Problem Set 2

Daniel Murray (13303981)

Due: October 3, 2022

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 Sunday October 16, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1 (40 points): 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).

Answer:

The χ^2 test statistic was calculated manually in R, using the code below. The value for the χ^2 test statistic was found to be 3.79.

```
# Input observed frequencies
o_uppernotstopped <- 14
o_upperbribe <- 6
o_upperstopped <- 7
o_lowernotstopped <- 7
o_lowerbribe <- 7
o_lowerstopped <- 1

# Calculate row totals for the explanatory variable (class)
total_upper <- sum(o_uppernotstopped, o_upperbribe, o_upperstopped)
total_lower <- sum(o_lowernotstopped, o_lowerbribe, o_lowerstopped)

# Calculate column totals for the response variable (bribe outcome)
total_notstopped <- sum(o_uppernotstopped, o_lowernotstopped)
total_bribe <- sum(o_upperbribe, o_lowerbribe)
total_stopped <- sum(o_upperstopped, o_lowerstopped)

# Calculate overall sample size
total_sample <- sum(total_upper, total_lower)

# Calculate expected frequencies that would satisfy a null hypothesis
# of independence
e_uppernotstopped <- (total_upper*total_notstopped/total_sample)
e_upperbribe <- (total_upper*total_bribe/total_sample)
e_upperstopped <- (total_upper*total_stopped/total_sample)
e_lowernotstopped <- (total_lower*total_notstopped/total_sample)
e_lowerbribe <- (total_lower*total_bribe/total_sample)
e_lowerstopped <- (total_lower*total_stopped/total_sample)

# Calculate chi-squared test statistic by 1) for each cell, squaring the
# differences between the observed and expected frequencies and then
    dividing
# that square by the expected frequency, and 2) summing these values.
chi_sqrd <-
  (((o_uppernotstopped - e_uppernotstopped)^2)/e_uppernotstopped) +
```

```
  (((o_upperbribe - e_upperbribe)^2)/e_upperbribe) +
  (((o_upperstopped - e_upperstopped)^2)/e_upperstopped) +
  (((o_lowernotstopped - e_lowernotstopped)^2)/e_lowernotstopped) +
  (((o_lowerbribe - e_lowerbribe)^2)/e_lowerbribe) +
  (((o_lowerstopped - e_lowerstopped)^2)/e_lowerstopped)

chi_sqrd
```

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

Answer:

The degrees of freedom was first calculated by multiplying the number of rows minus one by the number of columns minus one, and was found to equal 2. The P-value was then calculated using the `pchisq` function. Its value was found to be approximately 0.15.

The code below illustrates this process:

```
# Calculate degrees of freedom (df)
df <- (2-1)*(3-1)
df

# Calculate P-value for the test statistic
p_value <- pchisq(chi_sqrd, df, lower.tail = FALSE)
p_value
```

As the p-value (≈ 0.15) is higher than our $\alpha = 0.1$ threshold, we do not find sufficient evidence to reject the null hypothesis that there is no relationship between a driver's class and the likelihood of an officer soliciting a bribe after committing a minor traffic offence.

[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.

Answer:

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.32	-1.64	1.52
Lower class	-0.32	1.64	-1.52

The code below was used to calculate the standardised residuals in R:

```
# Calculate the standardised residuals for each cell
sr_uppernotstopped <- (o_uppernotstopped - e_uppernotstopped) /
  sqrt(e_uppernotstopped*(1-(total_upper/total_sample))*(1-(total_
    notstopped/total_sample)))
sr_upperbribe <- (o_upperbribe - e_upperbribe) /
  sqrt(e_upperbribe*(1-(total_upper/total_sample))*(1-(total_bribe/total_
    sample)))
sr_upperstopped <- (o_upperstopped - e_upperstopped) /
  sqrt(e_upperstopped*(1-(total_upper/total_sample))*(1-(total_stopped/
    total_sample)))
sr_lowernotstopped <- (o_lowernotstopped - e_lowernotstopped) /
  sqrt(e_lowernotstopped*(1-(total_lower/total_sample))*(1-(total_
    notstopped/total_sample)))
sr_lowerbribe <- (o_lowerbribe - e_lowerbribe) /
  sqrt(e_lowerbribe*(1-(total_lower/total_sample))*(1-(total_bribe/total_
    sample)))
sr_lowerstopped <- (o_lowerstopped - e_lowerstopped) /
  sqrt(e_lowerstopped*(1-(total_lower/total_sample))*(1-(total_stopped/
    total_sample)))
```

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

Answer:

Whereas the χ^2 test statistic summarises how close the observed frequencies are to the expected frequencies over all cells in the contingency table, the standardised residuals help us to understand the pattern of association among the individual cells. When H_0 (independence) is true, the standardised residuals would have a normal standard distribution. This means that there is a very small probability (approximately 5%) of their values being larger than 2 in absolute value.

In the case of our data, the standardised residuals are all less than 2 in absolute value. This suggests that no individual cell has too many or too few observations to indicate a departure from independence between the variables.

Question 2 (40 points): 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.

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

Answer:

As we are using a bivariate linear regression, we want our null hypothesis to reflect "no effect" for our β coefficient estimate. In other words, it states that our β coefficient estimate is equal to zero; there is no statistically significant relationship between x and y. As we want our alternative hypothesis to be two-tailed, it will simply state that our β coefficient estimate is not equal to zero, rather than indicating any particular directionality. It states that there is a statistically significant relationship between x and y.

Below are our null and alternative (two-tailed) hypotheses:

$$H_0 : \beta = 0$$
$$H_a : \beta \neq 0$$

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

Answer:

We begin by creating a separate data frame to include only data for GPs reserved for female leaders and GPs with a male leader. This is to remove GPs not reserved for female leaders but which have a female leader, as these GPs are thought to have potential confounding problems. We do this by subsetting the dataset and combining the two subsets using the `rbind` function

Next we use the `lm` function to run a bivariate linear regression with "water" as the response variable (y) and "female" as the explanatory variable (x).

```
# Import data file
data <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv", header=T)

# Create a subsetted data frame to remove GPs not reserved for female leaders but which have a female leader
data2 <- rbind(data[data$reserved == "1",], data[data$female == "0",])

# Run bivariate linear regression with "water" as response variable and "female" as the explanatory variable
ols <- lm(water ~ female, data = data2)
summary(ols)
```

`summary(ols)` in the code above produces the below output:

Figure 2: Summary of lm() function output

```
Call:
lm(formula = water ~ female, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-23.991 -14.813  -7.991   2.187 316.009

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.813      2.429   6.099 3.24e-09 ***
female         9.178      4.088   2.245   0.0255 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.18 on 304 degrees of freedom
Multiple R-squared:  0.01631,   Adjusted R-squared:  0.01307
F-statistic: 5.039 on 1 and 304 DF,  p-value: 0.0255
```

We can also calculate our α and β coefficients by hand in `R` using the code below:

```r
# Calculate alpha and beta coefficients by hand
beta <- sum((data2$water - mean(data2$water))*
              (data2$female - mean(data2$female)))/
  sum((data2$female - mean(data2$female))^2)

alpha <- mean(data2$water) - beta*mean(data2$female)
```

This gives the same values as Figure 2 above, with $\alpha = 14.813$ and $\beta = 9.178$.

(c) Interpret the coefficient estimate for reservation policy.

Answer:

In order to draw conclusions about our coefficient estimate for reservation police ($\beta = 9.178$), we construct a 95% confidence interval for its values. This has the following formula:

$$\beta \pm t(se)$$

Using a T table, we find the t-value for two-tailed 0.05 significance level at 304 degrees of freedom (n-2). This value is 1.9678.

Next we calculate the standard error of beta by dividing the residual standard error of the regression by the square root of the total sum of squares for x. The code below shows these calculations:

```
# Calculate residual standard error
rse <- sqrt(sum(ols$residuals^2)/(length(data2$water)-2))
rse

# Calculate total sum of squares of x (female)
TSSx <- sum((data2$female-mean(data2$female))^2)
TSSx

# Calculate standard error for beta
SEb <- rse/(sqrt(TSSx))
SEb
```

This give us a value of 4.088 for the standard error for β. Using these values we can now calculate our confidence interval, as below:

$$\beta \pm t(se)$$

$$9.178 \pm 1.9678 * 4.088$$

$$9.178 \pm 8.044$$

Our 95% confidence interval for β is (1.13, 17.22). As this confidence interval does not contain 0, we reject the null hypothesis and conclude that at the 95% confidence level, there is a statistically significant positive association between reservation policy and the number of new or repaired drinking water facilities in villages. At a 95% confidence level, we find that villages with female leadership have on average between 1.1 and 17.2 more new or repaired drinking water facilities than villages with male leadership, with our best estimate being an additional 9.2 water facilities.