Problem Set 2

Applied Stats/Quant Methods 1
Minh Trinh (Student ID:24350478)

October 14, 2024

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`).

We will calculate the χ^2 statistic in R by hand

```
#import observed data into matrix
observed_val_tab <- matrix( c(14,7,6,7,7,1),2,3 )
#add sum of row and columns to the observed value matrix
observed_val_tab <- addmargins(observed_val_tab)
```

Observed value table with total value:

[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.

```
            Sum
    14  6 7  27
     7  7 1  15
Sum 21 13 8  42
```

Under the assumption of independence between row and column variables we can start calculate each cell probability and expected value

```
#calculating the row and column probability
probability_col <- observed_val_tab[3,]/observed_val_tab[3,4]
probability_row <- observed_val_tab[,4]/observed_val_tab[3,4]
#doing matrix multiplication to calculate probability for each cell
probability_matrix <- matrix(probability_row,3,1)%*%matrix(probability_
    col,1,4)
```

Probability matrix of each cell:

```
          [,1]      [,2]       [,3]      [,4]
[1,] 0.3214286 0.1989796 0.12244898 0.6428571
[2,] 0.1785714 0.1105442 0.06802721 0.3571429
[3,] 0.5000000 0.3095238 0.19047619 1.0000000
```

```
#calculate expected value
expected_value_tab <- probability_matrix*observed_val_tab[3,4]
```

Matrix of expected values:

```
    [,1]      [,2]     [,3] [,4]
[1,] 13.5  8.357143 5.142857   27
[2,]  7.5  4.642857 2.857143   15
[3,] 21.0 13.000000 8.000000   42
```

```
#calculate the chi square stat
chi_square_stat <- sum((expected_value_tab[1:2,1:3] - observed_val_tab
    [1:2,1:3])^2/expected_value_tab[1:2,1:3])
```

The χ^2 statistic is 3.79

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

We check our assumptions: Variables are categorical, sample are randomly selected.
Lets state our hypothesis:
H_0: Police interaction in illegal encounter and class of drivers are statistically independent

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

H_a: Police interaction in illegal encounter and class of drivers are statistically dependent

Under the null hypothesis, we have calculated the χ^2 statistic in the last section, we can now calculate the p-value of the χ^2 test:

```
#calculate degree of freedom
df_num <- (3-1)*(2-1)
#calculate p-value of test
pchisq(chi_square_stat, df = df_num , lower.tail = FALSE)
```

p-value is 0.15 which is larger than the 0.1 significant level. We can conclude that there is not enough evidence to reject the null hypothesis that police interaction in illegal encounter and class of drivers are statistically independent.

(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	-1.641957	1.523026
Lower class	-0.3220306	1.641957	-1.523026

Detail work process done below:

```
#produce matrix of row probability and column probability for each cell
matrix_row <- matrix(probability_row[1:2],2,3)
matrix_col <- matrix(probability_col[1:3],2,3, byrow = TRUE)
```

Matrix of row probability for each cell:

```
          [,1]      [,2]      [,3]
[1,] 0.6428571 0.6428571 0.6428571
[2,] 0.3571429 0.3571429 0.3571429
```

Matrix of column probability for each cell:

```
     [,1]      [,2]      [,3]
[1,]  0.5 0.3095238 0.1904762
[2,]  0.5 0.3095238 0.1904762
```

```
#calculate standardize residual
standardized_residual <- (observed_val_tab[1:2,1:3] - expected_value_tab
    [1:2,1:3])/(sqrt(expected_value_tab[1:2,1:3]*(1-matrix_row)*(1-matrix_
    col)))
#add col and row names
rownames(standardized_residual) <- c("Upper class", "Lower class")
colnames(standardized_residual)<- c("Not Stopped", "Bribe requested" , "
    Stopped/Given warning" )
```

Table of standardized residual:

	Not Stopped	Bribe requested	Stopped/Given warning
Upper class	0.3220306	-1.641957	1.523026
Lower class	-0.3220306	1.641957	-1.523026

(d) How might the standardized residuals help you interpret the results?
Standardized residuals show the deviation in a standardized way from the expected value in each cell of the observed cells. If any absolute value of a cell larger than 3 then we have strong evidence that there is a true effect of a variable in that cell. Here we see that the biggest absolute value of all the cells is 1.64. We can guess loosely that bribe requested happened a little bit more for lower class than the expected value if independence holds true. But overall the standardize residuals suggest that there are no cells have large enough evidence against null hypothesis.

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.

If the relationship between the reserved policy for women and number of new or repaired drinking-water facilities in the village since the policy started is demonstrated by the equation:

$water = \alpha + \beta reserved + \epsilon$

where water and reserved are variables explain by the question prompt, β is the slope of the regression line, α is the intercept of the regression line, and ϵ is the error term, we can form our hypothesis:

H_0: $\beta = 0$
H_a: $\beta \neq 0$
And also:
H_0: $\alpha = 0$
H_a: $\alpha \neq 0$

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

We need to check out assumptions:

- Each observation is independent
- Observations are randomly generated
- There is a linear relationship between mean of the dependent variable and the value of independent variable
- The error term is normally distributed and has constant variable

This will be two-tail test and we will assume the significant level = 0.05 Next, lets import data and run the regression in R

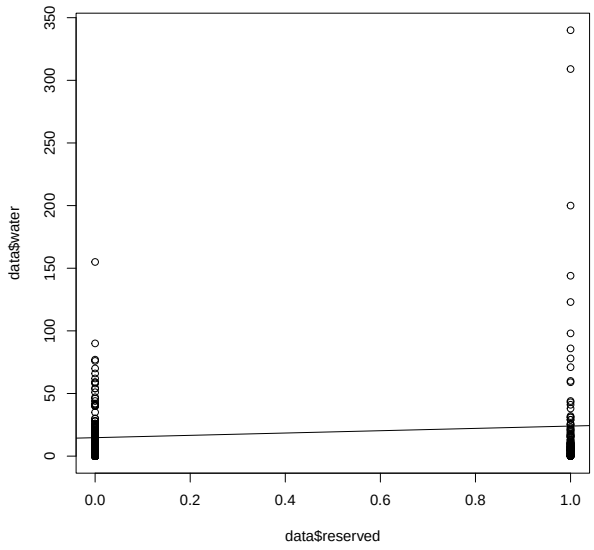
```r
#import data
data <- read.csv(url("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv"))
#run regression
model <- lm(water~reserved, data = data )
#get model result
summary(model)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-23.991 -14.738  -7.865   2.262 316.009

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.738      2.286   6.446 4.22e-10 ***
reserved       9.252      3.948   2.344   0.0197 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.45 on 320 degrees of freedom
Multiple R-squared:  0.01688, Adjusted R-squared:  0.0138
F-statistic: 5.493 on 1 and 320 DF,  p-value: 0.0197
```

We can see the p-value for the intercept = 4.22e-10 which is smaller than significant level of 0.05 so we can reject the null hypothesis that $\alpha = 0$
We also see the p-value for the coefficient of reserved = 0.0197 which is smaller than significant level of 0.05 so we can reject the null hypothesis that $\beta = 0$
We can plot the observed data and the regression line:

Figure 2: Scatterplot of reserved variable and water variable with regression line.



We can see in Figure 2 that the reserved variable only takes two value 1 and 0 while the water variable can take so much more values. If we check the goodness of fit measure R-squared, its only 0.017, which means the X variable only explains 1.7 percent the variation in Y.

(c) Interpret the coefficient estimate for reservation policy.
The coefficient estimate of the reservation policy is 9.252 which means that if the reservation policy is applied at a village then that village will have on average 9.252 more

new or repaired water facilities than a village have no policy applied since the policy started. The standard error is quite large at 3.948. Because n is large we can assume that sampling distribution of β is approximately normal. So the 95% CI for β is within 2 standard deviation:(1.356 , 17.148). This is a relatively big range.

The intercept coefficient estimate is 14.738 which means that if the policy is not applied at a village then that village will have an average of 14.738 more new or repaired water facilities since the policy started.