Problem Set 2

Applied Stats/Quant Methods 1

Due: October 14, 2024

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).

To calculate the test statistic χ^2, first we need f_e and f_o:

f_o is simply the observed frequency of the data. (which was named bribe_data in R)

```
fo <- bribe_data
```

Then f_e which is calculated by $f_e = \frac{\text{row total}}{\text{grand total}} * \text{column total}$

I computed this using for loops that cycle through the columns and rows performing this calculation and storing the answers in f_e.

```
for (i in 1:nrow(bribe_data)){

  for (j in 1:ncol(bribe_data)){

    fe_i <-   sum(bribe_data[,j]) / grand_total * sum(bribe_data[i,])

    fe <- append(fe, fe_i)
  }
}
```

This provides us with the following for f_o and f_e respectively:

Table 1: f_o

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	14	6	7
Lower class	7	7	1

Table 2: f_e

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	13.500	8.357	5.143
Lower class	7.500	4.643	2.857

*(Nb. these tables were extracted from R using the **Stargazer** package)*

The next stage is to calculate the test statistic χ^2 using:

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

Or in R:

```
test_stat <- sum( (fe - fo)^2 / fe)
# 3.791168
```

I was then able to verify this answer was correct via the **chisq.test()** function built into R.

```
chi_test <- chisq.test(bribe_data)
chi_test$statistic
#  3.791168
```

As both χ^2's match I can be confident my calculations were correct.

(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 I used the function **pchisq()**:

```
pchisq(test_stat ,
       df = degf,
       lower.tail=FALSE)
# 0.1502306
```

Where the degrees of freedom were calculated by:

```
degf <- (nrow(bribe_data) -1) * (ncol(bribe_data)-1)
# 2
```

Given that $\alpha = 0.1$ and our p-value = 0.1502306...

$$0.1502... > 0.1$$

$$\text{p-value} > \text{significance level}$$

Therefore, there is not sufficient evidence to reject the null hypothesis - that the two variables are dependant on each other. (class and bribe)

Hence, we cannot say whether being 'Upper' or 'Lower' Class are independent of being requested to solicit a bribe to an officer.

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

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

To calculate the standardised residuals I used the following equation:

$$z = \frac{f_o - f_e}{se}$$

where

$$\text{se} = \sqrt{f_e * (1 - \text{row proportion}) * (1 - \text{column proportion})}$$

In R this looked like:

```
for (i in 1:nrow(bribe_data)){

  for (j in 1:ncol(bribe_data)){

    se <- sqrt(fe[i,j] *
               (1 - sum(bribe_data[,j]) / grand_total) *
               (1 - sum(bribe_data[i,]) / grand_total)
               )

    z_i <- (fo[i,j]-fe[i,j]) / se

    z <- append(z, z_i)

    }
}
```

This produced to following table:

Table 3: Standardised Residuals

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.523

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

Standardised residuals provide insight into deviations from expected frequencies, helping to identify potential relationships between categorical variables. Residuals exceeding ± 1.96 indicate significant deviations at the 5% level, suggesting a possible lack of independence between variables in that cell. In this case, the residual for upper-class individuals "Bribe requested" (-1.642) suggests they are less likely to face bribe requests than expected, though not quite reaching significance. Similarly, the residual for upper-class individuals being "Stopped/given warning" (1.523) shows a slight, non-significant tendency to receive more warnings. For lower-class individuals, the residuals for "Bribe requested" (1.642) and 'Stopped/given warning" (-1.523) point to trends where they are more likely to face bribe requests and less likely to receive warnings, though these deviations remain suggestive rather than definitive, as they fall below the standard significance threshold.

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 $\beta = 0$**:** That there is no difference in the number of new or repaired drinking water facilities in the villages that had spaces reserved for women

Alternative $\beta \neq 0$**:** That there is a significant difference in the number of new or repaired drinking water facilities in the villages that had spaces reserved for women.

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

First I read in the data and inspected the variables to gain a better understanding of it's structure and layout.

```
# read in the data
econ <- read.csv(url("https://raw.githubusercontent.com/kosukeimai/qss/
    master/PREDICTION/women.csv"))

# have a look at the data
str(econ)
```

```
'data.frame': 322 obs. of  6 variables:
 $ GP        : int  1 1 2 2 3 3 4 4 5 5 ...
 $ village   : int  2 1 2 1 2 1 2 1 2 1 ...
 $ reserved  : int  1 1 1 1 0 0 0 0 0 0 ...
 $ female    : int  1 1 1 1 0 0 0 0 0 0 ...
 $ irrigation: int  0 5 2 4 0 0 4 0 0 0 ...
 $ water     : int  10 0 2 31 0 0 7 12 28 0 ...
```

To run a bivariate regression in R, I used the **lm()** function:

```
biv_reg <- lm(water ~ reserved, data = econ)
```

This produces:

```

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

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.

Looking at the Regression Coefficients:

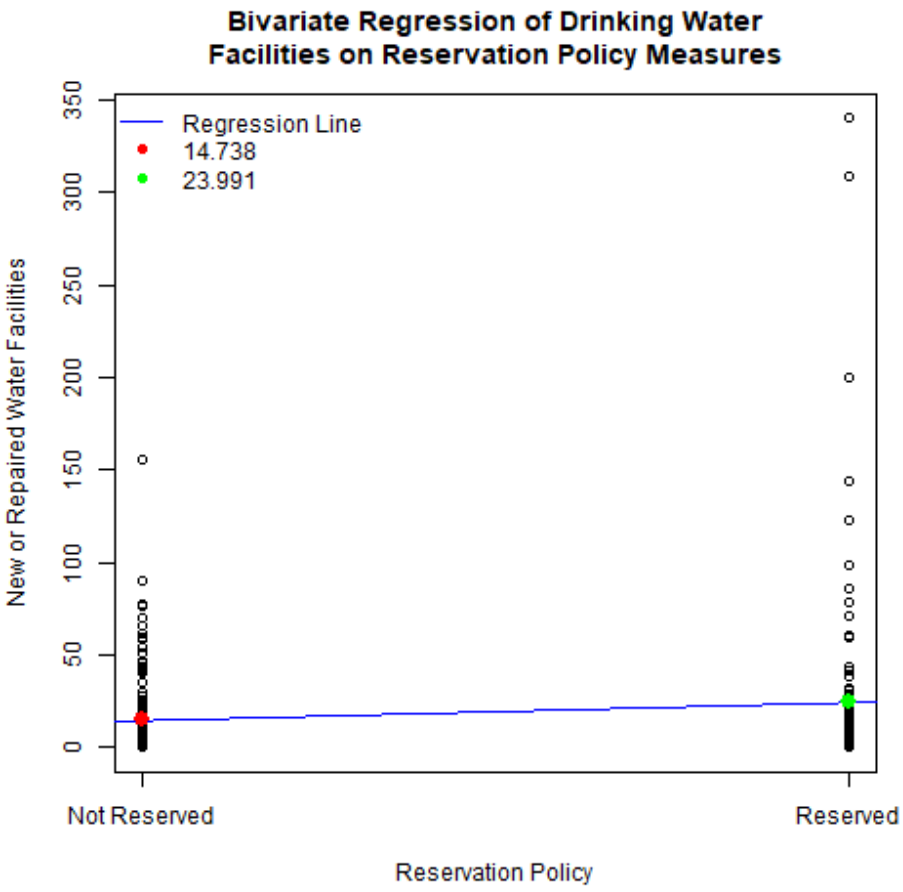
```
biv_reg$coefficients
```

```
(Intercept)    reserved
  14.738318    9.252423
```

The regression coefficients indicate a positive relationship between the variables. The slope of approximately 9 suggests that when a space was reserved for a female leader, there was a 9-fold increase in the likelihood of additional drinking water facilities being built. Specifically:

- For non reserved leadership, the expected number of facilities built is given by the intercept, 14.738.
- For reserved leadership, the expected number increases to 23.991 (intercept + slope).

This can be visualised by plotting the data with the regression line:


Bivariate Regression of Drinking Water Facilities on Reservation Policy Measures
Regression Line
14.738
23.991
New or Repaired Water Facilities
0
50
100
150
200
250
300
350
Not Reserved
Reserved
Reservation Policy