

Problem Set 2

Hanyu Li(Student ID:25346841)

October 19, 2025

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 Thursday October 23, 2025. 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.

¹Fried, 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.

	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).

This study examines the statistical independence between two categorical variables: the social class of employee drivers and whether police officers solicit bribes for traffic violations.

Thus, the null hypothesis (H_0) is: Officers' bribery solicitation and employee drivers' social class are statistically independent.

The alternative hypothesis (H_1) is: Officers' bribery solicitation and employee drivers' social class are statistically dependent.

Based on the hypotheses, the expected frequency for each cell in the table is first calculated using the formula:

$$E_{ij} = \frac{(\text{Row total}) \times (\text{Column total})}{\text{Grand total}}$$

Next, using the observed and expected frequencies, the chi-square statistic χ^2 is computed as:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

The R code lines are as below:

```

1 # Q1(a): calculate x^2 by hand
2 # Build contingency table shown in Q1
3 study_data <- matrix(c(14, 6, 7,
4                       7, 7, 1),
5                       nrow = 2, byrow = TRUE)
6
7 colnames(study_data) <- c("Not stopped", "Bribe requested", "Stopped
  Warning")
8 rownames(study_data) <- c("Upper calss", "Lower calss")
9
10 # Check data
11 study_data

```

```

12
13 # Calculate expected frequency for each cell
14 row_total <- rowSums(study_data)
15 col_total <- colSums(study_data)
16 grand_total <- sum(study_data)
17 expected_fre <- outer(row_total, col_total) / grand_total
18
19 # Calculate Chi-square t-statistic manually
20 chi_sq <- sum((study_data - expected_fre)^2 / expected_fre)
21 chi_sq

```

The manual calculation in R yields a chi-square value of 3.4125, indicating some discrepancy between the observed and expected frequency distributions. Further testing is required to draw a final conclusion.

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

To calculate the p-value, the degrees of freedom for this contingency table must first be determined using the formula:

$$df = (\text{rows} - 1) \times (\text{columns} - 1)$$

Subsequently, the p-value is computed using the `pchisq()` function in R.

The R code lines are as below:

```

1 # Q1(b): calculate the p-value
2 df = (nrow(study_data)-1)*(ncol(study_data)-1)
3 pchisq(chi_sq, df, lower.tail = FALSE)

```

The calculated p-value is 0.1502306. At a significance level of 0.1, this indicates that if the null hypothesis were true, the probability of observing the obtained frequency distribution would be approximately 15%. Since this is greater than a significance level of 10%, we do not have sufficient evidence to reject the null hypothesis. Thus, it can be concluded that the social class of employee drivers and bribery solicitation by police officers are statistically independent.

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

²Remember 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	0.3220306	-1.516426	1.649103
Lower class	-0.2740361	1.929528	-1.523026

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

The standardized residuals table shows that the absolute value of every residual is below 2, indicating no substantial discrepancies between any observed and expected frequencies. This finding further corroborates the modest chi-square statistic from Q1(a) and supports the decision in Q1(b) to retain the null hypothesis of statistical independence between a driver's social class and bribery solicitation by police officers.

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.

³Chattopadhyay and Duflo. (2004). 'Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica*. 72 (5), 1409-1443.

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

Figure 1: Names and description of variables from Chattopadhyay and Duflo (2004).

- (a) State a null and alternative (two-tailed) hypothesis.

To estimate the effect of female leaders' reservations in Gram Panchayats (GP) on the number of drinking water facilities, while controlling for other confounding factors, we use a simple linear regression model for hypothesis testing and prediction. The regression model is specified as:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where Y_i denotes the number of new or repaired drinking water facilities in village i , X_i is a binary variable indicating whether the GP leadership is reserved for a female leader, α is the intercept, β measures the effect of female leadership reservation, and ϵ_i is the error term.

The hypotheses are formulated as follows:

$$\text{Null hypothesis: } H_0 : \beta = 0$$

which states that female leaders' reservations in GP (X) have no effect on the number of drinking water facilities (Y), and

$$\text{Alternative hypothesis: } H_1 : \beta \neq 0$$

which states that female leaders' reservations in GP (X) do have an effect on the number of drinking water facilities (Y).

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

After importing the data, the structure and distribution were examined, and basic data cleaning was performed.

GP	village	reserved	female	irrigation
Min. : 1	Min. :1.0	NotReserved:214	Min. :0.0000	Min. : 0.000
1st Qu.: 41	1st Qu.:1.0	Reserved :108	1st Qu.:0.0000	1st Qu.: 0.000
Median : 81	Median :1.5		Median :0.0000	Median : 0.000
Mean : 81	Mean :1.5		Mean :0.3851	Mean : 3.264
3rd Qu.:121	3rd Qu.:2.0		3rd Qu.:1.0000	3rd Qu.: 2.000
Max. :161	Max. :2.0		Max. :1.0000	Max. :90.000

water
Min. : 0.00
1st Qu.: 3.00
Median : 9.00
Mean : 17.84
3rd Qu.: 20.00
Max. :340.00

A linear regression model was then fitted. The code and regression results are as follows:

```

1 #Q2(b):run a bivariate regression to test this hypothesis
2 # Read the data to be analyzed
3 data <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master
  /PREDICTION/women.csv")
4 # view data and grasp column names
5 summary(data)
6 # Transform reserved to binary variable
7 data$reserved <- factor(data$reserved,
8                           levels = c(0, 1),
9                           labels = c("NotReserved", "Reserved"))
10
11 # Build regression model and check the outcome

```

```

12 model <- lm(data$water ~ data$reserved)
13 summary(model)

```

Table 1: Regression Results

	<i>Dependent variable:</i>
	Number of Drinking Water Facilities
Intercept	9.252** (3.948)
Reserved (Female)	14.738*** (2.286)
Observations	322
R ²	0.017
Adjusted R ²	0.014
Residual Std. Error	33.446 (df = 320)
F Statistic	5.493** (df = 1; 320)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

The intercept ($\alpha = 14.738$) indicates that when the GP leadership is not reserved for a female leader, the average number of new or repaired drinking water facilities is about 14.7.

The p-value for the model is 0.0197. At a significance level of 0.05, this is less than 0.05, providing evidence to support the alternative hypothesis: female leaders' reservations in GP have a significant effect on the number of drinking water facilities. The positive coefficient indicates that the reservation of female leadership increases the number of facilities built or repaired in the villages.

However, the adjusted R -squared is only 0.014, very close to zero, indicating that the model does not fit the data well. The reservation of female leadership explains only a small part of the variation in the number of new or repaired drinking water facilities.

(c) Interpret the coefficient estimate for reservation policy.

The coefficient estimate of 9.252 in the model

$$Y_i = 14.738 + 9.252X_i$$

can be interpreted as follows: on average, when a GP changes from having no female leader (not reserved) to having a female leader (reserved), the number of new or repaired drinking water facilities in the village increases by approximately 9.