R PROJECT: Athlete Data Analysis, Modelling and Predicting

By: Rodolfo Croes
Date: December 18th 2023
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.

After reading in the data into R, I used the dim() function to see the dimensions of the data set and it shows that there are 178 rows and 3 columns of data. Calling the head() function displays the column names and the first few data in the column. From this I can see that there are 3 variables: Sex, BMI and LBM.
Using the n_distinct() function, I saw that there are only 2 unique variables in Sex, being 'male' and 'female'. Using this I summed the amount of appearances each had and added them together to see if the number of appearances in the variable Sex is equal to the amount of rows. This is the case thus there is no missing data on this column. Seeing that the Sex variables cannot be ranked, this means that Sex is a Nominal Categorical variable.
BMI and LBM are numerical and this variable that can range by person because it depends on their weight. This means that both BMI and LBM are Continuous Quantitative Variables. I checked for missing values in BMI and LBM using is.na() function to check for any NA's in the data set. This returned 0 for both thus there is no missing data for this dataset. Using the summary() function we can get the five number summary, the mean, variance and standard deviation of both the BMI and the LBM with these being displayed in the table below:

Numerical Summaries	BMI	LBM
Min	17.014	38.947
Q1	20.273	54.408
Median	21.998	59.557
Q3	23.185	68.920
Max	26.160	100.693
Mean	21.811	61.737
Variance	4.477	111.537
Standard Deviation	2.116	10.561

After this, I plotted the histograms and boxplots of the BMI and LBM to see if there are any outliers in the data. From the visualization, there are no outliers in the BMI data. However, in the LBM data there is an outlier present based on the boxplot, this being the max of LBM

Continued overleaf…

which is 100.693 kg's. Comparing the histogram and stem plots of BMI and LBM, we can see that the LBM is slightly more positively skewed than the BMI while the BMI has a less symmetric shape than the LBM. The steps that I used to perform the exploratory data analysis on this dataset were inspired by Statology and the Junyper notebooks given in session 6 of this course [1][2].

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

To find a difference between the mean LBM of males and females, I used the T-test with two independent groups with a 5% significance level. I started with the null hypothesis being that the mean LBM for males and the mean LBM for females are equal. The alternative hypothesis of this is that the mean LBM for males and females are not equal. Then I handled it as a two-tailed distribution given that we are looking for a difference between two means. Before I can perform the test, according to the lecture notes, I must test if the data is normally distributed using the Anderson Darling test [4]. For the Anderson Darling test, both the male and female populations had a p value of greater than 0.05 with male being 0.165 and female being 0.397 thus the data for both are normal. Because the data is normal, we do not need to apply the Levene's test given that this is used when the data comes from a non-normal distribution [3]. To start the T test, we calculate the mean, standard deviation and number of observations for the male and female present in our data. Then using this we find the observed test statistic which is 9.822. Then using this we find the p value for one side of the distribution and then double it at the end to account for this test being two-tailed. Seeing that the observed test statistic is a positive number, we can use this to call pt() when finding the p-value. The p-value for this statistic is 1.98e-18 which is less than 0.05. Thus, there is sufficient evidence to reject the null hypothesis at a 5% level and because of this there is a difference between the means of the males and females LBM.

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.

I started this by making a scatterplot of the relationship between the LBM and BMI of both the males and females separately. Seeing that these two datasets are normalized given the answer from question 2, we can use the Pearson corelation coefficient for this relationship. Doing this shows that both are positively related to one another. Finding the corelation coefficients we see that the males have a coefficient of 0.757 and the females have a coefficient of 0.602. These numbers indicate that the relationship between the LBM and BMI for both males and females are positive. But the relation of LBM and BMI for males are stronger than the relationship of LBM and BMI for the 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;

Continued overleaf…

(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.

I have chosen to use a linear regression model to test the relationship between the LBM and BMI because the assumptions found in our notebooks line up with the variables chosen for the model [5]. These assumptions being that the LBM is Normally distributed, the outlier(s) in the BMI are negligible given that the corelation coefficient is still valid, and the variance of LBM is constant. Seeing that the statistical summaries of the female's data is significantly different than the males, I have chosen to not include it when fitting my model. Also, the dependant variable is the LBM, and the independent variable is the BMI. Testing the model shows that the residuals is almost symmetrically listed around 0. The only minor exception is with the Min being -15.368 and the Max being 17.587. The R-squared value for this model is 0.573 which means that 57.3% of the variability in the BMI is explained by the regression on the LBM. The slope of this model is 3.570 and the p value of this slope is less than 2e-16. Because the p value is less than 0.05, then there is sufficient evidence to suggest that the slope is significantly different from 0. This means that there is a linear relationship between the LBM and the BMI.

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

With the model we developed and fitted in the previous question, it has predicted that a male sports person with a BMI of 25 has an LBM of 78.965 Kg's.

6. Assess the predictive performance of the model.

When assessing my model, I looked at the normality of it by using the Anderson-Darling test. The results for this are a p-value of 0.801. Because this value is greater than 0.05, the data that the model predicted is normally distributed. After this I plotted the fitted values compared to the actual values and from the plot, I can see that the data is a bit dispersed but still trending in the positive direction against the diagonal line. In conclusion, the model performs with good accuracy, but the precision of the predicted LBM could be better. A suggestion for this could be to use more data or to find a different prediction model.

Continued overleaf…

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] Zach, “How to Perform Exploratory Data Analysis in R (With Example),” Statology, https://www.statology.org/exploratory-data-analysis-in-r/ (accessed Dec. 16, 2023).

[2] University of Stirling FA23 MATPMDA Session 6 Jupyter R Notebooks (Chapter 7 & 8)

[3] “Levene’s test in R programming,” GeeksforGeeks, https://www.geeksforgeeks.org/levenes-test-in-r-programming/ (accessed Dec. 16, 2023).

[4] Zach, “How to conduct an Anderson-Darling test in R,” Statology, https://www.statology.org/anderson-darling-test-r/ (accessed Dec. 16, 2023).

[5] MATPMDA Session 8 Jupyter R Notebooks (Chapter 12 Section 12.2.2)

I declare that I did not use any kind of generative AI technology or online tool to produce this project.

Continued overleaf…