Problem Set 2/

Quant Methods 1/Due: October 14, 2024

Jia Lyu/ 23370062

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.

[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
#Observed frequencies
observed <- matrix(c(14, 6, 7,
                     7, 7, 1),
                   nrow = 2, byrow = TRUE)

# Calculate total frequency (grand total)
grand_total <- sum(observed)

# Calculate row totals and column totals
row_totals <- rowSums(observed)
col_totals <- colSums(observed)

# Calculate expected frequency for each cell
fe_11 <- (row_totals[1] / grand_total) * col_totals[1]  # Upper class, Not Stopped
fe_12 <- (row_totals[1] / grand_total) * col_totals[2]  # Upper class, Bribe requested
fe_13 <- (row_totals[1] / grand_total) * col_totals[3]  # Upper class, Stopped/given warning

fe_21 <- (row_totals[2] / grand_total) * col_totals[1]  # Lower class, Not Stopped
fe_22 <- (row_totals[2] / grand_total) * col_totals[2]  # Lower class, Bribe requested
fe_23 <- (row_totals[2] / grand_total) * col_totals[3]  # Lower class, Stopped/given warning

# Put the calculated expected frequencies into a matrix
expected <- matrix(c(fe_11, fe_12, fe_13,
                     fe_21, fe_22, fe_23),
                   nrow = 2, byrow = TRUE)

# Calculate the X^2 test statistic using the manually calculated expected frequencies
chi_square_stat <- sum((observed - expected)^2 / expected)
chi_square_stat
#chi_square_stat should be approximately 3.7912
```

(b) Now calculate the p-value from the test statistic you just created (in R).[2] What do you

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

conclude if $\alpha = 0.1$?

```r
# Chi-square statistic from part (a)
chi_square <- 3.7912

# Degrees of freedom
# (number of rows - 1) * (number of columns - 1)
df <- (2 - 1) * (3 - 1)

# Calculate p-value
p_value <- pchisq(chi_square, df = df, lower.tail = FALSE)
p_value
# p_value should be approximately 1.1502282

# Set significance level
alpha <- 0.1
```

The p-value(1.1502282) is greater than alpha(0.1),so we cannot reject the null hypothesis,and there is no significant association between a driver's social class and the police officer's behavior.

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

```r
#question c
# Perform chi-square test to get expected frequencies
chi_square_test <- chisq.test(observed)
expected <- chi_square_test$expected

# Calculate total frequency
grand_total <- sum(observed)

# Calculate standardized residuals
std_residuals <- (observed - expected) / sqrt(expected*(1-rowSums(
    observed)/grand_total)*(1-colSums(observed)/grand_total))
round(std_residuals, 3)

# Create a data frame for easier viewing
residuals_df <- as.data.frame(std_residuals)
residuals_df$Class <- rownames(residuals_df)
residuals_df <- residuals_df[, c(4, 1, 2, 3)]

residuals_df
```

	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 for each cell in the contingency table provide insight into how the observed frequencies compare to the expected frequency of police solicitation based on driver rank.

Upper level observations: No stops: The standardized residual is 0.3220306, which is positive and close to zero. This indicates that the observed number of upper class drivers not stopped is slightly higher than expected, but the deviation is small and not statistically significant. Solicited Bribes: The standardized residual is -1.516426, which is negative and relatively large in absolute value (greater than 1). This indicates that the observed number of upper class drivers being solicited for bribes is significantly lower than expected, suggesting that they are less likely to be solicited for bribes than the model predicts. Stopped/Warned: The standardized residual of 1.649103 is positive and greater than 1. This indicates that the observed number of upper class drivers stopped or warned is significantly higher than expected, suggesting that they are more likely to be stopped compared to predictions.

Lower Class Observations: Not Stopped: The standardized residual of -0.2740361 is negative and close to zero. This suggests that drivers in the lower strata are slightly less likely than expected to not be stopped, but this deviation is small and not statistically significant. Soliciting Bribes: The standardized residual is 1.929528, which is positive and relatively large (greater than 1). This suggests that lower-class drivers are solicited for bribes more frequently than expected, indicating a possible bias against them. Stop/Warning: The standardized residual is -1.523026, which is negative and significant (absolute value greater than 1). This indicates that lower tier drivers were stopped or warned less frequently than expected, suggesting that they may have been treated differently than upper tier drivers.

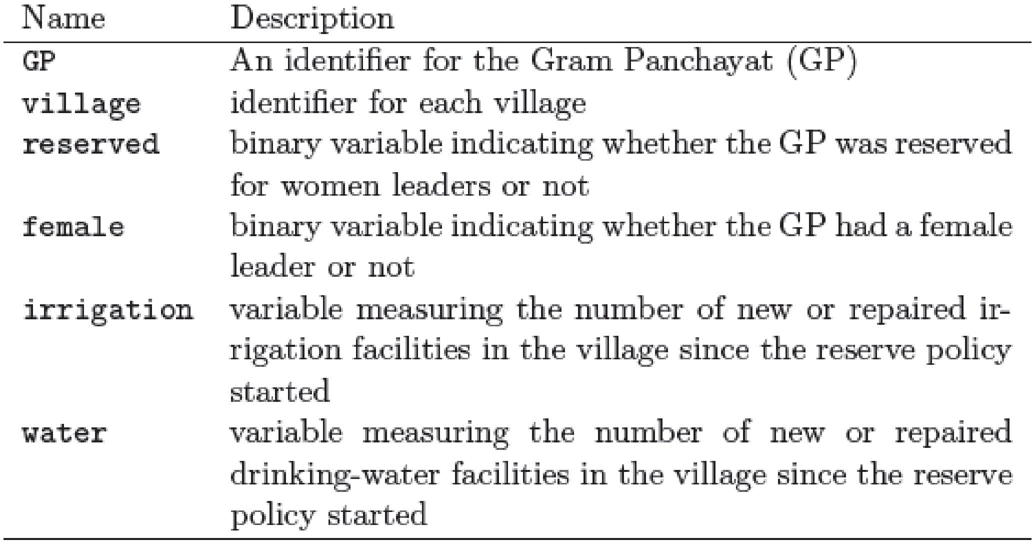
Conclusion: The standardized residual values for both the upper and lower strata are within the range of [-2, 2], indicating that there are no extreme deviations from what would be expected under the null hypothesis. Although there were significant differences in the way the strata interacted with police officers regarding bribery and stops, these differences do not reflect significant anomalies in the data. In summary, the standardized residuals provide insight into possible bias or discrimination in police behavior based on driver class, helping to inform discussions about corruption and inequality in law enforcement practices.

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 (H0): There is no difference in the number of new or repaired drinking water facilities between villages with reserved and unreserved council heads.

Alternative Hypothesis (Ha): There is a difference in the number of new or repaired drinking water facilities between villages with reserved and unreserved council heads.

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

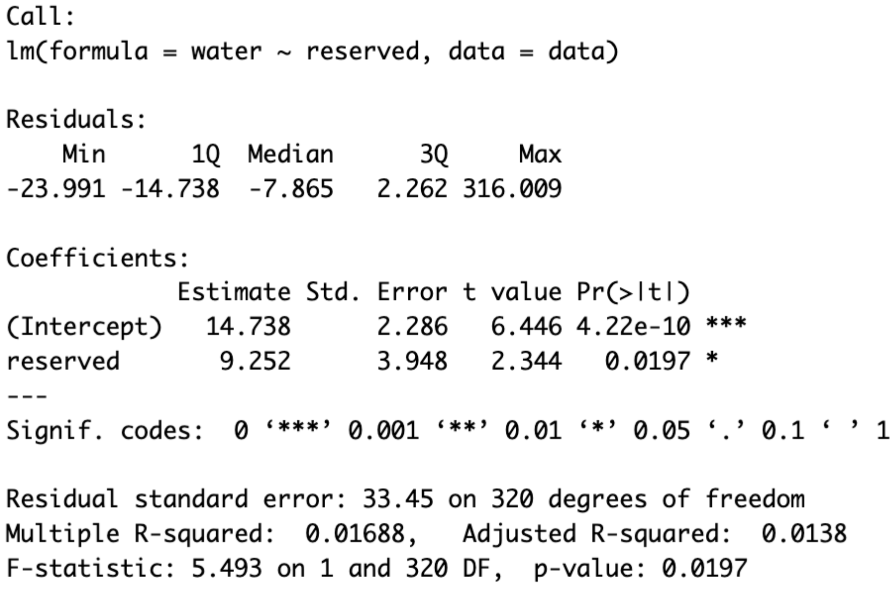
```r
# Load the dataset
data_url <- "https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv"
data <- read.csv(data_url)

# View the first few rows of the dataset
head(data)

# Run bivariate regression
model <- lm(water ~ reserved, data = data)

# Display summary of the regression model
summary(model)
```

Figure 2: results



```
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 Std.for the reservation policy is 9.252,meaning that there is a positive association between the reservation policy and the number of new or repaired drinking water facilities in villages. On average, villages that implemented the reservation policy had 9.252 more such facilities compared to those without the policy.