Problem Set 2

Applied Stats/Quant Methods 1
Zhuo Zhang/Student ID: 23346227

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 frequency matrix
observed_data <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)

# Perform a chi-squared test
test_result <- chisq.test(observed_data)

# Calculate row sums for the contingency table
row_sums <- rowSums(observed_data)

# Calculate column sums for the table
col_sums <- colSums(observed_data)

# Compute the overall total count
total_count <- sum(observed_data)

# Calculate the expected frequency table
expected_data <- outer(row_sums, col_sums) / total_count

# Calculate the chi-squared values
chi_squared_values <- (observed_data - expected_data)^2 / expected_data

# Calculate the chi-squared test statistic
chi_squared_statistic <- sum(chi_squared_values)
cat("Chi-Squared Test Statistic: ", chi_squared_statistic, "\n")
```

Result:
Chi-Squared Test Statistic: 3.79116

(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
# Degrees of freedom
# Calculate the degrees of freedom for the chi-squared test
df <- (nrow(observed_data) - 1) * (ncol(observed_data) - 1)

# Calculate the p-value for the chi-squared test
p_value <- 1 - pchisq(chi_squared_statistic, df)

# Set the significance level (alpha)
alpha <- 0.1

# Print the calculated p-value
cat("P-Value: ", p_value, "\n")

# Perform a hypothesis test based on the p-value and significance level
if (p_value <= alpha) {
  cat("Reject the null hypothesis. There is evidence of an association
   between variables.\n")
} else {
  cat("Fail to reject the null hypothesis. There is no strong evidence of
    an association between variables.\n")
}
```

Result:
Fail to reject the null hypothesis. There is no strong evidence of an association between variables.

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

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class			
Lower class			

```r
# Calculate the standardized residuals for each cell
standardized_residuals <- (observed_data - expected_data) / sqrt(expected_data)

# Create an empty result table
result_table <- matrix(NA, nrow = 2, ncol = 3)
colnames(result_table) <- c("Not Stopped", "Bribe requested", "Stopped/given warning")
rownames(result_table) <- c("Upper class", "Lower class")

# Format the standardized residuals with two decimal places
result_table <- format(standardized_residuals, digits = 2)

# Set row names for the result table
rownames(result_table) <- c("Upper class", "Lower class")

print(result_table)
```

result

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.14	-0.82	0.82
Lower class	-0.18	1.09	-1.10

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

- Looking at the cell where "Upper class" individuals are "Not Stopped," the standardized residual is 0.14. Since it is close to zero, it suggests that the observed and expected frequencies for this cell are reasonably close, and there may not be a strong association between being in the upper class and not getting stopped.
- For the "Upper class" group and "Bribe requested" cell, the standardized residual is -0.82. This negative value signals that the observed frequency of upper-class individuals requesting a bribe is lower than what would be expected. It suggests that being in the upper class might be associated with a lower likelihood of requesting a bribe.
- For the cell corresponding to "Upper class" and "Stopped/given warning," the standardized residual is 0.82, a positive value.This suggests that the observed frequency of upper-class individuals being stopped or given a warning is higher than expected. It implies that being in the upper class might be associated with a higher likelihood of being stopped or given a warning.
- For the cell corresponding to "Lower class" and "Not Stopped," the standardized residual is -0.18, which is close to zero. It indicates that there may not be a strong association between being in the lower class and not getting stopped.
- For the cell corresponding to "Lower class" and "Bribe requested," the standardized residual is 1.09, indicating that the observed frequency of lower-class individuals requesting a bribe is higher than expected. It suggests that being in the lower class might be associated with a higher likelihood of requesting a bribe.
- For the "Lower class" group and "Stopped/given warning," the standardized residual is -1.10, a negative value. It suggests that the observed frequency of lower-class individuals being stopped or given a warning is lower than expected. This implies that being in the lower class might be associated with a lower likelihood of experiencing these interactions.

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.

- Null Hypothesis : The retention policy for women among village committee leaders has no impact on the number of new or repaired drinking water facilities in the village. Mathematically speaking, this can be expressed as: where is the coefficient of the "reserved" variable in the regression model.
- Alternative Hypothesis: The retention policy of village committee leaders towards women has an impact on the number of new or repaired drinking water facilities in the village. Mathematically speaking, this can be expressed as: this is a double tailed substitution hypothesis, as we are testing whether the coefficients differ significantly from zero in any direction.

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

```r
data <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
#Run the bivariate regression
model <- lm(water ~ reserved, data = data)
## Summarize the regression results
summary(model)
```

Result:

```
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
```

(c) Interpret the coefficient estimate for reservation policy.

- The coefficient estimate for the "reserved" variable stands at 9.252, bearing a standard error of 3.948, along with an associated p-value of 0.0197. This coefficient signifies the alteration in the count of newly established or refurbished drinking water facilities within villages that correspond to a one-unit shift in the "reserved" variable. This variable determines whether the village council head's position is set aside for women.
  - Coefficient Value: The positive coefficient estimate of 9.252 implies that, on average, when the village council head's position is reserved for women (i.e., "reserved" = 1), there is an approximate increment of 9.252 in the number of fresh or fixed drinking water facilities in contrast to situations with no reservation policy (i.e., "reserved" = 0).
  - Statistical Significance: The coefficient bears statistical significance, as indicated by a p-value of 0.0197, which is below the conventional significance threshold of 0.05. This suggests that the implementation of a reservation policy for women in village council head positions has a noteworthy impact on the count of new or repaired drinking water facilities in villages.