Problem Set 2

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.

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).

```
# H0 = officers may not demand bribes from drivers, depending on their
    class
# H1 = officers may demand bribes from drivers, depending on their class
data <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)
data
#      [,1] [,2] [,3]
# [1,]   14    6    7
# [2,]    7    7    1
row_sums <- rowSums(data) # 27 15
col_sums <- colSums(data) # 21 13  8
total_sum <- sum(data) # 42

expected <- matrix(0, nrow = 2, ncol = 3)
for (i in 1:2) {
  for (j in 1:3) {
    expected[i, j] <- (row_sums[i] * col_sums[j]) / total_sum
  }
}
expected
#             Not stopped  Bribe requested Stopped/given warning
# Upper class        13.5         8.357143             5.142857
# Lower class         7.5         4.642857             2.857143
\chi^2_test_statistic <- sum((data - expected)^2 / expected)
\chi^2_test_statistic # 3.791168
```

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 $\alpha = 0.1$?

```
df <- (nrow(data) - 1) * (ncol(data) - 1)
df # 2
pchisq( 2 _test_statistic, df = 2, lower.tail = FALSE)
p_value # 0.1502306
p_value >    # 0.1502306 > 0.1
# we are failed to reject the null hypothesis, officers may not demand
    bribes from drivers, depending on their class.
```

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

```
7
8
9  # 2 What do you conclude if     = 0.1?
10 # calculate Quantile for the Chi-Squared Distribution
11 qchisq(0.1, df = 2, lower.tail = FALSE)
12 # 4.60517 > 3.791168, we are failed to reject the H0.
```

p-value is 0.1502306.
if $\alpha = 0.1$, we are failed to reject the H0, officers may not demand bribes 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	-1.641957	1.523026
Lower class	-0.3220306	1.641957	-1.523026

```r
expected_1 <- 13.5
expected_2 <- 8.357143
expected_3 <- 5.142857
expected_4 <- 7.5
expected_5 <- 4.642857
expectedz_6 <- 2.857143
SR_1 <- (14-13.5)/sqrt(13.5*(1-27/42)*(1-21/42))
SR_2 <- (6-8.357143)/sqrt(8.357143*(1-27/42)*(1-13/42))
SR_3 <- (7-5.142857)/sqrt(5.142857*(1-27/42)*(1-8/42))
SR_4 <- (7-7.5)/sqrt(7.5*(1-15/42)*(1-21/42))
SR_5 <- (7-4.642857)/sqrt(4.642857*(1-15/42)*(1-13/42))
SR_6 <- (1-2.857143)/sqrt(2.857143*(1-15/42)*(1-8/42))
SR_1 # 0.3220306
SR_2 # -1.641957
SR_3 # 1.523026
SR_4 # -0.3220306
SR_5 # 1.641957
SR_6 # -1.523026
SR_table <- matrix(c(SR_1, SR_2, SR_3, SR_4, SR_5, SR_6), nrow = 2, byrow
    = TRUE)
standardized_residuals_table
rownames(SR_table) <- c("Upper class", "Lower class")
colnames(SR_table) <- c("Not stopped", "Bribe requested", "Stopped/given
    warning")
SR_table
#             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?

```
# The maximum absolute value of standardized residuals is less than
    1.96(95%),
# So we are failed to reject the null hypothesis,
# officers may not demand bribes from drivers, depending on their 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.

```
install.packages("readr")
library(readr)
url <- "https://raw.githubusercontent.com/kosukeimai/qss/master/
    PREDICTION/women.csv"
data <- read_csv(url)
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 an 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!).

```
model <- lm(water ~ reserved, data=data)
summary(model)
p_value <- summary(model)$coefficients["reserved", "Pr(>|t|)"]
p_value
plot(data$reserved, data$water, main = "Scatterplot of water vs. reserved
    ",
     xlab = "Reserved", ylab = "Water")
```

```
Call:
lm(formula = water ~ reserved, data = data)

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
```

Figure 2: 2(b)model

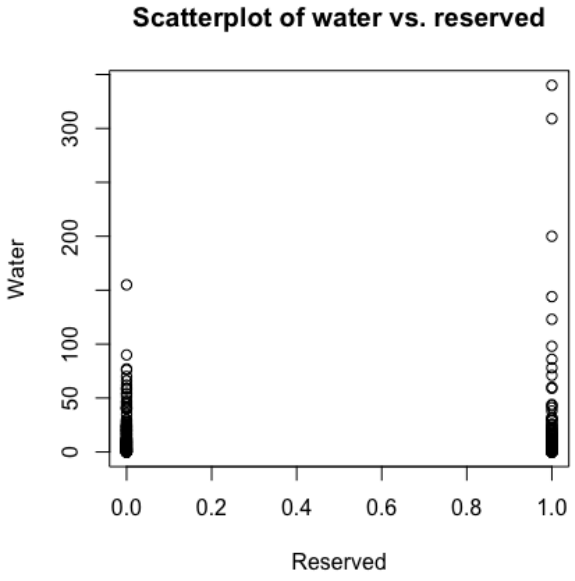

Figure 3: Rplot

(c) Interpret the coefficient estimate for reservation policy.

```r
table (data$reserved)
#   0   1
#  214 108
# reserved: binary variable ( 0 and 1 )
coefficients <- coef(model)
print(coefficients)
# (Intercept)    reserved
# 14.738318    9.252423
# yi = Intercept + slope * xi
# water_i = 14.738318 + 9.252423 * reserved_i

water_0 <- 14.738318 + 9.252423 * 0  # 14.73832
water_1 <- 14.738318 + 9.252423 * 1  # 23.99074
# the coefficient estimate is 14.73832, it means the GP was reserved for
    women leaders.
# the coefficient estimate is 23.99074, it means the GP was not reserved
    for women leaders.
```