Problem Set 2

QTM 200: Applied Regression Analysis

Due: February 10, 2020

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 the course GitHub page in `.pdf` form.
- This problem set is due at the beginning of class on Monday, February 10, 2020. 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).

```r
X = matrix(c(14, 7, 6, 7, 7, 1), nrow = 2, ncol = 3) #Creating matrix
    from problem

colnames(X) <- c("Not Stopped", "Bribe requested", "Stopped/given warning
    ")
rownames(X) <- c("Upper Class", "Lower Class")
X #Creation of Matrix

no_stop_sum <- sum(X[,1])
no_stop_sum #21

bribe_sum <- sum(X[,2])
bribe_sum #13

stopped_sum <- sum(X[,3])
stopped_sum #8

#getting total for each column

upper_sum <- sum(X[1,])
upper_sum #27

lower_sum <- sum(X[2,])
lower_sum #15

#Getting total for each row
```

```r
sum(upper_sum, lower_sum) #Matrix total is 42

#now remaking matrix, to have column and row totals for observed values
new_x = matrix(c(14, 7, 21, 6, 7, 13, 7, 1, 8, 27, 15, 42), nrow = 3,
    ncol = 4)

colnames(new_x) <- c("Not Stopped", "Bribe requested", "Stopped/given
    warning", "Column Total")
rownames(new_x) <- c("Upper Class", "Lower Class", "Row Total")
new_x #Creation of Matrix

#to calculate expected values for chi squared test = (row total/grand sum
    ) * column total
(matrix(new_x[3,1]) / matrix(new_x[3,4])) * matrix(new_x[1,4]) #13.5
(matrix(new_x[3,1]) / matrix(new_x[3,4])) * matrix(new_x[2,4]) #7.5
(matrix(new_x[3,2]) / matrix(new_x[3,4])) * matrix(new_x[1,4]) #8.36
(matrix(new_x[3,2]) / matrix(new_x[3,4])) * matrix(new_x[2,4]) #4.64
(matrix(new_x[3,3]) / matrix(new_x[3,4])) * matrix(new_x[1,4]) #5.14
(matrix(new_x[3,3]) / matrix(new_x[3,4])) * matrix(new_x[2,4]) #2.86

#creating matrix of expected values
expected_values = matrix(c(13.5, 7.5, 8.36, 4.64, 5.14, 2.86), nrow = 2,
    ncol = 3)
expected_values

#calculating chi squared directly
sum(((X - expected_values)^2) / expected_values) #Chi squared values is
    3.801141

#Using chi squared function in R to test if values are close, which they
    are with the function outputting 3.7912
chisq.test(X)
```

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

```r
#calculating p-value with chi squared value = 3.801141, df= (2-1)(3-1),
    and lower tail = false, as we only want the upper end due to
    distribution type
pchisq(3.801141, 2, lower.tail = F)
#p-value = 0.1495 meaning we fail to reject the null, that officers would
     not solicit a bribe depending on the class
```

[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.

```
#calculating standardized residuals and want to see the matrixes again
new_x
expected_values

A_one = ((X[1,1] - expected_values[1,1])/ (sqrt(expected_values[1,1])*
    (1-(27/42))* (1-(21/42))))
A_one
A_two = ((X[2,1] - expected_values[2,1])/ (sqrt(expected_values[2,1])*
    (1-(15/42))* (1-(21/42))))
A_two

B_one = ((X[1,2] - expected_values[1,2])/ (sqrt(expected_values[1,2])*
    (1-(27/42))* (1-(13/42))))
B_one
B_two = ((X[2,2] - expected_values[2,2])/ (sqrt(expected_values[2,2])*
    (1-(15/42))* (1-(13/42))))
B_two

C_one = ((X[1,3] - expected_values[1,3])/ (sqrt(expected_values[1,3])*
    (1-(27/42))* (1-(8/42))))
C_one
C_two = ((X[2,3] - expected_values[2,3])/ (sqrt(expected_values[2,3])*
    (1-(15/42))* (1-(8/42))))
C_two

#creating matrix for standardized residuals
standardized_matrix = matrix(c(0.762, -0.568, -3.310, 2.468, 2.838,
    -2.113), nrow=2, ncol=3)

colnames(standardized_matrix) <- c("Not Stopped", "Bribe requested", "
    Stopped/given warning")
rownames(standardized_matrix) <- c("Upper Class", "Lower Class")
standardized_matrix #Creation of Matrix
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class			
Lower class			

(d) How might the standardized residuals help you interpret the results? The standard residual data may help with the interpretation of the results because they show how far off each data point is from a normal projected value. Simply, a the standardized residuals illustrates the difference between the observed and expected results, and show which values are contributing most in either more/less bias or no bias ways.

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).

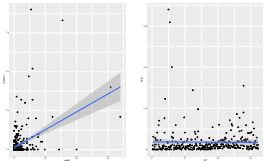
Name	Description
`GP`	An identifier for the Gram Panchayat (GP)
`village`	identifier for each village
`reserved`	binary variable indicating whether the GP was reserved for women leaders or not
`female`	binary variable indicating whether the GP had a female leader or not
`irrigation`	variable measuring 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.

(a) State a null and alternative (two-tailed) hypothesis. Null Hypothesis: There is no effect from the reservation policy on the number of new or repaired drinking water facilities in the villages Alternative Hypothesis: There is an effect from the reservation policy on the number of new or repaired drinking water facilities in the villages

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

```
ggplot(aes(water, irrigation), data = economics_ps) +
  geom_point() +
  geom_smooth(method = 'lm')

#can also plot against the identifier BP to see in each location how many
    water sites there are
ggplot(aes(GP, water), data = economics_ps) +
  geom_point() +
  geom_smooth(method = 'lm')
```



There seems to be very little change in the water situation as a whole in all GP locations with a near-zero slope, indicating a lack of change per GP. With irrigation however, as water increased, irrigation seemed to also increase. The difficult aspect of this interpretation however, is the lack of data points at the larger end of the scope.

(c) Interpret the coefficient estimate for reservation policy.

```
c(round(mean(economics_ps$water), 2), round(sd(economics_ps$water),2)) #x
    bar which is the calculated mean is 17.84, while the standard
    deviation is 33.68.

standardized_water <- (((economics_ps$water - (mean(economics_ps$water)))
    /sd(economics_ps$water)))
round(standardized_water, 2) #calculated standardized difference from the
    observed valued of lifespans

(1/(322-1)) * (sum(round(standardized_water, 2)))
```

The correlation coefficient is 0.00081. Meaning, the relationship from water as a whole in this data set, only has a relationship that is positive and extremely weak.

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

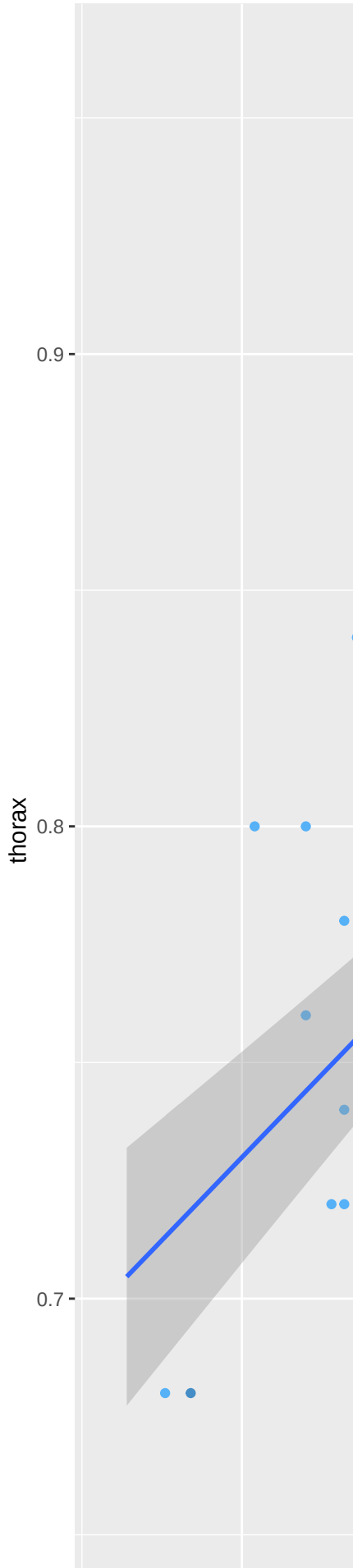
1. Import the data set and obtain summary statistiscs and examine the distribution of the

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

Fruitfly Lifespan
Lifespan (Days)
100
80
60
40
20
10

```r
library(tidyverse)
library(dplyr) #inputing library for filter function
fruitfly_type1 <- filter(fruitfly, type == 1)
fruitfly_type1 #isolating within the fruitfly dataset, each type

fruitfly_type2 <- filter(fruitfly, type == 2)
fruitfly_type2

fruitfly_type3 <- filter(fruitfly, type == 3)
fruitfly_type3

fruitfly_type4 <- filter(fruitfly, type == 4)
fruitfly_type4

fruitfly_type5 <- filter(fruitfly, type == 5)
fruitfly_type5
```

2. Plot `lifespan` vs `thorax`. Does it look like there is a linear relationship? Provide the


0.9
0.8
0.7
thorax

There does seem to be a positive, linear, relationship between lifespan and thorax length.

```
fruitfly
c(round(mean(fruitfly$lifespan), 2), round(sd(fruitfly$lifespan),2)) #x bar which is the calculated mean is 57.44, while the standard deviation is 17.56.

standardized_fly <- (((fruitfly$lifespan - (mean(fruitfly$lifespan)))/sd(fruitfly$lifespan)))
round(standardized_fly, 2) #calculated standardized difference from the observed valued of lifespans

standardized_thorax <- (((fruitfly$thorax - (mean(fruitfly$thorax)))/sd(fruitfly$thorax)))
round(standardized_thorax, 2) #calculated standardized difference from the observed valued of thorax length

standardized_fly_thorax <- standardized_fly * standardized_thorax #multiplying each standardarized values to then calculate correlation coefficient

(1/(125-1))*sum(round(standardized_fly_thorax, 2)) #the correlation coefficient is .636 which is a decently strong, and positive relationship that is explained between lifespan and thorax length.
```

0.70
0.75
0.80
0.85
0.90
y

The y-intercept is .66 mm of the thorax length when lifespan is theoretically 0 while the slope of the line is .002, as for every day gained in lifespan, the thorax length increases by .002 mm.

```
summary(lm(fruitfly$thorax ~ fruitfly$lifespan))

y <- .6597 + 0.0028*(fruitfly$lifespan) #creating the basic slope-intercept line with the gained intercept and slope from the summary.
plot(fruitfly$lifespan, y , type = "l") #creation of basic line graph to illustrate the intercept and slope of the relationship between lifespan and thorax length
```

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

```
t.test(fruitfly$lifespan, fruitfly$thorax, alternative = "two.sided", var.equal = FALSE)
summary(lm(fruitfly$thorax ~ fruitfly$lifespan))

thorax_lifespan <- lm(fruitfly$thorax ~ fruitfly$lifespan) #creation of one variable under the line relationship of thorax and lifespan
sigma(thorax_lifespan) #finding sigma of this relationship to be 0.06

beta_se <- sigma(thorax_lifespan)/sqrt(sum((fruitfly$lifespan - mean(fruitfly$lifespan))^2))
beta_se #finding the standard error of the lifespan quantity to be .0003

2*pt((.0028 - 0)/beta_se,dim(fruitfly)[1]-2, lower.tail = F) #p-value for beta-1 is 1.69e-15, which is very similar to checking through summary function
summary(thorax_lifespan)
```

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

   - Use the formula for typical confidence intervals to find the 90% confidence interval around the point estimate.

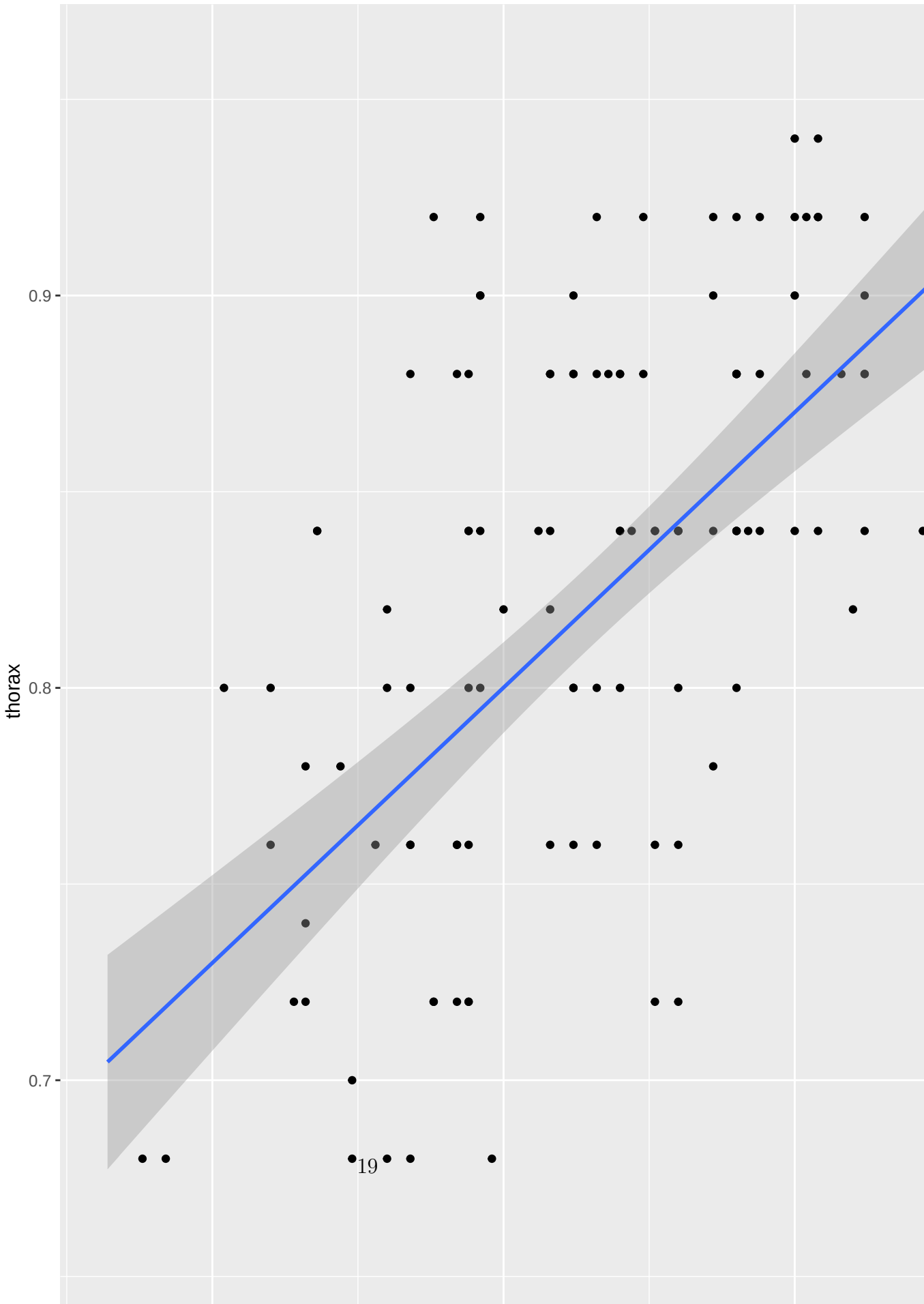
   - Now, try using the function `confint()` in `R`.

```
summary(lm(fruitfly$thorax ~ fruitfly$lifespan))

b1 <- .0028 #identification of the slope beta1
se <- .0003 #standard error of the graph between the relationship of
    thorax and lifespan
n <- 125 #sample size

t_test <- abs(qt(.1/2, df= n-2)) #t test that is two tailed
t_test

left <- b1 - se*t_test #obtaining the values of left and right
left
right <- b1 + se*t_test
right

thorax_lifespan <- lm(fruitfly$thorax ~ fruitfly$lifespan)
confint(thorax_lifespan, level = .9) #double checking the confidence
    interval which we obtain to be .0023 to .0033
```

4. 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?

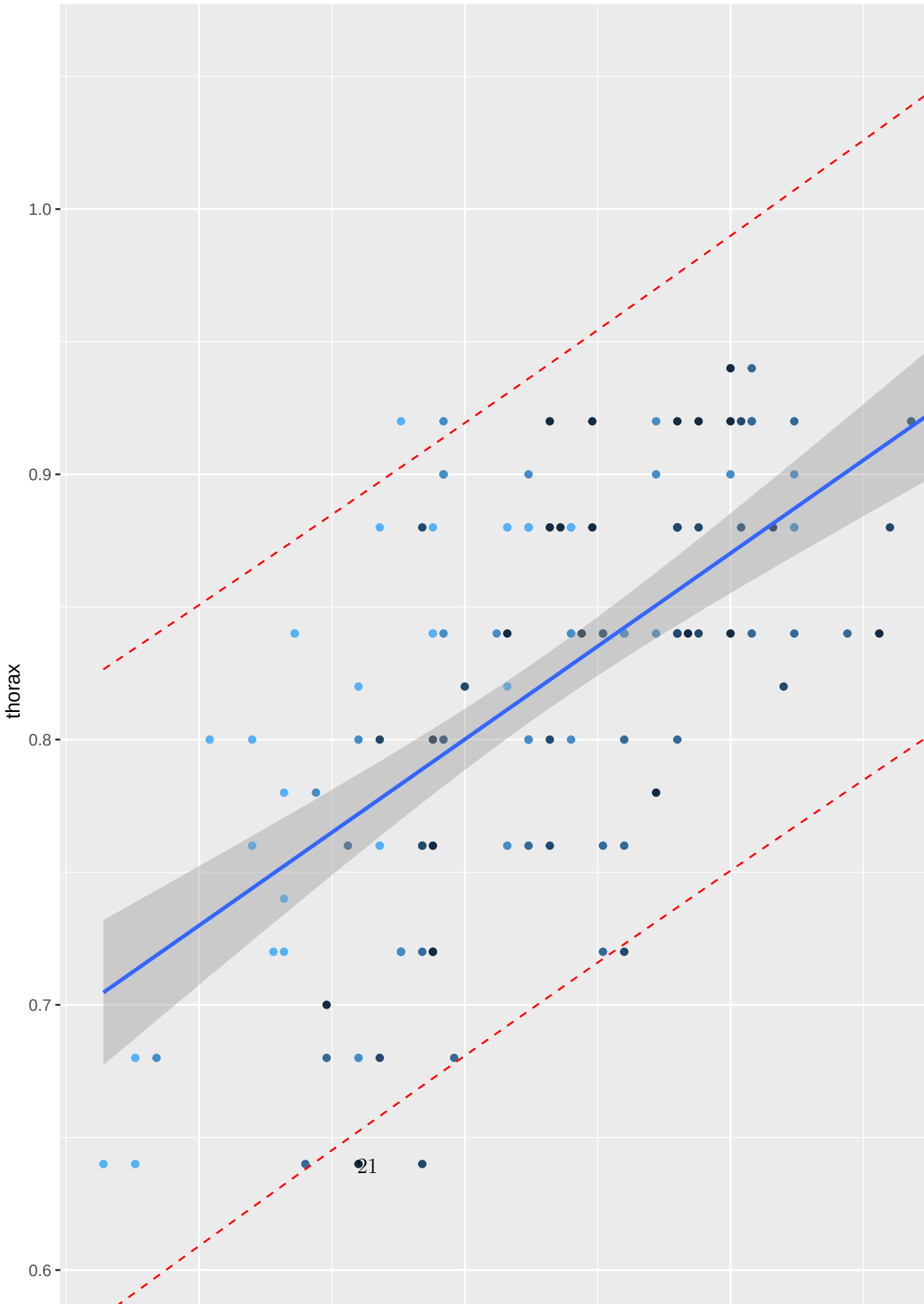
```
summary(lm(fruitfly$thorax ~ fruitfly$lifespan))
ggplot(fruitfly, aes(lifespan, thorax)) +
```

```
  geom_point() +
  stat_smooth(method = lm) #creating variable for graph with the grey
    color indicating the 95% confidence interval range from the regression
     line, whereas the red lines are the prediction lines, all under 95%.

new_fruitfly <- fruitfly
confidence <- as.data.frame(predict(thorax_lifespan, newdata = new_
    fruitfly, interval = "confidence")) #illustrates that on average per
    day in lifespan, the average thorax length ranges
mean(confidence$upr) #mean upper value in confidence interval is .8354
    for thorax length
mean(confidence$lwr) #mean lower value in confidence interval is .8065
    for thorax length

predict_data <- as.data.frame(predict(thorax_lifespan, newdata = new_
    fruitfly, interval = "prediction")) #illustrates that on average per
    day in lifespan, the average thorax length ranges
mean(predict_data$fit) #mean fitted value for predicted thorax length is
    .821
```


thorax
0.9
0.8
0.7

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

```
new_graph <- ggplot(aes(lifespan, thorax), data = fruitfly) +
  geom_point(aes(color = type))+
  geom_smooth(method = 'lm')

new_graph + geom_line(aes(y = predict_data $lwr), color = "red", linetype
    = "dashed")+
  geom_line(aes(y = predict_data$upr), color = "red", linetype = "dashed"
    )
```


thorax
1.0
0.9
0.8
0.7
0.6

Fitted values are existing on the blue straight regression line across the graph. The grey areas indicate the confidence interval of the fitted values along the regression line; the dashed, red lines, illustrate the prediction of 95