Problem Set 2

Liu Yuanyuan

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.

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).

```r
# Define the observed data
observed <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, ncol = 3,byrow = TRUE)
rownames(observed) <- c("Upper Class", "Lower Class")
colnames(observed) <- c("Not Stopped", "Bribe requesed","Stopped/given
    warnning")
observed
# Calculate the expected frequencies
row_sums <- rowSums(observed)
col_sums <- colSums(observed)
total_sum <- sum(observed)
for (i in 1:2) {
  for (j in 1:3) {
    expected[i, j] <- (row_sums[i] * col_sums[j]) / total_sum
  }
}
expected
# Run Chi square test
chi_squared <- sum(((observed - expected)^2) / expected)
chi_squared
```

- **Result**:
  the χ^2 test statistic is:3.791168

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

```r
p_value <- pchisq(chi_squared,df,lower.tail = FALSE)
p_value
if(p_value<0.1){
  cat("under the significance level a=0.1,we can reject null hypothsis,
    the officer were more likely to solicit a bribe from drivers depending
    on their class.\n")
}else{
  cat("under the significance level a=0.1,we fail to reject null
    hypothsis,the officer were less likely to solicit a bribe from drivers
    depending on their class.\n")
}
```

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

- **Result**:
  under the significance level a=0.1,we fail to reject null hypothesis,the officer were 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.

```
for (i in 1:2) {
  for (j in 1:3) {
    residuals[i, j] <- (observed[i, j] - expected[i, j]) / sqrt(expected[
   i, j]* (1 - row_sums[i]/total_sum) * (1 - col_sums[j]/total_sum) )
  }
}
residuals
```

	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?

- **Result**:
  under the significance level a=0.1,we fail to reject null hypothesis,the officer were less likely to solicit a bribe from drivers depending on their class.For data with positive standardized residuals, which indicates that the observed frequency is higher than the predetermined frequency, positive residual standard deviations indicate that police are more likely to obtain tickets from these drivers. For data with negative standardized residuals, this means that the observed frequency is lower than the predetermined frequency, in which case a negative residual means that the police are less likely to obtain tickets from these drivers. Based on the above data, the social class of drivers may have an impact on bribery behavior between police and drivers. Police are more likely to demand bribes from drivers of higher social class.

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.

- **Result**:
  H0:The reservation policy has no effect on the number of new or repaired drinking water facilities in the villages
  H1:The reservation policy has effect 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!).

```
data <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
X <- data$reserved
Y <- data$water
model <- lm(Y~X)
summary(model)
```

- **Result**:
  Call: lm(formula = Y X)
  Residuals:

Min	1Q r	Median	3Q	Max
-23.991	-14.738	-7.865	2.262	316.009

Coefficients:

	Estimate	Std. Error	t	Pr(¿—t—)
(Intercept)	14.738	2.286	6.446	4.22e-10
X	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

(c) Interpret the coefficient estimate for reservation policy.

- **Result**:
  1.On the confidence interval of 0.05, we can reject H0, the reservation policy has a significant effect on the number of new or repaired drinking water facilities in the villages.
  2.The coefficient of X is estimated to be positive, indicating that there is a positive correlation between the increase of new water or repaired drinking water facilities in the villages and reservation
  3.with the increase per unit of reservation, the number of new water or repaired drinking water facilities in the villages will increase by 9.252
  4.The t-value and p-value are statistical measures used to assess the significance of coefficient estimates. In this model, the t-value associated with the RESERVATION variable is 2.344, and the corresponding p-value is 0.0197. This indicates that the impact of the reservation policy on the number of water resource facilities is statistically significant. The p-value being below the commonly used significance level (usually 0.05) means that there is strong evidence to reject the null hypothesis, supporting the conclusion that the reservation policy has a meaningful effect on the number of water resource facilities.