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 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`).

	Not Stopped	Bribe requested	Stopped/given warning	Total
Upper class	$f_o = 14$ $f_e = 13.50$	$f_o = 6$ $f_e = 8.36$	$f_o = 7$ $f_e = 5.14$	27
Lower class	$f_o = 7$ $f_e = 7.50$	$f_o = 7$ $f_e = 3.69$	$f_o = 1$ $f_e = 3.81$	15
Total	21	13	8	42

$\chi^2 = \sum(f_0 - f_e)^2/f_e$
First, calculate the expected frequencies (f_e) for each cell:
Upper class:
Not stopped: f_e = (27 * 21) / 42 = 13.5
Bribe: f_e = (13 * 27) / 42 = 8.36
Warning: f_e = (8 * 27) / 42 = 5.14
Lower class:
Not stopped: f_e = (21 * 15) / 42 = 7.5
Bribe: f_e = (13 * 15) / 42 = 4.64
Warning: f_e = (8 * 15) / 42 = 2.86
Now calculate $(f_0 - f_e)^2/f_e$ for each cell:
Upper class:
Not stopped: $(14 - 13.5)^2/13.5 = 0.0185$
Bribe: $(6 - 8.36)^2/8.36 = 0.6662$
Warning: $(7 - 5.14)^2/5.14 = 0.6731$
Lower class:
Not stopped: $(7 - 7.5)^2/7.5 = 0.0333$
Bribe: $(7 - 3.69)^2/3.69 = 2.969$
Warning: $(1 - 3.81)^2/3.81 = 2.0725$
$\chi^2 = 0.0185 + 0.6662 + 0.6731 + 0.0333 + 2.969 + 2.0725 = 6.4326$
Therefore, the manually calculated χ^2 test statistic is 6.4326.

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

```
#df = (3-1) * (2-1) = 2
p_value = pchisq(6.4326, df=2, lower.tail=FALSE)
print(p_value)
#This gives a p-value of approximately 0.0401
```

At $\alpha = 0.1$ significance level, since p-value $< \alpha$, we reject the null hypothesis. This suggests there is a significant association between the driver's class and the officer's behavior.

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

	Not Stopped	Bribe requested	Stopped/given warning	Total
Upper class	$f_o = 14$ $f_e = 13.50$	$f_o = 6$ $f_e = 8.36$	$f_o = 7$ $f_e = 5.14$	27
Lower class	$f_o = 7$ $f_e = 7.50$	$f_o = 7$ $f_e = 3.69$	$f_o = 1$ $f_e = 3.81$	15
Total	21	13	8	42

$$Z_{11} = (14 - 13.5)/\sqrt{13.5 \cdot \left(1 - \tfrac{27}{42}\right) \cdot \left(1 - \tfrac{21}{42}\right)} = 0.322$$
$$Z_{12} = (6 - 8.36)/\sqrt{8.36 \cdot \left(1 - \tfrac{27}{42}\right) \cdot \left(1 - \tfrac{13}{42}\right)} = -1.9315$$
.
.
.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.9315	1.9414
Lower class	-0.322	3.0393	-2.5392

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

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

From our calculations, we observe:

Lower class drivers are requested bribes significantly more often than expected (3.0393). This is the largest positive residual, indicating this occurs far more frequently than random chance would predict.
Lower class drivers are stopped and given warnings significantly less often than expected (-2.5392). This is the largest negative residual, suggesting this occurs far less frequently than random chance would predict.
Upper class drivers are stopped and given warnings more often than expected (1.9414).
Upper class drivers are requested bribes less often than expected (-1.9315).
Both upper and lower class drivers are not stopped at rates close to expected (0.322 and -0.322 respectively), with relatively small deviations.

These results strongly suggest that police officers behave differently towards drivers of different classes. Lower class drivers are more likely to be requested bribes and less likely to be simply warned. In contrast, upper class drivers are more likely to be stopped and given warnings, and less likely to be requested bribes. This pattern may reflect discriminatory enforcement practices by the police based on class.

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.

Null Hypothesis
(H_0): There is no relationship between whether the GP (Gram Panchayat) has a female leader (female) and the number of new or repaired drinking-water facilities in the village since the reserve policy started (water).
$\beta_1 = 0$ (where β_1 iis the coefficient for the female variable)
Alternative Hypothesis
(H_a): There is a relationship between whether the GP has a female leader and the number of new or repaired drinking-water facilities in the village since the reserve policy started.
$\beta_1 \neq 0$

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

```r
# Read the data
data <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")

# Run the regression
lm <- lm(water ~ female, data=data)

# Display the results
summary(lm)
```

(c) Interpret the coefficient estimate for reservation policy.

```
Call:
lm(formula = water ~ female, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-22.68 -14.78  -7.81   2.29 317.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.813      2.382   6.220 1.56e-09 ***
female         7.864      3.838   2.049   0.0413 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.51 on 320 degrees of freedom
Multiple R-squared:  0.01295,   Adjusted R-squared:  0.009867
F-statistic: 4.199 on 1 and 320 DF,  p-value: 0.04126
```

Figure 2:

The coefficient estimate for the female variable is 7.864. Since female is a binary variable indicating whether the GP has a female leader or not, this coefficient can be interpreted as follows:
When a GP has a female leader (female=1), compared to GPs without a female leader (female=0), the number of new or repaired drinking-water facilities in the village since the reserve policy started increases by an average of 7.864.
This estimate is statistically significant (p-value = $0.0413 < 0.05$), indicating that we have sufficient evidence to reject the null hypothesis. There is a significant positive relationship between whether a GP has a female leader and the number of new or repaired drinking-water facilities in the village.