**STATISTICAL METHODS FOR DATA SCIENCE CS6313-001 FALL 2019**

**SSSSECTION-1: Answers to the specific questions asked**

Question 1(a):

[1] "summary for male data:"

body\_temperature heart\_rate

Min. : 96.40 Min. :57.00

1st Qu.: 98.00 1st Qu.:68.00

Median : 98.40 Median :76.00

Mean : 98.39 Mean :74.15

3rd Qu.: 98.80 3rd Qu.:80.00

Max. :100.80 Max. :89.00

[1] "summary for female data"

body\_temperature heart\_rate

Min. :96.3 Min. :58.00

1st Qu.:97.6 1st Qu.:70.00

Median :98.1 Median :73.00

Mean :98.1 Mean :73.37

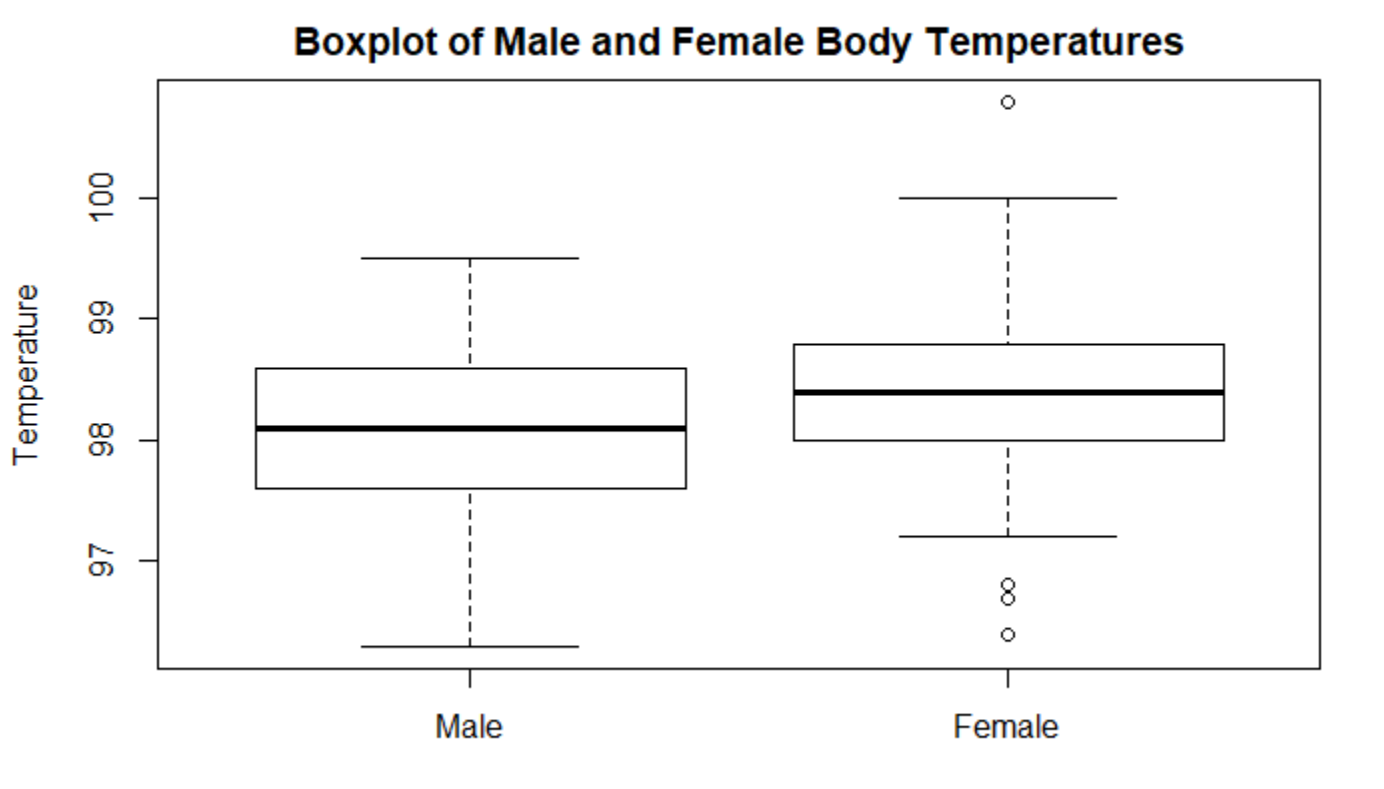
3rd Qu.:98.6 3rd Qu.:78.00

Max. :99.5 Max. :86.00

From the above summary, we can see that:

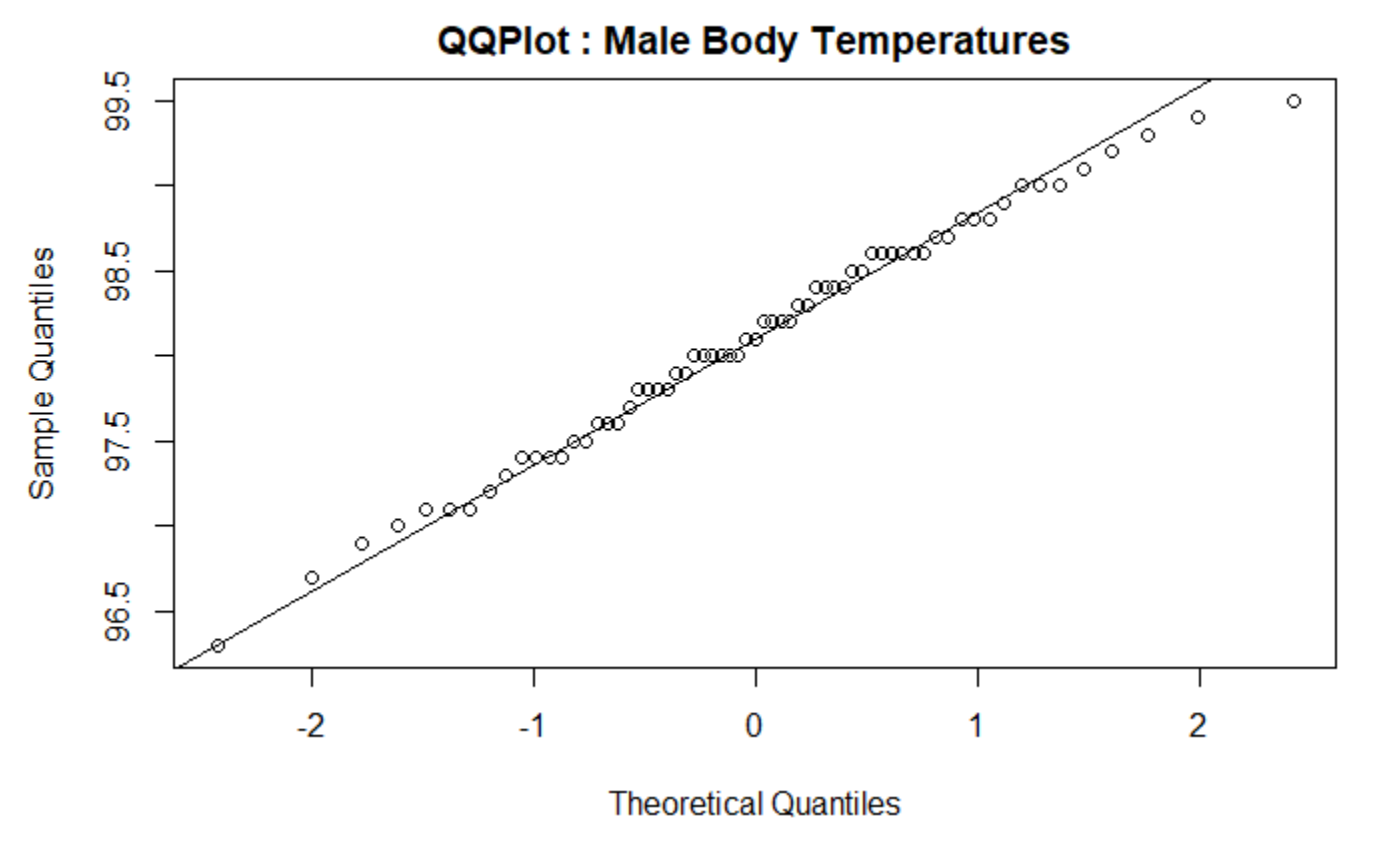
Mean body temperature of men is more than the mean body temperature of women.

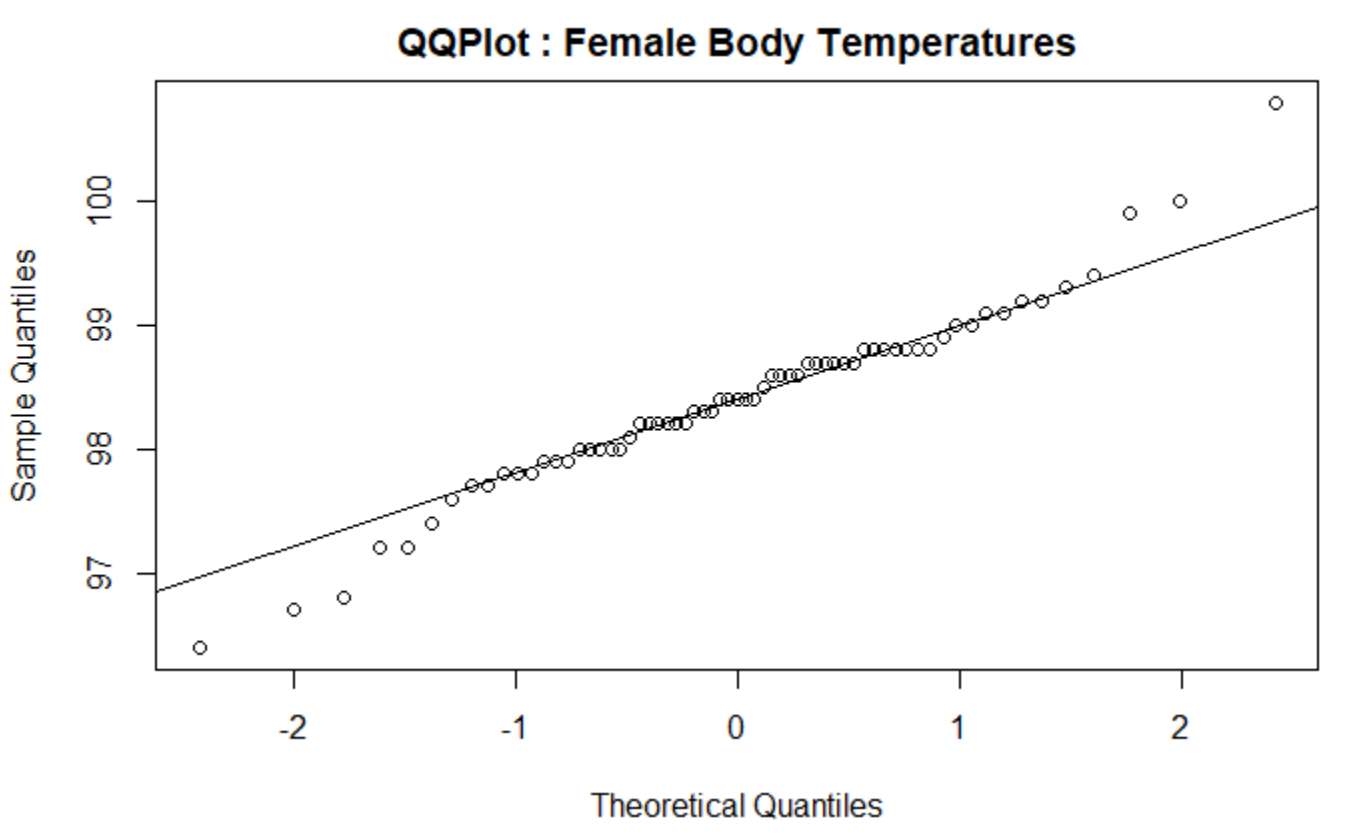
Mean heart rate of men is more than the mean heart rate of women.



From the above boxplot, we can conclude that:

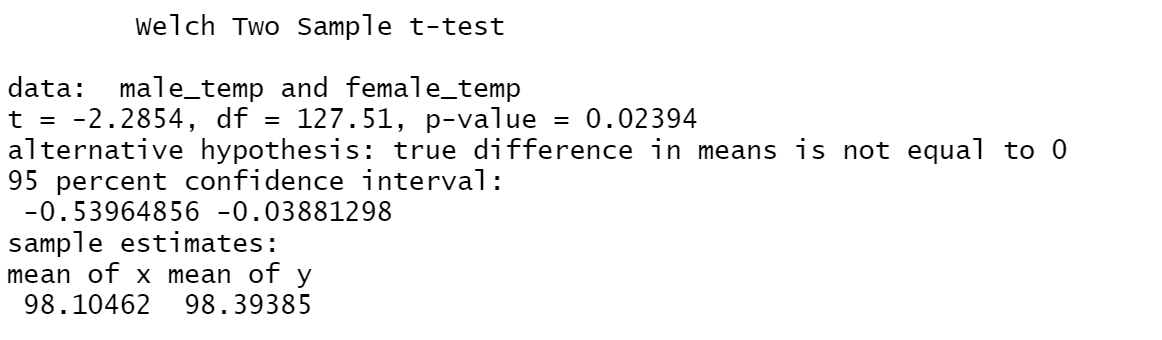
1. Median body temperature of men is less than the median body temperature of women.
2. There are four outliers in the body temperature boxplot for women which suggests more variability in values than men, hence, we cannot assume equal variances.
3. There are no outliers in the body temperature boxplot for men.
4. The Interquartile Range for the body temperature boxplot of men is more than the IQR of the body temperature boxplot of women.





Looking at the above QQ plots, we can make normality assumptions for the distribution of body temperatures among men and women.

We can now perform a T test to calculate the confidence interval for difference in mean body temperature of men and women.



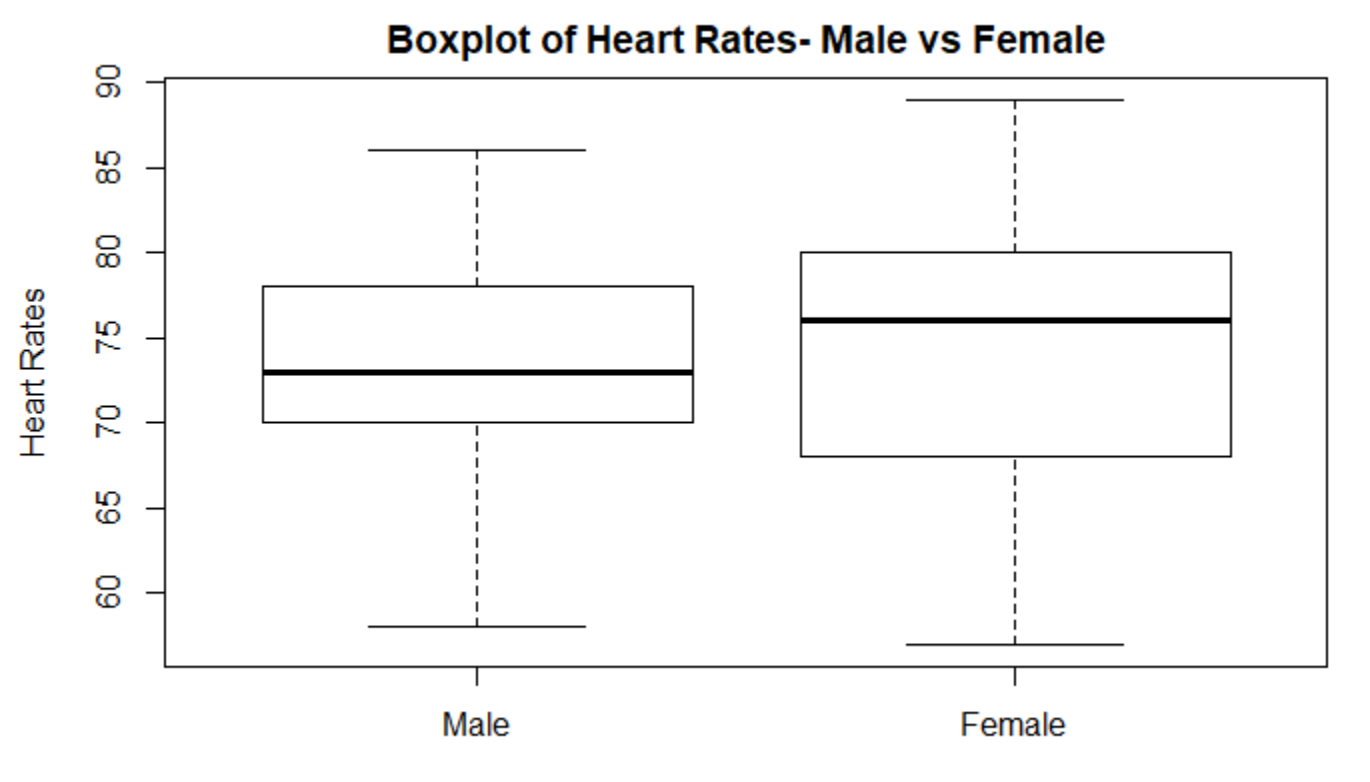
From the above Welch Two Sample T-Test, we can make the following observations:

1. Null Hypothesis : The difference in mean body temperatures among men and women is equal to zero. (µ1=µ2)
2. Alternative Hypothesis: There is a difference in the mean body temperatures of men and women. (µ1≠µ2)
3. The CI [-0.53964856, -0.03881298], calculated at 95% ( α = 0.05) is to the left of zero and the obtained p value 0.02394 is less than α.

Hence, we can conclude that that there is not enough evidence to prove the null hypothesis that the difference in mean body temperatures of men and women is zero.

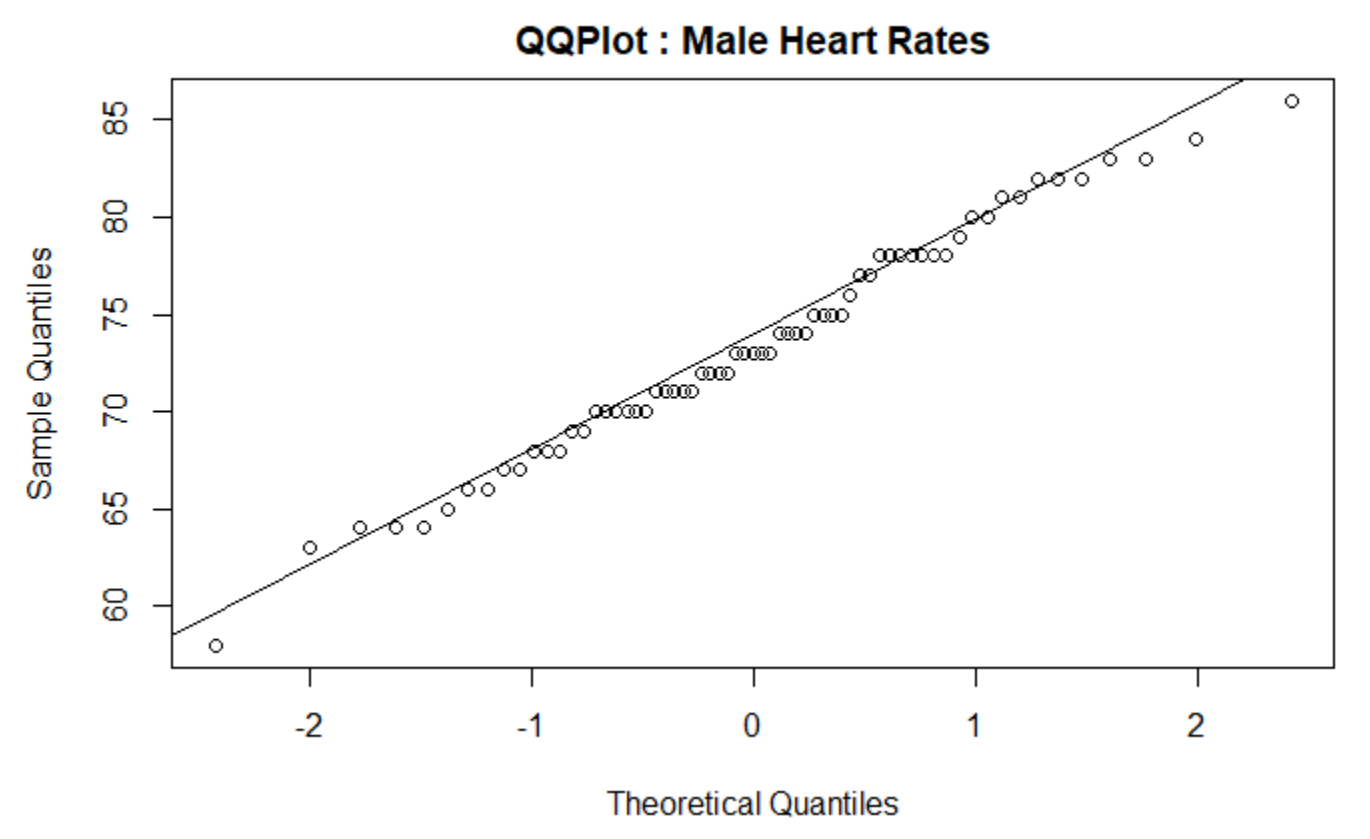
Thus, there is a difference between mean body temperatures of men and women.

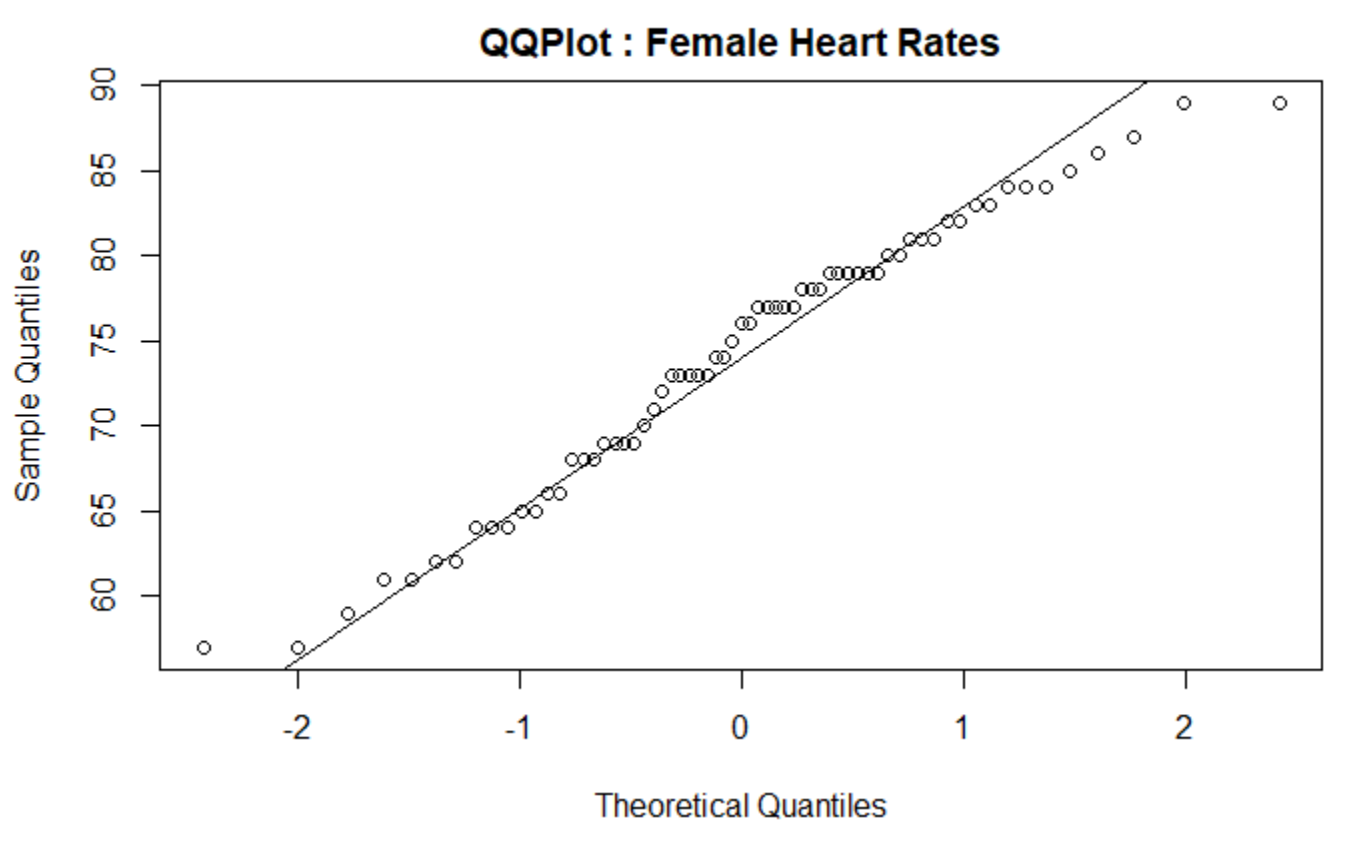
Question 1(b):



From the above boxplot, we can conclude that:

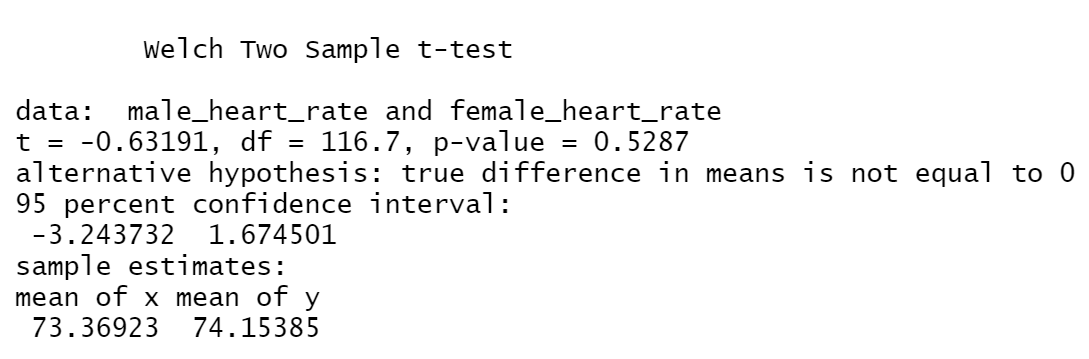
1. Median heart rate among men is less than the median heart rate among women.
2. There are no outliers for both the boxplots.
3. IQR for men is less than the IQR for women.
4. The values among women seem to be more stretched out than those of men, hence, we cannot assume equal variances.





From the above two QQ plots, we can make normality assumptions for the distribution of both male and female heart rates.

We can now perform a T test to calculate the confidence interval for difference in mean body temperature of men and women.



From the above Welch Two Sample T-Test, we can make the following observations:

Null Hypothesis : The difference in mean heart rates among men and women is equal to zero. (µ1=µ2)

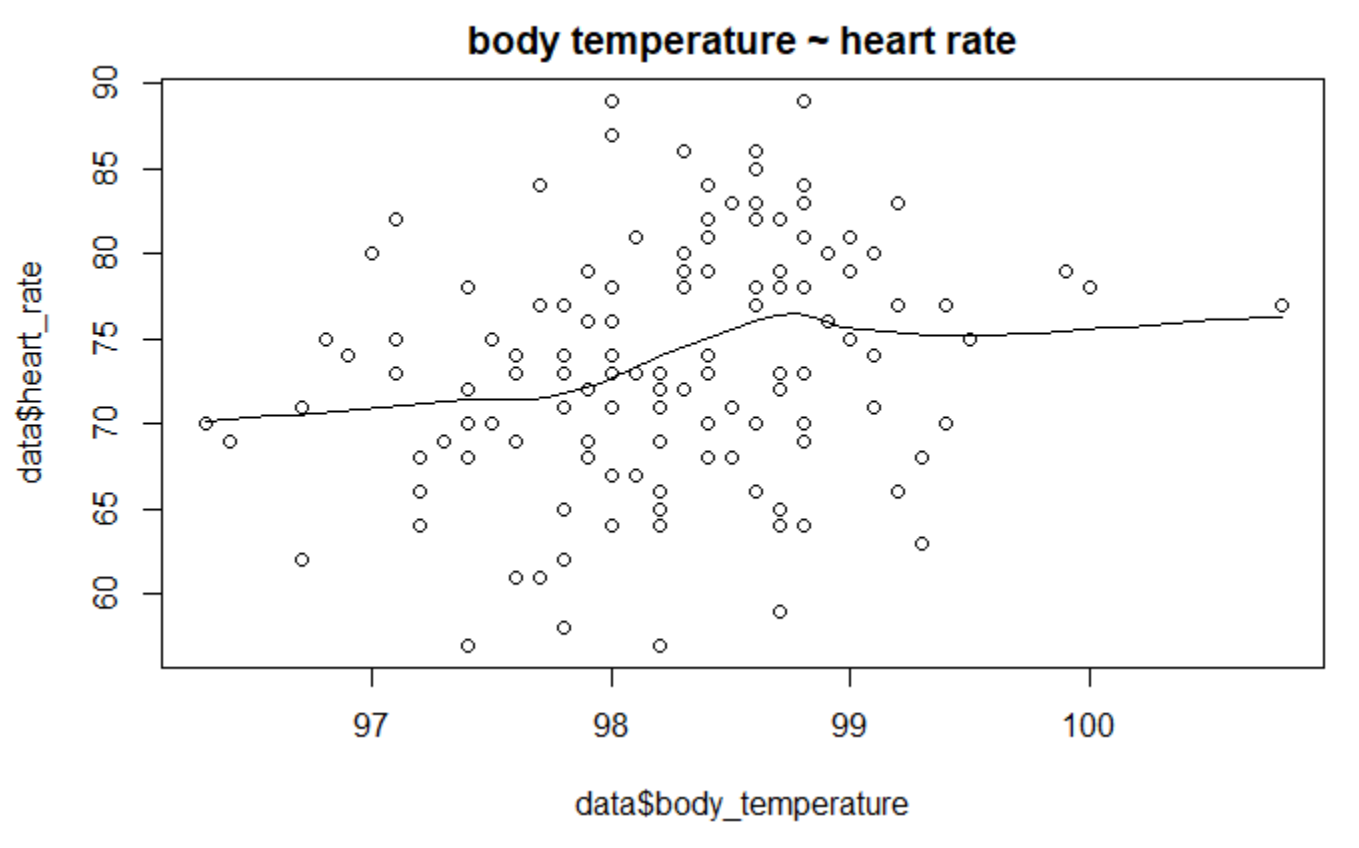
Alternative Hypothesis: There is a difference in the mean heart rates among men and women. (µ1≠µ2)

The CI [-3.243732, 1.674501], calculated at 95% ( α = 0.05) contains zero and the obtained p value 0.5287 is greater than α.

Hence, we can conclude that that there is enough evidence to prove the null hypothesis that the difference in mean heart rates among men and women is zero.

Thus, there is no difference between mean heart rates among men and women.

Question 1(c):

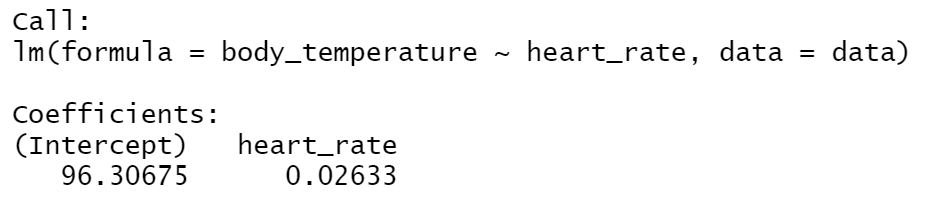


From the scatter plot above, there is a linear relationship between heart rate and body temperature upon considering all the male and female samples.

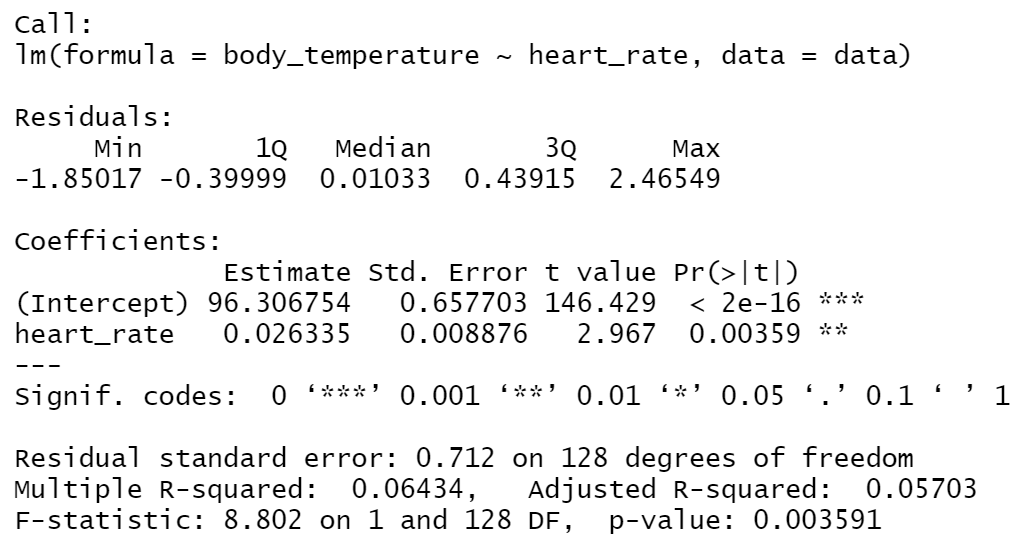
We can compute the linear relationship using a linear regression between body temperature and heart rate among both men and women; getting a linear line from the given data.

Correlation of combined men and women subjects’ body temperature and heart rate = 0.2536564

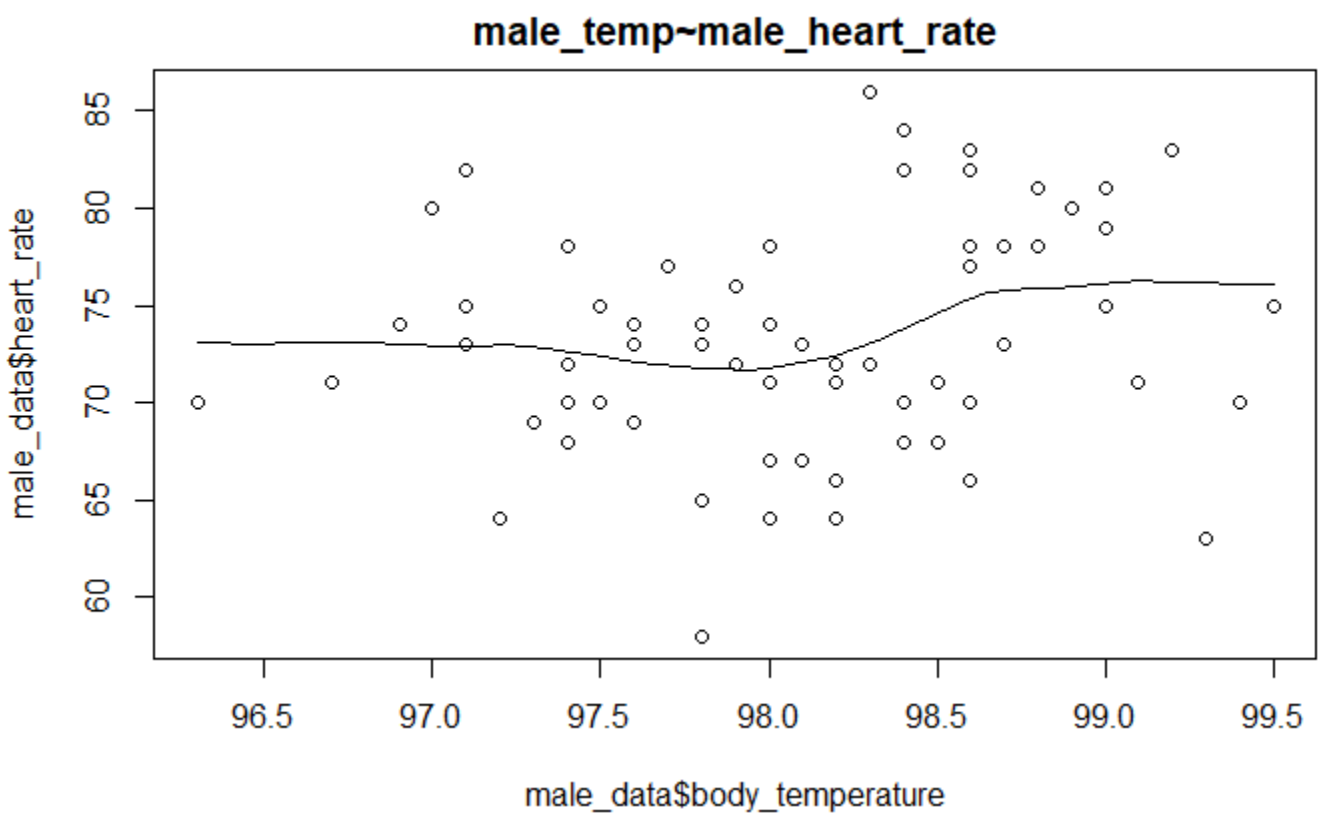
We could interpret the linear regression line as:



Overall summary of linear regression is given below:



From the scatterplot, we could conclude a linear relationship with a correlation of approx. 0.25 which is not a strong evidence that the parameters are perfectly linear.

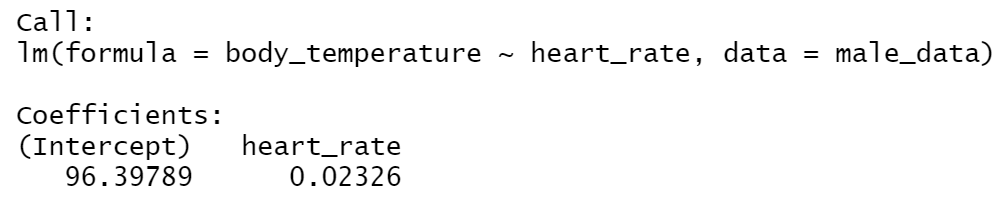


From the scatter plot above, there is a linear relationship between body temperature and heart rate when only male samples are considered.

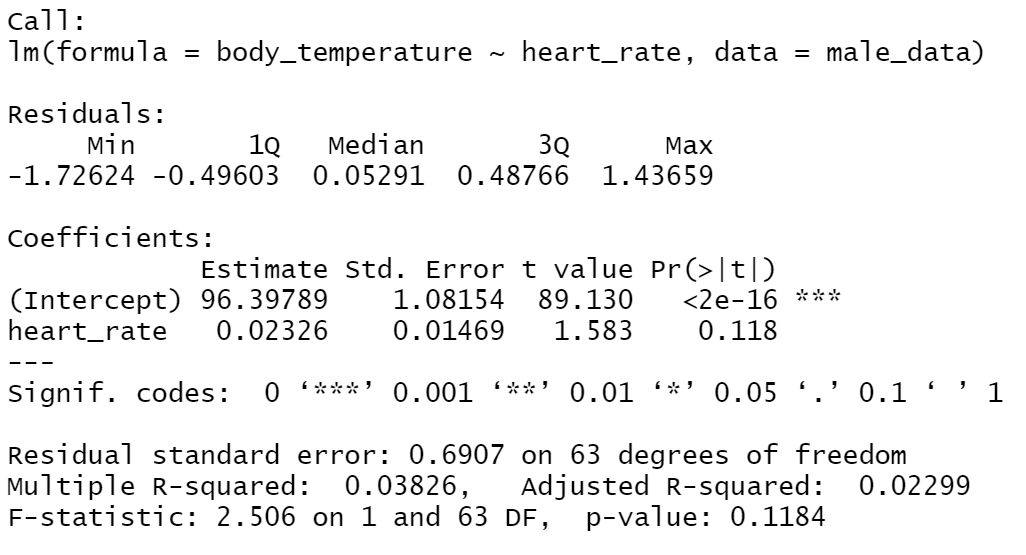
We could compute the linear relationship using a linear regression between body temperature and heart rate of male subjects and getting a linear line from the data.

Correlation of male subjects’ body temperature and heart rate = 0.1955894

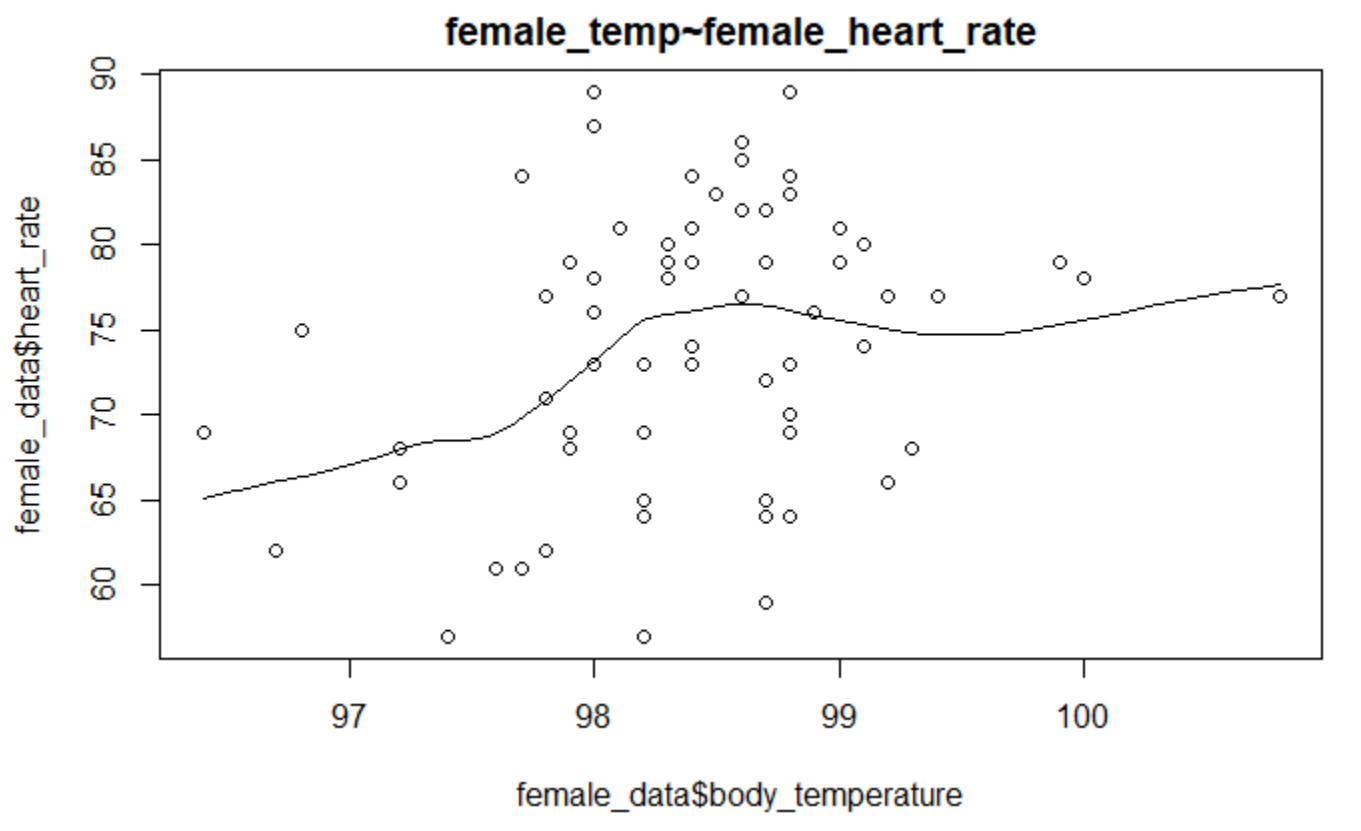
We could interpret the linear regression line as:



The overall summary of linear regression is given below:



From the scatter plot, we could conclude a linear relationship with correlation approx. 0.19 which is not a strong evidence that the parameters are perfectly linear.

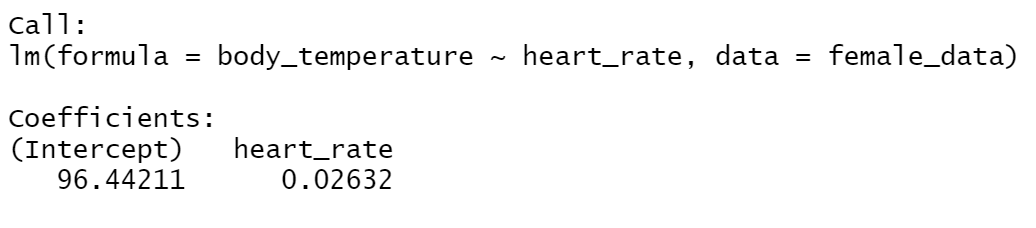


From the scatter plot above, there is a linear relationship between body temperature and heart rate when only female samples are considered.

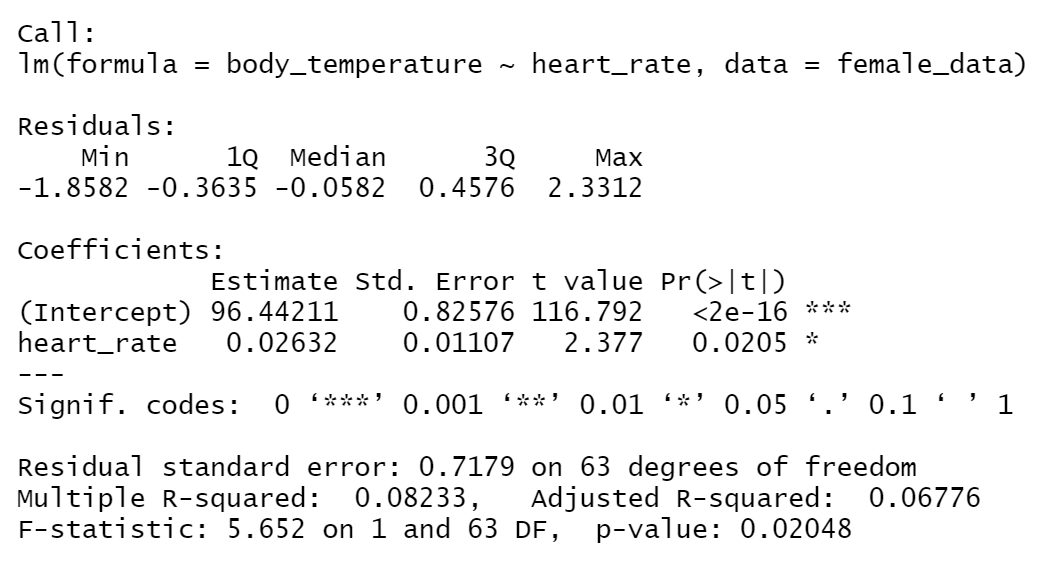
We could compute the linear relationship using a linear regression between body temperature and heart rate of female subjects and getting a linear line from the data.

Correlation of female subjects’ body temperature and heart rate = 0.2869312

We could interpret the linear regression line as:



The overall summary of linear regression is given below:



From the scatter plot, we could conclude a linear relationship with correlation approx. 0.28 which is not a strong evidence that the parameters are perfectly linear, but is better when compared to the male subjects.

The correlation between body temperature and heart rate for females is a little stronger when compared to that among males.

Question 2(a):

Using the functions given in Section 2, we have taken the value of n=10 and λ=0.1 to get the coverage probabilities as:

Z-Interval : 0.8740

Bootstrap Interval : 0.9198

Question 2(b):

Repeating the above process for remaining values of n and λ, we get the following values:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Z – Proportion | L = 0.01 | L = 0.1 | L = 1 | L = 10 |
| N = 5 | 0.8164 | 0.8052 | 0.8088 | 0.8048 |
| N = 10 | 0.8614 | 0.8746 | 0.8680 | 0.8672 |
| N = 30 | 0.9174 | 0.9210 | 0.9220 | 0.9242 |
| N = 100 | 0.9332 | 0.9384 | 0.9384 | 0.9376 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| B – Proportion | L = 0.01 | L = 0.1 | L = 1 | L = 10 |
| N = 5 | 0.8976 | 0.8974 | 0.8962 | 0.8982 |
| N = 10 | 0.9132 | 0.9166 | 0.9174 | 0.9186 |
| N = 30 | 0.9398 | 0.9324 | 0.9416 | 0.9358 |
| N = 100 | 0.9478 | 0.9472 | 0.9486 | 0.9472 |

Graphically representation of the above data is given below,

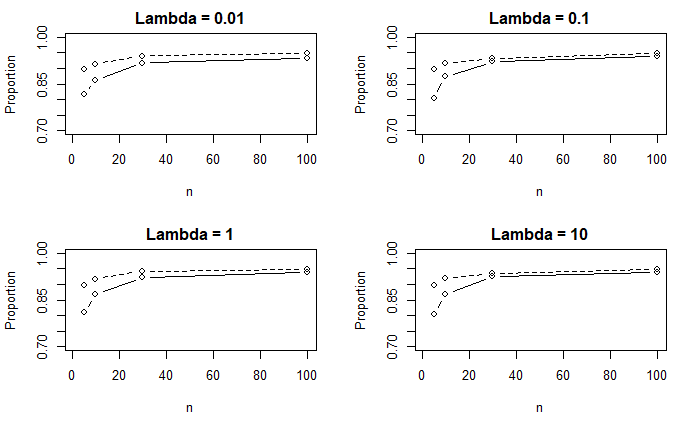


Figure 1 : The Solid line represents the Z-Proportion and the Dotted Line represents the Bootstrap Proportion. (Keeping λ fixed)

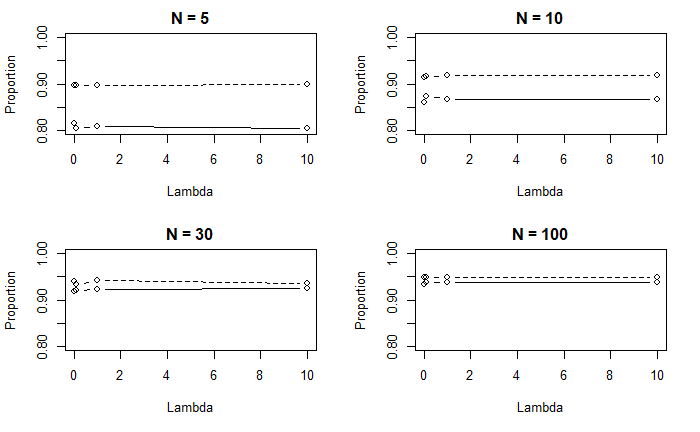


Figure 2 : The Solid line represents the Z-Proportion and the Dotted line represents the Bootstrap proportion. (Keeping λ fixed)

Question 2 (c) :

From graphs in Figure 1, we can see that:

The graphs do not change a lot when 𝜆 changes. Hence, we could conclude that the coverage probabilities do not depend on the value of 𝜆.

The coverage probabilities that we get from z interval method are lower than those we get from the bootstrap method.

From graphs in Figure 2, we can see that:

The coverage probabilities depend on the value of n.

For large sample z interval, we could observe that the coverage probabilities are as accurate as the coverage probabilities that we obtained from the bootstrap method (when n is large; n= 100)

N=30 onwards, the coverage probabilities for the bootstrap method are on the higher side.

Considering all the graphs, bootstrap method coverage probabilities are higher for any combination of n and 𝜆 than those of the large sample z interval method.

Thus, we could conclude that even for low values of n, the bootstrap method is more accurate.

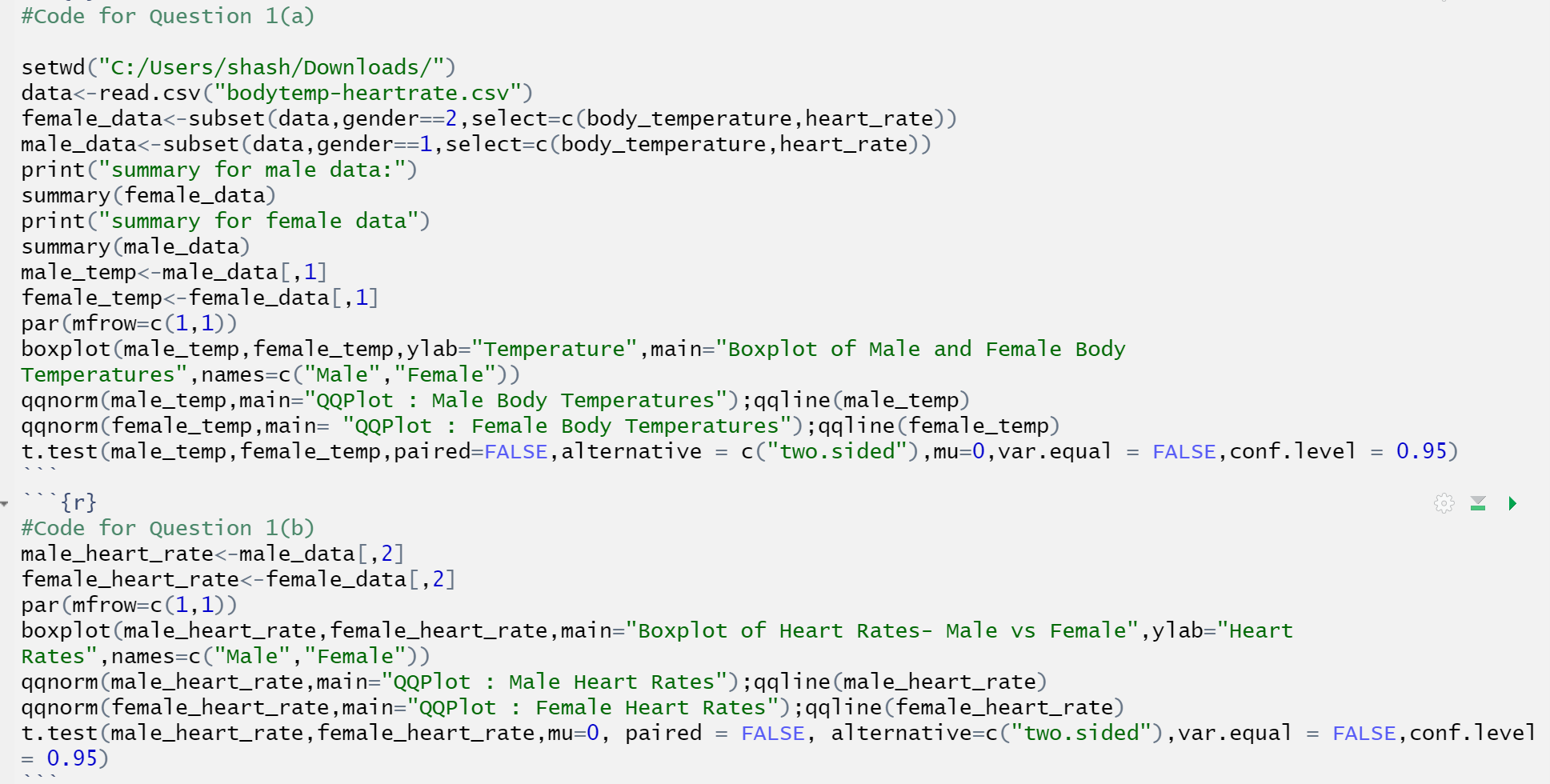
Hence, we recommend the bootstrap method.

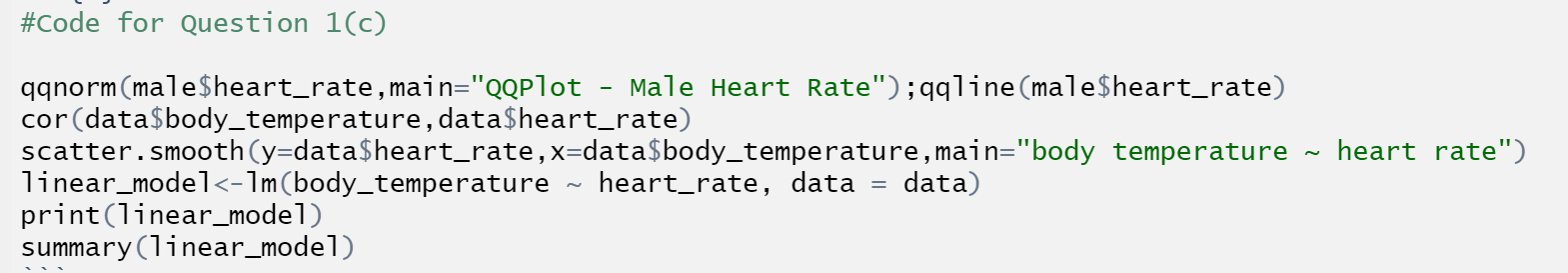
Question 2(d):

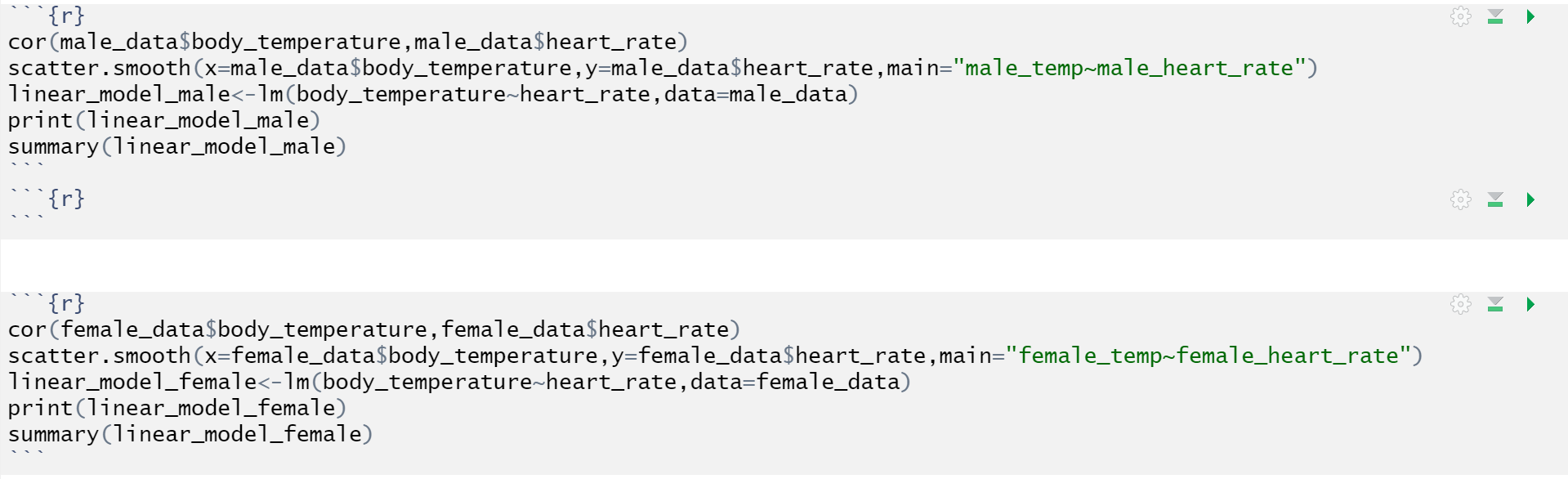
The conclusion in 2(c) does not depend upon values of λ because it is a parameter of the population distribution. Moreover, conclusions about the convergence of sampling methods shouldn’t depend on the population parameters.

**SECTION 2-R Code**

Question 1 :







Question 2:

