Problem Set 2

Applied Stats/Quant Methods 1 - 16327268 - Marcus Ó Faoláin

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 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).

(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.

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class			
Lower class			

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

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).

[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.

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

(c) Interpret the coefficient estimate for reservation policy.

Question 1 (40 points): Political Science

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

We start by inputting all of the observed frequencies into the code.

```
fo1 <- 14
fo2 <- 7
fo3 <- 6
fo4 <- 7
fo5 <- 7
fo6 <-1
```

We then calculate the row totals and the column totals by adding up the values in rows and columns respectively.

```
rt1 <- 14 + 6 + 7
rt1 # 27

rt2 <- 7+7+1
rt2 # 15

ct1 <- 14+7
ct1 # 21

ct2 <- 6+7
ct2 # 13

ct3 <- 7+1
ct3 # 8
```

We find out the total number of observed frequencies by adding up the two row totals.

```
total = rt1 + rt2
total #42
```

We calculate the expected frequencies for each cell by multiplying the row total by the column total and dividing by the total observed frequencies.

```
fe1 <- (rt1)*(ct1)/total
fe1 # 13.5

fe2 <- (rt2)*(ct1)/total
fe2 # 7.5

fe3 <- (rt1)*(ct2)/total
fe3 # 8.357143

fe4 <- (rt2)*(ct2)/total
fe4 # 4.642857

fe5 <- (rt1)*(ct3)/total
```

```
fe5 # 5.142857

fe6 <-(rt2)*(ct3)/total
fe6  # 2.857143
```

Through the use of the formula $\chi^2 = \sum \frac{(fo-fe)^2}{fe}$, we can calculate the χ^2 test statistic value.

```
# chi squared
chi_squared <- (((fo1-fe1)^2)/fe1) + (((fo2-fe2)^2)/fe2) + (((fo3-fe3)^2)/fe3) +
  (((fo4-fe4)^2)/fe4) + (((fo5-fe5)^2)/fe5) + (((fo6-fe6)^2)/fe6)

chi_squared
# 3.791168
```

The χ^2 value is equal to 3.791168

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

To calculate the p-value, we need to know the degrees of freedom. We obtain the degrees of freedom using the formula: df = (rows-1)(columns-1)

```r
# (b) degrees of freedom = (rows-1)*(columns -1)
df1 <- (2-1)*(3-1)
df1 # 2
```

The degrees of freedom is 2.

We can now calculate the p-values for chi-squared tests using the built in formula in R, inputting our χ^2 test statistic and our degrees of freedom.

```r
# Calculating the p-values for chi-squared tests:
pv1 <- pchisq(chi_squared, df = df1, lower.tail = FALSE)
pv1 # p-value = 0.1502306
```

We get a p-value of 0.1502306

Our null hypothesis (H0) states that police officers are neither more nor less likely to solicit a bribe from drivers depending on their class.

Our alternative hypothesis (Hα) states that officers likelihood of soliciting a bribe does depend on the class of the driver.

Since we have an α significance level of 0.1 and our p-value of 0.1502306 is greater than 0.1, we do not have enough evidence to reject the null hypothesis. We therefore accept the null hypothesis (H0).

We therefore cannot conclude from this study that the likelihood of an officer soliciting a bribe depends on the class of the driver committing the infraction.

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

To calculate the standardised residuals, we use the formula z $= \frac{fo-fe}{\sqrt{fe(1-rowprop.)(1-columnprop.)}}$.

```r
# (c) Calculating the standardised residuals
sr1 <- (fo1 -fe1)/(sqrt(fe1*(1-(rt1/total))*(1-(ct1/total))))
sr1 #0.3220306

sr2 <- (fo2 -fe2)/(sqrt(fe2*(1-(rt2/total))*(1-(ct1/total))))
sr2 # -0.3220306

sr3 <- (fo3 -fe3)/(sqrt(fe3*(1-(rt1/total))*(1-(ct2/total))))
sr3 # -1.642957

sr4 <- (fo4 -fe4)/(sqrt(fe4*(1-(rt2/total))*(1-(ct2/total))))
sr4 # 1.641957

sr5 <- (fo5 -fe5)/(sqrt(fe5*(1-(rt1/total))*(1-(ct3/total))))
sr5 # 1.523026

sr6 <- (fo6 -fe6)/(sqrt(fe6*(1-(rt2/total))*(1-(ct3/total))))
sr6 # -1.523026
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.3220306	-1.642957	1.523026
Lower class	-0.3220306	1.641957	-1.523026

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

Standardised residuals are used to measure the difference between the observed and expected values of a model.

In our case, we have used adjusted residuals to measure this difference.

In the same way that if a data point is more than two standard deviations away from the mean it is a highly unusual result, if a standardised residual is greater than a magnitude of two, it means that the residual is very unusual.

The sign of the standardised residual furthermore represents whether the observed value falls above or below the expected value based on the linear model developed. If the standardised residual is negative, the observed frequency is less than the expected frequency. If the standardised residual is positive, the observed frequency is greater than the expected frequency.

In the case of the 'Not Stopped' residuals, we can see that the standardised residuals are not that large at magnitudes around 0.322 each. This means that the expected and observed frequencies were not too far from each other and were therefore not too unexpected. Upper class people were a bit more likely to not be stopped, while lower class people were a little bit more likely to be stopped. It is reasonable to assume these results are not unusual.

In the cases of the 'Bribe requested' residuals, the magnitude of each residual is higher that the 'Not Stopped' residuals at around 1.64 each. The negative sign of the residual for the UC/BR residual suggests Upper Class people were slightly less likely to be requested for a bribe, while the positive sign of the LC/BR residual suggests lower class people were slightly more likely to be requested for a bribe. However since they are still below 2, they are not significant enough to show that this is an extremely unusal result.

In the case of the 'Stopped/given warning' residuals, the magnitude of the residuals is once more higher than the 'Not Stopped' residuals. The positive sign for the UC/SGW residual suggests that upper class people were slightly more likely to be given a warning, while the negative sign of the LC/SGW residual suggests lower class people were slightly less likely to be given a warning. Once more, as these residuals are less than 2, they are not significant enough to be extremely unusual.

Question 2 (40 points): Economics

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

Null hypothesis: The reservation policy has no impact on the number of new or repaired drinking water facilities in the villages.

H0: $\beta = 0$

Alternative hypothesis: The reservation policy does have an impact on the number of new or repaired drinking water facilities in the villages.

Hα: $\beta \neq 0$

Since this is a two tailed hypothesis, we must divide the α value for in order to get the critical values for both sides of the distribution

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

We start by importing the data into R Studio.

```r
# The first step is to import the data
bengal_data <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv", header=T)
```

We then use the lm() function in R to run the bivariate regression in R. We assign the 'reserved' data to be our explanatory variable, and our 'water' data to be our response variable.

```r
# We then carry out the bivariate regression
xreserved <- bengal_data$reserved
ywater <- bengal_data$water

bengalRegression <- lm(ywater ~ xreserved)
bengalRegression
```

Running the regression gives us an (α) intercept value of 14.738 and a (β) slope value of 9.252.

This β value of 9.252 is greater than the $\beta = 0$ value of the null hypothesis.

We use the summary() function on the regression to get the full results and to obtain values such as our standard errors, t-values and p-values.

```r
# To get the full results
summary(bengalRegression)
```

We wish to decide whether or not to accept or reject the null hypothesis. We use a significance level (α) of 0.05.

Since this is a two tailed hypothesis, we divide 0.05 by two and get a critical value is 0.025 on both sides of the distribution.

Since the p-value of the regression is 0.0197 and less than the critical value of 0.025, we can reject the null hypothesis (H0) that the reservation policy has no impact on the number of new or repaired drinking water facilities in the villages and accept the alternative hypothesis (Hα) that the reservation policy does have and impact on the number of new or repaired drinking water facilities. This means that there is a relationship between the reserved variable and the water variable.

(c) Interpret the coefficient estimate for reservation policy.

Our coefficient estimate of $\beta = 9.252$ means that for a one increase in our x variable ('reserved'), we expect on average a 9.252 increase in our y variable ('water').

In real terms, this would mean that villages that implement a reservation policy whereby one third of the village council heads have been reserved for one women on average have 9.252 more new or repaired drinking water facilities in the villages versus villages that don't have the reservation policy.

This would imply that there is a correlation between the reservation policy and the number of new or repaired drinking water facilities in the villages.