/42)

residuals <- c(0.165198209, -1.094519872, 1.100135481,
-0.201444062, 1.262233821,  -1.304459049)
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	.17	-1.09	1.10
Lower class	-.20	1.26	-1.30

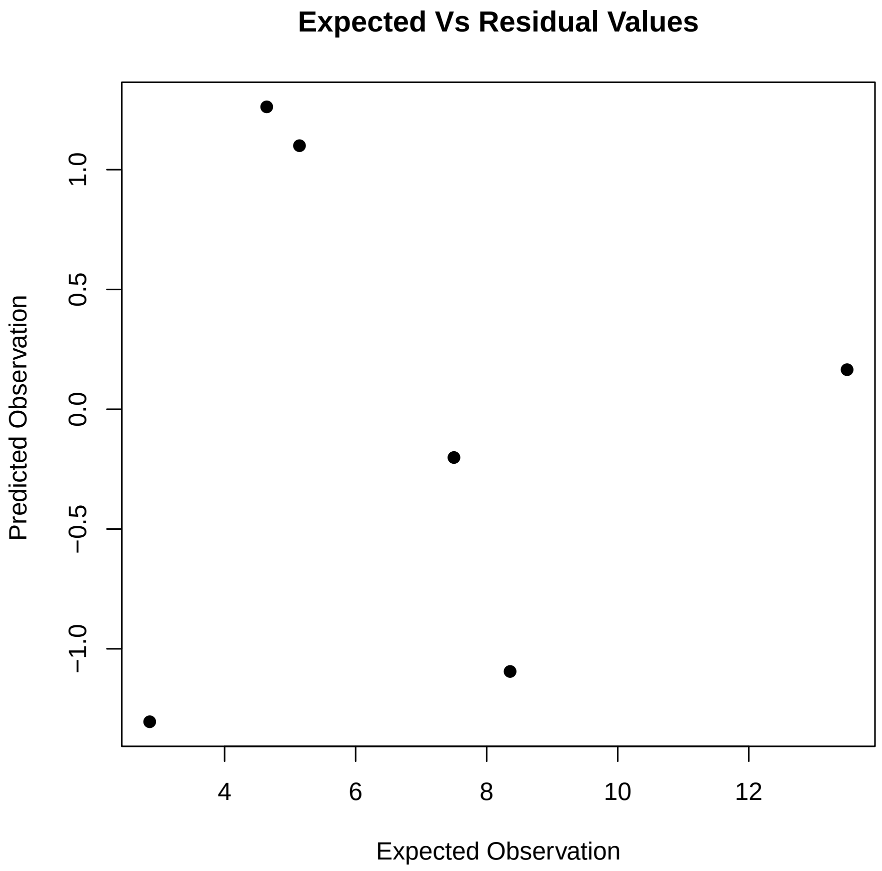
(d) How might the standardized residuals help you interpret the results?

A residual can help us determine the vertical distance. If the data point is above the line we receive a positive number and if it's below that line of best fit we get a negative number. If the residual is close to 0 we can see the model is a good fit for the data as 0 indicates our guess or predicted value was a match. The residuals for this data are not close to 0 indicating that the predicted values are not close to the expected values. I have plotted these to demonstrate how far the predicted values are from the 0 line. The further from the 0 line they are the less accurate the guess or prediction is.

```
residuals <- c(0.165198209, -1.094519872, 1.100135481,
-0.201444062, 1.262233821,  -1.304459049)

expected <- c(fe_1, fe_2, fe_3, fe_4, fe_5, fe_6)

plot(expected, residuals,
     main="Expected Vs Residual Values",
     xlab="Expected Observation",
     ylab="Predicted Observation", pch=19)
```


Expected Vs Residual Values
Predicted Observation
1.0
0.5
0.0
−0.5
−1.0
4
6
8
10
12
Expected Observation

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).

../../../../graphics/women_desc.png

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

Null Hypothesis: Female politicians are equally likely to support policies that female voters as male politicians.

Alternative Hypothesis: Female politicians are not equally likely to support policies female voters want as male politicians.

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

```
library(readr)

data <- read_csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
#First I must work out the mean and sum of my variables

mean(data$reserved)
# x -.34
sum(data$reserved)
# sumx = 108
mean(data$water)
#y - 17.84
sum(data$water)
# sumy = 5745

sum(data$water-mean(data$water))^2

## 2.47

s <- sum((data$water =
    mean(data$water))
    * (data$reserved=
      mean(data$reserved)))
s

## 5.98

B <- 5.98 / 2.47
B

#B = 2.42


2.42*.34

A <- 17.84 - (.82)

A

#A = 17.02

A-B


##Use the lm function to check your coefficient values A and B

```

```r
summary(lm(data$water ~ data$reserved , data = data))
```

```
 Output of Bivariate Regression:

     > summary(lm(data$water ~ data$reserved, data = data))

Call:
lm(formula = data$water ~ data$reserved, data = data)

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 ***
data$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 estimated coefficient value indicates that when x= 0 this is our y value. In this case if the reserved policy = 0 or is not present we see a score of 14.74 in our water measure. As the slope is not equal to zero this indicates there is a significant linear relationship between the two variables. We fail to reject the null hypothesis.