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`). The χ^2 is calculated using the following formula:

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$$

First we need to find the expected values, presented in the table below. Once these are found we plug them into our formula

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	27/42*21=13.5	27/42*13=8.35	27/42*8=5.14
Lower class	15/42*21=7.5	15/42*13=4.46	15/42*8=2.86

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$$

	Not Stopped	Bribe requested	Stopped/given warning	Chi Squared
Upper class	0.0185	0.0618	0.673	
Lower class	0.0333	1.2	1.207	
Chi Squared				3.1936

(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 the probability of obtaining a chi-square as large or larger than that in the current experiment and yet the data will still support the hypothesis. It is the probability of deviations from what was expected being due to mere chance. To find it in r we use the following line of code:

```
#get standardized residuals
chitest$stdres
```

With $\alpha = .1$?, when we reference the table we find a critical vale of 9.210. As our p value is 0.2849114 (less than the critical value) we would fail to reject our null hypotheses that the variables are statistically indepenent.

[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.

```
#get standardized residuals
chitest$stdres
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.32	-1.64	1.52
Lower class	-0.32	1.64	-1.52

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

Standardized residuals are useful in helping to interpret chi-square tables by providing information about which cells contribute to a significant chi-square.

If the chi squared test is significant we need to take a look at residuals.

If the value of standardized residual is lower than -2 it means that the cell contains fewer observations that it was expected (the case of variables independence). If the value of standardized residual is higher than 2 it means that the cell contains more observations that it was expected.

This is not the case for any of our cells.

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. Null Hypothesis: There is a relationship between reservation policies and drinking water facilities. Areas with reserved GP's for women will have higher amounts of new and repaired drinking water facilities

Alternative Hypothesis: There is not a relationship between reservation policies and drinking water facilities. Areas with reserved GP's for women will not have higher amounts of new and repaired drinking water facilities

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

```r
#run bivariate regression; Water and Reserved variables
lm1 <- lm(women$water ~ women$reserved)

summary(lm1)

 abline(lm(women$water ~ women$reserved), col = "red")

 plot(lm(women$water ~ women$reserved)
 abline(lm(women$water ~ women$reserved), col = "blue")

 ggplot(aes(water, reserved), data = women) +
   geom_point() +
   geom_smooth(method = "lm", formula = y ~ x)
```

(c) Interpret the coefficient estimate for reservation policy. We get a p value of 0.0197 for reservation policy, meaning that women's representation has a statistically significant effect on the presence of new or repaired water reservations.

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.

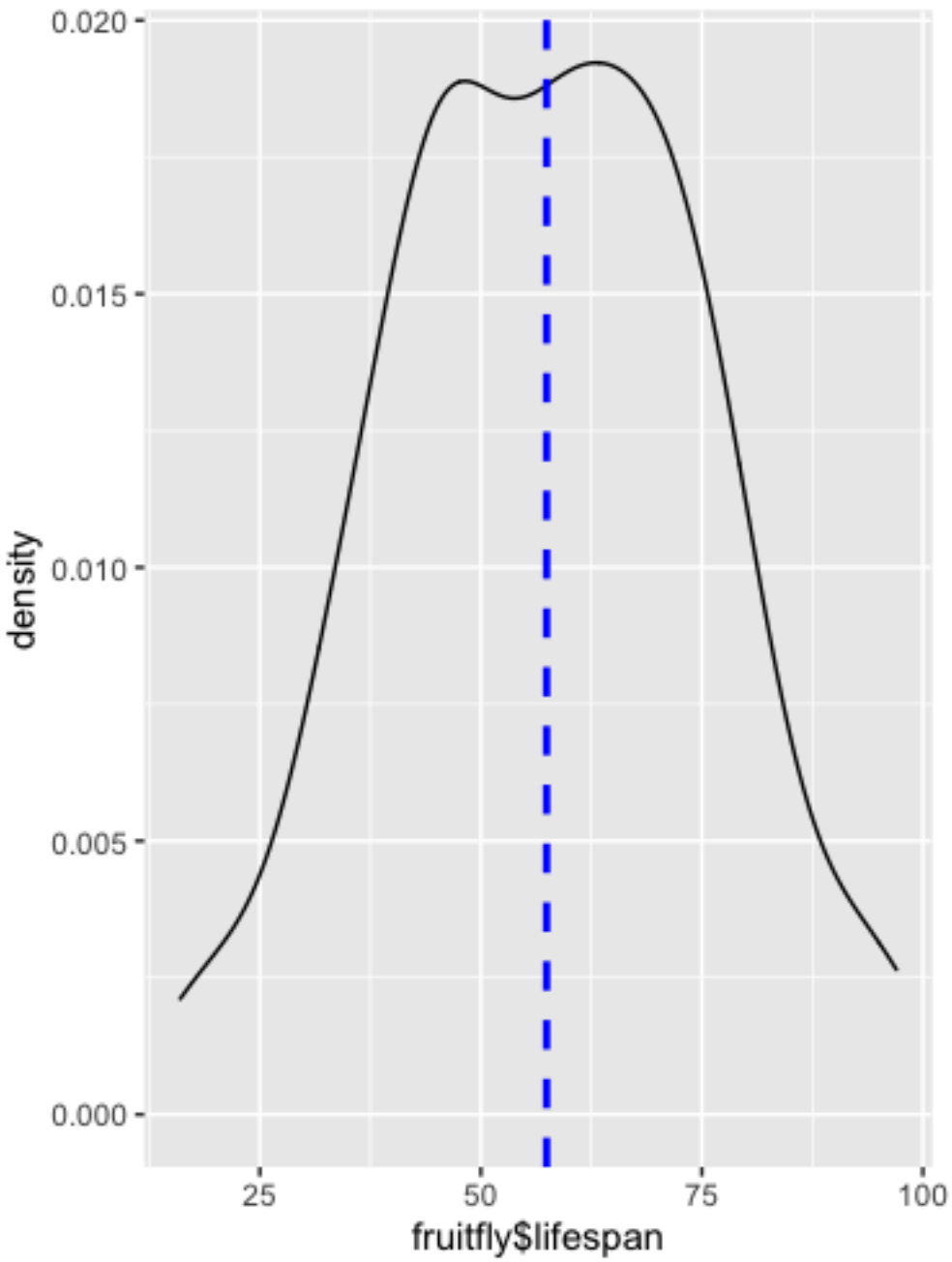
```r
#import the data set
fruitfly <- read.csv("https://raw.githubusercontent.com/jackoneillWHU/
    StatsI_Fall2021/main/datasets/fruitfly.csv")
summary(fruitfly)

library(ggplot2)

plot(fruitfly$type, fruitfly$lifespan,
     main = "Scatter Plot Lifespan by Type",
     xlab = "Type",
     ylab = "Lifespan")

ggplot(fruitfly, aes(x=fruitfly$lifespan)) +
  geom_density() + geom_vline(aes(xintercept=mean(lifespan)),
                              color="blue", linetype="dashed", size=1)
```

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


0.020
0.015
0.010
0.005
0.000
density
25
50
75
100
fruitfly$lifespan

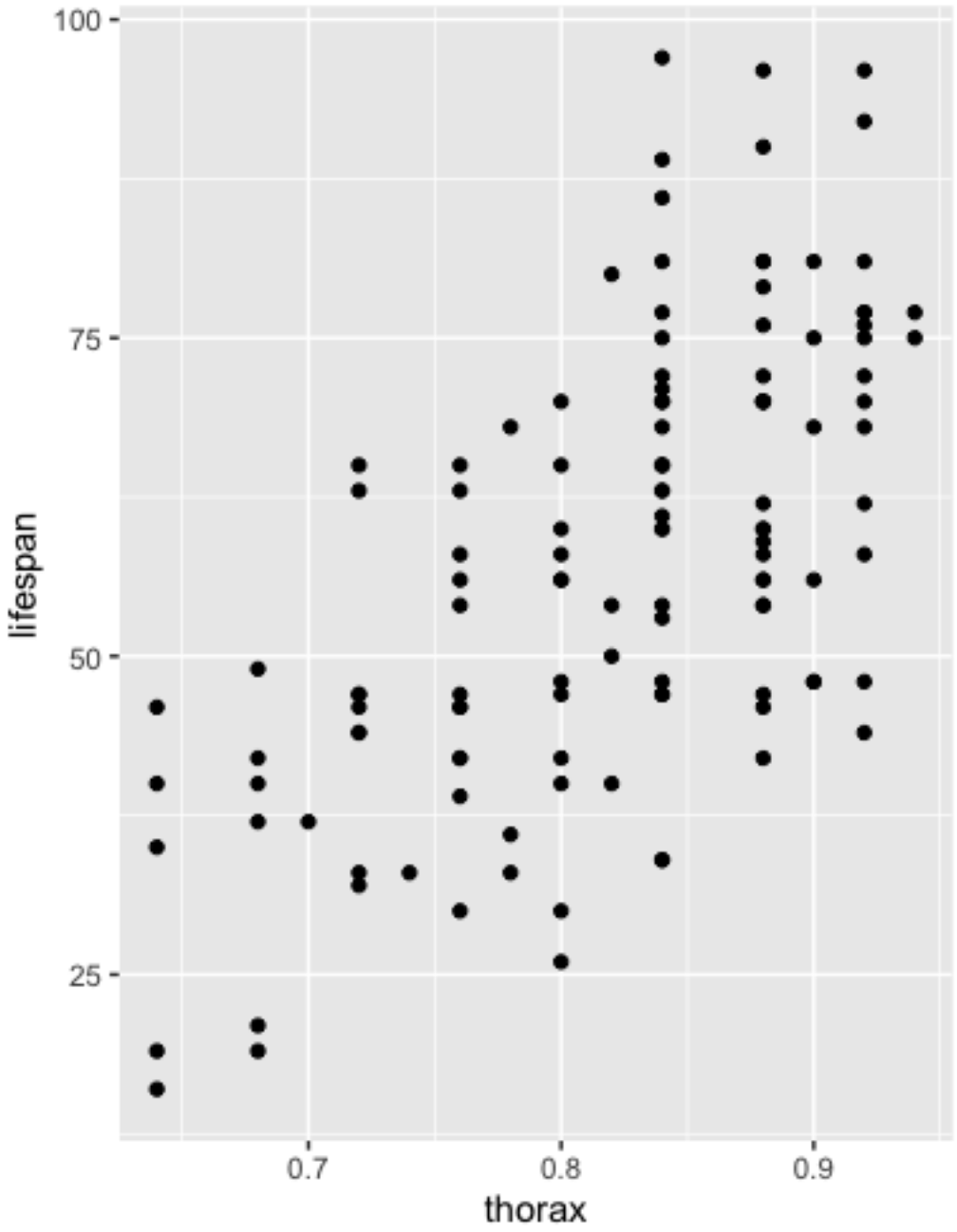
2. Plot `lifespan` vs `thorax`. Does it look like there is a linear relationship? Provide the plot. What is the correlation coefficient between these two variables?

```
#plot lifespan vs thorax
ggplot(fruitfly , aes(x=thorax, y= lifespan)) +geom_point()

#find correlation between the two
cor(fruitfly$thorax, fruitfly$lifespan)
```

When we look at the scatter plot of lifespan and thorax for interpretation it appears as if there is a strong positive relationship between the two variables, although there is some dispersion.

The correlation coefficient is 0.6365. This would be in line with out visual interpretation that there is a positive correlation between the two variables.

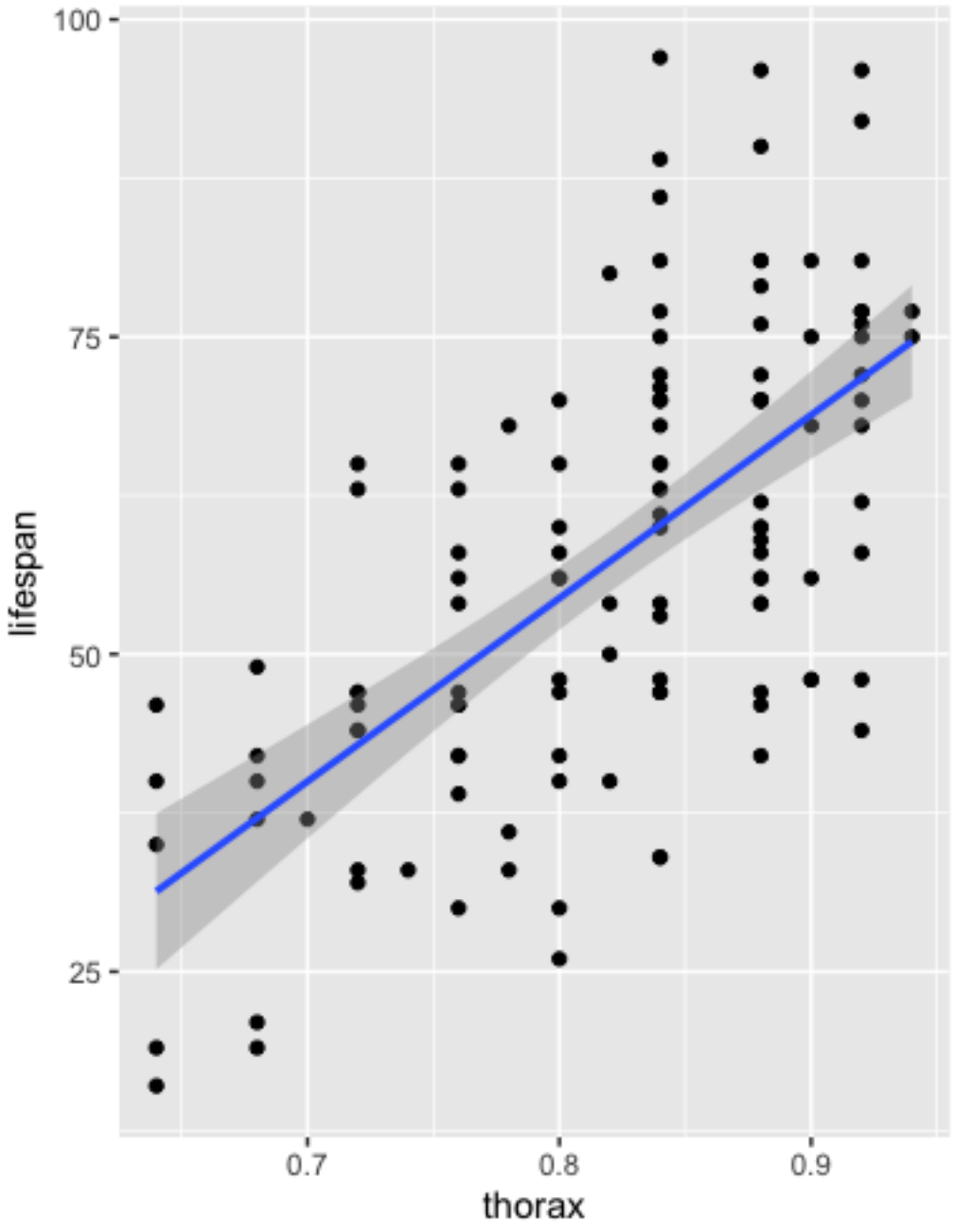


3. Regress `lifespan` on `thorax`. Interpret the slope of the fitted model.

```
#regression; lifespan and thorax
#run bivariate regression; Lifespan and thorax
lmfly <- lm(fruitfly$lifespan ~ fruitfly$thorax)

```

```r
summary(lmfly)
      ggplot(aes(thorax, lifespan), data = fruitfly) +
        geom_point() +
        geom_smooth(method = "lm", formula = y ~ x)
```



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

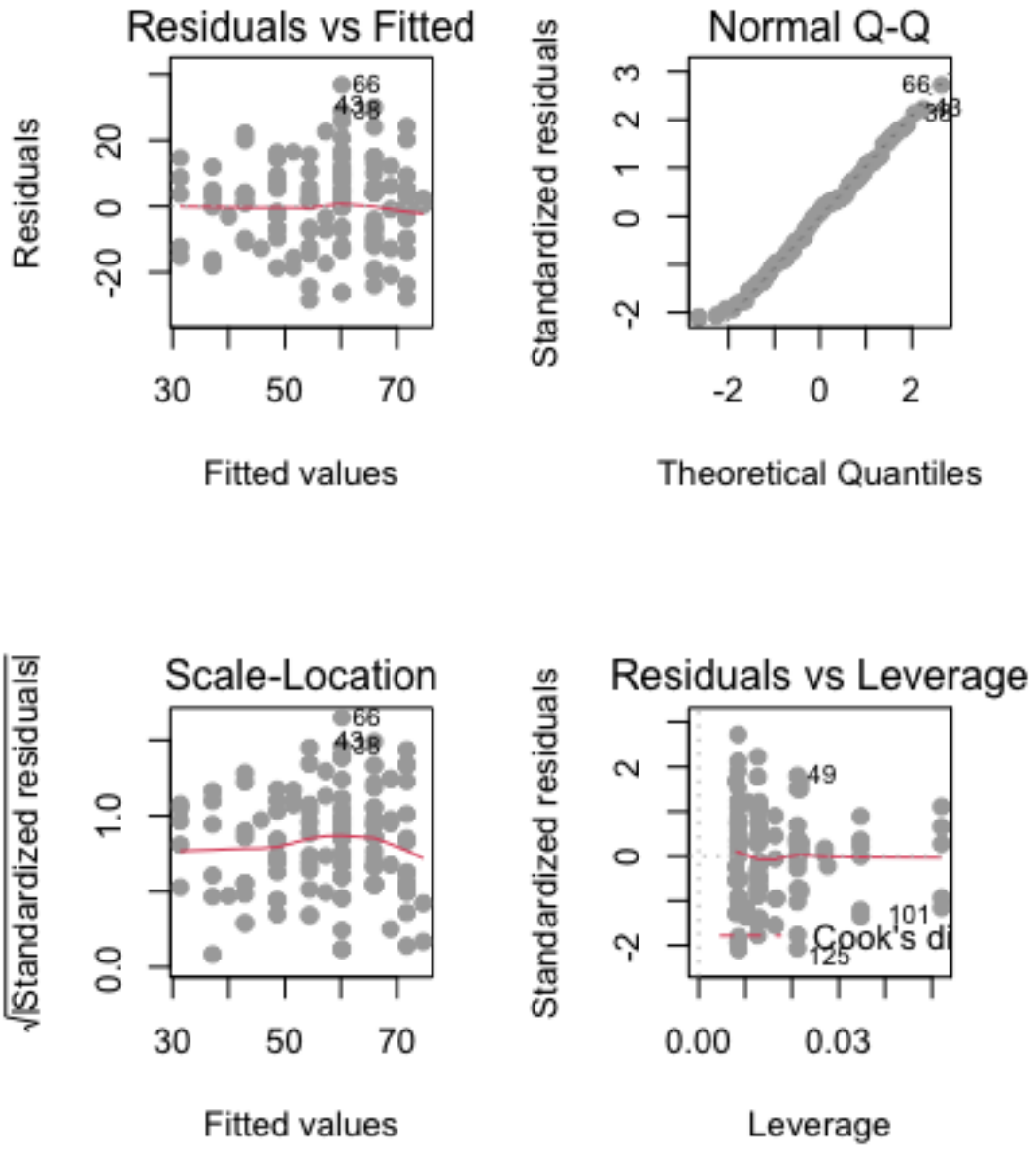
   If there is a significant linear relationship between the independent variable X and the dependent variable Y, the slope will not equal zero. The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

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

   - Use the formula of confidence interval.
   - Use the function `confint()` in `R` .

```r
#90% ci
      confint(lmfly, level=0.90)

      par(mfrow=c(2, 2))
      plot(lmfly, pch=19, col='darkgrey')
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's di

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?

```
## produce predicted lifespan for thorax = 0.8mm
 newdata <- data.frame(thorax = 0.8)
 predict(lmfly, newdata)


 newdata <- data.frame(thorax = 0.8)
 predict(lmfly, newdata, interval = "confidence", level = 0.97)


 plot(lifespan ~ thorax, data = fruitfly, pch = 19, col='darkgrey')


 newdata <- cbind(newdata, lifespan=predict(lmfly, newdata)
                  plot(lifespan ~ thorax, data = fruitfly, pch = 19, col=
   'darkgrey')
```

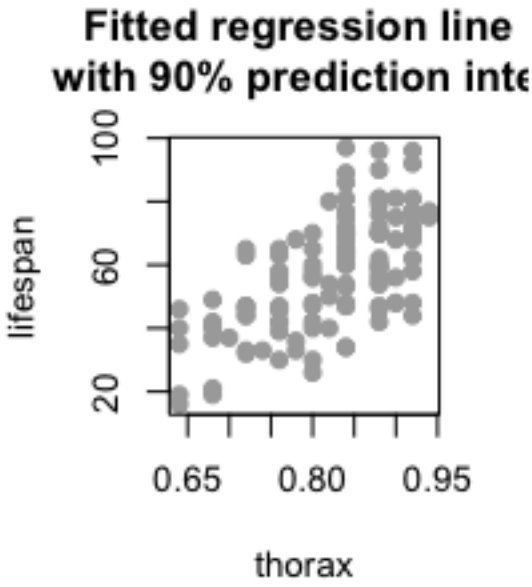
7. For a sequence of `thorax` values, draw a plot with their fitted values for `lifespan`, as well as the prediction intervals and confidence intervals.

```
      ## create new data frame to predict to
      newdata <- data.frame(
        thorax = seq(min(fruitfly$thorax),
                     max(fruitfly$thorax), length.out = 50)
      )

```

```
      ## produce predictions and intervals
      newdata <- cbind(newdata,
                       predict(lmfly, newdata,
                               interval = "prediction",
                               level = 0.90))
      newdata$lifespan <- newdata$lmfly
      newdata$lmfly <- NULL

      ## plot fitted line against the raw data
      plot(lifespan ~ thorax, data = fruitfly,
           pch = 19, col='darkgrey',
           main = "Fitted regression line
     with 90% prediction interval")
      lines(lifespan ~ thorax, data = newdata)
      lines(lwr ~ thorax, data = newdata,
            lty = 2)
      lines(upr ~ thorax, data = newdata,
            lty = 2)
```


Fitted regression line
with 90% prediction inte
lifespan
20
60
100
0.65
0.80
0.95
thorax