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).

```

bribe_data <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, ncol = 3, byrow = TRUE)
rownames(bribe_data) <- c("Upper class", "Lower class")
colnames(bribe_data) <- c("Not stopped", "Bribe requested", "Stopped")
bribe_data

r <- apply(bribe_data, 1, sum)
c <- apply(bribe_data, 2, sum)
r
fe_all <- numeric()
chisq <- 0
fe <- 0


for (i in seq(1:nrow(bribe_data))) {
  for (j in seq(1:ncol(bribe_data))) {
    fe <- (c[j] * r[i]) / sum(bribe_data)
    chisq <- chisq + ((bribe_data[i, j] - fe) ^ 2 / fe)
    fe_all <- c(fe_all, fe)
  }}
fe_all <- matrix(fe_all, nrow = 2, ncol = 3, byrow = TRUE)
print(fe_all)
print(chisq)

# check chi-squre value
chisq.test(bribe_data)
x_sq <- chisq.test(bribe_data)
chisq == x_sq$statistic
```

```
Pearson's Chi-squared
testdata:  bribe_data
X-squared = 3.7912, df = 2, p-value = 0.1502
```

(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
# (b) Now calculate the p-value from the test statistic you just created

p_value = pchisq(chisq, df = (nrow(bribe_data)-1) * (ncol(bribe_data)-1),
    lower.tail=FALSE)
p_value
p_value <= 0.1
```

```
> p_value
Not stopped   0.1502306

 > p_value <= 0.1
 Not stopped
 FALSE
```

In conclusion: Since p-values we calculated based on data is around 0.15 which is greater than$\alpha = 0.1$, we fail to reject null hypothesis, and the two categorical variables between bribe behaviors and class level are statistically independent.

[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	0.3220306	-1.641957	1.523026
Lower class	-0.3220306	1.641957	-1.523026

```

# (c) Calculate the standardized residuals for each cell a

z <- 0
z_all <- numeric()

for (i in seq(1:nrow(bribe_data))) {
  for (j in seq(1:ncol(bribe_data))) {
    z = (bribe_data[i,j] - fe_all[i, j]) /
          sqrt(fe_all[i,j] * (1- (r[i] / sum(bribe_data))) * ( 1- (c[j] / sum(bribe_data))))
    z_all <- c(z_all , z)
  }}
z_all <- matrix(z_all , nrow = 2, ncol = 3, byrow = TRUE)
print(z_all)

# check standardized residuals
x_sq$stdres
```

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

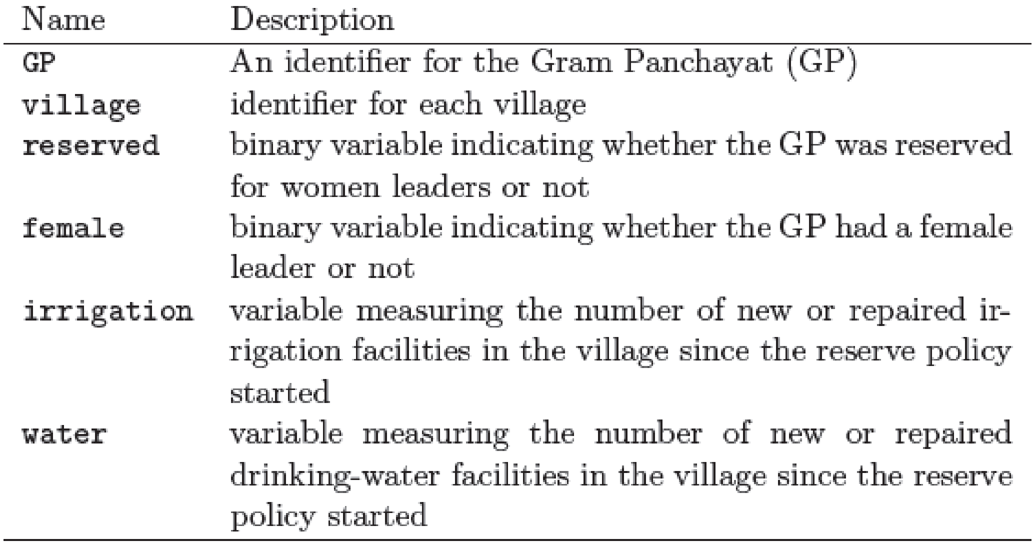
Since the null hypothesis was not rejected, the observed frequency in each cell don't significantly deviates from the expected frequency. These differences can be considered to be due to random sampling, not far enough to indicates that two variables are dependent.

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.

1. Null hypothesis: The reservation policy has no linear association on the number of new or repaired drinking water facilities in the vilages. $\beta_{\text{reserved}} = 0$

2. Alternative hypothesis: The reservation policy has linear association on the number of new or repaired drinking water facilities in the vilages. $\beta_{\text{reserved}} \neq 0$

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

```r
## (b) Run a bivariate regression to test this hypothesis

## Read data
df <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
head(df)
View(df)
reserved <- factor(df$reserved)

## Scatter plot
scatter <- ggplot(data = df,
                  mapping = aes(x = reserved, y = water)) +
  geom_point()
scatter

## Fit model
model <- lm(water~reserved, data = df)
summary(model)
```

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

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.

Linear regression model: $Y_{\text{water}} = 14.738 + 9.252 \times X_{\text{reserved}}$

The slope of this regression equation is about 9.252 with a p-value of 0.0197, which is less than the significant level of 0.05. This indicates that the relationship between the independent variable and the dependent variable is statistically significant. Therefore, the evidence can reject null hypothesis and supports the alternative hypothesis that there is a positive relationship between the reservation policy and the number of new or repaired drinking water facilities. As the value of X_{reserved} increases, the number of new or repaired drinking water facilities increases as well. One unit of increase in the variable of reservation policy corresponds to an increases of 9.252 units in the number of new or repaired drinking water facilities.

The Y-intercept is approximately equal to 14.738. This means when the policy is not reserved ()$X_{\text{reserved}} = 0$), the predicted number of new or repaired drinking water facilities is 14.738.