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).

We use the chi-square test as a way to test the statistical independence of our variables. Before conducting the chi-squared test, we state our null and alternative hypothesis:

H_0: These variables are statistically independent. H_1: These variables are statistically dependent.

I will first sum up the rows, the columns, and then everything all together.

```
# I'm building the given table
experiment <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)
rownames(experiment) <- c("Upper class", "Lower class")
colnames(experiment) <- c("Not Stopped", "Bribe requested", "Stopped/
    given warning")
experiment
```

Secondly, I will calculate the expected frequences following this formula:

$$f_{1e} = \frac{(\text{row total}_0 \times \text{column total}_e)}{\text{grand total}}$$

```
# I now calculate the expected frequencies
row_total <- rowSums(experiment) # I sum up the total rows
col_total <- colSums(experiment) # I sum up the total cols
total <- sum(experiment) # Sum up everything

f <- outer(row_total, col_total) / total
f
```

These are the expected frequences from the previous table:

```
             Not Stopped Bribe requested Stopped/given warning
Upper class         13.5        8.357143                5.142857
Lower class          7.5        4.642857                2.857143
```

Then, I proceed to calculate the chi-squared:

```
# Now that I have the expected frequencies, I calculate the chi-squared
chi_squared <- sum((experiment - f)^2 / f)
chi_squared
```

```
[1] 3.791168
```

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

I first calculate the degrees of freedom:

```
# Now I calculate the degrees of freedom
pchi <- (nrow(experiment) - 1) * (ncol(experiment) - 1)
```

And the p-value of the chi-squared:

```
# And now the p-value
p_value <- pchisq(chi_squared, pchi, lower.tail = FALSE)
p_value
```

Our p-value is:

```
[1] 0.1502306
```

If we take into consideration that $\alpha = 0.1$, and that our p-value is 0.15, we do not have enough evidence to reject H_0, which stated that the variables of our table ("Upper class" and "Lower class", and "Not Stopped", "Bribe requested", "Stopped/given warning") are independent. Therefore, we conclude that our variables are not independent.

[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	0.1360828	-0.8153742	0.818923
Lower class	-0.1825742	1.0939393	-1.098701

```r
# Calculate standardised residuals (how far are what we observe from what we expect)
stand_residuals <- (experiment - f) / sqrt(f)
stand_residuals
```

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

The standardized residuals help us understand how far the observed values are from the expected values. Values close to 0 (typically between -1 and +1) indicate that the observed values are relatively close to the expected ones. This means that the values we observe are similar to those we would expect to occur by chance. Conversely, values that range farther than this indicate a notable deviation from the expected counts, suggesting a significant association between the variables.

In our case, while some standardized residuals are relatively close to 0, indicating that observed values are close to expected values, others are farther from 0. For instance, the lower class has a significantly higher standardized residual for "Bribe Requested," while the upper class shows significant residuals for "Stopped/Given Warning." This suggests that there are notable differences in police interactions based on class. Therefore, we cannot conclude that there is no significant association; rather, the results indicate potential disparities in how different classes experience these outcomes.

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.

In this case, our null-hypothesis (H_0) is having the reserved policy does not affect the number of new drinking water facilities in the village. That is, villages with the reserved policy for female representatives did not present relevant differences in terms of new/repaired drinking water facilities.

$$H_0 : \beta_{\text{femcoun}} = 0$$

Our alternative hypothesis (H_1) states that there is a relevant association and effect between the having the reserved policy for female representatives in councils and the number of new/repaired water facilities in that village. That is, this hypothesis states that there is a noticeable difference between those village with the reserved policy and those villages with no reserved policy: villages with reserved policy could have either more or less new/repaired drinking water facilities than those villages with no reserved policy.

$$H_1 : \beta_{\text{femcoun}} \neq 0$$

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

Before running the bivariate regression, we make certain assumptions about our data. First, we consider there is a linear relationship between the variables. Second, we acknowledge there is a margin of error in the estimates and the relationship we are measuring. Third, we assume the data we have was collected through a randomised process. Furthermore, we also consider our observations to be independent. Lastly, we assume our data is normal and constant in variance: our Y values follow a normal distribution at X values with the same standard deviation σ at each X value (constant variance in Y for all X values).

We want to test whether this reservation policy is associated with an increase in new/repaired drinkinig water facilities. Thus, we need to reject (H_0) first.

In this case, our outcome variable (Y) is "water", which capsulates the number of water facilities opened/reformed in a village. The explanatory variable (X) is the reserved policy ("reserved").

The regression formula we are goiing to use is the following:

$$y = \alpha + \beta_1 x + \epsilon \tag{1}$$

```r
m1 <- lm(water ~ reserved, data = village)
summary(m1)
stargazer(m1, type = "latex")
```

(c) Interpret the coefficient estimate for reservation policy.

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

Our constant α represents the value Y ("water" variable) takes when X ("reserved") is 0. Considering that our variable "reserved" is a binary one which takes "1" for those villages that did apply the policy, and "0" for those that didn't, we interpret the estimate as follows: those villages that did not apply the reserved policy (therefore, are "0" in variable Y = "reserved") have, on average, 14.738 units of water facilities. This predicted value is statistically relevant, as its p-value is less than 0.01 in a confidence interval of 99

To know how this differs in those villages that did apply the "reserved" policy, we must look at our β_1 estimate. The β_1 is 9.252. Again, considering that our X variable is binary, we interpret the estimate as follows: those villages that applied the "reserved" policy experiment, on average, an increase of water facilities of 9.252 in comparision with those villages that did not apply this policy. That is, those villages that did apply the "reserved" policy have, on average, 14.738 + 9.252 = 23.99 water facilities (9.252

more than villages without the "reserved" policy). This estimate is also significantly relevant, as its p-value is less than 0.05 in a confidence interval of 95

With these estimates, we have enough evidence to reject H_0, which states that the "reserved" policy had no significant effect on the number of water facilities in a village.

If we take a look at R^2, we see that our explanatory variable only explains about 1,7% of the variance of variable "water". That is a very low value, that might indicate that we need more other explanatory variables to fully assess what factors determine the number of water facilities in a village.