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).

```
# Question 1a
# Expected = Row total*Column Total/Grand Total

row1_total <- 14 + 6 + 7
row2_total <- 7 + 7 + 1
column1_total <- 14 + 7
column2_total <- 6 + 7
column3_total <- 7 + 1
grand_total <- row1_total + row2_total

fe1 <- (row1_total/grand_total)*column1_total
fe2 <- (row1_total/grand_total)*column2_total
fe3 <- (row1_total/grand_total)*column3_total
fe4 <- (row2_total/grand_total)*column1_total
fe5 <- (row2_total/grand_total)*column2_total
fe6 <- (row2_total/grand_total)*column3_total

# Chi^2 Stat = (fo - fe)^2/fe

Chi2 <- (14-fe1)^2/fe1 + (6-fe2)^2/fe2 + (7-fe3)^2/fe3 + (7-fe4)^2/fe4 +
  (7-fe5)^2/fe5 + (1-fe6)^2/fe6
```

```
Chi-Squared Test Statistic = 3.79
```

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

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

```r
# Question 1b
# Number of rows = 2
# Number of columns = 3
# DF = (2-1)(3-1) = 2

pval_chi <- pchisq(Chi2, df = 2, lower.tail = FALSE)
```

P-value = 0.15. Since the p-value is greater than 0.1, there is not sufficient evidence to reject the null hypothesis that officers soliciting a bribe from drivers is statistically independent of class when a = 0.1. We therefore conclude the two variables are statistically independent of each other.

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

```r
# Question 1c
# Calculate the standardised residuals for every cell in the table using
    the
# given formula

# Standardised Residuals
z11 = (14-fe1)/sqrt(fe1*(1-row1_total/grand_total)*(1-column1_total/grand
    _total))
z12 = (6-fe2)/sqrt(fe2*(1-row1_total/grand_total)*(1-column2_total/grand_
    total))
z13 = (7-fe3)/sqrt(fe3*(1-row1_total/grand_total)*(1-column3_total/grand_
    total))
z21 = (7-fe4)/sqrt(fe4*(1-row2_total/grand_total)*(1-column1_total/grand_
    total))
z22 = (7-fe5)/sqrt(fe5*(1-row2_total/grand_total)*(1-column2_total/grand_
    total))
z23 = (1-fe6)/sqrt(fe6*(1-row2_total/grand_total)*(1-column3_total/grand_
    total))
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.642	1.523
Lower class	-0.322	1.641	-1.523

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

Cell-by-cell comparison of the residuals helps describe the pattern of association among the cells - thereby making it clearer exactly where any deviations from independence between variables may be taking place. A cell with a large standardized residual provides evidence against independence in that cell. When H0 is true, there is approximately a 5% chance that any cell has a standardized residual value that exceeds 2.

In the table above, all of the standardized residuals are below 2 in absolute value. This indicates that there is not strong evidence of dependence in any of the cells - further backing up the chi-squared test results which indicated the two variables are statistically independent.

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).

[3] Chattopadhyay and Duflo. (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica*. 72 (5), 1409-1443.

```r

# Question 2
# Load in and Inspect the Data
df <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/
    PREDICTION/women.csv")
```

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

Step 1 - Assumptions:

1. Data are randomly generated and quantitative.
2. Observations are independent.
3. There is a linear relationship between the explanatory and outcome variables.
4. The errors are normally distributed with zero mean and a constant variance.

Step 2 - Hypotheses:

H0: The reservation policy has had no effect on the number of new or repaired drinking water facilities in the villages

H1: The reservation policy has had an effect on the number of new or repaired drinking water facilities in the villages

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

```
1
2 # Question 2b
3 # Linear Regression Model
4 model1 <- lm(df$water ~ df$reserved)
5 summary(model1)
```

Step 3: T-Value for X variable (reserved) = 2.344
Step 4: P-Value = 0.0197

P-value is less than 0.05, therefore the variable, reserved, is statistically significant at the 5% significance level.

Table 1:

	Dependent variable:
	water
reserved	9.252**
	(3.948)
Constant	14.738***
	(2.286)
Observations	322
R^2	0.017
Adjusted R^2	0.014
Residual Std. Error	33.446 (df = 320)
F Statistic	5.493** (df = 1; 320)
Note:	*p<0.1; **p<0.05; ***p<0.01

Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

(c) Interpret the coefficient estimate for reservation policy.

Step 5 - Conclusions:
The coefficient for the reservation variable indicates that on average, GPs that were reserved for female leaders are associated with a 9.252 unit increase in the number of new or repaired water facilities in the village relative to GPs that do not have the reservation policy in place. Assuming a significance level of a = 0.05, this variable is statistically significant at the 5% significance level (since the p-value is less than 0.05). Hence we can reject the null hypothesis that on average there is no difference in the number of new or repaired drinking water facilities in villages that have the reservation policy relative to villages that don't have the reservation policy.

The constant indicates that on average, when there is no reservation policy in the GP, the village will have 14.738 new or repaired water facilities since the reserve policy started. On average, villages with the reservation policy in place will have 23.99 (14.738 + 9.252) new or repaired water facilities since the reserve policy started.