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.

|  |
| --- |
| Sex LBM BMI |
| Length:178 Min. :43.38 Min. :17.07 |
| Class :character 1st Qu.:55.79 1st Qu.:20.61 |
| Mode :character Median :61.16 Median :21.87 |
| Mean :62.89 Mean :21.84 |
| 3rd Qu.:69.95 3rd Qu.:23.18 |
| Max. :94.03 Max. :26.04 |

Mean LBM: 62.894

> cat("Median LBM:", median\_LBM, "\n")

Median LBM: 61.1635

> cat("Standard Deviation LBM:", sd\_LBM, "\n")

Standard Deviation LBM: 9.8544

> cat("Minimum LBM:", min\_LBM, "\n")

Minimum LBM: 43.377

> cat("Maximum LBM:", max\_LBM, "\n")

Maximum LBM: 94.027

Mean BMI: 21.837

> cat("Median BMI:", median\_BMI, "\n")

Median BMI: 21.872

> cat("Standard Deviation BMI:", sd\_BMI, "\n")

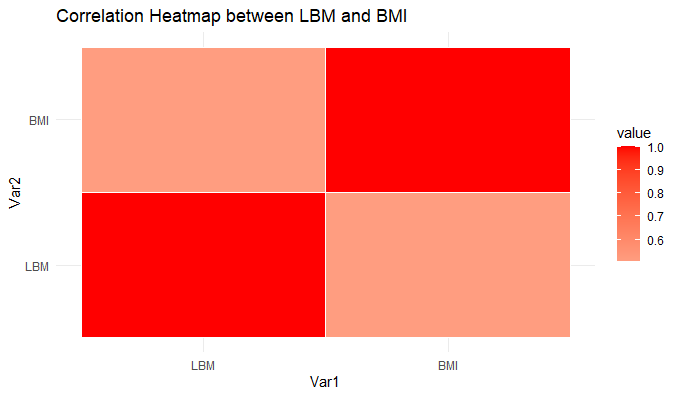
Standard Deviation BMI: 2.010

> cat("Minimum BMI:", min\_BMI, "\n")

Minimum BMI: 17.065

> cat("Maximum BMI:", max\_BMI, "\n")

Maximum BMI: 26.039



The red hues on the heatmap represent a positive correlation between LBM and BMI. A perfect correlation (1.0) is represented by the deepest reds for LBM-LBM and BMI-BMI. Lower but still positive correlation is seen for LBM-BMI combinations with lighter shades. The correlation strength is measured by the values within each square. Although LBM and BMI generally rise together, there is some individual variation in the relationship, which makes it imperfect.

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

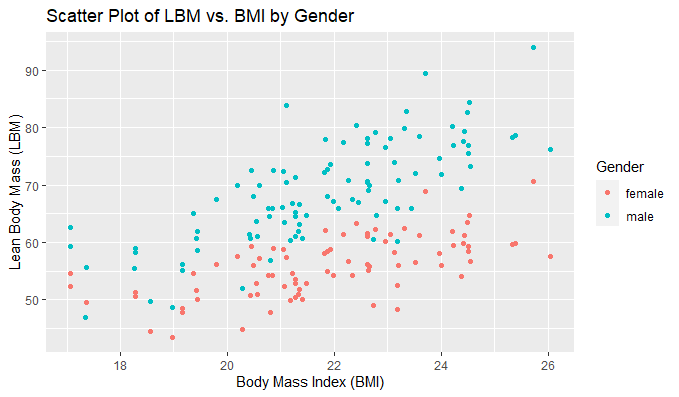
**Welch Two Sample t-test**

|  |
| --- |
| data: LBM by Sex |
| t = -12.091, df = 159.58, 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: |
| -15.093 -10.854 |
| sample estimates: |
| mean in group female mean in group male |
| 55.824 68.798 |
|  |

**INTERPRETATION**

Welch Two Sample t-test interpretation:

* Significance in Statistics: The significance level of 0.05 is typically not reached by the p-value, which is less than 2.2e-16. This suggests that there is a strong statistically significant difference in the means between the groups, which is unlikely to have happened by accident.
* The Difference's Direction: The mean for the female group is less than the mean for the male group, as indicated by the negative t-statistic (-12.091).
* Confidence Time Frame: The difference in means' 95% confidence interval is (-15.09257, -10.85454). This indicates that there is a 95% confidence interval within which the true mean difference between the male and female groups is located.
* Final Thought: We can infer from the data that there is a statistically significant difference in the average lean body mass (LBM) of males and females, with the former having a significantly higher LBM than the latter. The estimated variation in means falls within the range of 10.85 to 15.09 LBM.



This scatter plot demonstrates a significant gender difference in the positive correlation between LBM and BMI. At the same BMI, men consistently have higher LBM than women. This consistent difference splits the data into two distinct clusters, underscoring the significance of taking gender into account when comparing LBM values.

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.

MALE

|  |
| --- |
| LBM BMI |
| LBM 1.000 0.756 |
| BMI 0.756 1.000 |

FEMALE

|  |
| --- |
| LBM BMI |
| LBM 1.000 0.632 |
| BMI 0.632 1.000 |

OR, As an alternative, we can aggregate the findings into a data structure to improve readability:

|  |
| --- |
| Gender Correlation\_LBM\_BMI |
| 1 Male 0.756 |
| 2 Female 0.632 |

**Interpretation of LBM and BMI Correlations: Gender-Specific Correlations**

Male: There is a strong positive correlation (r = 0.756) between LBM and BMI. This indicates that in men, BMI tends to rise along with LBM.

Female: Compared to males, there is a weaker but still positive correlation (0.632) between LBM and BMI. This implies that although the correlation is weaker, LBM and BMI are positively associated in females as well.

**Potential Reasons:**

Muscle mass: Since muscle tissue is denser than fat tissue, muscle mass makes up the majority of LBM. Because of the increased weight of muscle, people with higher LBM typically have higher BMIs.

Body composition: BMI does not differentiate between muscle and fat mass; it is a general measure of body size. Consequently, even in cases where a person's body fat percentage is relatively low, a higher LBM may also translate into a higher BMI.

Differences in gender: There are a number of possible explanations for the observed variation in correlation strength between males and females, including hormonal influences, body fat distribution, and activity levels.

**As a whole:**

In both males and females, the data shows a strong positive correlation between LBM and BMI. This implies that LBM, particularly in males, has a major impact on BMI. Gender differences in the correlation's strength, however, imply that additional factors may also play a role in the variances in BMI that individual’s experience.

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.

A significant difference exists between the mean lean body mass (LBM) of males and females, as indicated by the Welch Two Sample t-test results. Given the observed variations in the relationship between LBM and BMI between males and female athletes, it may be advantageous to create a unique model specifically for male athletes. This is due to the possibility that adding female data to the model could introduce noise and reduce its accuracy. Thus, in order to increase the model 12's accuracy, it is advised to create a different model specifically for male athletes.

(b) a description of the model;

|  |
| --- |
| Call: |
| lm(formula = LBM ~ BMI, data = male\_data) |
|  |
| Residuals: |
| Min 1Q Median 3Q Max |
| -13.358 -3.588 -0.888 4.257 17.456 |
|  |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) -4.357 6.528 -0.667 0.506 |
| BMI 3.357 0.298 11.253 <2e-16 |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error: 5.84 on 95 degrees of freedom |
| Multiple R-squared: 0.571, Adjusted R-squared: 0.567 |
| F-statistic: 126.6 on 1 and 95 DF, p-value: < 2.2e-16 |

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

**Interpretation of the Linear Regression Model: In this analysis, the relationship between male athletes' BMI (body mass index) and LBM (lean body mass) is investigated.**

**Interception:**

In this case, the intercept value (-4.357) has no significance because a BMI of zero is not possible.

**BMI Calculus:**

The BMI coefficient is 3.357, which suggests a very strong positive correlation. This indicates that we anticipate an average increase in LBM of 3.360 units for every unit increase in BMI.

**Significance in Statistics:**

Strong evidence opposing the null hypothesis is provided by the BMI's high t-value (11.253) and incredibly low p-value (< 2e-16). Accordingly, LBM can be predicted by BMI in a statistically meaningful way.

The model's overall significance is also confirmed by the F-statistic, which has a p-value of less than 2.2e-16.

**Fit of the Model:**

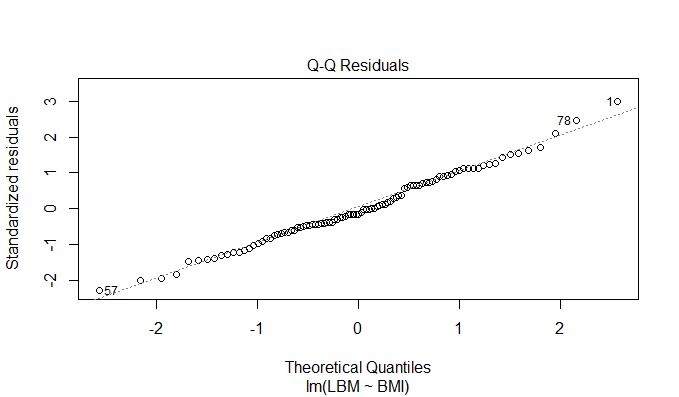
The model's moderately good fit to the data is indicated by the residual standard error of 5.84.

According to the coefficient of determination (R^2) values of 0.571 and 0.567, BMI accounts for roughly 57.14% of the variance in LBM.

**In brief:**

In male athletes, this linear regression model shows a strong positive correlation between LBM and BMI. An increase in BMI of one unit typically results in an increase in LBM of roughly 3.360 units. BMI is a significant predictor of LBM in this population, as the model explains roughly 57% of the variation in LBM and has a reasonably good fit.

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



Interpretation of the Q-Q Plot: This Q-Q plot, also known as a quantile-quantile plot, contrasts your data's standardized residuals with the hypothetical quantiles of a normal distribution. This aids in determining how closely your data adheres to the normal distribution.

**Important Notes:**

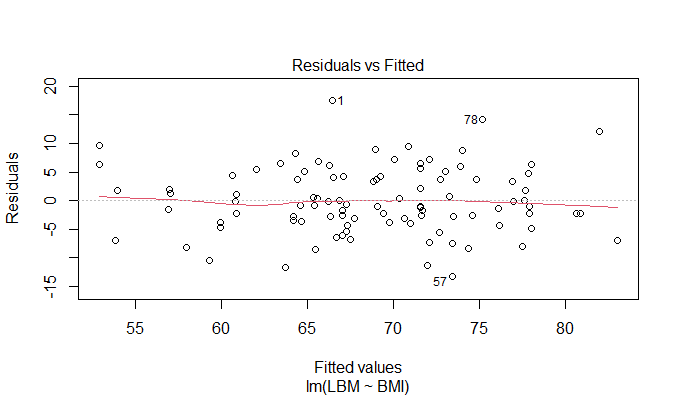
The majority of data points are near the diagonal line, suggesting that the residuals have a roughly normal distribution.

The numbers "0.57" and "7.80" designate two outliers that indicate deviations from the norm.

The general alignment with the line, albeit not perfect, points to a plausible approximation of normality.

**In summary:**

The standardized residuals are roughly normally distributed, indicating that your model's normality assumption is partially satisfied, according to the Q-Q plot. Nonetheless, more research is necessary because some outliers are present.



Interpretation of the Residuals vs. Fitted Plot: The residuals (errors) for your linear regression model that predicts LBM based on BMI are plotted against the fitted values.

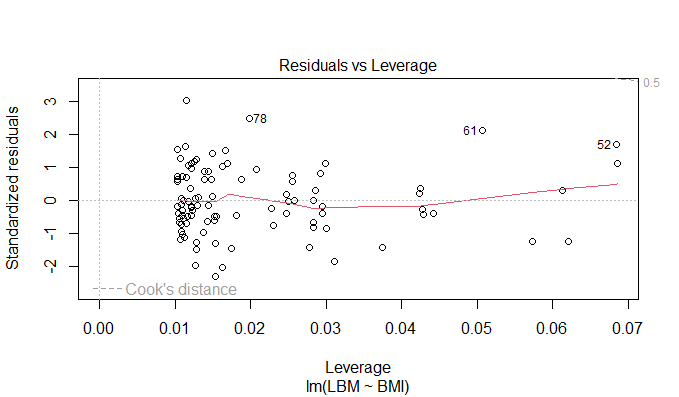
**Important Notes:**

The residuals show no discernible pattern, suggesting that homoscedasticity and linearity are probably satisfied.

The fact that residuals are dispersed throughout the plot and do not follow any particular pattern indicates that the model fits the data.

**In summary:**

The "Residuals vs. Fitted" plot suggests that the linearity and homoscedasticity model assumptions are probably met. This shows that your data are a good fit for the linear regression model, which yields accurate predictions.



The plot that displays the residuals (errors) for your linear regression model that predicts LBM based on BMI is interpreted by comparing it to the fitted values.

**Principal Findings:**

Since the residuals show no discernible pattern, linearity and homoscedasticity are most likely satisfied.

The absence of a clear trend in the residuals across the plot indicates that the model fits the data.

**In conclusion:**

The linearity and homoscedasticity assumptions of the model seem to be fairly met, based on the "Residuals vs. Fitted" plot. This shows that the linear regression model produces accurate predictions and is appropriate for your data.

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

between LBM and BMI.

|  |
| --- |
| Estimate Std. Error t value Pr(>|t|) |
| 3.356907e+00 2.983059e-01 1.125324e+01 3.582592e-19 |

Summary of the Linear Relationship between LBM and BMI: In male athletes, the analysis demonstrates a significant linear relationship between LBM and BMI.

**Important Points:**

* Estimate: LBM is predicted to rise by roughly 3.357 units for every unit increase in BMI.
* Standard Error: This shows that the estimated relationship has a moderate level of precision. The presence of a significant relationship is strongly supported by both the t-value and the p-value.

In summary, there is substantial evidence from the analysis indicating a linear relationship between BMI and LBM in this population.

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

1

79.565

Estimated LBM for Male with BMI 25: The model projects an LBM of roughly 79.57 units for a male athlete with a BMI of 25. This indicates that a male having a BMI of 25 is likely to have a lean body mass of about 79.57 units based on the relationship between LBM and BMI seen in the data used to construct the model. Keep in mind that these are estimates, and there might be individual differences.

6. Assess the predictive performance of the model.

[1] 80.142

**Meaning of Mean Squared Error (MSE)**

The model's ability to predict LBM based on BMI for male athletes is indicated by its MSE value of 80.14182. Here's a quick summary:

Better is a lower MSE: Better prediction accuracy is indicated by a lower MSE. To put it another way, the model's predictions tend to match the real LBM values. Interpretation is scale-dependent. Since MSE is the average squared difference between values that were predicted and those that were actual, how it is interpreted relies on the LBM units. It's in units squared here. 80.14182 In relation to Without knowing the typical range of LBM values, we are unable to interpret this number directly. A higher MSE could result from a wider range. Think about your objective: Whether this MSE is appropriate for you will depend on your needs. A lower mean square error (MSE) indicates a more accurate prediction.

**As a whole:**

While 80.14182 indicates that the model's predictions are fairly close to actual LBM values, the precise interpretation is contingent upon the desired level of accuracy and the context.

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.

Data <- X3343162SportsPeople\_Data

> head(Data)

# A tibble: 6 × 3

|  |
| --- |
| Sex LBM BMI |
| *<chr>* *<dbl>* *<dbl>* |
| 1 male 83.9 21.1 |
| 2 female 57.6 26.0 |
| 3 female 70.7 25.7 |
| 4 male 56.9 20.8 |
| 5 male 60.7 20.4 |
| 6 female 54.2 22.0 |

**1. EDA**

# Summary statistics

summary(Data)

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

#Convert sex to a factor

Data$Sex <- as.factor(Data$Sex)

# Assuming 'Sex' is a variable in the dataset

t\_test\_result <- t.test(LBM ~ Sex, data = Data)

# Print the results

print(t\_test\_result)

**3. Correlation Analysis for Males and Females:**

cor\_male <- cor(Data[Data$Sex == "male", c("LBM", "BMI")])

cor\_female <- cor(Data[Data$Sex == "female", c("LBM", "BMI")])

# Create a data frame for correlation coefficients

cor\_results <- data.frame(

Gender = c("Male", "Female"),

Correlation\_LBM\_BMI = c(cor\_male[1,2], cor\_female[1,2])

)

# Print the results

print(cor\_results)

**4. Model Development for Male Sportspeople:**

**(b) Description of the model:**

# Create a subset for male sportspeople

male\_data <- Data[Data$Sex == "male", ]

# Fit a linear regression model

model\_male <- lm(LBM ~ BMI, data = male\_data)

**(c) Summary of the fitted model:**

# Summary of the fitted model

summary(model\_male)

**(d) Evidence of assumptions:**

# Check assumptions (e.g., residuals)

plot(model\_male)

**(e) Conduct a formal test for a significant linear relationship:**

# Formal test for linear relationship

linear\_test\_male <- summary(model\_male)$coefficients["BMI", c("Estimate", "Std. Error", "t value", "Pr(>|t|)")]

linear\_test\_male

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

# New data for prediction

new\_data\_male <- data.frame(BMI = 25)

# Predict LBM for a male with BMI 25

predicted\_LBM\_male <- predict(model\_male, newdata = new\_data\_male)

**6. Assess the predictive performance of the model.**

# Assuming Data is your original dataset

set.seed(123) # Set seed for reproducibility

indices <- sample(1:nrow(Data), 0.8 \* nrow(Data)) # 80% for training, 20% for testing

# Create training and testing sets

train\_data <- Data[indices, ]

test\_data <- Data[-indices, ]

# Fit the model on the training set

model\_train <- lm(LBM ~ BMI, data = train\_data)

# Predict LBM on the testing set

predicted\_LBM\_test <- predict(model\_train, newdata = test\_data)

# Calculate Mean Squared Error (MSE)

mse\_test <- mean((test\_data$LBM - predicted\_LBM\_test)^2)

# Print or view the MSE for testing set

print(mse\_test)

References.

* R Programming Tutorial - Learn the Basics of Statistical Computing from freeCodeCamp.org
* Youtube Video: R package reviews | sjPlot | Easily Visualize Data And Model Results yuzaR Data Science