Problem Set 2

QTM 200: Applied Regression Analysis

February 10, 2020

Question 1 (40 points): Political Science

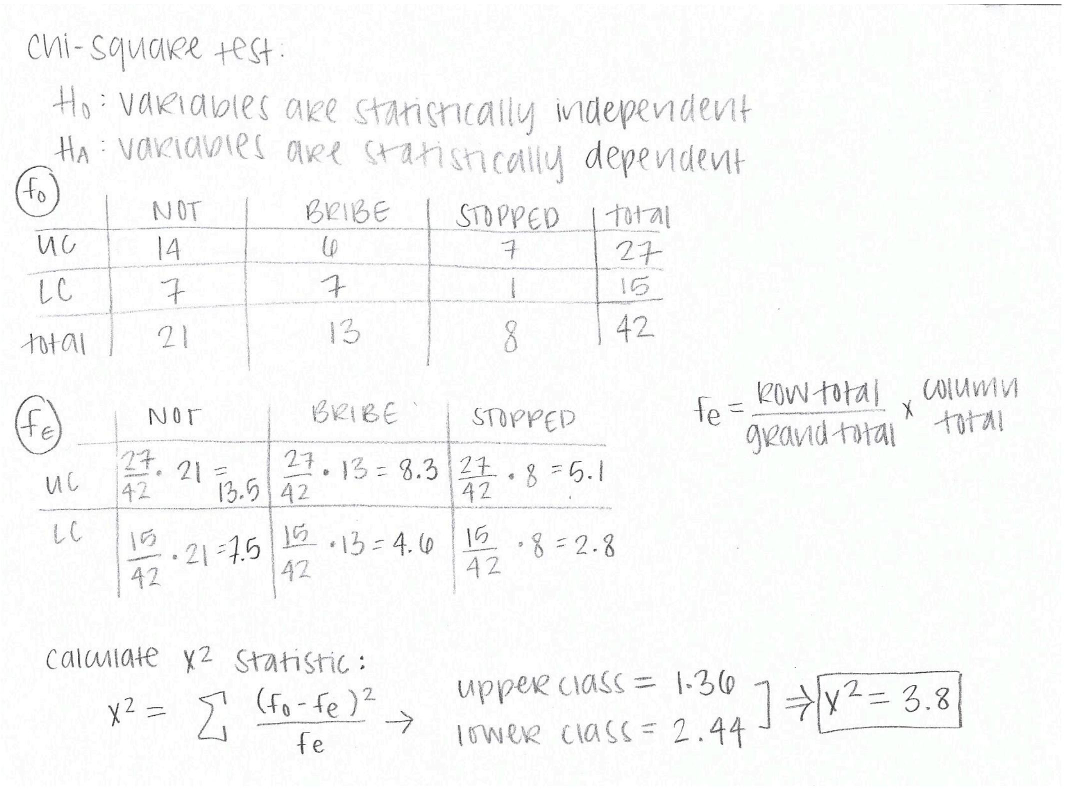
The following table was created using the data from a study run in a major Latin American city.[1] As part of the experimental treatment in the study, one employee of the research team was chosen to make illegal left turns across traffic to draw the attention of the police officers on shift. Two employee 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.

	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] 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.

Figure 1: χ^2 by hand:



chi-square test:

H_0: variables are statistically independent

H_A: variables are statistically dependent

f_o

	NOT	BRIBE	STOPPED	total
UC	14	6	7	27
LC	7	7	1	15
total	21	13	8	42

f_e

	NOT	BRIBE	STOPPED
UC	$\frac{27}{42} \cdot 21 = 13.5$	$\frac{27}{42} \cdot 13 = 8.3$	$\frac{27}{42} \cdot 8 = 5.1$
LC	$\frac{15}{42} \cdot 21 = 7.5$	$\frac{15}{42} \cdot 13 = 4.6$	$\frac{15}{42} \cdot 8 = 2.8$

$$f_e = \frac{\text{Row total}}{\text{grand total}} \times \text{column total}$$

calculate χ^2 statistic:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \rightarrow \left.\begin{array}{l}\text{upper class} = 1.36 \\ \text{lower class} = 2.44\end{array}\right] \Rightarrow \boxed{\chi^2 = 3.8}$$

$\chi^2 = 3.8$

Check χ^2 in R:

```
# Check by-hand answer for chi-square
chisq.test(classbribe) #3.79
```

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

```
# find df and calculate p value
(nrow(classbribe)-1)*(ncol(classbribe)-1)
pval <- pchisq(3.8, df=2, lower.tail=FALSE)
```

P-value = 0.15 At $\alpha = .1$, we fail to reject the null hypothesis that the variables are statistically independent.

[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.

```
(14-13.5)/sqrt(13.5*(1-(27/42))*(1-(21/42))) # 0.32
(6-8.3)/sqrt(8.3*(1-(27/42))*(1-(13/42))) # -1.60
(7-5.1)/sqrt(5.1*(1-(27/42))*(1-(8/42))) # 1.56
(7-7.5)/sqrt(7.5*(1-(15/42))*(1-(21/42))) # -0.32
(7-4.6)/sqrt(4.6*(1-(15/42))*(1-(13/42))) # 1.67
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.32	-1.60	1.56
Lower class	-0.32	1.67	-1.49

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

The researchers were interested in whether officers were more or less likely to solicit a bribe from drivers depending on their class, either upper class or lower. In response, the chi-square test revealed that the two variables were statistically independent. The standardized residuals help interpret this result, as they reveals where the deviation from independence, or lack thereof took place. The proportion of ”not stopped,” for example, deviated the least from the expected value for both upper and lower class officers.

Question 2 (20 points): Economics

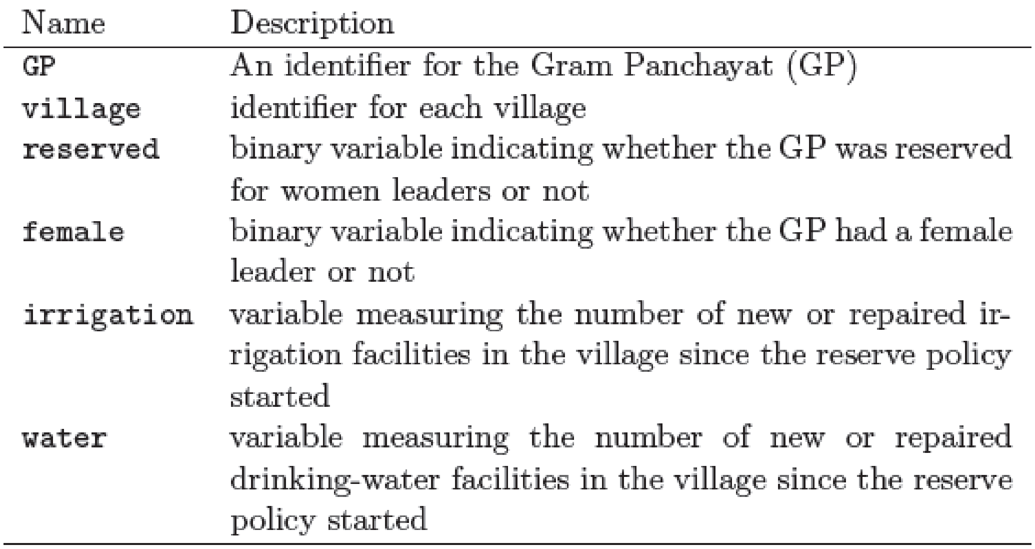
Chattopadhyay and Duflo were interested in whether women promote different policies than men.[3] Answering this question with observational data is pretty difficult due to potential confounding problems (e.g. the districts that choose female politicians are likely to systematically differ in other aspects 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 2 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

[3] Chattopadhyay and Duflo. (2004). “Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica*. 72 (5), 1409-1443.

of the reservation policy on the number of new or repaired drinking water facilities in the villages.

Figure 2: 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

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

$H_0 = 0$ (there is no association)

$H_a > 0$ (there is an association)

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

```

mean(women$reserved)
mean(women$water)
sum(women$reserved)
sum(women$water)
sum(((women$reserved)-mean(women$reserved))*((women$water)-mean(women$
    water)))
sum((women$water-mean(women$water))^2)

regress <- lm(data=women, water~reserved)
regress
# (Intercept)      reserved
# 14.738          9.252
```

Slope = 9.252 & Intercept = 14.74

(c) Interpret the coefficient estimate for reservation policy.

$$Y = \alpha + \ X$$

$$Y = 14.738 + 9.252X$$

When reservation policy is 0, the number of repaired drinking water fountains is 14.738. A one unit increase in the reserved variable is associated with a 9.252 unit increase in the number of new or repaired drinking water facilities in the village since the reserve policy started. We can thus reject the null hypothesis, since there is an effect between the two variables.

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]

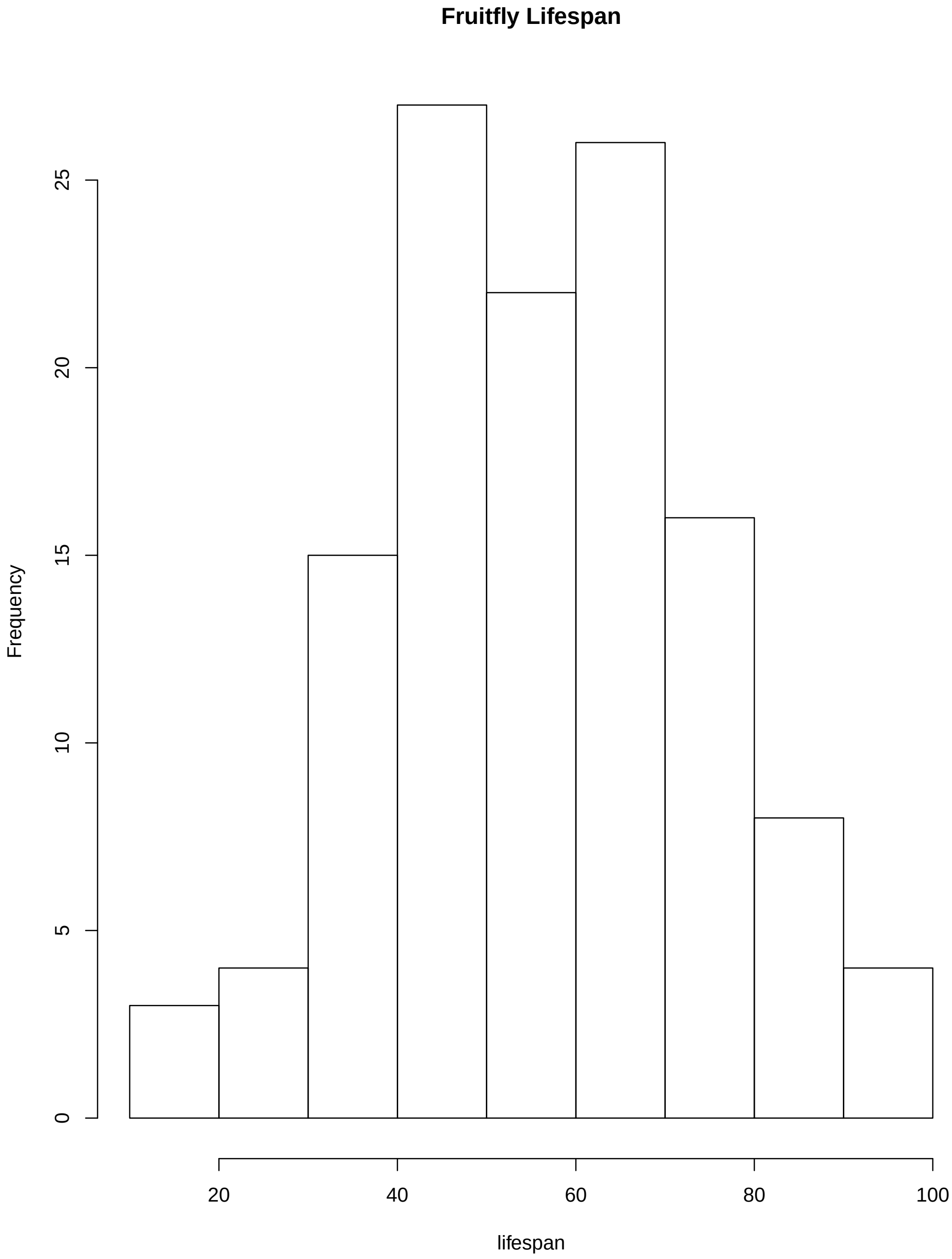
`No`	serial number (1-25) within each group of 25
`type`	Type of experimental assignment
	1 = no females
	2 = 1 newly pregnant female
	3 = 8 newly pregnant females
	4 = 1 virgin female
	5 = 8 virgin females
`lifespan`	lifespan (days)
`thorax`	length of thorax (mm)
`sleep`	percentage of each day spent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

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

```r
summary(fruitfly)
hist(fruitfly$lifespan, xlab="lifespan", main="Fruitfly Lifespan")
```

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

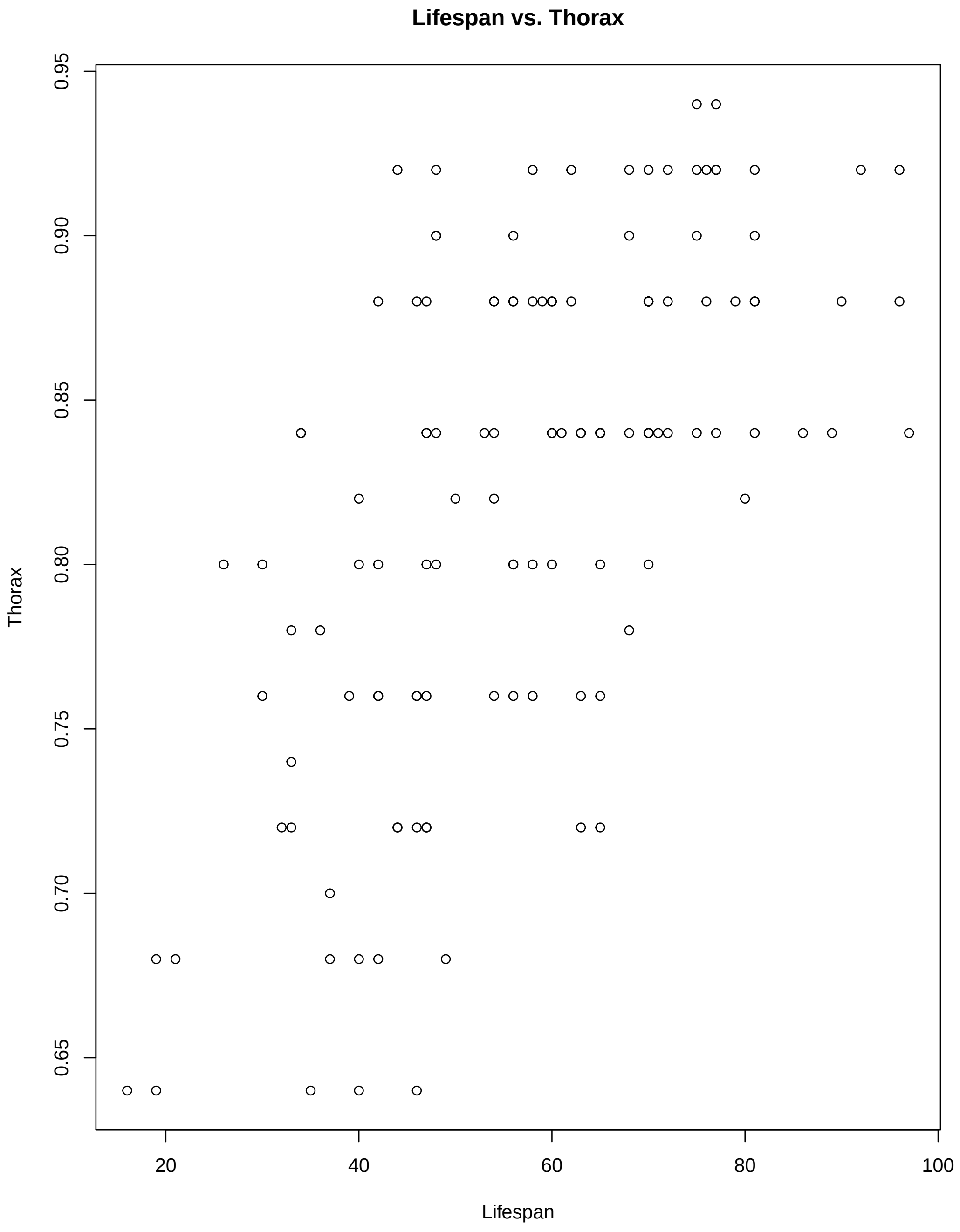
Figure 3: Distribution of Fruitfly Lifespan



The lifespan variable is normally distributed with the highest frequency between 40 and 80 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?

Figure 4: Scatterplot of Lifespan and Thorax length



It looks like there is a positive linear relationship between lifespan and thorax.

```
plot(fruitfly$lifespan, fruitfly$thorax, xlab="Lifespan", ylab="Thorax",
   main="Lifespan vs. Thorax")
cov(fruitfly$lifespan,fruitfly$thorax)
```

```
sd(fruitfly$lifespan)
sd(fruitfly$thorax)
0.8658645/(17.56389*0.07745367) # 0.636
```

The correlation coefficient is 0.636.

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

```
regress2 <- lm(fruitfly$lifespan~fruitfly$thorax)
regress2
```

Y = α + X

Y = -61.05 + 144.33X

For every 1 unit increase in lifespan, the length of the thorax will increase by 144.33 units.

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

```
anova(regress2)
15497/(15497+22756) #0.405
```

The r-squared = 0.405, meaning 40.5% of the variation in lifespan of a fruitfly is explained by the length of the thorax.

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

```
# Use the formula
summary(regress2)
144.33 + 1.657*15.77
144.33 - 1.657*15.77

# Use the function confint() in R
confint(regress2, level=.90)
```

90% CI: 118.19,170.47

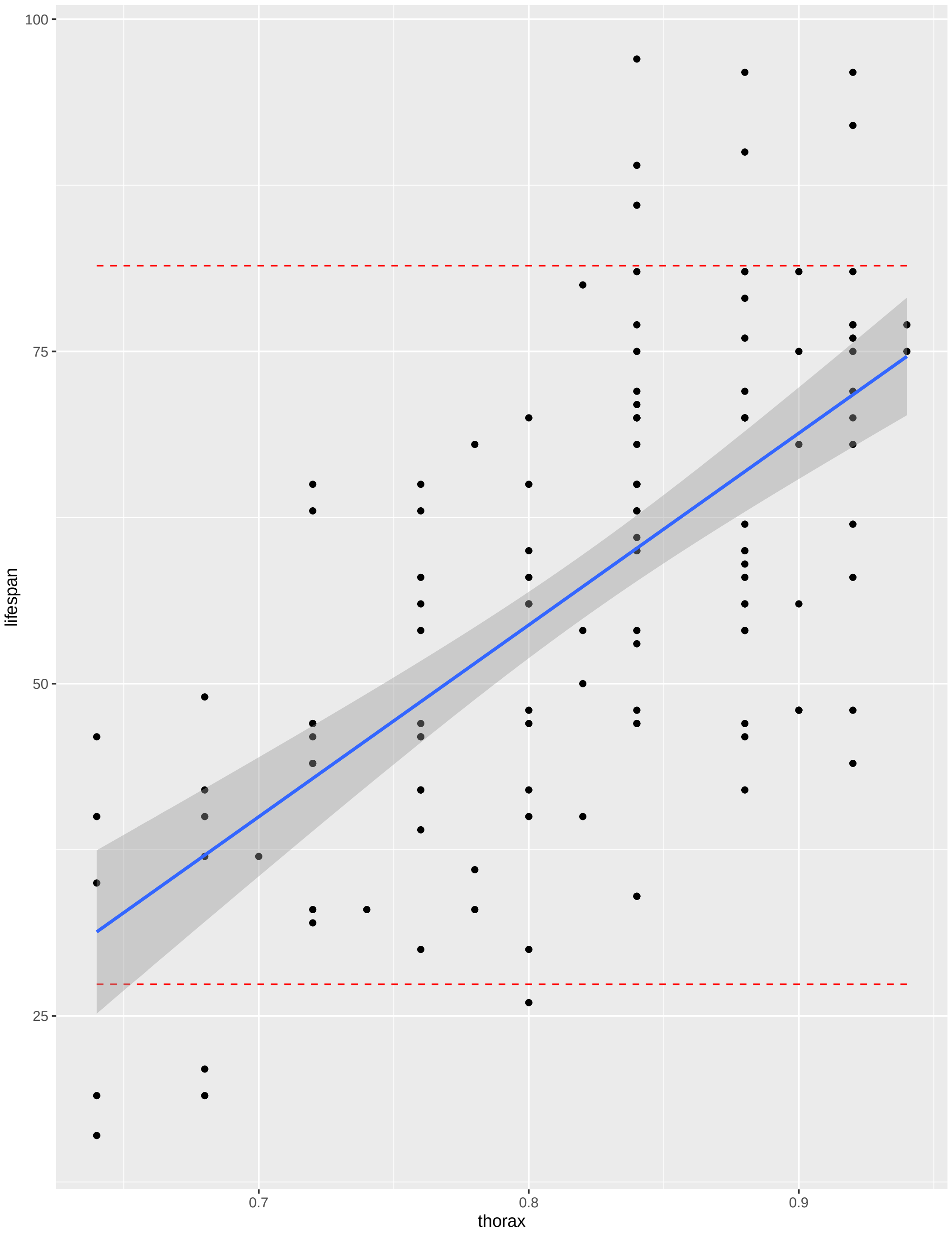
6. Use the `predict()` function in `R` to (1) predict an individual fruitfly's lifespan when `thorax`=0.8 and (2) the average `lifespan` of fruitflies when `thorax`=0.8 by the fitted model. This requires that you compute prediction and confidence intervals. What are the expected values of lifespan? What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

```r
# 6. Use the predict() function in R to
# (1) predict an individual fruitflys lifespan (Y) when thorax (X) = 0.8
new_fruitfly <- fruitfly; new_fruitfly$thorax <- .8
predict(lm(lifespan~thorax), newdata=new_fruitfly, se.fit=T)
predict(lm(lifespan~thorax), newdata=new_fruitfly, interval="prediction", level=0.95)
# 54.41478 years. Prediciton interval: (27.37, 81.45)

# (2) predict the average lifespan of fruitflies when thorax = 0.8 by the fitted model.
predict(lm(lifespan~thorax), newdata=new_fruitfly, interval="confidence", level=0.95)
# 54.41478 years. Confidence interval: (51.91, 56.91)
```

The expected values of lifespan are 54.41 years. Prediction interval: (27.37, 81.45) Confidence interval: (51.91, 56.91)

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.


100
75
50
25
lifespan
0.7
0.8
0.9
thorax