Problem Set 2

Applied Stats/Quant Methods 1

Due: October 14, 2024

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 Monday October 14, 2024. 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.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	14	6	7
Lower class	7	7	1

[1] Fried, Lagunes, and Venkataramani (2010). "Corruption and Inequality at the Crossroad: A Multi-method Study of Bribery and Discrimination in Latin America. *Latin American Research Review.* 45 (1): 76-97.

(a) Calculate the χ^2 test statistic by hand/manually (even better if you can do "by hand" in `R`).

`Answer: The chi-squared test statistic is 3.791168, df = 2`

```r
######1.1######
# create data
class_bribery <- matrix(c(14,6,7,7,7,1), nrow = 2, ncol = 3, byrow = TRUE)
# rename
rownames(class_bribery) <- c("Upper class", "Lower class")
colnames(class_bribery) <- c("Not Stopped", "Bribe requested", " Stopped/given warning")

# sum
sum_class_bribery <- cbind(class_bribery, row_sum = rowSums(class_bribery))
sum_class_bribery <- rbind(sum_class_bribery,col_sum = colSums(sum_class_bribery))

# calculate frequency
expected_values <- class_bribery*0
for (i in 1:2) {
  for (j in 1:3) {
    expected_values[i,j] <- (sum_class_bribery[i, "row_sum"]*sum_class_bribery[-(1:2),j])/sum(class_bribery)
  }
}

# chi-squared statistic
chi_statistic <- sum((class_bribery-expected_values)^2/expected_values)
print(chi_statistic)

# df
df <- (nrow(class_bribery)-1) * (ncol(class_bribery)-1)
print(df)

# test
chisq.test(class_bribery)
```

```
> print(chi_statistic)
[1] 3.791168
> print(df)
[1] 2
```

(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: Since the p-value (0.15) > 0.10, we don't have sufficient evidence
to reject the null hypothesis at the alpha = 0.1 significance level.
This suggests that there is no significant difference in the likelihood
of police officers soliciting a bribe from drivers based on their class.
```

```r
######12######
pvalue <- pchisq(chi_statistic, df, lower.tail=FALSE)
print(pvalue)
print("Since the p-value (0.15) > 0.10, we don't have sufficient evidence
    to reject the null hypothesis at the alpha = 0.10 significance level.
    This suggests that there is no significant difference in the
   likelihood of police officers soliciting a bribe from drivers based on
    their class.")
```

```
> print(pvalue)
[1] 0.1502306
```

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

```r
######13######
standard_residuals_values <- class_bribery*0
se_values <- class_bribery*0
for (i in 1:2) {
  for (j in 1:3) {
    se_values[i,j] <- sqrt(expected_values[i,j]*
                           (1-sum(class_bribery[i,])/sum(class_bribery))*
                           (1-sum(class_bribery[,j])/sum(class_bribery)))}
}
standard_residuals_values <- (class_bribery-expected_values)/se_values

# test
chisq.test(class_bribery)$stdres

# list
print(standard_residuals_values)
```

```
> print(standard_residuals_values)
Not Stopped Bribe requested  Stopped/given warning
Upper class   0.3220306       -1.641957              1.523026
Lower class  -0.3220306        1.641957             -1.523026
```

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.523

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

```
Answer:

1.Standardized residuals with absolute values greater than 2 or less than -2
are typically considered outliers. According to the table, no observations
show significant deviations from the model predictions.

2.The standardized residual is very small, which to some extent suggests that
there is no dependency relationship present.

3.To draw conclusions, it is also necessary to consider the resualt of chi-squared
test and the p-value.
```

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.

```
Answer:

Ho = Having reserved village council heads for female does not affect the number
drinking water facilities in the villages.

Ha = Having reserved village council heads for female affects the number drinking
water facilities in the villages.
```

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

```r
######2.2######
data_problem2 <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv", header=T)
regression_model <- lm(water ~ reserved, data = data_problem2)
summary(regression_model)
```

```
>summary(regression_model)
Call:
lm(formula = water ~ reserved, data = data_problem2)

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
```

(c) Interpret the coefficient estimate for reservation policy.

```
Answer:

Since the p-value (approximately 0.02) < 0.05, we have sufficient evidence to
reject the null hypothesis at the alpha = 0.05 significance level. This suggests
that Having reserved seats for female politicians increase the number drinking
water facilities in the villages.
```