Problem Set 2: QTM 200: Applied Regression Analysis

Farris Sabir

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). A matrix was first created in R to represent the above table.

```
# create a matrix of the data from which the chi-squared test can be conducted
bribematrix <- matrix(c(14, 6, 7, 7, 7, 1), byrow=T, nrow=2)
```

Then, a new matrix was filled given expected values derived from the sum of each row in the original matrix and the sume of each column. Each expected value was calculated as the product of the sum of its row's values and the sum of its column's values divided by the total number of observations.

```
# create a new matrix of expected values using the sums of each row and sums of each column
rowsum <- rowSums(bribematrix)
columnsum <- colSums(bribematrix)
totalsum <- sum(rowsum)
expectedmatrix <- outer(rowsum, columnsum, "*") / totalsum
rownames(expectedmatrix) <- c("Upper class", "Lower class")
colnames(expectedmatrix) <- c("Not Stopped", "Bribe requested", "Stopped/given warning")
#some expected values are less than 5, meaning chi-squared test may not be appropriate
```

Note: because two of the six expected values are less than 5, a chi-squared test may not be appropriate to draw conclusions about this data, so this should be noted when interpreting the final results. Then, the differences between the matrix of observed values and the matrix of expected values were input to another matrix, and that matrix's sum was equivalent to the test statistic, which was 3.79.

```
matrixdifference <- (bribematrix - expectedmatrix)^2/(expectedmatrix)
teststatistic <- sum(matrixdifference)
```

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

[2]Remember frequency should be > 5 for all cells, but let's calculate the p-value here anyway.

The test-statistic of 3.79 can be used, as well as the degrees of freedom calculated from the number of rows and columns, to provide the p-value. The p-value of 0.1502 is greater than the alpha of 0.1, which leads to failing to reject the null hypothesis that the class of the employee and the response of the police are statistically independent. R's chisq.test function supported this lack of a statistically significant relationship or dependence.

```
# use the test-statistic and degrees of freedom to find the p-value
df <- (ncol(bribematrix)-1)*(nrow(bribematrix)-1) #df stands for degrees
    of freedom
pvalue <- pchisq(teststatistic, df = df, lower.tail=FALSE)
chisq.test(bribematrix)
# the p-value of 0.1502 is greater than the alpha of 0.1
# this leads us to fail to reject the null hypothesis that the class of
    the employee and the police's response are statistically independent
# the chisq.test function in R verified the results by hand
```

(c) Calculate the standardized residuals for each cell and put them in the table below. After creating an empty matrix similar to the original one to input the residual values, the standardized residual for each value of the original table was calculated and input. The coding and the eventual table of residuals are shown below:

```
# create an empty matrix to put in the residual values
residualmatrix <- matrix(data=NA, nrow = 2, ncol = 3)
rownames(residualmatrix) <- c("Upper class", "Lower class")
colnames(residualmatrix) <- c("Not Stopped", "Bribe requested", "Stopped/
    given warning")

# calculate the standardized residual for each value and place in the
    matrix
residualmatrix[1] <- (bribematrix[1]-expectedmatrix[1])/sqrt(
    expectedmatrix[1]*(1-columnsum[1]/totalsum)*(1-rowsum[1]/totalsum)) #
    upper class x not stopped
residualmatrix[2] <- (bribematrix[2]-expectedmatrix[2])/sqrt(
    expectedmatrix[2]*(1-columnsum[1]/totalsum)*(1-rowsum[2]/totalsum)) #
    lower class x not stopped
residualmatrix[3] <- (bribematrix[3]-expectedmatrix[3])/sqrt(
    expectedmatrix[3]*(1-columnsum[2]/totalsum)*(1-rowsum[1]/totalsum)) #
    upper class x bribe requested
residualmatrix[4] <- (bribematrix[4]-expectedmatrix[4])/sqrt(
    expectedmatrix[4]*(1-columnsum[2]/totalsum)*(1-rowsum[2]/totalsum)) #
    lower class x bribe request
residualmatrix[5] <- (bribematrix[5]-expectedmatrix[5])/sqrt(
    expectedmatrix[5]*(1-columnsum[3]/totalsum)*(1-rowsum[1]/totalsum)) #
    upper class x stopped/given warning
residualmatrix[6] <- (bribematrix[6]-expectedmatrix[6])/sqrt(
    expectedmatrix[6]*(1-columnsum[3]/totalsum)*(1-rowsum[2]/totalsum)) #
    lower class x stopped/given warning
residualmatrix
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.3220306	-1.641957	1.523026
Lower class	-0.3220306	1.641957	-1.523026

(d) How might the standardized residuals help you interpret the results? After the conclusion to fail to reject the null hypothesis that the class of employees and response of police was independent, the relationship between the two variables can be further explored. The test led to failing to reject the null hypothesis, which is consistent with residuals whose absolute values are not very large. The expected and observed values are not greatly different, in other words. The conditions where the employees were not stopped had the least deviation from independence whereas the toher conditions relatively experienced the most. However, the residuals were never different enough from each other to conclude any statistically significant dependence.

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.

The null hypothesis is that the slope of the linear approximation of the relationship between the average number of new and repaired drinking-water facilities in villages and the number of villages with the reservation policy implemented is equal to zero. Formally, $H_o : \beta_1 = 0$. The alternative hypothesis is that the slope of the linear approximation of the relationship between the average number of new and repaired drinking-water facilities in villages and the number of villages with the reservation policy implemented is unequal to zero. Formally, $H_a : \beta_1 \neq 0$.

(b) Run a bivariate regression to test this hypothesis in R (include your code!). First, the data needed to be input.

```
women <- read.csv("WomenasPolicyMakers.csv")
women <- women[c(1:322),]
```

Before conducting the bivariate regression, some assumptions need to be addressed. First, the data must be assumed to have been generated randomly, which is reasonable given the policy was randomly assigned to some villages versus others. Second, the observations have to be assumed to be independent, like one village's choice to change the number of facilities did not influence another's choice. Third, a linear relationship must be assumed between the variables of reservation policy and the number of water-drinking facilities in the villages. Finally, the population values of number of water-drinking facilities at both reserved and non-reserved villages must be assumed to follow a normal distribution with the same standard deviation for both reserved and non-reserved villages. Because the residuals center around zero, which can be shown in Figure 2 below, this last assumption can be reasonably met.

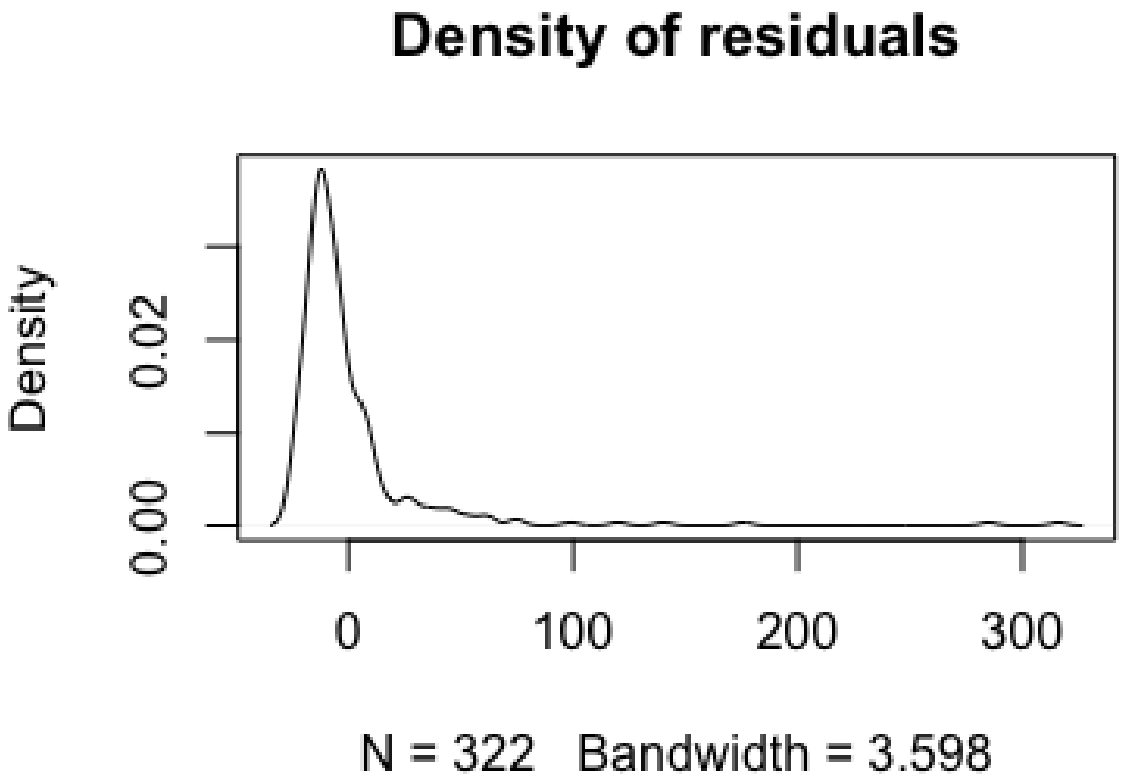
```
resid = women$water - predict(lm(water ~ reserved, data=women))
plot(density(resid), main="Density of residuals")
```

In seeking the $\hat{\beta}_1$, the means for the number of villages with the reservation policy and for the number of water-drinking facilities are calculated, and therefrom, $\hat{\beta}_1$ can be found using the equation $\hat{\beta}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$, which produced a $\hat{\beta}_1$ of 9.252.

```
water_hat <- mean(women$water)
reserved_hat <- mean(women$reserved)
beta_hat <- (sum((women$water - water_hat)*(women$reserved - reserved_hat
    )))/(sum((women$reserved - reserved_hat)^2))
beta_hat
```

Meanwhile, the $\hat{\beta}_0$ or $\hat{\alpha}$ is calculated using the mean number of water-drinking facilities, the mean number of villages with the reservation policy, and the $\hat{\beta}_1$, and it proves to be 14.738.

Figure 2: Density of Residuals




```
alpha_hat <- water_hat - reserved_hat*beta_hat
alpha_hat
```

The null hypothesis is tested for by calculating the standard deviation, standard error of the slope, the t-statistic, and the p-value in that order. Using the standard error, the t-statistic proved to be 2.344, and this helped lead to the resulting p-value of 0.0197. This was verified by R's linear model fitting function.

```
# testing the null hypothesis that beta = 0
stdv_water_hat <- sqrt(sum(resid^2)/(length(women$water)-2))
se_water_hat <- stdv_water_hat/sqrt(sum((women$reserved - reserved_hat)
    ^2))
tstat <- (beta_hat - 0)/se_water_hat
pval <- 2*pt(tstat, df = (length(women$water)), lower.tail=F)
pval
#p-value = 0.0197
```

```r

# checking with R's linear model fitting function
womenmodel <- lm(water ~ reserved, data = women)
summary(womenmodel)
```

(c) Interpret the coefficient estimate for reservation policy.

The coefficient estimate $\hat{\beta}_1$ of 9.252 for the slope is associated with a t-value of 2.344 and p-value of 0.0197, which with an alpha of 0.05, leads to rejecting the null hypothesis. Therefore, the reservation policy led to a significant change in the number of new or repaired drinking-water facilities since the reserve policy started. This can be interpreted so that a 1 unit increase in the number of villages with reservation policy leads to, on average, 9.252 more new or repaired drinking-water facilities. On the other hand, the $\hat{\beta}_0$ or $\hat{\alpha}$ of 14.738 can be interpreted so that, with no reservation policies being implemented in any village, the number of new or repaired drinking-water facilities is approximately 14.738.

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

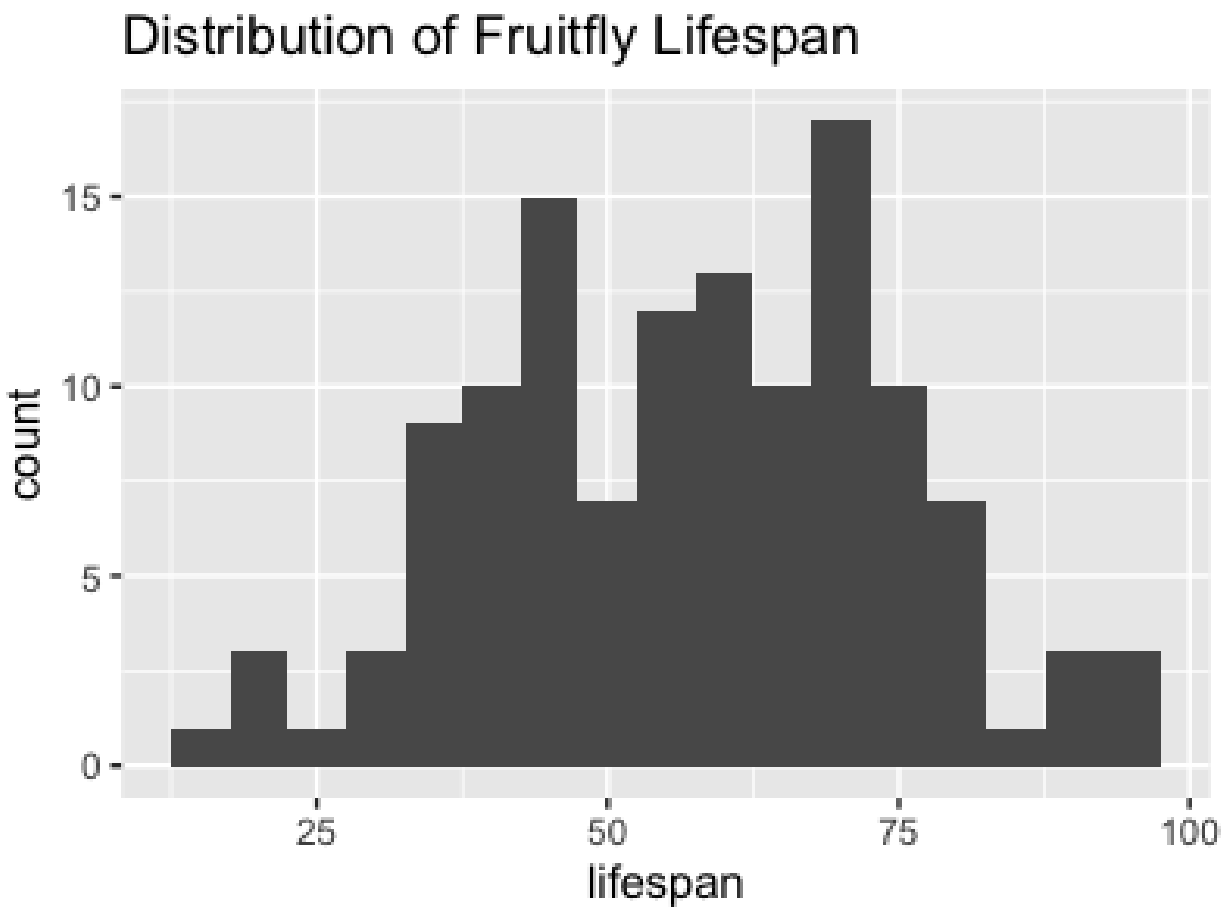
1. Import the data set and obtain summary statistiscs and examine the distribution of the overall lifespan of the fruitflies.

```
# importing the data set
fruitfly <- read.csv("fruitfly.csv")

# obtaining the summary statistics
summary(fruitfly)

# examining the distribution of the overall lifespan of the fruitflies
ggplot(fruitfly, aes(lifespan)) + geom_histogram(binwidth = 5) + ggtitle(
    "Distribution of Fruitfly Lifespan")
```

```
       No          type      lifespan        thorax          sleep
 Min.   : 1   Min.   :1   Min.   :16.00   Min.   :0.640   Min.   : 1.00
 1st Qu.: 7   1st Qu.:2   1st Qu.:46.00   1st Qu.:0.760   1st Qu.:13.00
 Median :13   Median :3   Median :58.00   Median :0.840   Median :20.00
 Mean   :13   Mean   :3   Mean   :57.44   Mean   :0.821   Mean   :23.46
 3rd Qu.:19   3rd Qu.:4   3rd Qu.:70.00   3rd Qu.:0.880   3rd Qu.:29.00
 Max.   :25   Max.   :5   Max.   :97.00   Max.   :0.940   Max.   :83.00
```

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

Figure 3: Density of the Lifespans of Fruitflies

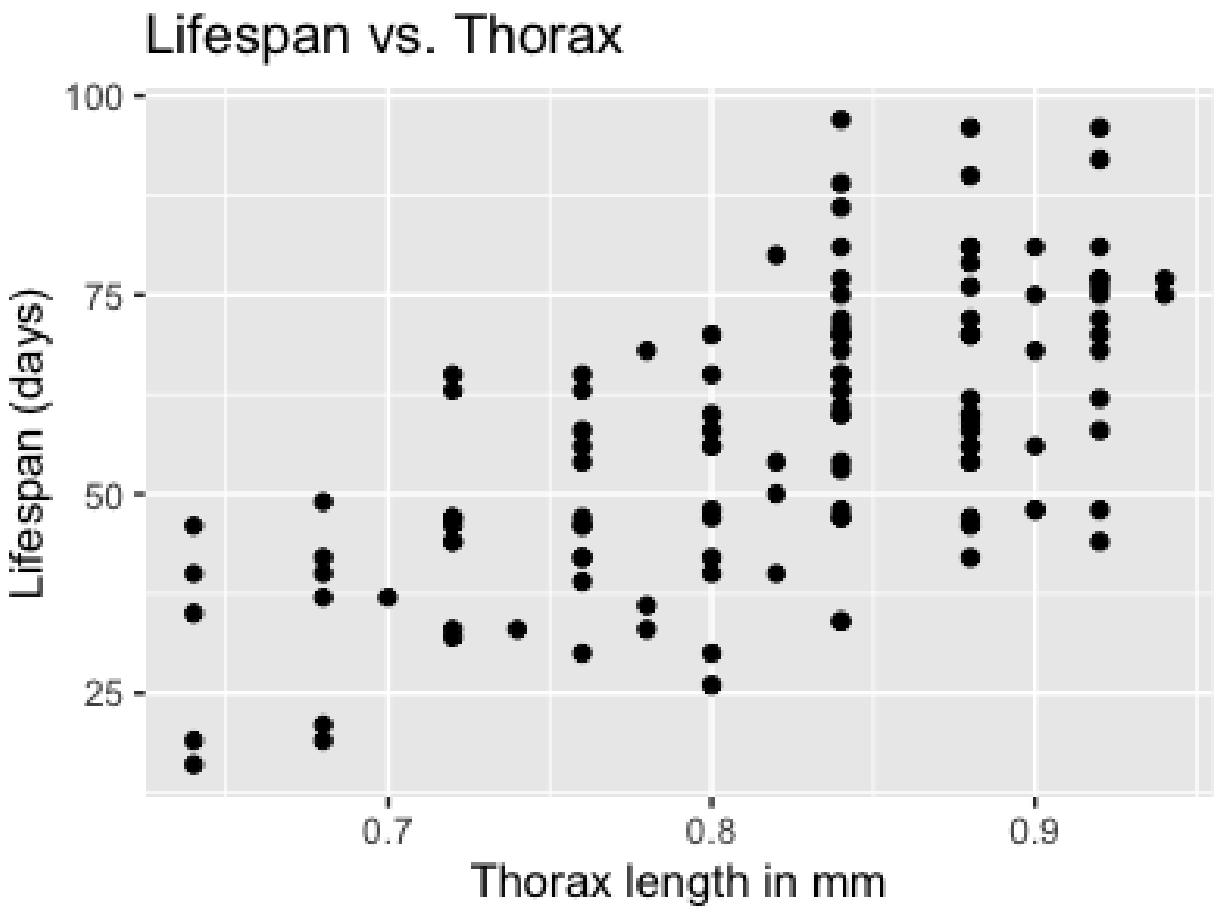


The data is imported in the above R code and the summary statistics are listed as well. In examining the distribution in Figure 3, the lifespans of fruitflies are approximately normally distributed in this sample.

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?

```
# plotting lifespan vs. thorax
ggplot(fruitfly, aes(x=thorax, y=lifespan)) + geom_point() + ggtitle("
    Lifespan vs. Thorax") + xlab("Thorax length in mm") + ylab("Lifespan (
    days)")
```

Figure 4: Fruitflies' Thorax Length vs. Lifespan



The above plot in Figure 4 looks as though it shows a linear relationship between thorax length and lifespan.

The correlation coefficient is calculated below using the mean of fruitfly lifespan and mean of fruitfly thorax length as well as their standard deviations. The exact formula is that the coefficient is equal to the covariance between thorax length and lifespan divided by the product of both the standard deviation in thorax length and the standard

deviation in lifespan. This proved to be 0.6365 and was verified by R's covariation and correlation functions.

```
# calculating the correlation coefficient
lifespan_bar <- mean(fruitfly$lifespan)
lifespan_sd <- sd(fruitfly$lifespan)
thorax_bar <- mean(fruitfly$thorax)
thorax_sd <- sd(fruitfly$thorax)
r <- (1/(length(fruitfly$thorax)-1))*sum(((fruitfly$lifespan - lifespan_
    bar)/lifespan_sd)*((fruitfly$thorax - thorax_bar)/thorax_sd))
r
#The correlation coefficient is 0.6365 between lifespan and thorax

# checking with R's covariation and correlation functions
cov(fruitfly[,c(4,3)])[1,2]/(sd(fruitfly[,3]))/(sd(fruitfly[,4]))
cor(fruitfly$thorax, fruitfly$lifespan, method = "pearson")
```

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

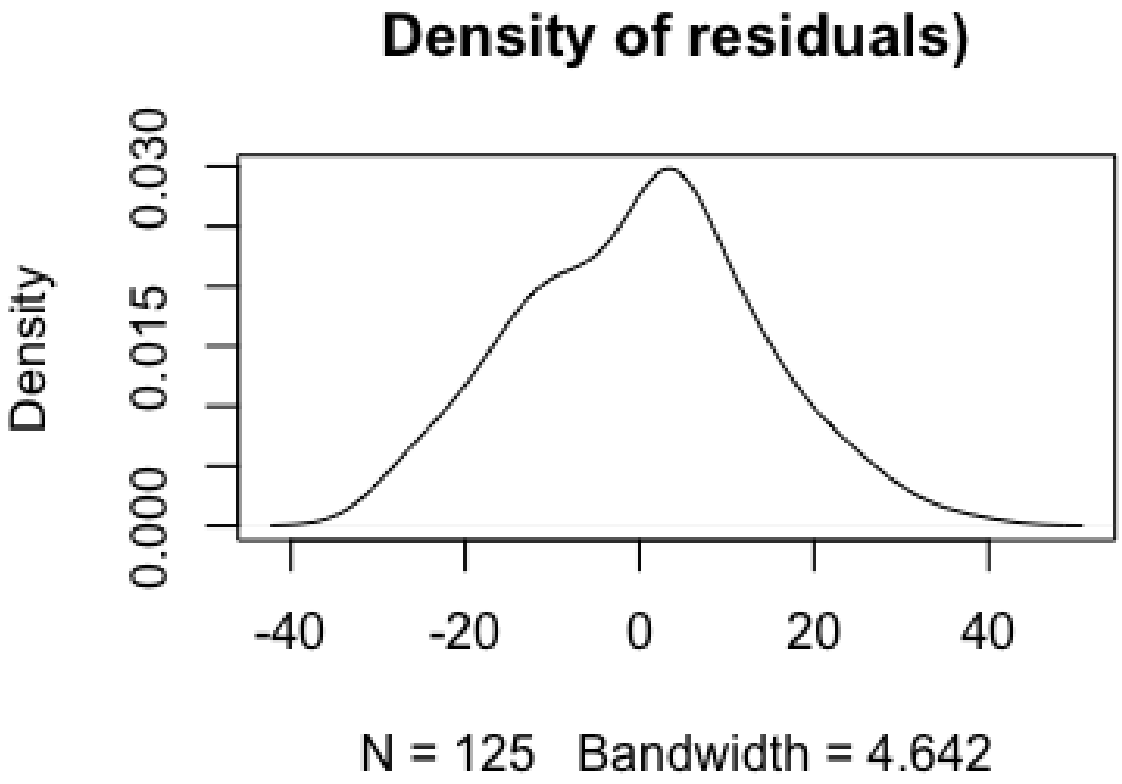
Before conducting a linear regression, some assumption must be considered. First, the data must be assumed to have been generated randomly, and the types at least were likely randomized. Second, the observations must be assumed to be independent, meaning the thorax length or lifespan of different flies did not influence each other. Third, the relationship between the variables thorax length and lifespan must be assumed to be independent. Finally, the population values of length of thorax at each value of lifespan must be assumed to follow a normal distribution with the same standard deivation for all values of lifespan. This is verified by the residual plot below in Figure 5.

```
residf = fruitfly$lifespan - predict(lm(lifespan~thorax, data=fruitfly))
plot(density(residf), main="Density of residuals)")
```

Thereafter, the $\hat{\beta}_1$ can be calculated, using the aforementioned means. This ended up being 144.333. This also helps find $\hat{\beta}_0$ or $\hat{\alpha}$ from the mean lifespan and mean thorax length, and $\hat{\beta}_0$ turned out to be -61.052. These were verified with the R's lienar model fitting function.

```
# calculating beta_hat
beta_hatf <- (sum((fruitfly$lifespan - lifespan_bar)*(fruitfly$thorax -
    thorax_bar)))/(sum((fruitfly$thorax - thorax_bar)^2))
beta_hatf
#beta hat = 144.333

# calculating alpha_hat
alpha_hatf <- lifespan_bar - thorax_bar*beta_hatf
alpha_hatf
#alpha hat = -61.052
```

Figure 5: Density of Residuals




```
10
11 # checking with R's linear model fitting function
12 fruitflymodel <- lm(lifespan~thorax, data=fruitfly)
13 summary(fruitflymodel)
```

The slope of this regression indicates that for every 1 millimeter increase in thorax length, a fruitfly's lifespan is predicted to increase by 144.333 days on average.

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

   The null hypothesis is that there is no linear relationship between fruitfly lifespan and thorax length, meaning its correlation coefficient or $\rho = 0$. The alternative hypothesis

is that there is a linear relationship between fruitfly lifespan and thorax length, meaning its correlation coefficient of $\rho \neq 0$.

To test the null hypothesis that $\rho = 0$, the test statistic was calculated and therefrom, the p-value was calculated. The test statistic using the correlation coefficient that was previously calculated proved to be 9.152 and the p-value proved to be $1.497 * 10^{-}15$. This was verified by R's correlation test function.

```
# testing the null hypothesis that rho = 0
TS <- (r*sqrt(length(fruitfly$lifespan)-2))/sqrt(1-r^2)
2*pt(TS, length(fruitfly$lifespan) - 2, lower.tail=F)
#p-value = 1.497*10^-15

# checking with R's correlation test function
cor.test(fruitfly$thorax, fruitfly$lifespan)
```

Because the r value of 0.636 is associated with a t-value of 9.152 and a p-value of $1.497 * 10^{-}15$, with an alpha of 0.05, the null hypothesis is rejected. As a result, there is a statistically significant linear relationship between the thorax length and lifespan of fruitflies.

5. 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`.

   The confidence interval was calculated by finding the t-value that reflects 90% confidence interval at degrees of freedom of 2, and the margin of error factored in this and the standard error that was calculated from before. The lower and upper limit of the confidence interval was simply the margin of error added to and subtracted from the $\hat{\beta}_1$. The confidence interval proved to be 118 to 170, which was verified by using the function `confint()` in `R`.

```
# calculate confidence interval using equation
tfci <- qt(0.9, df = length(fruitfly) - 2)
me <- tfci*se_lifespan_hat
lowerlim <- beta_hatf - me
upperlim <- beta_hatf + me
ci <- c(lowerlim, upperlim)
ci
```

```r
# The confidence interval is 118.196 to 170.470.

# checking with R's confidence interval function
confint(fruitflymodel, level=0.9)
```

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?

   First, the predict function was used to predict an individual fruitfly's lifespan at thorax of 0.8 mm, which output 54.415 days. The standard error of this prediction was also calculated and proved to be 13.660, and the confidence limits for the predicted value of this lifespan turned out to be 22.268 to 86.562.

```r
# (1) predict an individual fruitfly's lifespan at thorax = 0.8
thoraxi <- data.frame(thorax = 0.8)
lifespani <- predict(fruitflymodel, thoraxi)
lifespani
# The fruitfly's lifespan, given its thorax is 0.8, is predicted to be
    54.415 days.
sei <- stdv_lifespan_hat*sqrt(1+1/length(fruitfly$lifespan)+((thoraxi-
    thorax_bar)^2)/(sum((fruitfly$thorax - thorax_bar)^2)))
sei
# The standard error of this prediction is 13.660.
mei <- qt(0.95, df = length(fruitfly)-2)*sei
cii <- c(lifespani - mei, lifespani + mei)
cii
#The confidence limits for the predicted value of lifespan 54.415 days
    with confidence coefficient 95% are 22.268 to 86.562.
```

   Second, the predict function was used to predict the average lifespan of fruitflies when the thorax length is 0.8 by the model, which again output 51.415 days. However, what changed was the standard error of the prediction, which was estimated at 1.261, and the confidence interval, which was calculated at 51.448 to 57.382.

```r
# (2) average lifespan of fruitflies when thorax = 0.8 by the model
mui <- predict(fruitflymodel, data.frame(thorax = 0.8))
mui
# The fruitfly's average lifespan when thorax = 0.8 is 54.415.
se_mui <- stdv_lifespan_hat*sqrt(1/length(fruitfly$lifespan) + ((0.8-
    thorax_bar)^2)/(sum((fruitfly$thorax - thorax_bar)^2)))
se_mui
# The standard error of this estimate is 1.261.
me_mui <- qt(0.95, df = length(fruitfly)-2)*se_mui
ci_mui <- c(mui - me_mui, mui + me_mui)
```

```
ci_mui
#The confidence limits for the average lifespan at thorax = 0.8 with
    confidence coefficient 95% are 51.448 to 57.382.
```

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.

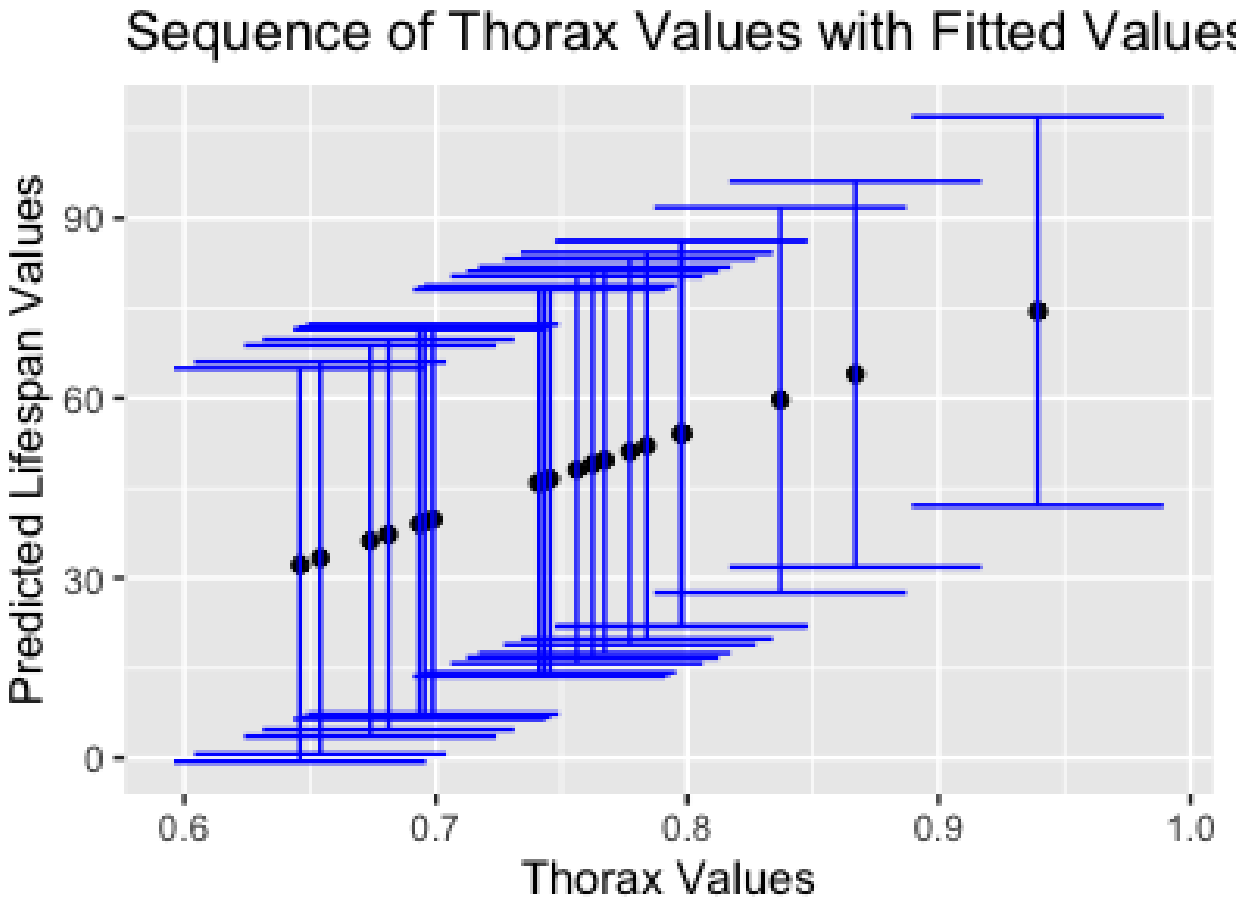
   First, a plot for the fitted values for lifespan with prediction intervals given a sequence of thorac values was depicted in Figure 6. A random sample of 20 numbers within the domain of thorax length values was drawn to provide a sequence of thorax values from which the y values or lifespan values could be predicted. This was plotted in a dot plot with the sequence of thorax values against the predicted lifespan values. The standard error and prediction intervals were determined from this and were also taken into account in the plot in the form of error bars.

```
# plot with fitted values for lifespan with prediction intervals
xi <- runif(20, min(fruitfly$thorax), max(fruitfly$thorax))
yi <- alpha_hatf + beta_hatf*xi
sexi <- stdv_lifespan_hat*sqrt(1+1/length(fruitfly$lifespan)+((xi -
    thorax_bar)^2)/sum((fruitfly$thorax - thorax_bar)^2))
yimin <- yi - qt(0.95, df = length(fruitfly)-2)*sexi
yimax <- yi + qt(0.95, df = length(fruitfly)-2)*sexi
fitted <- data.frame(xi, yi, yimin, yimax)
ggplot(fitted, aes(x=xi, y=yi)) +
  geom_point() +
  geom_errorbar(width=.1, aes(ymin=yimin, ymax=yimax), color="blue") +
  ggtitle("Sequence of Thorax Values with Fitted Values for Lifespan and
    Prediction Interval") +
  labs(x = "Thorax Values", y = "Predicted Lifespan Values")
```

   Second, a plot was depicted in Figure 7 for average values of lifespan given a sequence of thorax lengths, the same randomly distributed sample used in the previous step. From this sequence of thorax values, average y values or lifespan values were estimated. This was plotted in a dot plot with the sequence of thorax values against the average lifespan values. The standard error and confidence intervals were determined from this and were also taken into account in the plot in the form of error bars.

```
# plot with fitted values for lifespan with confidence intervals
mu_xi <- alpha_hatf + beta_hatf*xi
semu_xi <- stdv_lifespan_hat*sqrt(1/length(fruitfly$lifespan) + ((xi-
    thorax_bar)^2)/(sum((fruitfly$thorax - thorax_bar)^2)))
muimin <- mu_xi - qt(0.95, df = length(fruitfly)-2)*semu_xi
muimax <- mu_xi + qt(0.95, df = length(fruitfly)-2)*semu_xi
fittedmu <- data.frame(xi, mu_xi, muimin, muimax)
ggplot(fittedmu, aes(x=xi, y=mu_xi)) +
  geom_point() +
  geom_errorbar(width=.1, aes(ymin=muimin, ymax=muimax), color="red") +
```

Figure 6: Sequence of Thorax Values with Fitted Values for Lifespan and Prediction Interval




```
  ggtitle("Average Lifespan Given Sequence of Thorax Values and
    Confidence Interval") +
  labs(x="Thorax Values", y = "Average Lifespan Values")
```

Figure 7: Average Lifespan Given Sequence of Thorax Values and Confidence Interval