Problem Set 2

Applied Stats/Quant Methods 1

Due: October 15, 2021

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 in `.pdf` form.
- This problem set is due before class on Friday October 15, 2021. No late assignments will be accepted.
- Total available points for this homework is 100.

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 χ^2 test statistic by hand (even better if you can do "by hand" in `R`).**

1. Here we will be using the Chi Square test to determine if two variables in a contingency table are related. Seeing whether the distributions are different.

   To calculate the χ^2 test statistic by hand we first need to calculate *the expected frequencies*.

   This involves multiplying the total number of observations per row by the total number of observations for the column. This is then all divided by the total number of observations:

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	(27*21)/42=13.5	(27*13)/42=8.36	(27*8)/42=5.14
Lower class	(15*21)/42=7.5	(15*13)/42=4.64	(15*8)/42=2.86

   Note: A chi square should only be done when the expected frequencies are above 5, as this is not the case a warning message is produced in R. (the assignment has said to calculate the p value anyways)

   Now that we have the observed and expected frequencies. We can now compare the two groups to see if they differ significantly. This is calculated by finding the square of the observed frequencies minus the expected frequencies *(showing this difference)*, then dividing this by the expected frequencies. The sum of all of these are the test statistic.

	subgroup	calculation	to be summed:
1	upperclass*notstopped	((14-13.5)$\hat{2}$)/13.5 =	0.0185
2	upperclass*briberequested	(6-8.36)$\hat{2}$)/8.36 =	0.6662
3	upperclass*stopped/givenwarning	(7-5.14)$\hat{2}$)/5.14 =	0.6731
4	lowerclass*notstopped	((7-7.5)$\hat{2}$)/7.5 =	0.0333
5	lowerclass*briberequested	(7-4.64)$\hat{2}$)/4.64 =	1.2003
6	lowerclass*stopped/givenwarning	(1-2.86)$\hat{2}$)/2.86 =	1.21

   Sum of the above = 3.79 . This is the χ^2 test statistic.

(b) **Now calculate the p-value from the test statistic you just created (in R).[2] What do you conclude if $\alpha = .1$?**

We can input the above χ^2 test statistic into R code.

```
bribe_chisq <- pchisq(3.79, df = 2, lower.tail = FALSE)
bribe_chisq
```

First we need to figure out the null and alternative hypothesis: As the Chi square is a test of Independence we are looking at whether the two distributions are statistically different.

H0 = the variables are statistically independent

Ha = the variables are statistically dependant.

This produces a p-value of **0.1503183**, as this is above our $\alpha = .1$, we fail to reject the null hypothesis. This means that we fail to reject that the variables are statistically independent.

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

The following code computes the Chi Square and the standardized residuals. *Note:* the χ^2 test statistic is the same as the calucaltion done by hand. *Standardized residual output is in the table*

```
datatable <- matrix(c(14,6,7,7,7,1),nrow=3,ncol=2)
datatable

chisq <- chisq.test(datatable,correct=FALSE)
chisq
chisq$stdres
```

This outputs the following:

```
Pearson's Chi-squared test

data:  datatable
X-squared = 3.7912, df = 2, p-value = 0.1502
```

Standardised residuals

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.532

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

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

Standardized residuals help you to interpret the results as they work on the individual cells and they can therefore show information on which cell contributes to the significance. The standardized residuals are a measure of the strength of the difference between observed and expected values.

The standardized residuals make it easier to tell which off the cells are contributing to the most and least to the value. A general rule states that a standardised residual of **below -2** means the value is less the expected frequency and **above 2** is above the expected frequency.

Question 2 (20 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).

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.

1. **(a)State a null and alternative (two-tailed) hypothesis.**

   The null hypothesis is that the reservation policy has *no* effect on the number of new or repaired drinking water facilities in the village.

   The alternative hypothesis is that the reservation policy has an effect on the number of new or repaired drinking water facilities in the village.

   H0 = the slope = 0

   H1 = the slope != 0 (slope is significantly different to zero)

   As we are looking at a two tailed test we are interested in whether the slope is positive and negative. The null hypothesis would fail to be rejected if the observation was no different than would be expected by chance alone and therefore having a slope of zero.

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

   To run a bivariate regression I used the following code:

```
1
2    lm <- lm(economicsdata$irrigation ~ economicsdata$reserved)
3    summary(lm)
4
```

   This outputs the following data:

```
1      Call:
2 lm(formula = economicsdata$irrigation ~ economicsdata$reserved)
3
4 Residuals:
5    Min      1Q  Median      3Q     Max
6 -3.388  -3.388  -3.019  -1.019  86.612
7
8 Coefficients:
9                          Estimate Std. Error t value Pr(>|t|)
10 (Intercept)               3.3879     0.6498   5.214 3.33e-07 ***
11 economicsdata$reserved   -0.3693     1.1220  -0.329    0.742
12 ---
13 Signif. codes:  0     ***     0.001     **     0.01     *     0.05     .     0.1
              1
14
15 Residual standard error: 9.506 on 320 degrees of freedom
16 Multiple R-squared:  0.0003385, Adjusted R-squared:  -0.002785
17 F-statistic: 0.1084 on 1 and 320 DF,  p-value: 0.7422
18
```

3. **(c) Interpret the coefficient estimate for reservation policy.**

   The coefficient estimate for the reservation policy is **-0.3693**. This means there is a negative weak relationship. As the reserved variable was a binary variable saying wether the GP was reserved or not (1 being reserved, 0 being not), we can say that

there are less new or repaired irrigation facilities in relation to GP's that were not reserved for women leaders; as the reserved variable gets closer to 0 there is a decrease in the number of irrigation facilities.

However, when we are interpreting the bivariate regression output the p-value 0.7422 is greater than the significance level 0.05. This indicates that there is insufficient evidence within the data to conclude that a non-zero correlation exists. We therefore fail to reject the null hypothesis. This relationship is too weak to make any real conclusions about.

Question 3 (40 points): Biology

There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity for male fruit flies. The five groups are: males forced to live alone, males assigned to live with one or eight newly pregnant females (non-receptive females), and males assigned to live with one or eight virgin females (interested females). The name of the data set is `fruitfly.csv`.[4]

`No`	serial number (1-25) within each group of 25
`type`	Type of experimental assignment 1 = no females 2 = 1 newly pregnant female 3 = 8 newly pregnant females 4 = 1 virgin female 5 = 8 virgin females
`lifespan`	lifespan (days)
`thorax`	length of thorax (mm)
`sleep`	percentage of each day spent sleeping

1. **Import the data set and obtain summary statistiscs and examine the distribution of the overall lifespan of the fruitflies.**

   I imported the data set and then used the following code to look at summary statistics:

```
summary(fruitfly)
```

   This summary function in R gave some basic statistics about the data.

```
    thorax          longevity          activity
 Min.   :0.6400   Min.   :16.00   isolated:25
 1st Qu.:0.7600   1st Qu.:46.00   one      :25
 Median :0.8400   Median :58.00   low      :25
 Mean   :0.8224   Mean   :57.62   many     :24
 3rd Qu.:0.8800   3rd Qu.:70.00   high     :25
 Max.   :0.9400   Max.   :97.00
```

   This data shows that the *mean* thorax length of fruit flies in mm is **0.8224** and the *mean* lifespan in days is **57.62**.

   To examine the distribution of the lifespan of the fruitflies I used a histogram and input the following code in R:

[4] Partridge and Farquhar (1981). "Sexual Activity and the Lifespan of Male Fruitflies". *Nature*. 294, 580-581.

```
p<-ggplot(fruitfly , aes(x=longevity)) +
  geom_histogram(color="black", fill="lightgreen", binwidth = 7)+
  labs(title="Fruitfly Lifespan Distibution",x="Days", y = "Count")
p
```

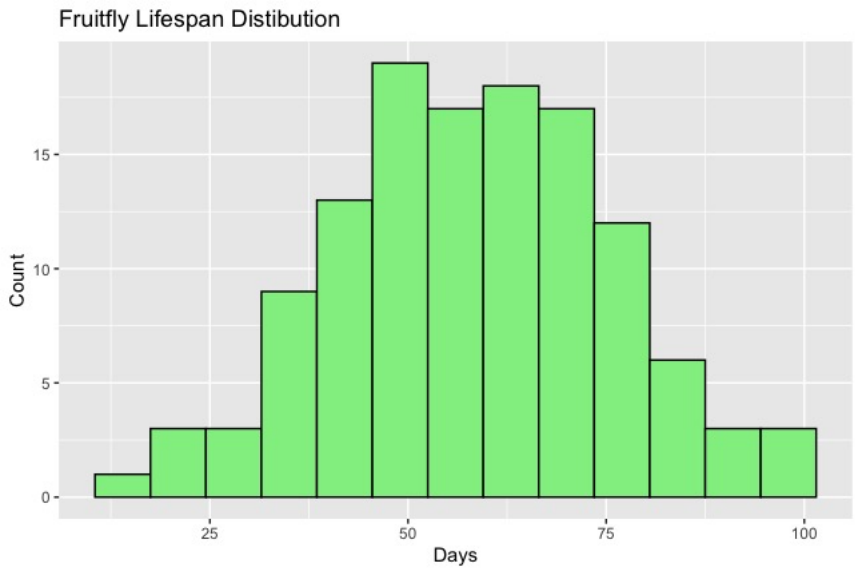

Figure 2: Histogram of the distibution of the lifespan of fruitflies in the dataset

2. **Plot `lifespan` vs `thorax`. Does it look like there is a linear relationship? Provide the plot. What is the correlation coefficient between these two variables?**

   To examine the linear relationship I visually inspected a scatterplot which I used to following code to produce:

```
ggplot(aes(longevity, thorax), data = fruitfly) +
  geom_point(size=2, shape=16) +
  ggtitle("Scatterplot of Thorax and lifespan") +
  labs(y="Thorax Size in mm")+
  labs(x = "Number of Days")
```

   This produced the following graph:

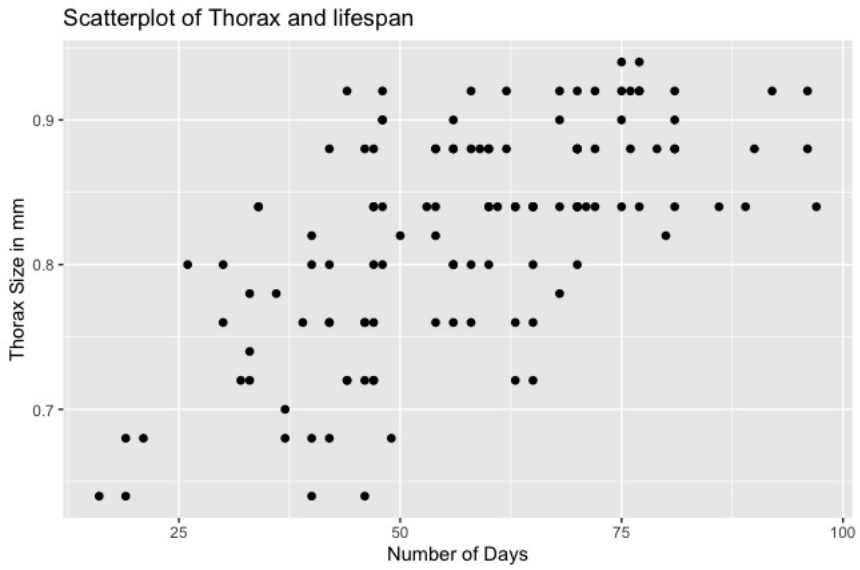

Figure 3: Scatterplot to show the relationship between Thorax size and Lifespan

From examining the above scatterplot it suggests that there is a weak linear relationship.

To find the correlation coefficient between the two variables I used a Pearsons product moment correlation in R using the `cor.test` function in R:

```
cor.test(fruitfly$longevity, fruitfly$thorax)
```

The following output is produced:

```
        Pearson's product-moment correlation

data:  fruitfly$longevity and fruitfly$thorax
t = 9.1521, df = 123, p-value = 1.497e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5188709 0.7304479
sample estimates:
      cor
0.6364835

```

This shows that the lifespan of the fly and the size of the thorax have a correlation of **0.6364835** which is positive linear relationship. This suggests there is a correlation but is not saying that one is causing the other. In addition, the data points are quite spaced out, suggesting it is not a very string relationship. A strong relationship would usually be defined has having a correlation above 0.8.

3. **Regress `lifespan` on `thorax`. Interpret the slope of the fitted model.**

   To regress `lifespan` on `thorax` I used the following code to run a regression:

```
   Call:
lm(formula = fruitfly$longevity ~ fruitfly$thorax)
```

   This code is written as such because the first variable *depends* on the second. And in this case we are interested on the lifespan *depending* on the thorax size. The lifespan is the **dependant variable** and the thorax is the **independent variable**.

```
Coefficients:
    (Intercept)  fruitfly$thorax
         -61.86           145.28
```

   This shows that the slope of the line is 145.28. This slope means that for every one unit increase in thorax size there is a 145.8 day increase in longevity. As we are dealing with millimeters we can divide that by 10 which shows that for every .1mm increase in thorax size there is a 14.5 day increase.

4. **Test for a significant linear relationship between `lifespan` and `thorax`. Provide and interpret your results of your test.**

   The following R code can be used to use a linear regression to set for a significant linear relationship.

```
   lm <- lm(fruitfly$longevity~ fruitfly$thorax)
summary(lm)
```

   Provide: This outputs the following values:

```
    Call:
lm(formula = fruitfly$longevity ~ fruitfly$thorax)

Residuals:
    Min      1Q  Median      3Q     Max
-28.364  -9.986   1.258   9.264  36.825
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -61.86      13.37  -4.625 9.39e-06 ***
fruitfly$thorax  145.28      16.19   8.971 4.27e-15 ***
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
          1
Residual standard error: 13.65 on 122 degrees of freedom
Multiple R-squared:  0.3975,  Adjusted R-squared:  0.3926
F-statistic: 80.49 on 1 and 122 DF,  p-value: 4.275e-15
```

   Interpret:This shows that the p value for the test is **4.275e-15**. Using a 95% confidence interval with a alpha value of 0.05, and the p vlaue of regression is less that this so we can conclude that there is a significant relationship between `lifespan` and `thorax`.

5. **Provide the 90% confidence interval for the slope of the fitted model.**

- **Use the formula of confidence interval.**

Here I have used the critical value of 1.645 to represent the 90% confidence interval. And input the slope of the line and standard error from the summary of the regression. The upper limits and lower limits of the confidence interval are found by plus and minusing the margin of error (critical value by standard error) from the slope of the limit. See code below for the *by hand* version in R:

```
slopeoftheline <- 145.28
criticalvalue <- 1.645
standarderror <- 16.19

marginoferror <- criticalvalue*standarderror

upperlimit = slopeoftheline + marginoferror
lowerlimit = slopeoftheline - marginoferror

upperlimit
lowerlimit
```

This outputs the following confidence interval: *[118.6474, 171.9126]*

- **Use the function `confint()` in `R`**

Using the following function:

```
    confint(lm, level = 0.9)

```

The outputs the following:

```
                5 %       95 %
(Intercept)      -84.02438 -39.69101
fruitfly$thorax 118.43754 172.11657

```

7. **Use the `predict()` function in `R` to (1) predict an individual fruitfly's lifespan when `thorax`=0.8 and (2) the average `lifespan` of fruitflies when `thorax`=0.8 by the fitted model. This requires that you compute prediction and confidence intervals. What are the expected values of lifespan? What are the prediction and confidence intervals around the expected values?**

   First I changed my data from a dataframe to lm. This allows the `predict()` function to work in R.

```
fruitflylm <- lm(longevity ~ thorax, data = fruitfly)
```

   From here I was able to input the new data of `thorax`=0.8 to preform a prediction.

```
newdataframe <- data.frame(thorax = 0.8)
```

   For (1) *to predict an individual fruitfly's lifespan when* ***thorax****=0.8* I was able to compute the prediction and confidence interval using the following code:

```
predict(fruitflylm, newdata = newdataframe, interval = "predict")
```

   This output the following:

```
fit       lwr       upr
1 54.36395 27.21787 81.51003
```

   This shows that an individuals expected lifespan is 54.36395 when the thorax size is 0.8, and that we are 95% confident that the data will fall between [27.21787, 81.51003] for an individual.

   For (2) *the average* ***lifespan*** *of fruitflies when* ***thorax****=0.8* I used the following code to compute for the average lifespan:

```
predict(fruitflylm, newdata = newdataframe, interval = "confidence")
```

   This output the following:

```
fit       lwr       upr
1 54.36395 51.83262 56.89528
```

   This shows that the average expected lifespan is 54.36395 when the thorax size is 0.8mm, and that we are 95% confident that the data will fall between [51.83262, 56.89528] on average.

8. **For a sequence of `thorax` values, draw a plot with their fitted values for `lifespan`, as well as the prediction intervals and confidence intervals.**

   The code below allowed the confidence and prediction intervals to be plotted on a scatter plot of thorax values.

```
fruitflylm <- lm(longevity ~ thorax, data = fruitfly)
predictions <- predict(fruitflylm, interval = "predict")
alldata <- cbind(fruitfly, predictions)

```

```
ggplot(aes(thorax, longevity), data = alldata) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x) +
  geom_line(aes(y = lwr), col = "coral2", linetype = "dashed") +
  geom_line(aes(y = upr), col = "coral2", linetype = "dashed") +
  labs(y="Days") +
  labs(x ="Thorax size mm")

```

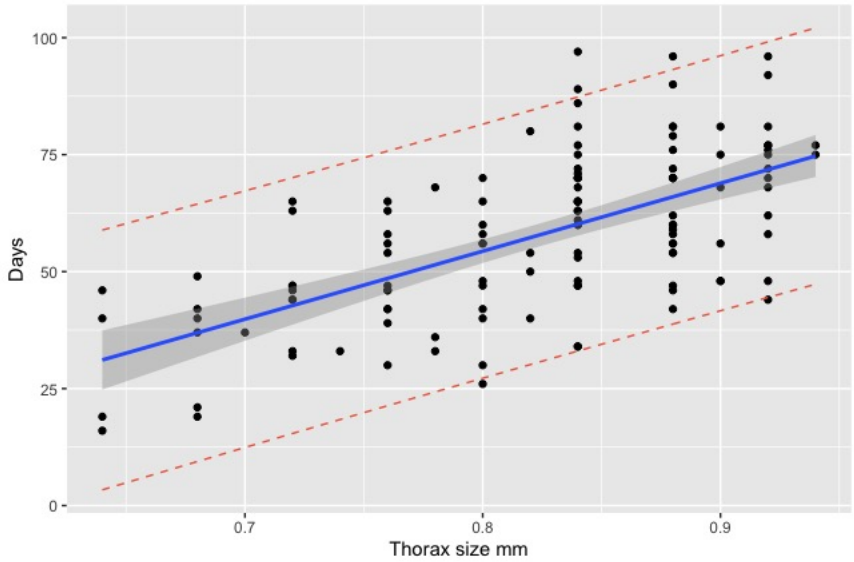

Figure 4: Scatterplot showing the prediction and confidence intervals around the Regression line

This visually portrays that the shaded in grey area is the confidence interval of [51.83262, 56.89528] and the coral dotted line shows the prediction interval between [27.21787, 81.51003].