# Problem Set 2

Applied Stats/Quant Methods 1

Due: October 16, 2022

## 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 16, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

## Question 1 (40 points): 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 $\chi^2$ test statistic by hand/manually (even better if you can do "by hand" in R).

First, the totals of each column and row was created, with an overall total being recorded also

| | Not Stopped | Bribe requested | Stopped/given warning | Sum |
|---|---|---|---|---|
| Upper class | 14 | 6 | 7 | 27 |
| Lower class | 7 | 7 | 1 | 15 |
| Sum | 21 | 13 | 8 | 42 |

After this, the expected values were then calculated, as shown in the following table:

| | Not Stopped | Bribe requested | Stopped/given warning |
|---|---|---|---|
| Upper class | 14(13.5) | 6(8.36) | 7(5.14) |
| Lower class | 7(7.5) | 7(4.64) | 1(2.86) |

The degrees of freedom was then found, using the formula (1- number of columns)(1 - number of rows), which was 2. Using the formula, the test statistic was then calculated, with $\chi^2 = 3.80352207$. This calculation was also checked against the 'chisq()' function in R, which gave the following output:

```
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 $\alpha = 0.1$?
Using the following code, the following p-value was produced:

```r
pchisq(3.80342207, df=2, lower.tail=F)
```

```
[1] 0.1493129
```

With the p-value being greater than .10, we fail to reject the null hypothesis that there was no significant difference between the likelihood of a bribe being solicited depending on whether the drivers were in the upper or lower-class groups.

[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 | 0.322 | -1.643 | 1.526 |
| Lower class | -0.322 | 1.644 | -1.525 |

(d) How might the standardized residuals help you interpret the results?
Standardised residuals help to interpret the results as it normalises the observed residuals. This means that the residuals are converted to Z scores, where it can be easily observed above that all values fall in the interval $-2 \leq Z \leq 2$, as any score outside this interval would be determined as significant.

# Question 2 (40 points): 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).

[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.
H0 = The gender of a leader in a GP will not have an effect on the relationship between the number of irrigation facilities created or repaired and the number of drinking water facilities repaired.
H1 = The gender of a leader in a GP will have an effect on the relationship between the number of irrigation facilities created or repaired and the number of drinking water facilities repaired.

(b) Run a bivariate regression to test this hypothesis in R (include your code!). The following code was used to run the regression:

```
reg1 <- lm(water ~ irrigation*female, data = dat3)
```

From this, the following output was created

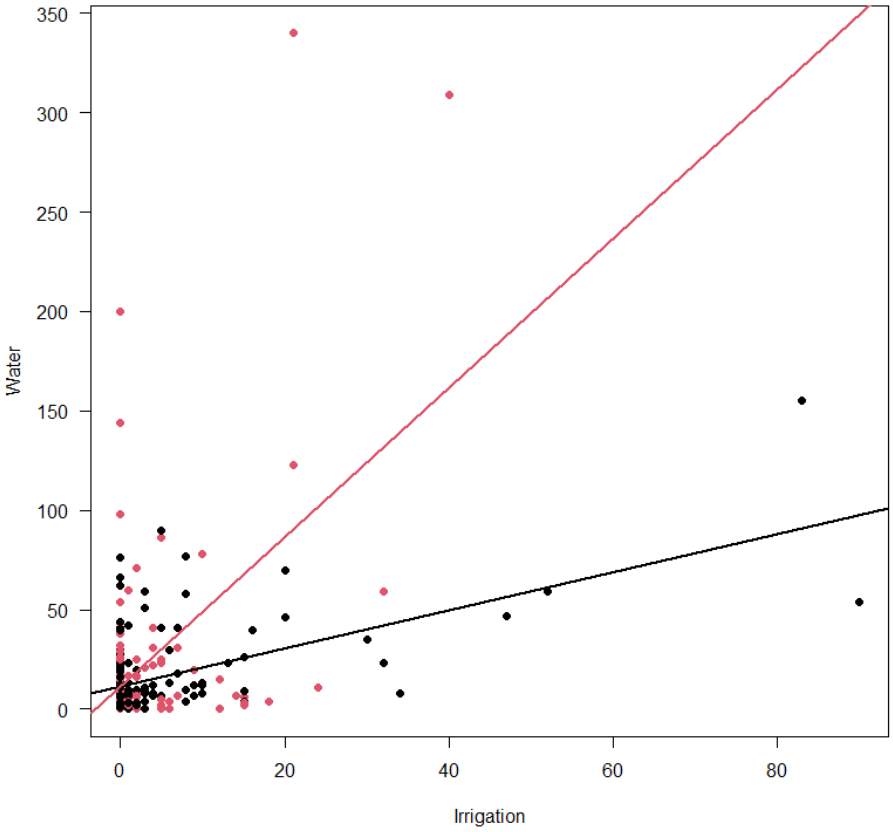
```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)         11.5474     2.1497   5.372 1.51e-07 ***
irrigation           0.9579     0.1873   5.116 5.42e-07 ***
female1             -0.2583     3.5843  -0.072    0.943
irrigation:female1   2.7978     0.4449   6.288 1.06e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.88 on 318 degrees of freedom
Multiple R-squared:  0.2713,Adjusted R-squared:  0.2645
F-statistic: 39.47 on 3 and 318 DF,  p-value: < 2.2e-16
```

As shown above, the regression itself is significant, however, the R-squared figure is low at .27. This means that the model only accounts for 27 percent of the variance of the dependent variable, being the water facilities created/repaired.
However, there does appear to be a significant interaction effect between gender and

the amount of irrigation facilities created/repaired, as displayed by this scatterplot, where the black dots indicate male leaders and the red dots indicate female leaders:

Figure 2: Scatterplot




(c) Interpret the coefficient estimate for reservation policy.

In the form of y = mx + b, the coefficients give the following formula:
no. of water facilities = 11.5474 + 0.9579(no. of irrigation facilities) - 0.2583(female) + 2.7978(irrigation facilities * female)

This can actually be separated out into two different regression formulas for both male and female leaders, which give the following coefficients:

```
gp_male <- dat3[dat3$female==0,]
gp_female <- dat3[dat3$female==1,]
reg_male <- lm(water ~ irrigation, data=gp_male)
reg_female <- lm(water ~ irrigation, data=gp_female)
```

```
> reg_male$coefficients
(Intercept)  irrigation
11.5474327   0.9579383
> reg_female$coefficients
(Intercept)  irrigation
11.289121    3.755716
```

This displays that while the Y intercepts for the two genders are similar, the slopes are quite different, with the female regression having a much steeper slope, as shown above in Figure 2. However, as previously stated, neither have a strong R-squared statistic, with the female regression model only accounting for 25.3 % of the variance, and with the male regression accounting for 29.5 % of the variance in the outcome variable