University of Stirling MATPMDA

Computing Science & Mathematics 2023

**MATPMDA MATHEMATICAL AND STATISTICAL FOUNDATIONS**

**PROJECT : AUTUMN SEMESTER 2023**

**Submission due 18th December 17:00**

**Student Number: <Student id>**

**Declaration: In submitting this project I declare that this is all my own work and I did not seek help to complete it.**

**For each project question, insert answers below.**

1. Perform an exploratory data analysis, taking care to describe the type of variables in the data set.

**Exploratory Data Analysis**

* The data will be exploratory analyzed in this question

**Descriptive Statistics**

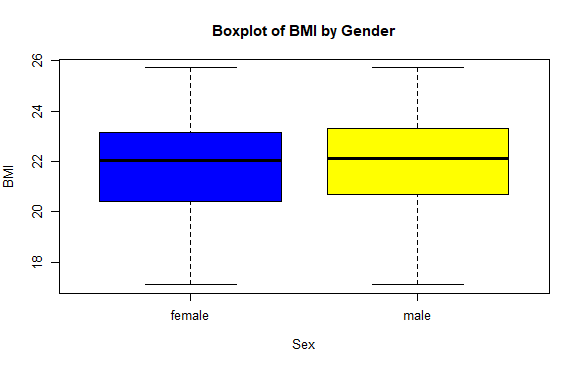
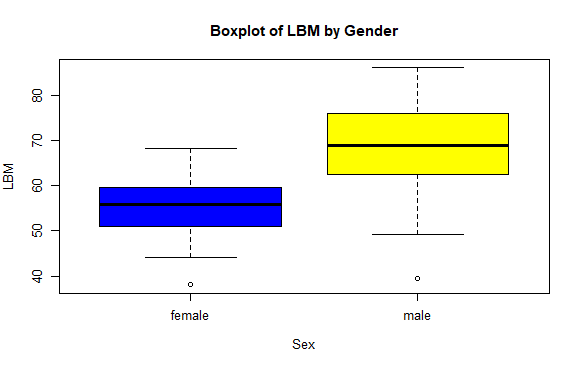
Given is a dataset of 178 individuals of their body mass index & lean body max, 81 being females while 97 males:

Sex LBM BMI   
 Length:178 Min. :38.06 Min. :17.11   
 Class :character 1st Qu.:54.90 1st Qu.:20.62   
 Mode :character Median :61.12 Median :22.07   
 Mean :62.47 Mean :21.91   
 3rd Qu.:69.71 3rd Qu.:23.28   
 Max. :86.07 Max. :25.73

Male LBM BMI   
 Length:97 Min. :39.56 Min. :17.11   
 Class :character 1st Qu.:62.44 1st Qu.:20.70   
 Mode :character Median :68.92 Median :22.11   
 Mean :68.62 Mean :21.92   
 3rd Qu.:76.05 3rd Qu.:23.32   
 Max. :86.07 Max. :25.73

Female LBM BMI   
 Length:81 Min. :38.06 Min. :17.11   
 Class :character 1st Qu.:51.03 1st Qu.:20.41   
 Mode :character Median :55.86 Median :22.04   
 Mean :55.10 Mean :21.89   
 3rd Qu.:59.60 3rd Qu.:23.17   
 Max. :68.23 Max. :25.73

We’ll use boxplot to analyze the shape of the data distribution of our sample:

the difference is greater in terms of lean body mass for males and females.

IQRs: LBM=14.81, BMI=2.66 (both sexes)

LBM=13.601, BMI=2,62 (male)

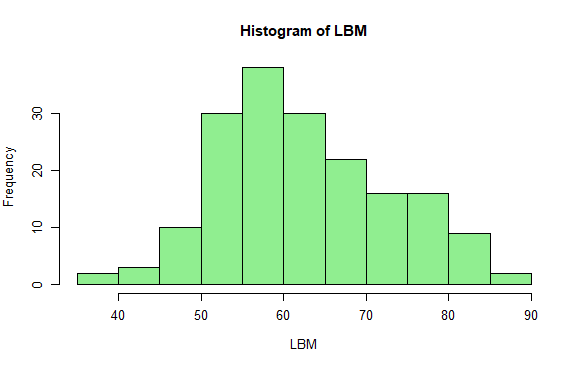
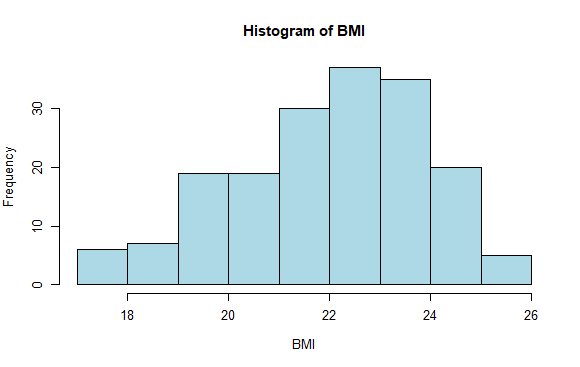
LBM=8.57, BMI=2.76 (female)

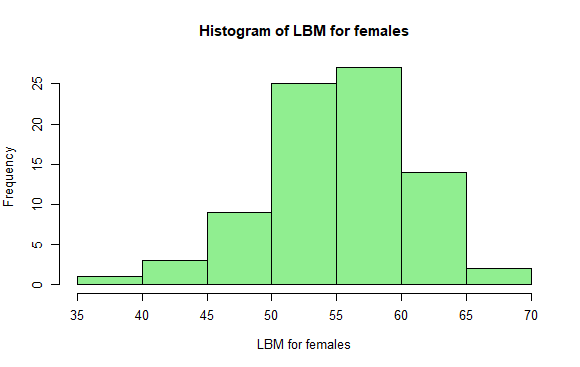
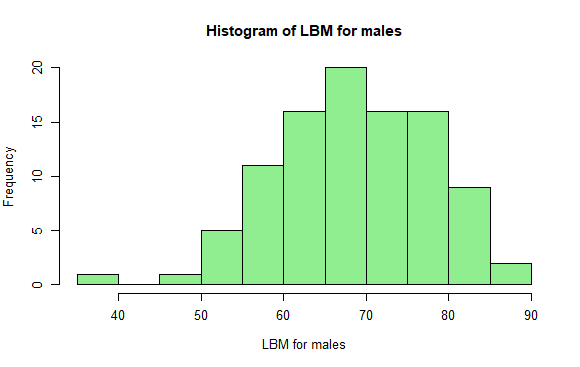
SKEWNESS: LBM=0.283 (p.), BMI=-0.360(n.)

LBM=-0.332 (n.), BMI=-0.353(n.) (male)

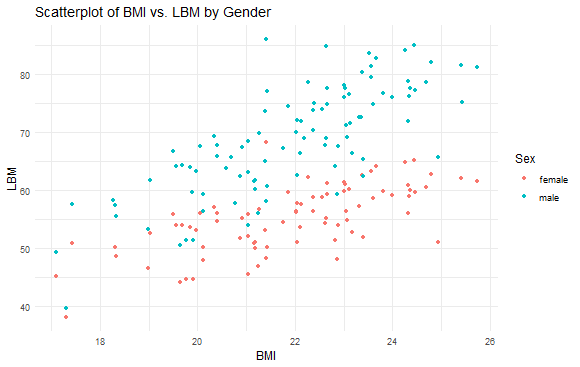
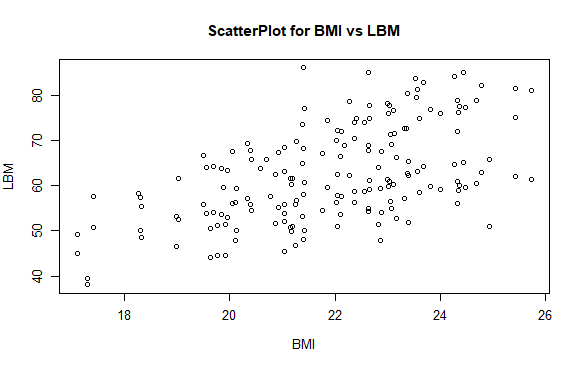
LBM=-0.315 (n.), BMI=-0.358(n.) (female)

Histogram for better shape of distribution:



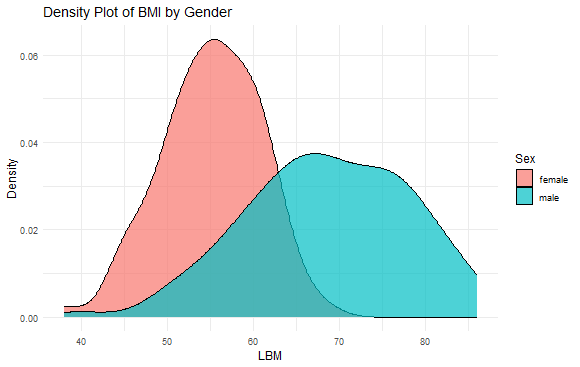


The scatterplot for the whole data will show the overall relationship between BMI’s and LBMs:



The scatters depict higher correlation positivity for male data.

Finally, we analyse density plot to look at the differences in LBM values:



2. Using an appropriate statistical test, investigate whether there is a difference in mean LBM between males and females.

**Testing the significance of difference of mean LBMs**

* The values of population standard deviations aren’t given, we may test by using student’s distribution, the “t” distribution, or t-test. (Richmond, n.d.)

data: LBM by Sex  
t = -11.801, df = 162.09, p-value < 2.2e-16  
alternative hypothesis: true difference in means between group female and group male is not equal to 0  
95 percent confidence interval:  
 -15.773 -11.251

sample estimates:  
mean in group female mean in group male   
 55.103 68.615

We reject the null hypothesis, the confidence interval includes not the value “0” in it, the difference in mean LBM’s is significant, as suggested by the box plots.

3. For male and female sports people separately, calculate the correlation coefficient for LBM and BMI given and comment on the relationship between LBM and BMI.

**Correlation test for BMI & LBM**

Testing the correlations for male, female and both combined:

Both:

BMI LBM  
BMI 1.000 0.546  
LBM 0.546 1.000

Male

BMI LBM  
BMI 1.000 0.764  
LBM 0.764 1.000

Female

BMI LBM  
BMI 1.000 0.675  
LBM 0.675 1.000

74.6% for males shows high positive correlation, whereas moderate (67.5%) for females.

4. We would like to investigate a model to test the relationship between LBM and BMI for male sportspeople. You must include output from R to support your findings.

Details you should include are:

(a) using your previous results comment on whether there would be any value in

including the data for females in this model.

(b) a description of the model;

(c) a summary of the fitted model with interpretation of test statistics and parameter estimates;

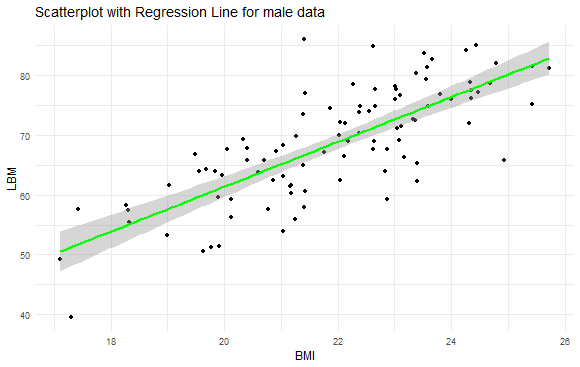
(d) evidence as to whether assumptions of the model have been met;

(e) conduct a formal test to question whether there is a significant linear relationship

between LBM and BMI.

**THE PROPOSED MODEL**

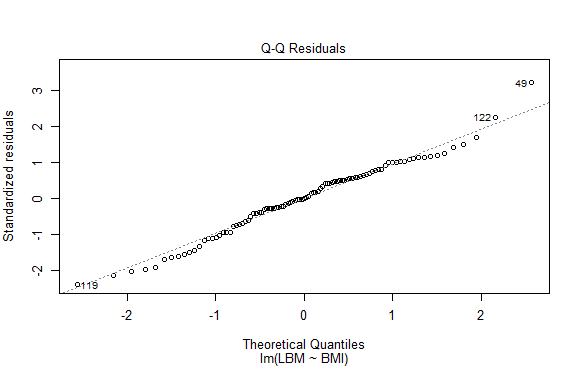
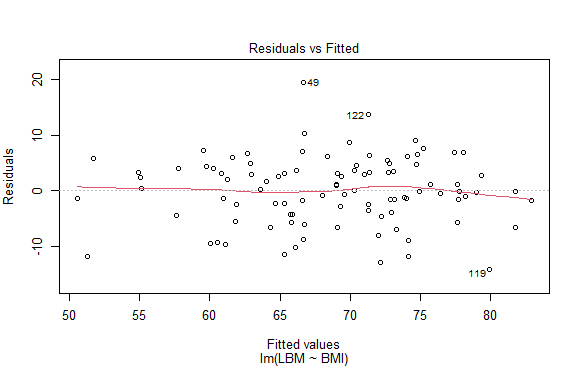
Beginning with scatterplot and analysing the relationship of LBM ~ BMI values for males.

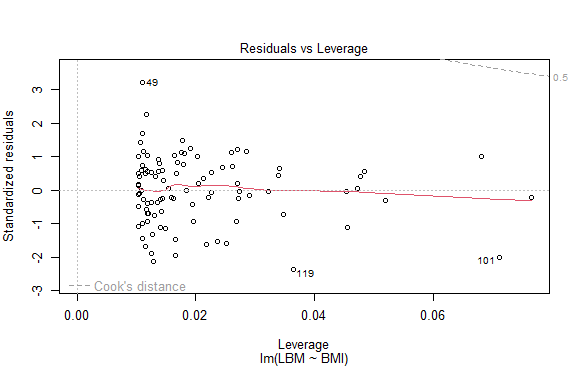
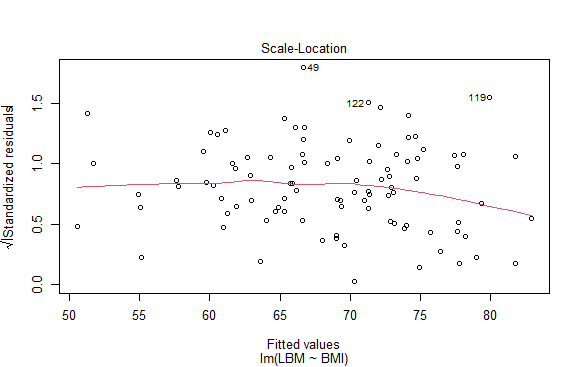


We note a good positive relation, the summary of the model suggests:

Residuals:  
 Min 1Q Median 3Q Max   
-14.236 -3.858 0.004 3.913 19.399   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -13.725 7.162 -1.916 0.0583 .   
BMI 3.757 0.326 11.540 <2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 6.082 on 95 degrees of freedom  
Multiple R-squared: 0.584, Adjusted R-squared: 0.579   
F-statistic: 133.2 on 1 and 95 DF, p-value: < 2.2e-16

Hence as there’s an increase of 1 unit in BMI, on average, there’s an increase of 3.757 in LBM, showing a significant relationship.





Now we see the model summary after adding female data

Residuals:  
 Min 1Q Median 3Q Max   
-20.404 -6.990 -1.052 7.332 25.103   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -2.329 7.524 -0.310 0.757   
BMI 2.958 0.342 8.645 3.24e-15 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 8.733 on 176 degrees of freedom  
Multiple R-squared: 0.298, Adjusted R-squared: 0.294   
F-statistic: 74.73 on 1 and 176 DF, p-value: 3.244e-15

The relationship slope decreased from “3.757” to “2.958” of BMI and LBM, hence adding the female data would not be of much use.

5. Use the model developed in Question 4 to predict the LBM for a male whose BMI is 25.

**Predicting Value of LBM**

Our model suggests, LBM=-13.725+3.757BMI,

Predicted LBM for BMI = 25: 80.19958

6. Assess the predictive performance of the model.

**PERFROMANCE**

**R Squared:** The r2 score varies between 0 and 100%, So if it is 100%, the two variables are perfectly correlated, i.e., with no variance at all. A low value would show a low level of correlation, meaning a regression model that is not valid, but not in all cases. (BMC blogs.)

**MSE:** the average of the square of the errors. The larger the number the larger the error. (BMC blogs)

Mean SQUARED Error: 36.227, suggests a moderate level of prediction error,

R2: 58.4%, approximately 58.4% of the total variation in LBM is defined by that in BMI in our model for male data.

7. In this final section include all R code that you have used for this project verbatim. Ensure that:

* the code for each question can be easily found;
* all code is adequately commented;
* variable names are sensible.

1. summary(), boxplot(), IQR(), e1071 package: skewness(), hist(), plot(), ggplot()+geom\_point(), ggplot()+geom\_density #exploratory data analysis
2. t.test(), #testing hypothesis
3. cor(), #correlation matrix,
4. ggplot()+geom\_point+geom\_smooth, lm(), summary(), #regression model
5. predict() #predicting values
6. mean() #mse, summary() #coefficient of variation, #performance

References.

Include here references to statistical methods you have used in the module notes, or any online resources you have used to produce this project.

Richmond, S. B., n.d. Statistical Analysis. Second ed. s.l.:s.n.

[Mean Square Error & R2 Score Clearly Explained – BMC Software | Blogs](https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/)