Problem Set 2

Applied Stats/Quant Methods 1

Due: October 14, 2024

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 Monday October 14, 2024. 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`).

```
totalUClass<-sum(table[1,-1])
totalLClass<-sum(table[2,-1])
totalnStopped<-sum(table$Not_Stopped)
totalBribe<-sum(table$Bribe_Requested)
totalStopped<-sum(table$Stopped)

print("The total presentations from upper class across all results was:")
print(totalUClass)

print("The total presentations from lower class across all results was:")
print(totalLClass)

print("The summary of the values in the columns 'Not Stopped', 'Bribe Requested' and 'Stopped or Given Warning' respectively were:")
print(totalnStopped)
print(totalBribe)
print(totalStopped)

total1<-totalLClass+totalUClass
total2<-totalBribe+totalnStopped+totalStopped
print(total1)
print(total2)

print("Having found all row and column totals, next the expected frequency for each cell must be found")

nStopped1<-c((totalUClass*totalnStopped/total1))
nStopped2<-c((totalLClass*totalnStopped/total1))
bribe1<-c((totalUClass*totalBribe/total1))
bribe2<-c((totalLClass*totalBribe/total1))
stop1<-c((totalUClass*totalStopped/total1))
stop2<-c((totalLClass*totalStopped/total1))
Expected_Frequency<-c("Upper Class", "Lower Class")
Not_StoppedF<-c(nStopped1,nStopped2)
BribeF<-c(bribe1,bribe2)
StoppedF<-c(stop1,stop2)
EFtable<-data.frame(Expected_Frequency,Not_StoppedF,BribeF,StoppedF)
print(EFtable)

print("Expected_Frequency Not_StoppedF    BribeF  StoppedF
1        Upper Class          13.5 8.357143 5.142857
```

```r
2        Lower Class           7.5 4.642857 2.857143")
print("Now to find x^2 statistic for each cell, take the expected frequency of that cell from the actual value, square it and divide by the expected frequency of the cell:")

ns1<-((Not_Stopped[1]-nStopped1)^2)/nStopped1
ns2<-((Not_Stopped[2]-nStopped2)^2)/nStopped2
brb1<-((Bribe_Requested[1]-bribe1)^2)/bribe1
brb2<-((Bribe_Requested[2]-bribe2)^2)/bribe2
stp1<-((Stopped[1]-stop1)^2)/stop1
stp2<-((Stopped[2]-stop2)^2)/stop2

Chi_Table<-c("Upper Class", "Lower Class")
Not_StoppedC<-c(ns1,ns2)
BribeC<-c(brb1,brb2)
StoppedC<-c(stp1,stp2)

chi.table<-data.frame(Chi_Table,Not_StoppedC,BribeC,StoppedC)
print(chi.table)

print("Summing all cell values from chi-table to find X^ statistic:")

XSquared<-sum(chi.table[,-1])

print(XSquared)

print("The X^2 statistic is: 3.791168")
```

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

```r
print("(b) Calculating the p-value:")

df<-(2-1)*(3-1)
print(df)
pValue<-pchisq(XSquared,df=2,lower.tail = FALSE)
print(pValue)

print("The p-value is: 0.1502306")

print("The p-value for this dataset is greater than 0.1.  Therefore, at the 10% significance level we fail to reject the null hypothesis - that is, from the evidence provided by this sample, you cannot determine if police officers are more or less likely to solicit a bribe from drivers 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.

```
StandRes<-c("Upper Class", "Lower Class")
srnotstop1<-((Not_Stopped[1]-nStopped1)/(nStopped1^(1/2)))
print(srnotstop1)

srnotstop2<-((Not_Stopped[2]-nStopped2)/(nStopped2^(1/2)))
srbribe1<-((Bribe_Requested[1]-bribe1)/(bribe1^(1/2)))
srbribe2<-((Bribe_Requested[2]-bribe2)/(bribe2^(1/2)))
srstop1<-((Stopped[1]-stop1)/(stop1^(1/2)))
srstop2<-((Stopped[2]-stop2)/(stop2^(1/2)))
Not_StoppedSR<-c(srnotstop1,srnotstop2)
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.1360828	-0.8153742	0.818923
Lower class	-0.1825742	1.0939393	-1.098701

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

```
print("Looking at the standardised residuals from this dataset, the first
  consideration of note is that the values all reside between the
  values of -2 and +2 non-inclusive, indicating that there are no
  significant deviations from the expected values. They are mixed
  positive and negative values, showing that the observed frequencies
  while not moving remarkably beyond what was expected, the observed
  frequencies were both higher and lower than what was statistically
  anticipated. The observations categorised as being 'not stopped' lay
  closest to what the expected observations would have been, with the
  observations of bribes being requested and individuals stopped or
  given warnings being equivalently close to 1 unit above or below the
  expected frequency (Lower class observations of bribes or being
  stopped/given warnings holding the strongest deviation from expected
  at over 1.09 positive and negative difference from the expected counts
  respectively. The equal split of positive and negative values across
  the standardised residuals with no discernable relation to class or
  outcome create difficultly in ascertaining if the outcomes or classes
  were more or less represented than expected. Overall, the
  standardised residuals of the combinations in this dataset, were not
  conspciously large enough or consistently positive or negative enough
  to infer that the observed frequency was notably different from the
  expected frequency.")
```

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.

```r
print("Null Hypothesis: There is no relationship between the presence of female leaders and the quality of drinking water in a village.")
print("Alternate Hypothesis: The level of female leaders in a village has an impact on the quality of drinking water in that village")
```

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

```r
print("Testing the hypothesis with a bivarate regression model:")


waterQual<-lm(water~GP+village+female+irrigation+reserved, data = bengalData)
summary(waterQual)

print("lm(formula = water ~ GP + village + female + irrigation + reserved, data = bengalData)
Residuals: Min =-89.140; 1Q = -12.628; Median = -5.277; 3Q = 5.035; Max = 285.749.")
print("For the coefficients, the Estimate, Std. Error, t value, and Pr(>|t|) for wach variable were as follows:")
print("(Intercept): 4.30135, 6.29921, 0.683, 0.495")
print("GP: -0.02493, 0.03684, -0.676, 0.499")
print("village: 5.00291, 3.40251, 1.470, 0.142")
print("female: -0.08429, 7.96740, -0.011, 0.992")
print("irrigation: 1.46776, 0.17995, 8.157, 8.21e-15")
print("reserved: 9.82875, 8.21005, 1.197, 0.232")


print("Residual standard error: 30.52 on 316 degrees of freedom, Multiple R-squared:  0.1915, Adjusted R-squared:  0.1787
F-statistic: 14.97 on 5 and 316 DF,  p-value: 3.406e-13")


print("From the model summary, we can see that only the presence of any new or repaired irrigation system has any statisticaly important bearing on the quality of drinking water in a village.  When all other variables are held constant, the increase of each additional unit of a repaired or new irrigation system has a positive correlation of increasing the quality of drinking water by 1.46776 units.  As the p-value for this variable is 8.21e-15, this is statistically meaningful at the 5% level (p<0.05).  No other variable has any statistically significant impact on the dependent variable (the quality of drinking water).  Considering first the lack of statistical significance of the presence of female leaders in a village, and the lack of statistical significance of the reservation of female places in government at the level of GP, we fail to reject the null hypothesis that the presence of female leaders has any impact on the quality of drinking water in a village.  Moreover, the Multiple R-Squared value of 0.1915 indicates that only 19.15% of variablility in the quality of drinking water is explained by this model, thus the model is clearly missing one or more
```

```
variables which will contribute to the explanation of variability in
drinking water quality by village.")
```

(c) Interpret the coefficient estimate for reservation policy.

```
print("The reservation policy (denoted in the model by 'reserved') infers
that there should be a substantial positive relationship between the
GP being reserved for women leaders and the quality of drinking water
in the area (that is, with all other variables held constant, for each
additional unit GP being reserved for female leaders, the coefficient
of reserved indicates that there should be a positive increase in the
units of quality of drinking water by 9.82875). However, the p-value
of 0.232 highlights that this relationship holds no statistically
significant correlation at the 10% level of significance, as p>0.1, so
no meaningful relationship between the two variables can be inferred
from the data provided in this model.")
```