Problem Set 2

Applied Stats/Quant Methods 1

Due: October 23, 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.

[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).

I will calculate the χ^2 test statistic by hand to determine whether the difference between both categorical variables is statistically significant.

I will do it in R and the first step will be creating the table there.

```r
data <- matrix(c(14, 6, 7,7, 7, 1), nrow = 2,byrow = TRUE)
rownames(data) <- c("Upper", "Lower")
colnames(data) <- c("Not_Stopped", "Bribe_Requested", "Stopped_Warning")
data
#Now I will calculate the expected values under the assumption of independence.
#We will start by summing all the rows, columns, and the table as a whole.
row_totals <- rowSums(data)
column_totals <- colSums(data)
table_total <- sum(data)
#I will use the function "outer" to calculate the Outer Product of the two
#vectors "column_totals" and "row_totals" and then divide it by "table_total"
expected <- outer(row_totals,column_totals)/table_total
expected
#Now I've a table with the expected values under the assumption of independence.
#With this, I can apply the chi square statistic formula:
chi_square_stat <- sum((data-expected)^2/expected)
chi_square_stat
#Now, to check if I did it right, I will ask R to calculate the value too:
chisq.test(data)
#In effect, the value remains 3.7912, which is the same I obtained manually.
```

The χ^2 test statistic of these categorical variables has a value of 3.7912.

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

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

```r
#To calculate the p-value, I already have the chi_square_stat, but I need to
#calculate the degrees of freedom too.
#The DF are just a multiplication of the n of rows - 1 x the n of columns - 1:
df <- (nrow(data) - 1) * (ncol(data) - 1)
df
#The DF are therefore 2.
#Now, I will use the pchisq function:
p_value <- 1 - pchisq(chi_square_stat, df)
#I wil calculate the critical value for alfa = 0.1
qchisq(0.9, df = 2)
#The value is 4.6052. Thus, for my CI, the chi_square_state is still
#smaller than the critical value (3.8 < 4.6), meaning that I can't reject
#the null hypothesis.
```

As I can’t reject the null hypothesis, I can’t affirm that the officers are more or less likely to solicit a bribe from drivers based on their social class.

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

```
#Now, I will calculate the standardised residuals for each cell.
#I already have the matrix ”data” and the matrix ”expected”.
#I can calculate the simple ones like that:
std_res <- (data - expected) / sqrt(expected)
std_res
#But the adjusted ones correct biases and are thus better for interpretation:
row_prop <- row_totals/table_total
column_prop <- column_totals/table_total
adj_std_res <- (data - expected) / sqrt(expected * (1 - row_prop[row(data)]) * (1 - column_prop[col(data)]))
adj_std_res
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.32	-1.64	1.52
Lower class	-0.32	1.64	-1.52

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

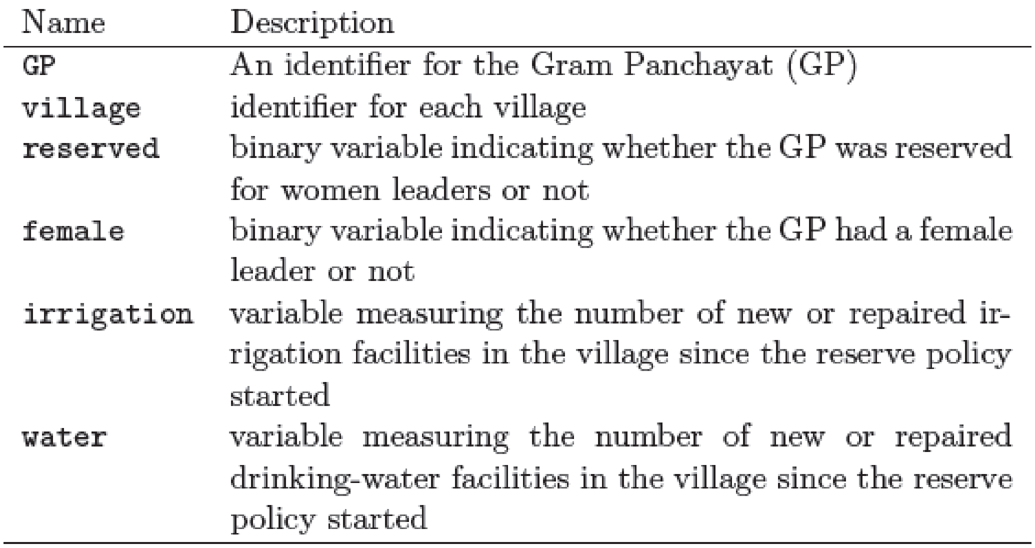
As any of these standardized residual values is over 2 or below -2, there is no strong evidence of association between both variables (class and police officer treatment).

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.

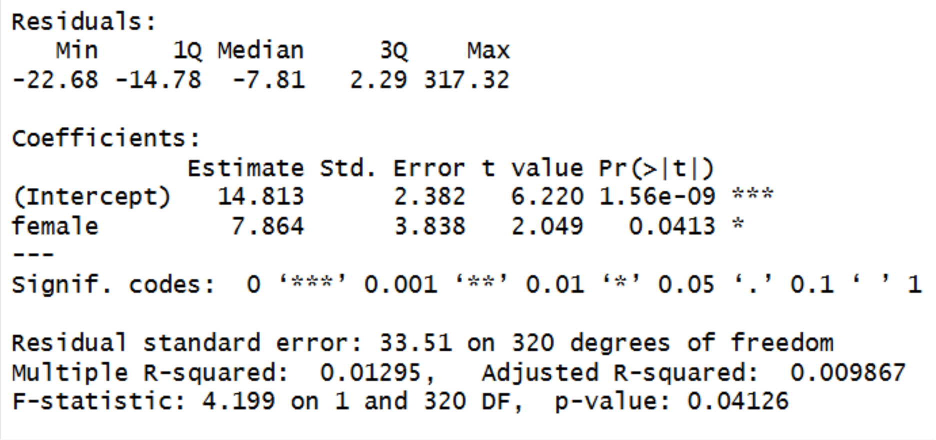
A two-tailed hypothesis posits that there is a significant difference or relationship between two groups or variables without specifying the direction of the difference. Therefore, following Chattopadhyay and Duflo (2004), I will hypothesise (H1) that the amount of new or repaired drinking-water facilities in the village is different when between female- and male-led Gram Panchayats. However, I will not specify–in my hypothesis–whether the amount will be bigger or smaller in each case. By the same token, the null hypothesis (H0) will suggest that there is no difference.

H0: there is no difference in the number of new or repaired drinking-water facilities between female- and male-led Gram Panchayats.

H1: there is a difference in the number of new or repaired drinking-water facilities between female- and male-led Gram Panchayats.

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

```
View(data_India)

#To run a bivariate regression, I need to create a model that crosses the two
#variables (”water” and ”female”).
model1 <- lm(water ~ female, data = data_India)
summary(model1)
```

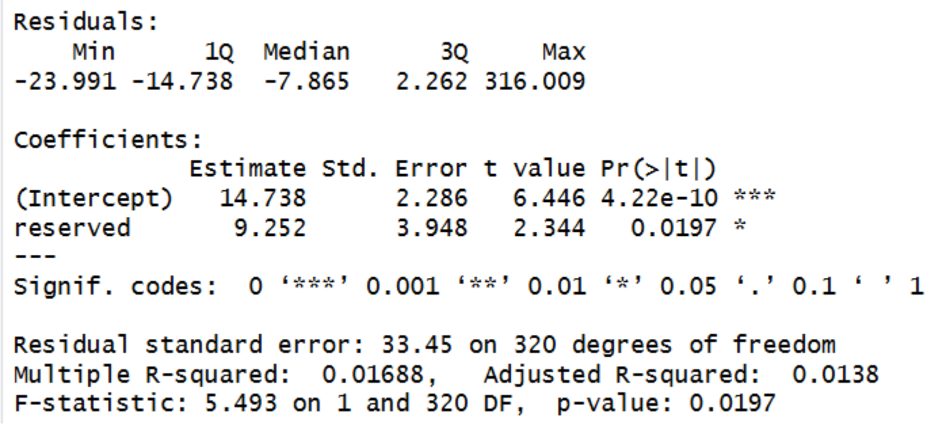


```
Residuals:
   Min     1Q Median     3Q    Max 
-22.68 -14.78  -7.81   2.29 317.32 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   14.813      2.382   6.220 1.56e-09 ***
female         7.864      3.838   2.049   0.0413 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.51 on 320 degrees of freedom
Multiple R-squared:  0.01295,	Adjusted R-squared:  0.009867 
F-statistic: 4.199 on 1 and 320 DF,  p-value: 0.04126
```

The coefficient for female is positive (7.86) and statistically significant (p = 0.041), indicating that Gram Panchayats headed by women have, on average, built 7.86 more drinking-water facilities than those headed by men. This result supports hypothesis (H1) that the amount of new or repaired drinking-water facilities in the village is different when between female- and male-led Gram Panchayats. As we can identify a statistically significant difference in the number of drinking-water facilities between male- and female-headed Gram Panchayats, we can reject the null hypothesis (H0).

(c) Interpret the coefficient estimate for reservation policy.

```
#Now, I will run a bivariate regression to crosses the two variables (”
    water”
#and ”reserved”).
model2 <- lm(water ~ reserved, data = data_India)
print(summary(model2))
```



```
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
```

The coefficient for reservation policy is positive (9.252) and statistically significant (p = 0.0197), indicating that Gram Panchayats affected by the reservation policy have, on average, built 9.252 more drinking-water facilities than those that were not affected by the policy.