Problem Set 2

Applied Stats/Quant Methods 1

Due: October 15, 2021

Name: Gareth Moen

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).

χ^2 is 3.7861

*See images for hand calculation.

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

The p-value is less than 0.25 and greater than 0.1 based on the tables so we might say that it's around 0.177 which isn't strong enough to reject H0.

[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	0.24	1.63	1.52
Lower class	-0.32	1.65	1.52

*See images for hand calculations.

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

Standardized residulas in an individual cell provides evidence against independence in that cell. Values between -3 and 3 are convincing evidence of a true effect in that cell.

By hand calculation for Question 1 (page 1)

Q1

		Not stopped	Bribe	warning	Total
a)	up.	14 (13.5)	6 (8.35)	7 (5.14)	27
	low	7 (7.5)	7 (4.64)	1 (2.85)	15
		21	13	8	42

$$\frac{(14-13.5)^2}{13.5} + \frac{(6-8.35)^2}{8.35} + \frac{(7-5.14)^2}{5.14} +$$

$$\frac{(7-7.5)^2}{7.5} + \frac{(7-4.64)^2}{4.64} + \frac{(1-2.85)^2}{2.85}$$

$$= .0185 + .6613 + .673$$

$$+ .0333 + 1.2 + 1.2 = \boxed{3.7861} = \chi^2$$

b) P-value from test statistic

$$df = (r-1)(c-1)$$
$$= (2-1)(3-1)$$
$$= 2$$

$p \approx .25 > p > .1$ → Not strong enough to reject H_0

(.177 more or less)

if $\alpha = .1$ H_0 can't be rejected, variables are probably independent

By hand calculation for Question 1 (page 2)

c) Standardized residuals (blue)

f_e = Red

	Not stopped		Bribe		Warning		
Up	① (.24) 14	13.5	③ 1.63 6	8.35	⑤ 1.52 7	5.14	27
Low	② (-.32) 7	7.5	④ 1.65 7	4.64	⑥ 1.52 1	2.85	15
	21		13		8		42

$$z = \frac{(f_o) - (f_e)}{\sqrt{(f_e)(1 - \text{row prop})(1 - \text{col prop})}}$$

① $\frac{14 - 13.5}{\sqrt{13.5(1 - .64)(1 - .5)}} = \frac{.5}{.208} = .24$

② $\frac{7 - 7.5}{\sqrt{7.5(1 - .35)(1 - .5)}} \quad \frac{-.5}{1.56} = .32$

③ $\frac{6 - 8.35}{\sqrt{8.35(1 - .64)(1 - .31)}} = \frac{-2.35}{1.44} = 1.63$

④ $\frac{7 - 4.64}{\sqrt{4.64(1 - .36)(1 - .31)}} = \frac{2.36}{1.43} = 1.65$

⑤ $\frac{7 - 5.14}{\sqrt{5.14(1 - .64)(1 - .19)}} = \frac{1.86}{1.22} = 1.52$

⑥ $\frac{1 - 2.85}{\sqrt{2.85(1 - .36)(1 - .19)}} = \frac{-1.85}{1.22} = -1.52$

d Standardized residuals in an individual cell provides evidence against independence in that cell. Values of -3 to 3 are convincing evidence of a true effect in that cell.

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).

[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.

The null hypothesis would be that the reservation policy has had NO EFFECT on the number of new or repaired drinking water facilities. The alternative hypothesis would be that the reservation policy HAS HAD an effect on the number of new or repaired drinking water facilities.

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

Subsetting the data into two groups, group (1) being the group were the GPs where positions were reserved for women and (2) where they weren't reserved. In group 1 the mean number of number of new or repaired drinking water facilities is 23.99 while in 2 its 14.75 using the code...

```
round(mean(reserved_women_yes$water), 2)
round(mean(reserved_women_no$water), 2)
```

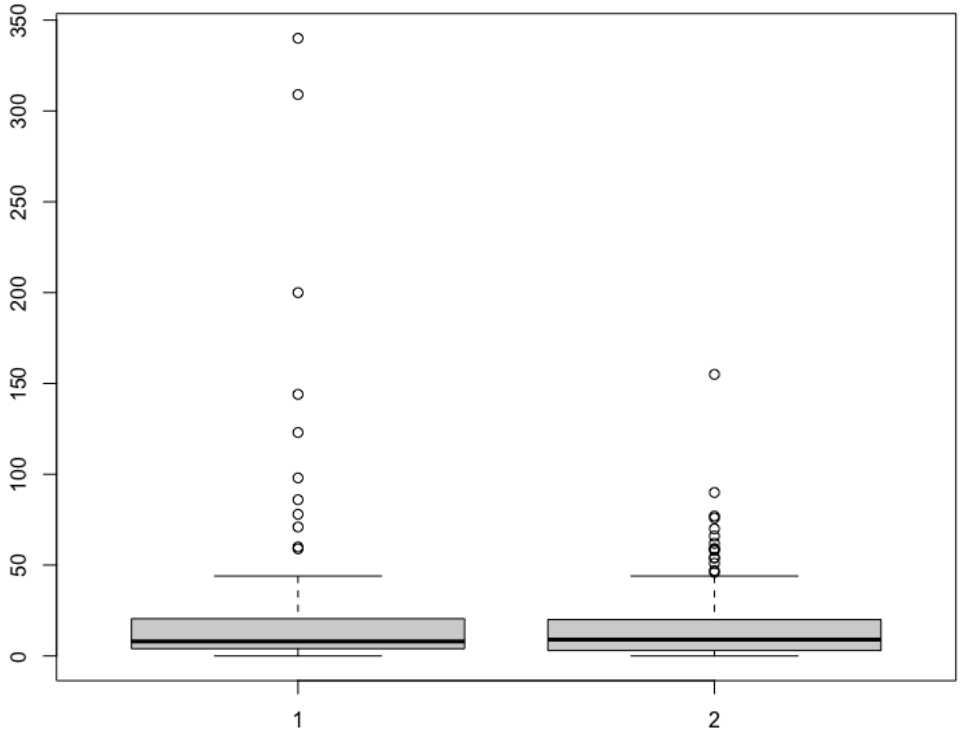
This suggests that the number of new or repaired drinking-water facilities in the villages is higher due to the presence of women.

The sd of group 1 is 51.26 and for group 2 it's 18.99 using the code...

```
round(sd(reserved_women_yes$water), 2)
round(sd(reserved_women_no$water), 2)
```

This suggests that the data is more spread out in group 1 than 2.

A boxplot shows the means and outliers.



It seems that a few outliers may be greatly distorting the data.

Here's an attempted bivarate regression of means.

```
#Bivariate regressoin of means
lm(reserved_women_yes$water ~ reserved_women_no$water, data=regressMat)
#this doesn't work as regressMat isn't defined and I don't know where
#it's supposed to come from.
```

(c) Interpret the coefficient estimate for reservation policy.

Seeing as I didn't conduct a bivariate regression in (b) I can't do this. I haven't got to bivariate regression yet in my catch-up reading of AF. On the plus side I have managed to use LaTex.

Question 3 (40 points): Biology

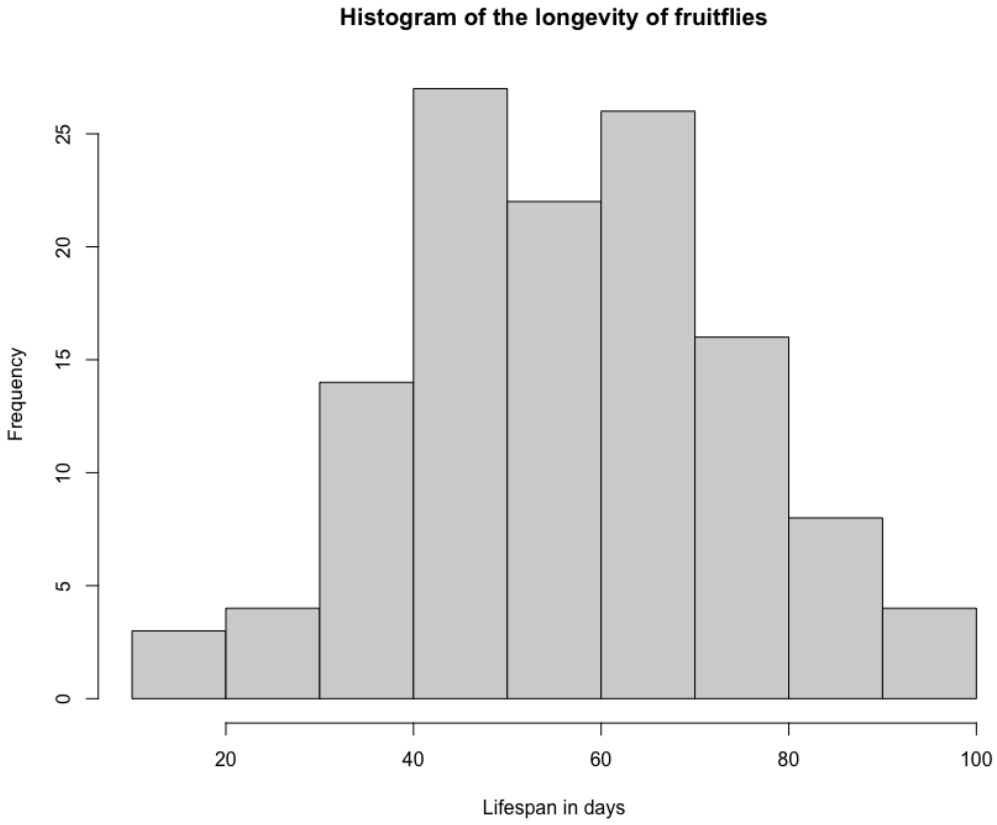
There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity 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

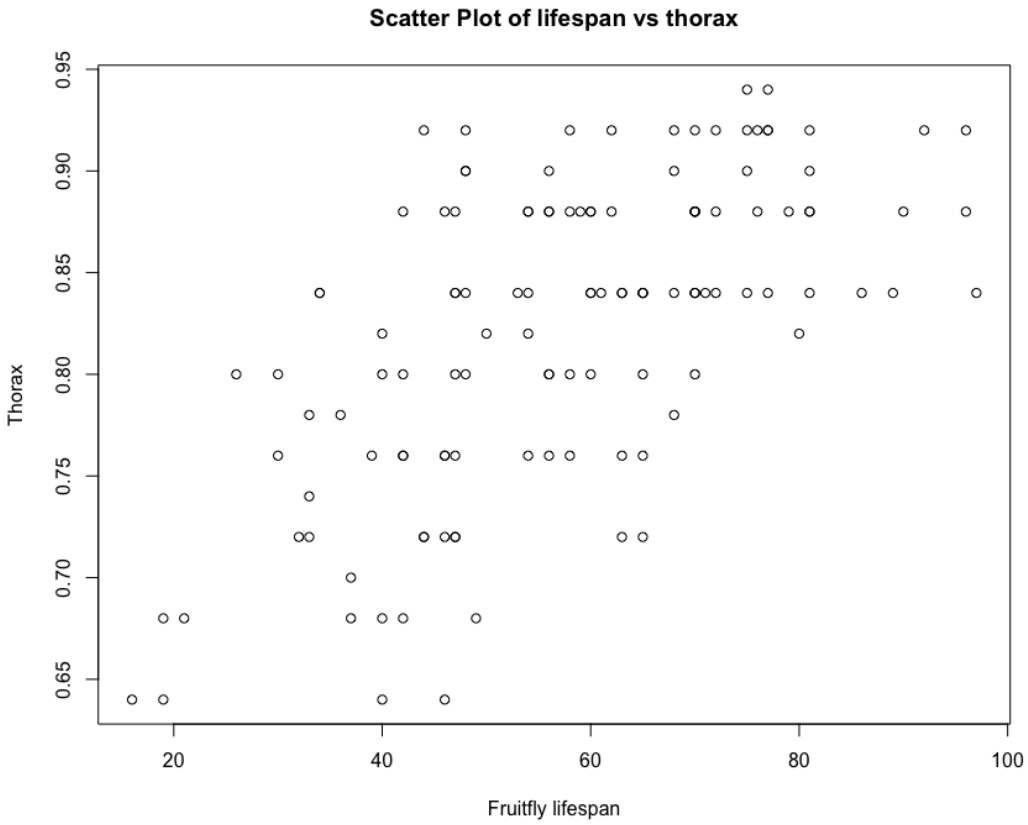
1. Import the data set and obtain summary statistiscs and examine the distribution of the overall lifespan of the fruitflies.

```
data(fruitfly)
?fruitfly # explanation of the dataset 'fruitfly'
str(fruitfly) #overview of dataset
summary(fruitfly) #overview of dataset
View(fruitfly)
hist(fruitfly$longevity) # mapping a histogram of the longevity in a barchart
```

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


Histogram of the longevity of fruitflies
Frequency
0
5
10
15
20
25
20
40
60
80
100
Lifespan in days

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?



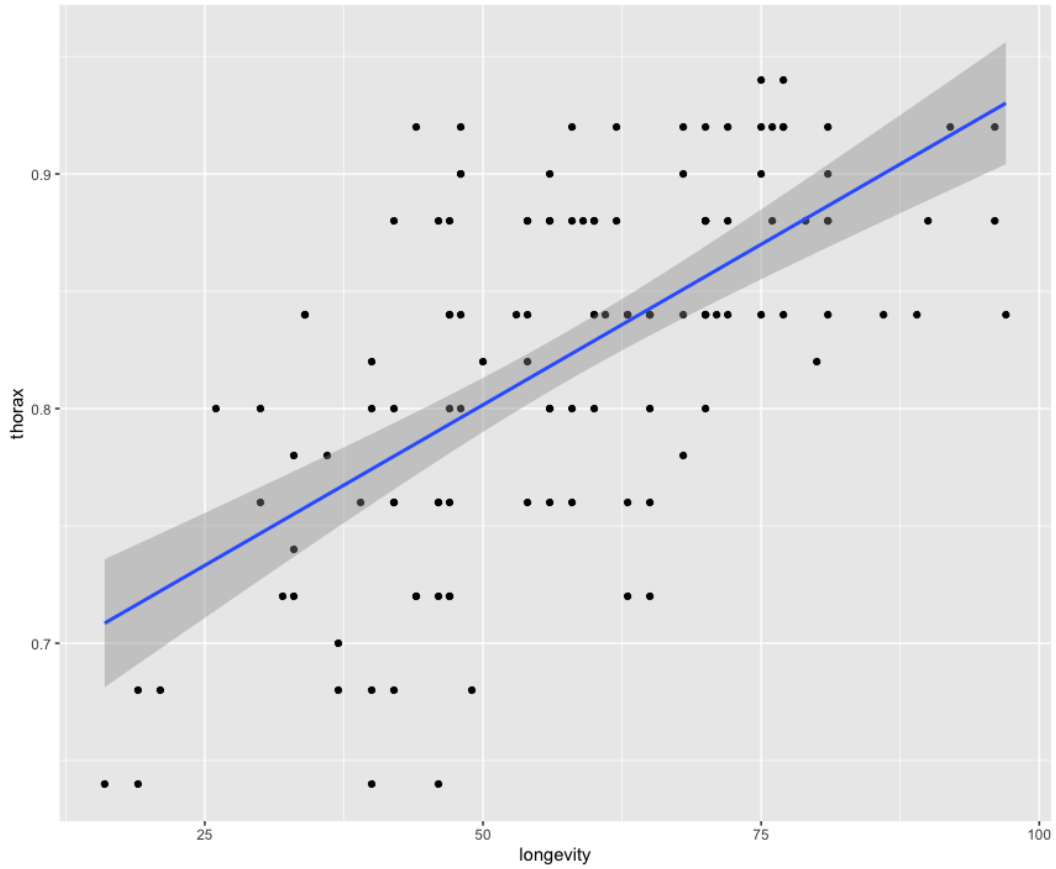
There appears to be a linear relationship between lifespan and thorax.

Using the code below to assess the correlation...

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

...gives us a correlation coefficient of 0.6304674

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



The line intercepts the x-axis at around 0.7. There is a positive relationship between the x an y variables, longevity and thorax in this case.

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

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

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 ***
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
```

The p-value is a lot less than .05 so there is a significant relationshiop between the variables.

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

   - Use the formula of confidence interval.

     Formula...

     β_0: $\hat{\beta}_0 \pm t \times se$

     Gives the result -61.86 + 1.66 x 13.37 = -39.67 and -61.86 - 1.66 x 13.37 = -84.05

   - Use the function `confint()` in `R` . Using the function...

     ```
     confint(fruitfly_lm, level = .9)
     ```

     Provides the result...

     ```
     > confint(fruitfly_lm, level = .9)
                       5 %      95 %
     (Intercept) -84.02438 -39.69101
     thorax      118.43754 172.11657
     ```

6. 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?

```
(1) fruitfly_ndf <- data.frame(thorax = c(runif(.94, .64, 30)))
#trying to make a random vector
#with similar characteristics to thorax to make predicitons,
#but something is wrong :-(
```

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

   I've got nothing for this.

   The formatting is a bit rough in this document, especially with the positioning of images, but one step at a time I suppose.