Problem Set 2

Applied Stats/Quant Methods 1

Yu Fan

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).

```
#f_observed
a <- 14
b <- 6
c <- 7
d <- 7
e <- 7
f <- 1
N <- a + b + c + d + e + f
# N=42

# f_expected = Row total/Grand total    Column total
expected_a= (a+b+c)*(a+d)/N
expected_b= (a+b+c)*(b+e)/N
expected_c= (a+b+c)*(c+f)/N
expected_d= (d+f+e)*(a+d)/N
expected_e= (d+f+e)*(b+e)/N
expected_f= (d+f+e)*(f+e)/N

#Chi-square
chi_squared <- ((a - expected_a)^2 / expected_a) +
  ((b - expected_b)^2 / expected_b) +
  ((c - expected_c)^2 / expected_c) +
  ((d - expected_d)^2 / expected_d) +
  ((e - expected_e)^2 / expected_e) +
  ((f - expected_f)^2 / expected_f)
print(chi_squared)
```

chi-squared=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
#(b)
# df
df <- 2
# p-value
p_value <- pchisq(chi_squared, df, lower.tail = FALSE)
p_value
```

P-value is 0.1502306, greater than 0.1. So we can't reject the Null. There is no significant association between a driver's social class and the police 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
Upper class			
Lower class			

```r
O <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE, dimnames = list
    (c("Upper class", "Lower class"), c("Not Stopped", "Bribe requested",
    "Stopped/given warning")))
E <- outer(rowSums(O), colSums(O)) / sum(O)
# props
row_props <- rowSums(O) / sum(O)
col_props <- colSums(O) / sum(O)
row_props
col_props
s_residuals <- (O - E) / sqrt(E*(1-row_props) %o% (1-col_props))

print(s_residuals)
```

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

[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?

Upper class: Not Stopped: The observed frequencies are slightly higher than the expected frequencies, but the difference is not significant (absolute value less than 1).

Upper class: Bribe requested: the number of observations is lower than the expected number of frequencies and the difference is close to significant (absolute value between 1 and 2).

Upper class: Stopped/given warning: The observed frequency is higher than the expected frequency and the difference is close to significant (absolute value between 1 and 2).

Lower class: Not Stopped: The observed frequency is lower than the expected frequency, but the difference is not significant.

Lower class: Bribe requested: The observed frequency is higher than the expected frequency and the difference is nearly significant.

Lower class: Stopped/given warning: the observed frequency is lower than the expected frequency and the difference is also close to significant.

Significant differences: upper class drivers were more likely to be given stops/warnings, while lower class drivers were more likely to be asked for bribes. These differences are significant, suggesting that social class may influence police behaviour.

Non-significant difference: the difference between the observed and expected frequencies of non-stopping behaviour across the two classes is not significant. These results suggest that there are significant differences in police behaviour towards drivers from different social classes, which may reflect inequalities in social structure and law enforcement practices.

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 (H0): There is no significant difference in the number of new or repaired drinking water facilities between villages with reserved and unreserved council heads.

Alternative Hypothesis (Ha): There is a significant difference iin the number of new or repaired drinking water facilities between villages with reserved and unreserved council heads.

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

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

head(data)
# Run bivariate regression
model <- lm(water ~ reserved, data = data)
summary(model)
# p-value=0.0197
```

Figure 2: bivariate regression results

```
Call:
lm(formula = water ~ 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 ***
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 coefficient estimate for the reservation policy is 9.252, indicating a positive correlation between the reservation policy and the number of new or repaired drinking water facilities in villages. Villages that adopted the reservation policy had 9.252 more drinking water facilities compared to those that did not implement the policy.