University of Stirling
Computing Science & Mathematics

MATPMDA
2023

# MATPMDA MATHEMATICAL AND STATISTICAL FOUNDATIONS

**PROJECT : AUTUMN SEMESTER 2023**
**Submission due 18th December 17:00**

**Student Number: 3074863**

**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.

**My Dataset consist of Three Variable in two type of Data.**

**Sex:** This is a Categorical Data, making up of both male and female which can be refer as Nominal Variable.
**LBM:** This fall on Quantitative Data, which can be Continuous Variable, they are in the form of Numerical Variable.
**BMI:** This also fall on Quantitative type of Data, which are Continuous Variable, that are in form of a Numerical Variable.

**STEP For the exploratory analysis I took**
I install the necessary packages and load my Dataset
View the head and bottom, check the column, dimension and structure of my Dataset.
I check if there is any missing value.
I checked for the summary statistics to know it mean, median standard deviation and range
I visualize with histogram to know it distribution and boxplot to visualize the relationship between the sex variable of both LBM & BMI.

**RESULT:**
**Statistical Summary**

**LBM (Lean Body Mass):**
Count: 178
Mean: 62.40

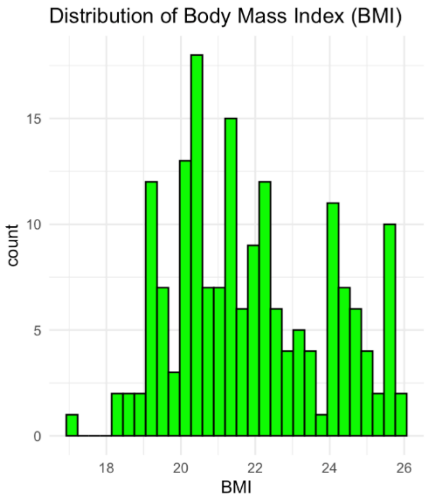
Continued overleaf…

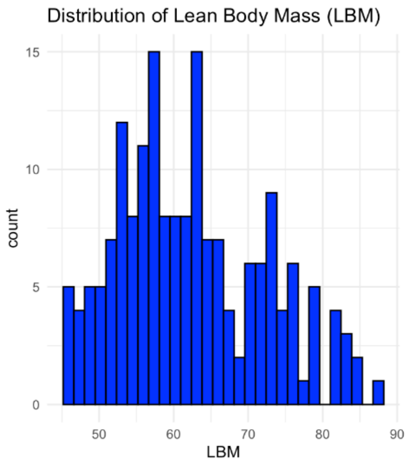
Standard Deviation: 9.85
Minimum: 45.47
25th Percentile: 55.01
Median: 60.98
75th Percentile: 70.10
Maximum: 87.06

**BMI (Body Mass Index):**
Count: 178
Mean: 21.89
Standard Deviation: 2.03
Minimum: 17.22
25th Percentile: 20.31
Median: 21.51
75th Percentile: 23.54
Maximum: 26.07

**Sex Variable Count**
Males is 97 and females is 81 Sport people.

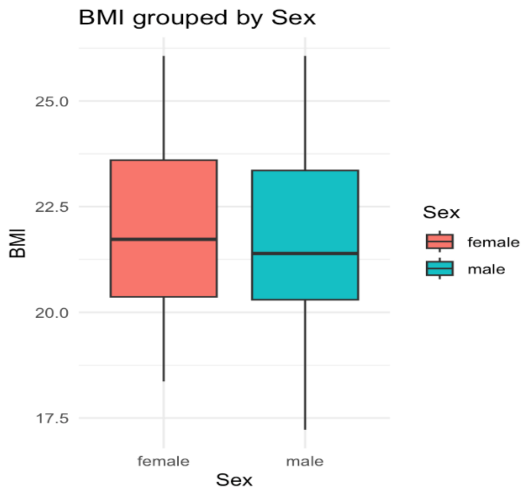
**Histogram Distribution**

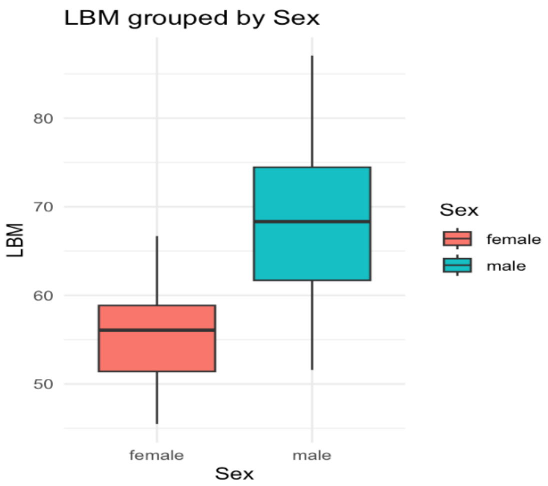




The distribution of BMI showed to be slightly normally distributed, it zenith point around 21, while the LBM tend to also be normally slightly skewed in it distribution and the visualization shows a zenith point around 60.

**Boxplot**





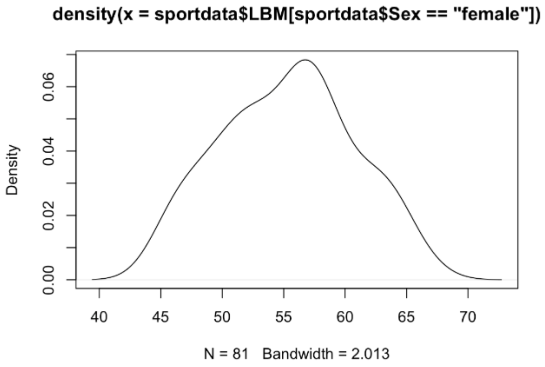
The Boxplot shows that BMI for male and female are closely related with little differences, whereas for LBM is not closely related for male and female

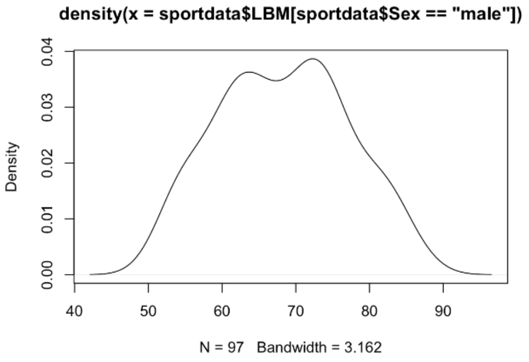
2. Using an appropriate statistical test, investigate whether there is a difference in mean LBM between males and females.

**Answer:**

I used Anderson-Darling test for normality of LBM in males and also in female.

The p-value of LBM male is 0.1995, so there is no evidence to suggest it isn't normally distributed.





The plot density in both male and female shows there is no evidence to suggest that this samples does not come from a Normal distribution. The plot density help in visualize in line that it is normally distributed.

I then Test if the variances of LBM for males and females are equal.

F-statistics giving me my degree of freedom F = 2.651, degree of freedom 96 is numerator and 80 in denominator and the p-value = 0.00001.141, which is less than 0.05. Assumption on equality on variance are not correct and we suggest the variance are not equal.

**We then perform Welch Two Sample t-test**

## The sample estimates:

Mean in group female is 55.308
Mean in group male is 68.328

The Welch t-test result shows a statistically significant difference in LBM between female and male sport people, it shows that male have higher average LBM than females. This let out know that it is substantial and statistically significant.

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.

**Answer**

Correlation of males is 0.710
Correlation of females is 0.532
This shows that BMI and LBM are positively correlated in both male and female, this also suggest that individual with higher BMI tend to have higher LBM. However with male having a stronger correlation of 0.710 and female having a moderate correlation of 0.532. This can indicate that BMI might be a better predictor of LBM in male compared to female sport person.

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.

## Answer

4 (a) The correlation of male 0.710 and for females 0.532, if we include data for females in the same model as males might not be advisable because the relationship dynamic between LBM and BMI are different in both males and females.

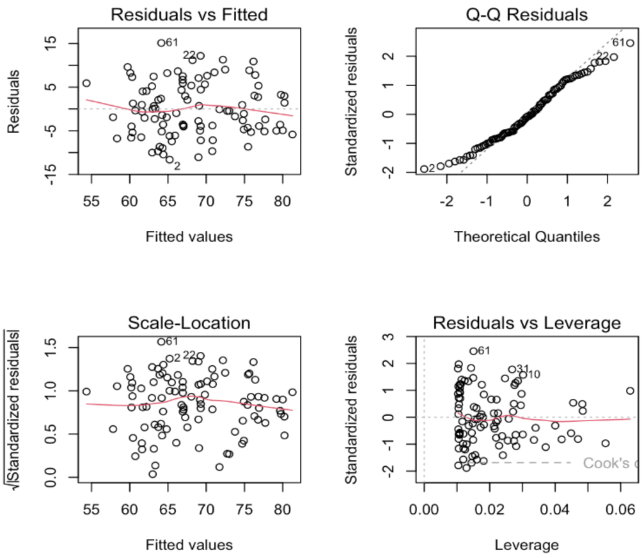
4 (b) A simple Linear Regression
$y = \alpha + \beta x + \epsilon$

The description of my model of linear regression are
$LBM=\beta 0 +\beta 1 \times BMI + \epsilon$
Where;
LBM is the dependent or response variable.
β0 = Intercept
β1 is the coefficient of BMI.

BMI is the independent, explanatory or predictor variable.
ϵ is the error

4 (c) the output of the F-statistic is 96.35 and the p- value of (4.088e-16) which is lesser than 0.05, indicating that the model is statistically significant.
This model suggest that there is a significant relationship between BMI and LBM in male sport person and it also shows that BMI is a strong predictor of LBM.

4 (d)



QQ Plot Residuals shows that the residual are normally distributed.
and the Scale location and Residual vs fitted both shows no clear patterns suggesting the homoscedasticity & linearity assumption is reasonable.
Residual vs leverage plot shows no point of high leverage or residuals.
All this suggest that linear regression model of LBM & BMI for the male sport person is well fitted with residual of the assumption underlying linear regression.

4(e)
My ANOVA table for the model shows that there is a statistically significant relationship between BMI and LBM. The high F value (96.351) suggests that BMI is a strong predictor of LBM and the p-value (4.088e-16) which is less than 0.05 indicate evidence against the null hypothesis

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

**Answer: 78.026 kilograms**
If I enter a BMI of 25 into the model. My model prediction for the sport person LBM would be 73.026, this my prediction is based on the relationship that exist between BMI and LBM in my determined model.

6. Assess the predictive performance of the model.

**Answer**
RMSE is 6.148
MAE is 5.172
RMSE and MAE provide measures of the model's predictive accuracy, and a Lower values indicate better performance.
This value help me in know my model accuracy, based on my value of RMSE of 6.148kg and MAE of 5.172kg, my model seems to have a reasonable performance in predicting LBM from BMI for male sportspeople.

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.

```
# We are getting some tools to help us read and draw.
# Install necessary packages and libraries

install.packages("readxl")
install.packages("tidyverse")
install.packages("nortest")
install.packages("Metrics")
library(Metrics)
library(nortest)
library(dplyr)
library(ggplot2)
library(readxl)

# Load our SportData dataset
sportdata <- read_excel("3074863SportsPeople Data.xlsx")

#To view the top 6 of our dataset
head(sportdata)

# To view the bottom 6 of our dataset
tail(sportdata)

# To see all our column name in our dataset
colnames(sportdata)

# To show the Shape our dataset both row and column
dim(sportdata)

# To view the Structure of our dataset
str(sportdata)

# Check if there is missing values in each column
```

```r
missing_value <- sapply(sportdata, function(x) sum(is.na(x)))
# Printing out to check if there is any missing value
cat("Number of missing values in each of the column:", missing_value, "\n")

# To know the number of male and female sportpeople in our distribution
sex_distribution <- sportdata %>% count(Sex)
print(sex_distribution)

# View the statistical Summary of our Dataset
summary(sportdata)

# Histogram for LBM
ggplot(sportdata, aes(x = LBM)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  theme_minimal() +
  ggtitle("Distribution of Lean Body Mass (LBM)")

# Histogram for BMI
ggplot(sportdata, aes(x = BMI)) +
  geom_histogram(bins = 30, fill = "green", color = "black") +
  theme_minimal() +
  ggtitle("Distribution of Body Mass Index (BMI)")

# Boxplot for LBM by Sex
ggplot(sportdata, aes(x = Sex, y = LBM, fill = Sex)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("LBM grouped by Sex")

# Boxplot for BMI by Sex
ggplot(sportdata, aes(x = Sex, y = BMI, fill = Sex)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("BMI grouped by Sex")

# Anderson-Darling test for normality of LBM in males
ad.test(sportdata$LBM[sportdata$Sex == "male"])

# Histogram, QQnorm and QQline to show that it is normally distributed
hist(sportdata$LBM[sportdata$Sex == "male"])
qqnorm(sportdata$LBM[sportdata$Sex == "male"])
plot(density(sportdata$LBM[sportdata$Sex == "male"]))

# Anderson-Darling test for normality of LBM in females
ad.test(density(sportdata$LBM[sportdata$Sex == "female"]))

# Histogram, QQnorm and plot density to show that it is normally distributed
hist(sportdata$LBM[sportdata$Sex == "female"])
qqnorm(sportdata$LBM[sportdata$Sex == "female"])
plot(density(sportdata$LBM[sportdata$Sex == "female"]))
```

```
# Test if the variances of LBM for males and females are equal
var.test(sportdata$LBM[sportdata$Sex == "male"], sportdata$LBM[sportdata$Sex ==
"female"])

# Welch's t-test to compare mean LBM between males and females
t.test(LBM ~ Sex, data = sportdata, var.equal = FALSE)


# Correlation for males both LBM & BMI
corr_male <- cor(sportdata$LBM[sportdata$Sex == "male"],
              sportdata$BMI[sportdata$Sex == "male"])

# Correlation for females both LBM & BMI
corr_female <- cor(sportdata$LBM[sportdata$Sex == "female"],
                sportdata$BMI[sportdata$Sex == "female"])

# Print out the correlation using cat()
cat("Correlation coefficient for males:", corr_male, "\n")
# Printing out the female correlation coefficient
cat("Correlation coefficient for females:", corr_female, "\n")



# Modeling relationship between LBM and BMI for Males
model_male <- lm(LBM ~ BMI, data = sportdata[sportdata$Sex == "male",])
# The summary of the model is called using the summary()
summary(model_male)

#we can plot the residuals against the fitted values.
par(mfrow = c(2, 2))
plot(model_male)


#To conduct a formal test to question whether there is a significant linear relationship
between LBM and BMI,
# we can use an F-test. The results of the F-test are shown below:
# Formal test for significance of the overall model
anova(model_male)

# To predict the LBM for a male with a BMI of 25, we can use the model we fitted in
Question 4
predicted_lbm <- predict(model_male, newdata = data.frame(BMI = 25))
predicted_lbm


# Separating data by gender
male_data <- subset(sportdata, Sex == 'male')
female_data <- subset(sportdata, Sex == 'female')
# Predictions
```

```
predictions <- predict(model_male, newdata = male_data)

# RMSE and MAE provide measures of the model's predictive accuracy.
# Lower values indicate better performance.
# RMSE
rmse_value <- rmse(male_data$LBM, predictions)
rmse_value

# MAE
mae_value <- mae(male_data$LBM, predictions)
mae_value
```

**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.

1) How to Perform Exploratory Data Analysis in R (With Example) in R (n.d.) Available at: https://www.statology.org/exploratory-data-analysis-in-r/ (Accessed: 14 December 2023).

2) Mindrila, D. and Balentyne, P. (n.d.) Linear Regression Notes. Available at: https://www.westga.edu/academics/research/vrc/assets/docs/linear_regression_notes.pdf (Accessed: 18 December 2023).

3) R Programming A-Z™: R for Data Science with Real Exercises', taught by Kirill Eremenko, available on Udemy. Accessed on December 16, 2023. URL: https://www.udemy.com/course/r-programming/learn/lecture/4652008#overview.

4) Nora, T. (2023) 'Correlation and Linear Regression', MATPMDA, University of Stirling. Available at: https://canvas.stir.ac.uk/courses/14154/pages/session-9 (Accessed: 15 December 2023).

5) Nora, T. (2023) 'Chi-squared test', MATPMDA, University of Stirling. Available at: https://canvas.stir.ac.uk/courses/14154/pages/session-10 (Accessed: 15 December 2023).

6) Nora, T. (2023) 'Hypothesis testing; t-tests', MATPMDA, University of Stirling. Available at: https://canvas.stir.ac.uk/courses/14154/pages/session-7 (Accessed: 15 December 2023).

7) Nora, T. (2023) 'Visualising Data', MATPMDA, University of Stirling. Available at: https://canvas.stir.ac.uk/courses/14154/pages/session-6 (Accessed: 15 December 2023).

8) Nora, T. (2023) 'One-way ANOVA; correlation', MATPMDA, University of Stirling. Available at: https://canvas.stir.ac.uk/courses/14154/pages/session-8 (Accessed: 15 December 2023).