Question 1 (40 points): Political Science

(a) Calculate the χ^2 test statistic by hand (even better if you can do ”by hand” in R).

```
getwd()
setwd(”//Users/lucykinnear/Desktop/ASDS/Statistics”)
mydata = read.csv(”Users/lucykinnear/Desktop/ASDS/Statistics\\02_Stats_Data_Orig.csv”)

#Assign Observed Frequencies
Fo_1 <- 14
Fo_2 <- 6
Fo_3 <- 7
Fo_4 <- 7
Fo_5 <- 7
Fo_6 <- 1

#Assign Expected Frequencies
Fe_1 <- ((27/42) * 21)
Fe_2 <- ((27/42) * 13)
Fe_3 <- ((27/42) * 8)
Fe_4 <- ((15/42) * 21)
Fe_5 <- ((15/42) * 13)
Fe_6 <- ((15/42) * 8)

#Calculate chisq
chisq <- (
  ((Fo_1 - Fe_1)^2/Fe_1) + ((Fo_2 - Fe_2)^2/Fe_2) +
    ((Fo_3 - Fe_3)^2/Fe_3) +  ((Fo_4 - Fe_4)^2/Fe_4) +
    ((Fo_5 - Fe_5)^2/Fe_5) + ((Fo_6 - Fe_6)^2/Fe_6)
)
chisq
```

$\chi^2 = 3.7911...$

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

```
# df = (rows -1)(columns-1) = 1(2) = 2
pchisq <- pchisq(3.791168, df = 2, lower.tail = FALSE)
pchisq
```

pchisq = 0.15023... The p-value is bigger than alpha, therefore we cannot reject the null hypothesis. We should not be surprised to observe these results. Class did not meaningfully impact whether participants were asked for bribes.

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

```r
table <- matrix(c(14, 6, 7, 7, 7, 1), ncol=3, byrow=TRUE)
colnames(table) <- c('Not Stopped', 'Bribed', 'Warned')
rownames(table) <- c('Upper Class', 'Lower Class')
table <- as.table(table)
chisq_test <- chisq.test(table)
chisq_test
chisq_test$stdres

              Not Stopped     Bribed      Warned
Upper Class    0.3220306 -1.6419565   1.5230259
Lower Class   -0.3220306  1.6419565  -1.5230
```

(d) How might the standardized residuals help you interpret the results?
We can compare the standardized residuals in the table to see which category of variables have the largest difference between the expected and the actual counts relative to size. Standardized residuals are the raw residuals divided by the square root of the expected counts. Positive standardized residuals indicate that there were more occurrences of this outcome than expected. The negative standardized residuals indicate that there were less occurrences of this outcome than expected. Small standardised residuals tell us that the prediction line is a good fit for the data.

Question 2 (20 points): Economics

(a) State a null and alternative (two-tailed) hypothesis.
Null hypothesis: whether or not there is a reservation policy does not impact the number of new or repaired drinking water facilities in the villages. Alternative hypothesis: whether or not there is a reservation policy does impact the number of new or repaired drinking water facilities in the villages. Two-Tailed Tests:

```
z   /2 =   qnorm  (1          /2)
```

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

```
Q2_data <- read.csv(file="~/Desktop/women.csv")
Q2_data_of_interest = lm(Q2_data$water ~ Q2_data$reserved)
Q2_data_of_interest
summary(Q2_data_of_interest)
Call:
lm(formula = Q2_data$water ~ Q2_data$reserved)

Coefficients:
     (Intercept)  Q2_data$reserved
          14.738             9.252

> summary(Q2_data_of_interest)

Call:
lm(formula = Q2_data$water ~ Q2_data$reserved)

Residuals:
    Min      1Q   Median      3Q     Max
-23.991 -14.738  -7.865   2.262 316.009

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        14.738      2.286   6.446 4.22e-10 ***
Q2_data$reserved    9.252      3.948   2.344   0.0197 *
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1
          1

Residual standard error: 33.45 on 320 degrees of freedom
Multiple R-squared:  0.01688, Adjusted R-squared:  0.0138
F-statistic: 5.493 on 1 and 320 DF,  p-value: 0.0197
```

Our p-value is 0.0197. We can reject the null, if alpha is 0.05, but not if alpha is 0.01 ! This means, we have evidence to support the alternative hypothesis. Although, the determinant coefficient (multiple r2) is quite low.

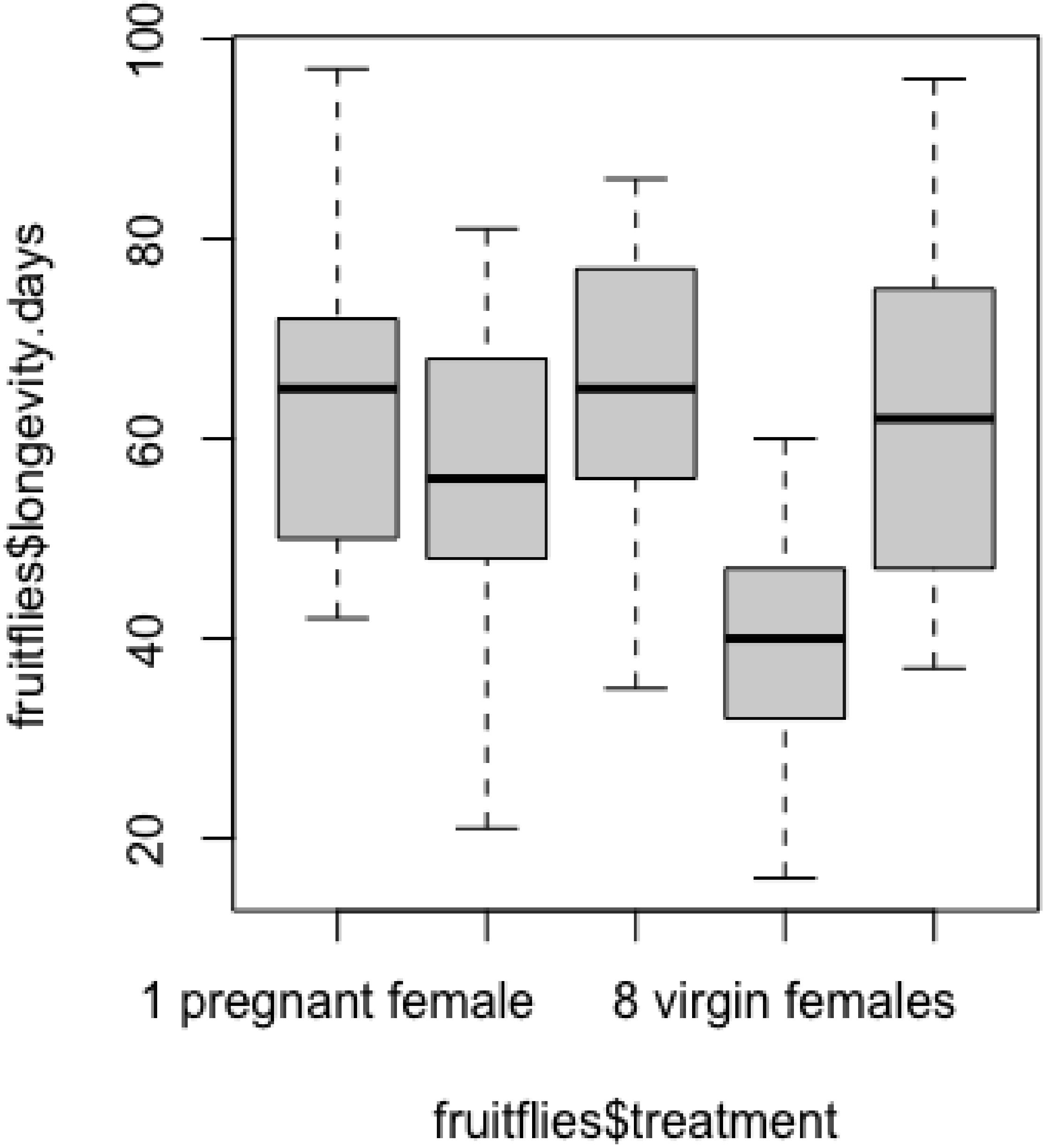
(c) Interpret the coefficient estimate for reservation policy.
The coefficient of is 9.252 for reservation policy. Changes in the independent variable are associated with changes in the dependent variable at the population level. This variable is statistically significant and probably a worthwhile addition to your regression model. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant. For our data, if women are reserved places on the council, water outcomes improve.

Question 3 (40 points): Biology

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

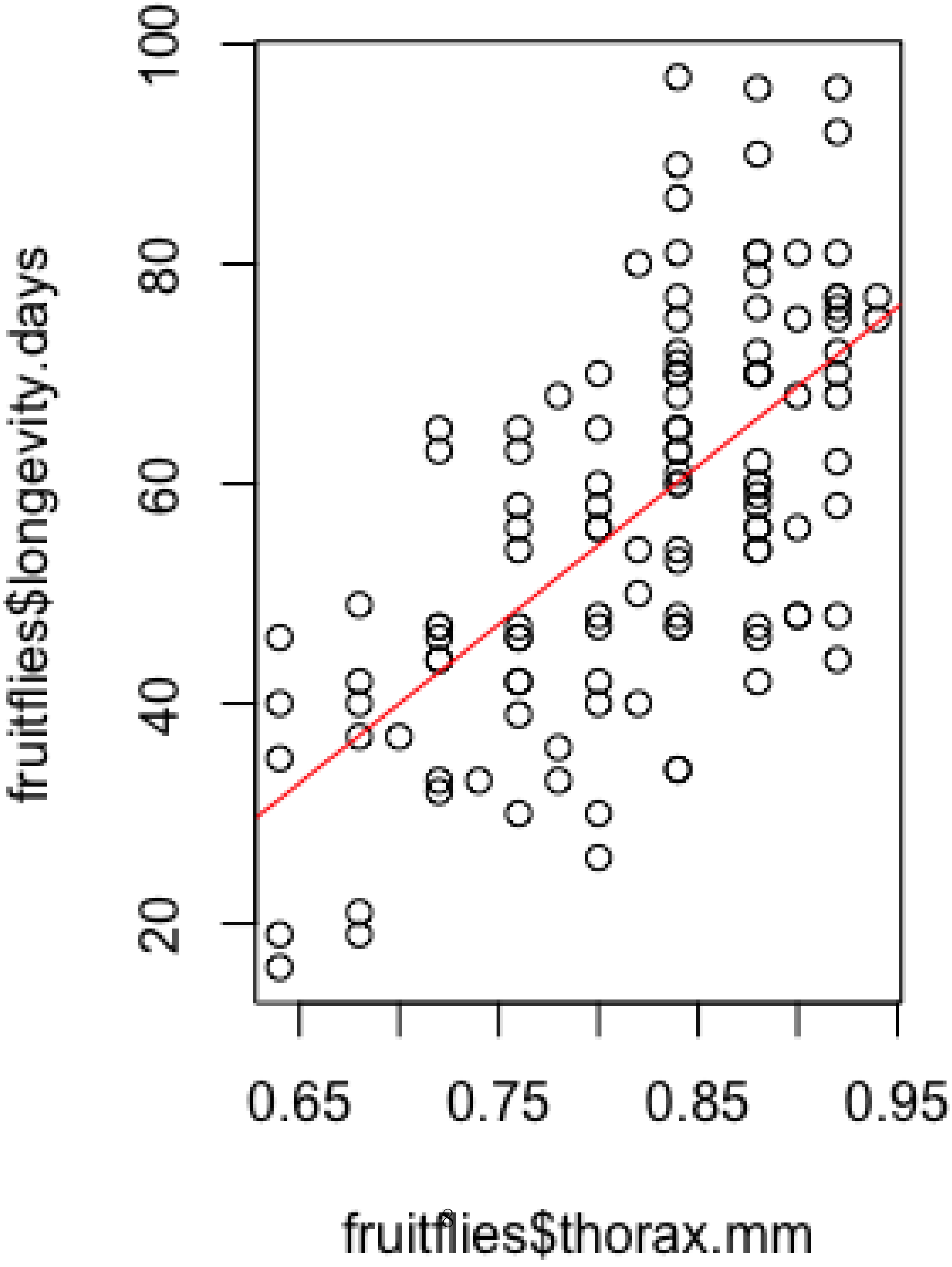
```
fruitflies <- read.csv(file=”~/Desktop/fruitflies.csv”)
summary(fruitflies)
hist(fruitflies$longevity.days)
means
 1 pregnant female    1 virgin female 8 pregnant females    8 virgin
   females    no females added
              64.80                56.76                63.36
   38.72                63.56

boxplot(fruitflies$longevity.days ~ fruitflies$treatment)
```

Fruitflies lived a minimum of 16 days and a maximum of 97 days, with a median life span of 58 days. Noteably, this is close to the average lifespan of a fruitfly, 57.4 days - potentially indicating a normal distribution. Groups with 8 virgin females had shortest average lifespan, groups with 1 pregnant female the longest.


fruitflies$longevity.days
100
80
60
40
20
1 pregnant female
8 virgin females
fruitflies$treatment

02_fruitflies_his.png

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?


fruitflies$longevity.days
20
40
60
80
100
0.65
0.75
0.85
0.95
fruitflies$thorax.mm

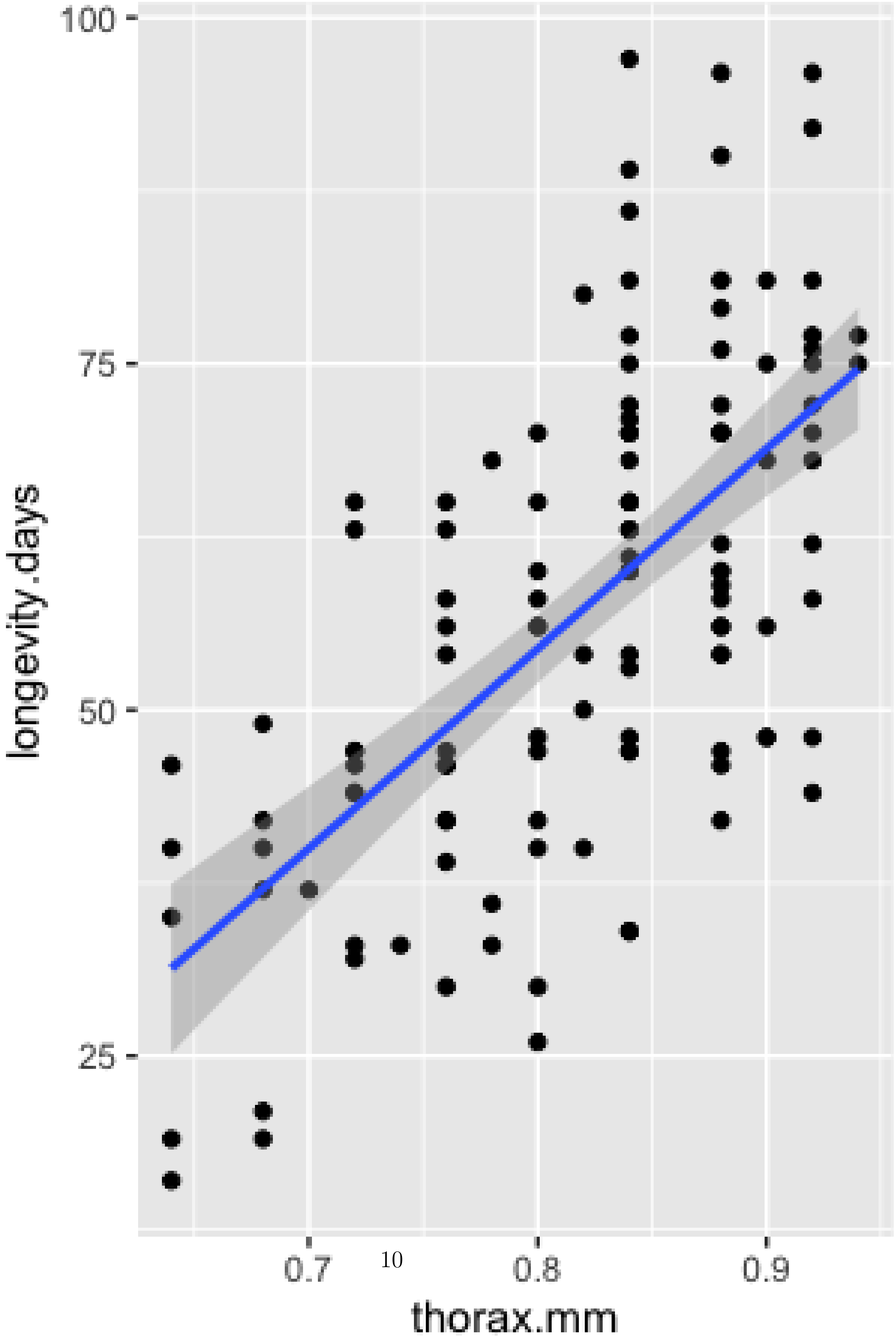
There does appear to be a positive linear relationship.

```
cor(fruitflies$longevity.days, fruitflies$thorax.mm)
```

The correlation coefficient is 0.63640...

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

```
q3_lifespan_thorax = lm(fruitflies$longevity.days ~ fruitflies$thorax.mm)
summary(q3_lifespan_thorax)
plot(fruitflies$longevity.days ~ fruitflies$thorax.mm)
abline(lm(fruitflies$longevity.days ~ fruitflies$thorax.mm), col = ”red”)
library(ggplot2)
ggplot(aes(longevity.days, thorax.mm), data = fruitflies) +
  geom_point(position = ”jitter”)
abline(lm(fruitflies$longevity.days ~ fruitflies$thorax.mm), col = ”red”)

ggplot(aes(thorax.mm, longevity.days), data = fruitflies) +
  geom_point() +
  geom_smooth(method = ”lm”, formula = y ~ x)
```


100
75
50
25
longevity.days
0.7
0.8
0.9
thorax.mm

The thorax.mm coefficient is 144.33 with tiny p-value.

4. Test for a significant linear relationship between `lifespan` and `thorax`. Provide and interpret your results of your test.
   Extract p-values Null: slope of regression is 0 From summary function:

```
  summary(q3_lifespan_thorax)

Call:
lm(formula = fruitflies$longevity.days ~ fruitflies$thorax.mm)

Residuals:
    Min      1Q  Median      3Q     Max
-28.415  -9.961   1.132   9.265  36.812

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            -61.05      13.00  -4.695  7.0e-06 ***
fruitflies$thorax.mm   144.33      15.77   9.152  1.5e-15 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1
       1

Residual standard error: 13.6 on 123 degrees of freedom
Multiple R-squared:  0.4051,  Adjusted R-squared:  0.4003
F-statistic: 83.76 on 1 and 123 DF,  p-value: 1.497e-15
```

   p-value way below alpha of 0.05 or 0.01, can reject the null - there is a significant linear relationship here.

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

   Using the formula of confidence interval did not allow me to find the confidence interval because argument was not numeric or logical. Using the function `confint()` in `R` was fruitful.

```
    confint(q3_lifespan_thorax, level = 0.9)
                           5 %      95 %
(Intercept)          -82.60361  -39.4998
fruitflies$thorax.mm 118.19616  170.4700
```

   The 90% confidence interval is 118.2 - 170.5.

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. predict

```
new.data = data.frame(thorax.mm = 0.8)
predict.lifespan = predict(lm(fruitflies$longevity.days ~
  fruitflies$thorax.mm), newdata = new.data)
predict.lifespan

```

   Lifespan prediction is 54.41.

   2. fitted

```
fitted = fitted(lm(fruitflies$longevity.days ~ fruitflies$thorax.mm))
data = cbind(fruitflies$thorax.mm, unname(fitted))
colnames(data) = c(”Thorax in mm”, ”FittedModel”)
data = as.data.frame(data)
data.08 = data[which(data$thorax.mm == 0.8]

mean(data.08$FittedModel)


```

   The mean is 54.41 - i.e. the same as the predicted from part 1.

```
predict(lm(fruitflies$longevity.days ~ fruitflies$thorax.mm), newdata =
   new.data, interval = ’confidence’)

predict(lm(fruitflies$longevity.days ~ fruitflies$thorax.mm), newdata =
   new.data, interval = ’prediction’)

```

   fit = 54.41, lower = 59.92, upper = 56.91
   fit = 54.41, lower = 27.38, upper = 81.45
   Prediction interval is wider because it accounts for variability and error of the estimates. Average lifespan is the same for both, and expected lifespan in prediction is wider, between 27 and 8 days.

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 have no clue how to do this !