Problem Set 2: Solution

Student: Shekhar Kedia (23351315)
Applied Stats/Quant Methods 1

Due: October 15, 2023

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 15, 2023. No late assignments will be accepted.

Notes:

- The responses are nested right after each problem in blue colour. For instance- `For problem 1.a`, the responses are mentioned right after the question for `1.a`.
- The format for response is in the following manner:
  - Steps to be followed
  - The `R` script
  - Interpretation of result
- The example `.R` file and `.tex` file that was provided, is modified to develop this document. My knowledge of library packages in both `R` and `LaTeX` is limited and so, I might have attached packages which are either redundant or unnecessary in doing this assignment.
- Many `R` functions used in the solutions are from the stats tutorial classes or Programming module taught to us as part of ASDS Term 1
- Links are mentioned in the `R` script if any external sources are referred for solution.

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.

	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).

Steps followed to calculate the χ^2 test statistic by hand using R:

- The expected frequencies are calculated if the two variables were independent. A function was created in R to avoid repetition of steps.
- The expected frequencies are calculated using the formula:

  $f_e = \frac{Rowtotal}{Grandtotal} * Columntotal$
- The χ^2 test statistic is then calculated using the formula:

  $\chi^2 = \Sigma \frac{(f_0 - f_e)^2}{f_e}$
- The built-in chisq.test() function in R is used to cross-verify the findings.

R script used for calculations:

```
# Creating the table for use
tab <- matrix(c(14, 6, 7, 7, 7, 1), ncol=3, byrow=TRUE)
colnames(tab) <- c('Not stopped','Bribe requested','Stopped/given warning
    ')
rownames(tab) <- c('Upper class','Lower class')
tab <- as.table(tab)
tab #Displaying the crosstab with observed frequencies

```

[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.

```r
# Creating a function to calculate the expected frequencies if the variables were independent
f_exp <- function (input_table, row, col) {
  total_sum <- sum(input_table)
  col_sum <- colSums(input_table)
  row_sum <- rowSums(input_table)
  return(col_sum[col:col]*row_sum[row:row]/total_sum)
}

# Creating the expected frequency table for use
f_expected <- c() #Creating an empty vector
for (row in 1:nrow(tab)) {
  for (col in 1:ncol(tab)) {
    temp = f_exp(tab,row,col) #Temporary variable to capture the expected frequency
    f_expected = c(f_expected, temp) #Appending values after end of each iteration
  }
}
tab_exp <- matrix(f_expected, ncol = 3, byrow = TRUE)
tab_exp <- as.table(tab_exp)
tab_exp #Displaying the crosstab with expected frequencies

# Calculating the chisquare statistics by hand using the formula taught in stats class
chisqr <- 0
for (row in 1:nrow(tab)) {
  for (col in 1:ncol(tab)) {
    chisqr = chisqr + ((tab[row,col]-tab_exp[row,col])^2/tab_exp[row,col])
  }
}
chisqr #Displaying the chisquare value calculated by hand

# Using the built-in function to cross verify the findings
chisq.test(tab) #We get the same chisquare value = 3.7912
```

Interpretation of results:
The χ^2 test statistic by hand is equal to 3.7912. This was also cross-verified using the built-in chisq.text() function which gives the same result.

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

Steps followed to calculate the p-value using R:

- The in-built pchisq() function is used to calculate the p-value
- The degrees of freedom (df) is calculated using the following formula:
  df = (rows-1)(columns-1)
- We then compare the p-value with the given significance level $\alpha = 0.1$.

R script used for calculations:

```
# Using the built-in function to calculate the p-value
p_val <- pchisq(chisqr, df= ((nrow(tab)-1)*(ncol(tab)-1)), lower.tail =
    FALSE)
p_val #We get the p-value = 0.1502
```

Interpretation of results:
The p-value calculated is equal to 0.1502 which is $>$ than given $\alpha = 0.1$. We therefore do not have sufficient evidence to reject the null hypothesis or we fail to reject the null hypothesis that the two variables are statistically independent.

In other words we conclude that, we do not have sufficient evidence to say that the officers were more or less likely to solicit a bribe from drivers depending on their class.

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.3220306	-1.6419565	1.5230259
Lower class	-0.3220306	1.6419565	-1.5230259

R script used for calculations:

```
(chisq.test(tab))$stdres
```

N.B: The R script file has calculations by hand as well as use of built-in function.

[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?

The standardized residuals (adjusted) give a standardized or unit-less measure of how far each observed value is from expected value if the two variables were independent i.e. the null hypothesis was true.

The magnitude of all the adjusted standardized residuals in this case is not very high i.e. < 2. This is also an indication that the observed value do not vary highly than the expected value (which is why we also see that the chisquare test fails to reject the null hypothesis).

Again, The sign (positive or negative) of the adjusted standardized residuals also indicate whether we have more or less observations than expected. For instance, the adjusted standardized residual value for 'bribe requested' category for Upper class is -1.64 i.e. the observed number of cases are fewer than what we would expect if the two variables were 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.

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

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

R script used for importing the data and viewing/exploring it:

```
# Reading the csv data and importing to R environment
ind_eco <- read.table("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv", header = TRUE, sep = ",")

View(ind_eco) #Viewing the data to visually understand the structure and spread

summary(ind_eco) #Inspecting the data through summary
```

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

The null hypothesis (H_o) = There is no effect of reservation policy on the number of new or repaired drinking water facilities in the villages.

The alternate hypothesis (H_a) = There is effect of reservation policy 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!).

Steps followed to run a bi-variate regression model using R:

- The hypothesis helps us identify the variables we need to consider. We then check the data type of the two variables (as it is a bi-variate regression, we are only bothered about one predictor/independent and one outcome/dependent variable in the model).

- The independent variable (reserved) in the data is binary and refers to whether or not the village was reserved for women leaders. Note, we assume that it is 1 if the village was reserved for women leaders and 0 otherwise.
- The dependent variable (water) in the data is continuous and refers to the number of new or repaired drinking water facilities in the village since the reserve policy started.
- We then plot a scatter and check the correlation value to understand if there is an association between the two variables
- We then run the linear bi-variate regression using the built-in lm() function in R to identify the parameters (α and β) which explains the best linear relationship between the two variables using Ordinary Least Square (O.L.S.) technique by minimizing the error term.
  N.B. *We are not testing the assumptions of linear regression as it currently excludes the scope of the assignment. But ideally, all the assumption tests should be performed and only then should the linear regression model be built.*
- We check the regression parameters i.e., the intercept value (α) and the slope (β). We also check the p-value, especially the one corresponding to the slope (β), to understand if there is sufficient evidence to establish the linear relationship between the independent (reserved) and dependent (water) variable.
- We then also look at the R^2 value to understand the goodness of fit of the model or to understand the proportion of variance in the dependent variable that is explained by the independent variable.

R script used for calculations:

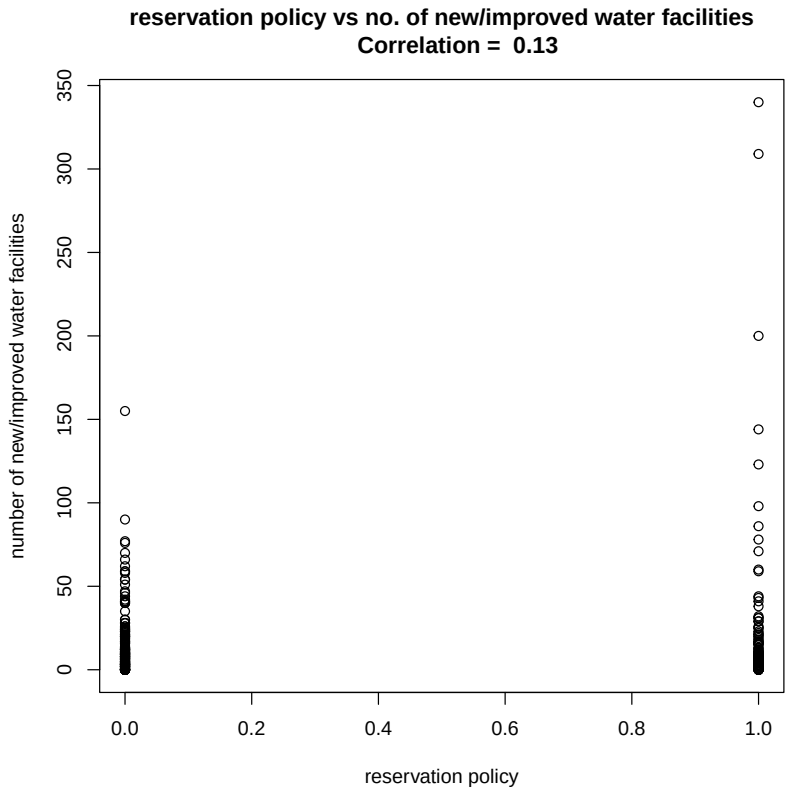
```
# Creating scatterplot & calculating correlation value to show relationship between the two variables
cor_Y_X <- round(cor(ind_eco$water, ind_eco$reserved),2) #Shows the default pearson correlation value rounded upto 2 decimals

pdf("plot_Y_X.pdf") #Creates a pdf file where we can then input/write the scatter plot
plot(ind_eco$reserved, ind_eco$water,
     xlab = "reservation policy",
     ylab = "number of new/improved water facilities",
     main = paste("reservation policy vs no. of new/improved water facilities
     Correlation = ", cor_Y_X)) #The scatter plot also has the correlation value mentioned in the title
dev.off()

# Bivariate regression function
reg <- lm(water~reserved, data = ind_eco)

summary(reg) #Displaying the summarized results of the linear regression model
```

The following graph shows the relationship between reservation policy in a village and number of new or improved drinking-water facilities:



Interpretation of results:

While the scatter plot is not very intuitive to show that there is an association between the reservation policy in a village and number of new or improved drinking-water facilities, we do see there are higher number of new or improved drinking-water facilities in villages where there is reservation policy i.e. reserved variable = 1.

We can also see some **outliers** and that may influence the regression output as well. However, we are not performing any data manipulation as it is beyond the scope of the assignment.

Next, the regression output shows that the y-intercept or $\alpha = 14.738$ and slope or $\beta = 9.252^*$.

Lastly, the goodness of fit of the model i.e., R^2 (adjusted) = 0.0138. This shows that the model explains about 14% of the relationship or 14% of the variance in the dependent variable (water) is explained by the independent variable (reserved). There remains a lot of unexplained variance and therefore a need for including other relevant independent variables.

(c) Interpret the coefficient estimate for reservation policy.

The regression output shows that the y-intercept or $\alpha = 14.738$ and coefficient estimate or $\beta = 9.252^*$.

This shows that we have sufficient evidence (significance level 5%) that villages having reservation policy for women leaders have on average about nine new or repaired drinking-water facilities more compared to villages where there are no reservation policy for women.
N.B.: We can also say that one unit increase in reservation policy or when the reservation policy switches from no reservation for women leaders in the village to having reservation, there is sufficient evidence that it leads to on an average increase of nine new or repaired drinking-water facilities in the village.

Furthermore, the intercept value indicates that irrespective of reservation policy for women leaders, a village will have on average about 15 new or repaired drinking-water facilities.

End of document

Last edits made on: October 15, 2023