Problem Set 2

Applied Stats / Quant Methods 1

Lucas de Melo Prado / Due: October 16, 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).

To calculate χ^2 in R, we first import the data and create a contingency table:

```
# Import data:
notStopped <- c(14, 7)
bribeRequest <- c(6, 7)
warning <- c(7, 1)

# Create contingency table:
data <- as.data.frame(cbind(notStopped, bribeRequest, warning),
                      row.names= c("upperClass", "lowerClass"))
```

Then, we follow the steps to perform the test.

1. Assumption. Data collection was randomized and most frequencies ($> 75\%$) are ≥ 5

2. Hypotheses. For the *null hypothesis*, we consider that variables are statistically independent:

$$H_0 : f_0 = f_e \text{ or } H_0 : f_0 - f_e = 0.$$

The *alternative hypothesis* is that variables are statistically dependent:

$$H_a : f_0 \neq f_e.$$

3. Chi-squared test statistic. To perform the chi-squared test, we first need to find the expected frequencies (f_e) for each cell were the null hypothesis true:

$$f_e = \frac{(\text{row total}) \cdot (\text{column total})}{\text{overall sample size}}.$$

```
# Table of expected frequencies (fe):
notStopped_fe <- c(rowSums(data)*colSums(data)[1]/sum(data))
bribeRequest_fe <- c(rowSums(data)*colSums(data)[2]/sum(data))
warning_fe <- c(rowSums(data)*colSums(data)[3]/sum(data))

data_fe <- as.data.frame(cbind(notStopped_fe, bribeRequest_fe, warning_fe
    ),
                         row.names = c("upperClass", "lowerClass"))
```

We can summarize the values of observed frequencies (f_0) and expected frequencies (f_e) in the following table:

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	$f_0 = 14$ $f_e = 13.5$	$f_0 = 6$ $f_e = 8.36$	$f_0 = 7$ $f_e = 5.14$
Lower class	$f_0 = 7$ $f_e = 7.5$	$f_0 = 7$ $f_e = 4.64$	$f_0 = 1$ $f_e = 2.86$

Finally, we calculate χ^2:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(14 - 13.5)^2}{13.5} + \frac{(6 - 8.36)^2}{8.36} + \frac{(7 - 5.14)^2}{5.14} + \frac{(7 - 7.5)^2}{7.5} + \frac{(7 - 4.64)^2}{4.64} + \frac{(1 - 2.86)^2}{2.86}$$

$$\chi^2 = 3.79$$

```
# Chi-squared test statistics:
X2 <- sum(((data - data_fe)^2)/data_fe)
```

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

4. P-value.

```
# P-value:
df <- (nrow(data)-1)*(ncol(data)-1)
p <- pchisq(X2, df, lower.tail = FALSE)
```

$$df = (\text{number of rows - 1}) \cdot (\text{number of columns - 1}) = (2 - 1) \cdot (3 - 1) = 2$$

$$p = 0.15$$

5. Conclusion. For $\alpha = 0.1$, *we do not have enough evidence to reject the null hypothesis*, according to which variables are statistically 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.

To calculate the standardized residuals (z) for each cell, we use this formula:

$$z = \frac{f_0 - f_e}{se} = \frac{f_0 - f_e}{\sqrt{f_e \cdot (1 - \text{row proportion}) \cdot (1 - \text{column proportion})}}$$

```
# Standardized residuals:
notStopped_res <- round(c((data$notStopped - data_fe$notStopped_fe)/
                          sqrt(data_fe$notStopped_fe*
                               (1-(rowSums(data)/sum(data)))*
                               (1-(colSums(data)[1]/sum(data))))),
                          digits = 2)
bribeRequest_res <- round(c((data$bribeRequest-data_fe$bribeRequest_fe)/
                            sqrt(data_fe$bribeRequest_fe*
                                 (1-(rowSums(data)/sum(data)))*
                                 (1-(colSums(data)[2]/sum(data))))),
                            digits = 2)
warning_res <- round(c((data$warning - data_fe$warning_fe)/
                       sqrt(data_fe$warning_fe*
                              (1-(rowSums(data)/sum(data)))*
                              (1-(colSums(data)[3]/sum(data))))),
                       digits = 2)

data_res <- as.data.frame(cbind(notStopped_res, bribeRequest_res, warning
    _res),
                                row.names = c("upperClass", "lowerClass"))
```

These are the results we obtained:

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

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

Standardized residuals represent the number of standard errors the residuals $(f_0 - f_e)$ fall from the expected value if the null hypothesis is true, i.e., $f_0 - f_e = 0$. As standard residuals behave like a normal distribution, large standard residuals (below -3 or above 3) would be strong evidence against H_0.

However, in the present case, all standard residuals are between -2 and 2. This reinforces the high p-value we have already found. Therefore, standard residuals do not allow us to reject the null hypothesis either.

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).

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.

For the *null hypothesis*, we assume that there is no association between the two variables, *i.e.*, between the reservation policy and the number of new or repaired drinking water facilities in the village:

$$H_0 : \beta = 0$$

For the *alternative hypothesis*, we assume that there is some association between these two variables:

$$H_a : \beta \neq 0$$

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

1. Assumptions. We work with *four* assumptions: 1) data were randomly collected; 2) the association between variables is linear; 3) the values of the response variable for each value of the explanatory variable follow a normal distribution; and 4) all distributions of the response variable have the same standard deviation.

2. Hypothesis. $H_0 : \beta = 0$; $H_a : \beta \neq 0$.

3. Test statistic. For the test statistic of β, we perform a t-test. This test depends on the prediction equation for the linear function ($\hat{y} = a + bx$), as well as on the standard error of β (se_β), which, in turn, depends on the standard deviation of β (s_β).

First, we establish the *prediction equation* ($\hat{y} = a + bx$):

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{664.10}{71.78} = 9.25$$

$$a = \bar{y} - b\bar{x} = 17.84 - 9.25 \cdot 0.33 = 14.74$$

$$\hat{y} = 14.74 + 9.25x$$

```r
# Import data:
women <- read_csv("women.csv")
x <- women$reserved
y <- women$water

# Linear function:
b <- sum((x - mean(x))*(y - mean(y)))/sum((x - mean(x))^2)
a <- mean(y) - b*mean(x)
y_hat <- a+b*x
```

Then, we calculate the *standard deviation* (s_β) and the *standard error* (se_β) of β:

$$s_\beta = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}} = \sqrt{\frac{357,956.3}{320}} = \sqrt{1,118.61} = 33.44$$

$$se_\beta = \frac{s_\beta}{\sqrt{\sum(x-\bar{x})^2}} = \frac{33.44}{8.47} = 3.95$$

```
# Standard deviation and standard error of beta:
s_beta <- sqrt((sum((y-y_hat)^2))/(length(y)-2))
se_beta <- s_beta/sqrt(sum((x-mean(x))^2))
```

Finally, we perform the *t-test* (t):

$$t = \frac{b-\beta_0}{se} = \frac{9.25}{3.95} = 2.34$$

```
# T-test:
t <- (b-0)/se_beta
```

4. P-value. For a two-tailed test, $p = 0.019$.

```
# P-value (two-tail):
p <- 2*pt(t, df=length(y)-2, lower.tail=FALSE)
```

Finally, we check our results:

```
# Checking:
summary(lm(y~x))
```

(c) Interpret the coefficient estimate for reservation policy.

Depending on the established significance level (α), we can draw different conclusions from the test. For a more demanding $\alpha = 0.01$, $p = 0.019$ does not provide enough evidence to reject the null hypothesis. However, for $\alpha = 0.05$, the test allows the rejection of the null hypothesis.

If we assume $\alpha = 0.05$ and, therefore, reject H_0, then the association between the reservation policy (x) and the number of new or repaired drinking water facilities (y) is positive ($b > 0$). The regression line intercepts the y-axis at $a = 14.74$, and y increases $b = 9.25$ for each one-unit increase in x.