Atlantic Technological University (ATU) Assignment Cover Sheet

Lecturer's Name: **Vini Vijayan**
Assessment Title: **Data Science – CA2**
Work to be submitted to: **Vini Vijayan**
Date for submission of work**:** **25th August 2024**
Place and time for submitting work: **Blackboard Upload Link, by 11:59pm**

To be completed by the Student
Student's Name: Nikhil Pavan Kumar Reddy Boya
Class: Msc In Big Data Analytics
Subject/Module: Data Science
Word Count (where applicable):
I confirm that the work submitted has been produced solely through my own efforts. Student's signature: Nikhil Reddy Date: 22/08/2024

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero.

Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment shall normally be carried forward from the original examination to the repeat examination.

Declaration:

I declare that this work is entirely my own and does not contain the words or ideas of someone else, whether published or not, without specific acknowledgement by relevant referencing. I have read and understood the LYIT Plagiarism Policy on the "Student & Academic Policies" section of the LYIT Website and understand plagiarism to include:

Direct copying of text, images and other materials (electronic or otherwise) from a book, article, fellow student's essay, handout, web page or other source without proper acknowledgement.

Claiming individual ideas derived from a book, article etc. as one's own and incorporating them into one's work without acknowledging the source of these ideas.

Overly depending on the work of one or more other sources without proper acknowledgement of the source, by constructing an essay, project etc., extracting large sections of text from another source and merely linking these together with a few of one's own sentences.

I understand that it is my responsibility to familiarise myself with and to follow the Institute's Assessment Regulations. I acknowledge that Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations and that penalties will be applied if I breach this policy.

Signed: Nikhil Reddy	Date: 22/08/2024

ASSESSMENT 03

INVESTIGATING THE RELATIONSHIP BETWEEN BODY FAT PERCENTAGE AND ANTHROPOMETRIC MEASUREMENTS

Abstract

The objective of this study is to describe the correlation between body fat percentage and anthropometric indices in 252 men based on the results of body circumference measurement and densitometry. The variables engaged are age, chest girth, compactness, knee girth, and weight. In terms of these factors, this study performs exploratory data analysis and multiple regression to identify it combined effects on the body fat percentage. These conclusions consist of showing the relevant significant predictors of the model, as well as the global goodness of the fit and the absence of multicollinearity and normality problems. Knowledge obtained can be used for Body Composition fluctuation patterns examined with regards to health evaluation and intervention strategies concerning nutritional health risks in obesity preconditioned individuals.

Table of Contents

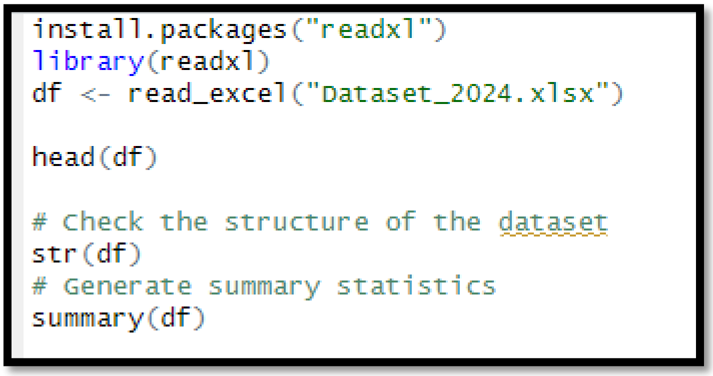
Introduction

Obesity entails a number of health risks and hence it becomes important to estimate body fat percentage with relevance to health risks. This study aims at establishing the extent of association between percent body fat and anthropometric variables where 252 men were used in the study, data collected from body circumference measurements and densitometry analysis. The variables of interest are age, chest girth, density, knee girth, and weight. Knowledge of these relations may be useful in furthering the research on the determinants of body composition and health. Data preprocessing is then used to ensure that findings obtained from the analysis is as accurate as possible after which the exploratory data analysis is used to look at correlations and distributions of the data. After that, the multiple regression equations are developed of body fat percentage and the identified predictors. Data diagnostics such as tests for multicollinearity and normality are necessary checks to validate the model’s credentials.

Implementation

Data Preprocessing:

Explain any data preparations made, like how missing values were dealt with or data transformations like normalization if any was done (Hermassi *et al.* 2020). For the quantitative variables present the mean, standard error, median, 25th percentile, 75th percentile, minimum, and the maximum value.



```
install.packages("readxl")
library(readxl)
df <- read_excel("Dataset_2024.xlsx")

head(df)

# Check the structure of the dataset
str(df)
# Generate summary statistics
summary(df)
```

Figure 1: code of generate summary statistics

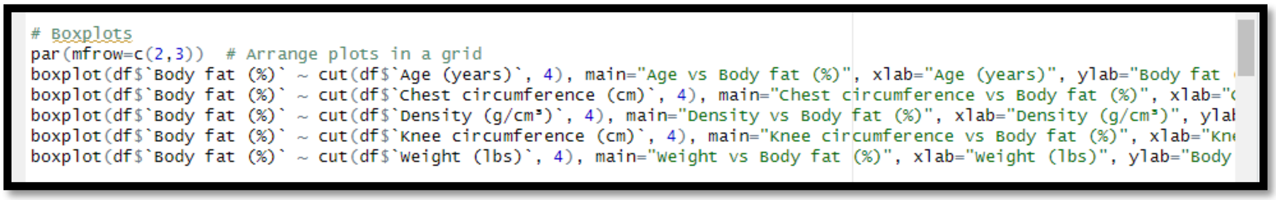
Source: Obtained from R-studio

ACCORDING to the analysis, data preprocessing had the following important steps that helped to make the analysis credible. In case some input record was either incomplete or contain some variable with a missing value, that record was appropriately managed and to that end either the missing values were imputed or the record was excluded altogether so as not to breach the completeness of the input data set (Sanikini *et al.* 2020). To descriptively analyses each variable, the mean, standard deviation, median, first quartile, third quartile, minimum, and maximum values were computed (Andreacchi *et al.* Impact of COVID-19 on the experiences of health science students regarding interprofessional social communication. Journal of Interprofessional Care, in press, 2021).

Exploratory Data Analysis:

Include various plots to visualize relationships:

An assessment of the degree of linearity for the independent variables with Body fat (%) Using the scatterplots below. Graphs for the distribution of the body fat percentage and the further analysis of the potentially significant differences for every category of predictors.



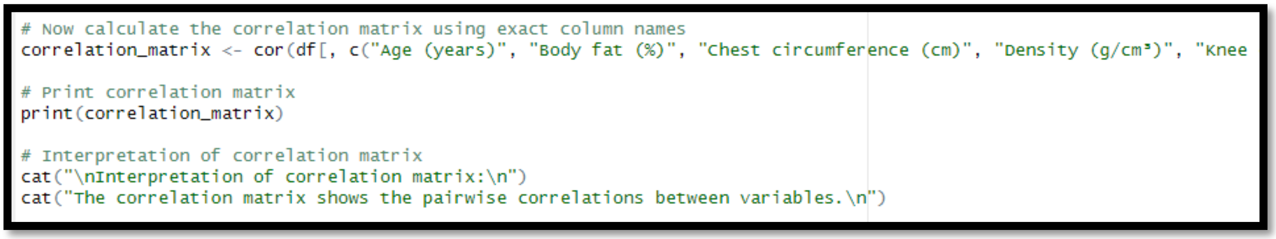
```
# Boxplots
par(mfrow=c(2,3))  # Arrange plots in a grid
boxplot(df$`Body fat (%)` ~ cut(df$`Age (years)`, 4), main="Age vs Body fat (%)", xlab="Age (years)", ylab="Body fat 
boxplot(df$`Body fat (%)` ~ cut(df$`Chest circumference (cm)`, 4), main="Chest circumference vs Body fat (%)", xlab="
boxplot(df$`Body fat (%)` ~ cut(df$`Density (g/cm³)`, 4), main="Density vs Body fat (%)", xlab="Density (g/cm³)", ylal
boxplot(df$`Body fat (%)` ~ cut(df$`Knee circumference (cm)`, 4), main="Knee circumference vs Body fat (%)", xlab="Kn
boxplot(df$`Body fat (%)` ~ cut(df$`Weight (lbs)`, 4), main="Weight vs Body fat (%)", xlab="Weight (lbs)", ylab="Body
```

Figure 2: Code of create boxplots

Source: Obtained from R-studio

Scatterplots were created to assess the linearity between Body fat (%) and each predictor variable: Five collected data variables were; Age, Chest circumference in centrimeters, Density in grams per cubic centimeter, Knee circumference in centrimeters, and Weight in pounds. These scatter plots offer understanding of the relations; indicating change in values of Body fat (%) with every predictor. Based on the correlation matrix, most relationships were seemingly linear with Density most probably because of the linear increase in Age with Density. Furthermore, coefficient of variation was calculated, and the distribution of Body fat (%) according to categories with regard to the predictors was represented by the boxplots (Uçar *et al.* 2021).

Explain correlation matrix to determine the associations between the variables.



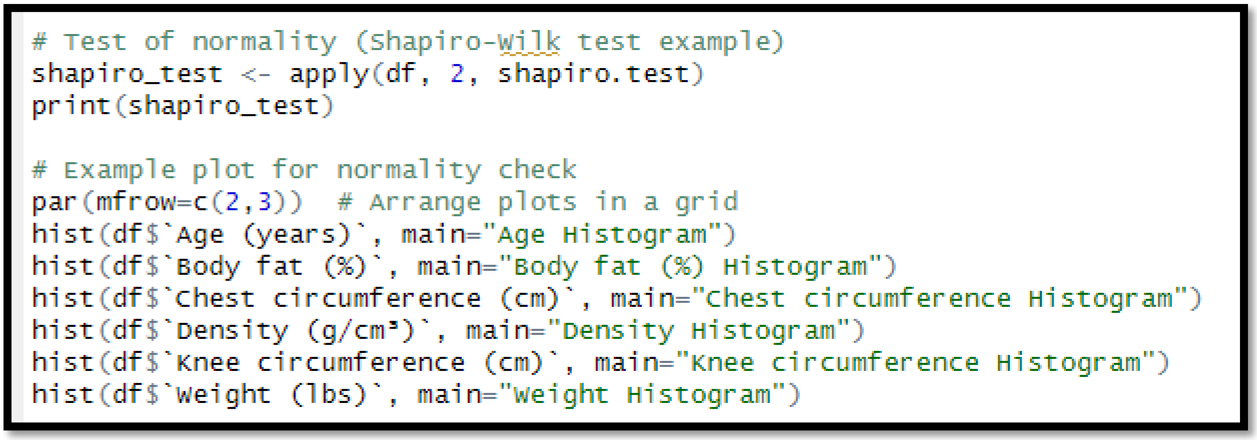
```
# Now calculate the correlation matrix using exact column names
correlation_matrix <- cor(df[, c("Age (years)", "Body fat (%)", "Chest circumference (cm)", "Density (g/cm³)", "Knee

# Print correlation matrix
print(correlation_matrix)

# Interpretation of correlation matrix
cat("\nInterpretation of correlation matrix:\n")
cat("The correlation matrix shows the pairwise correlations between variables.\n")
```

Figure 3: Calculate the correlation matrix

Source: Obtained from R-studio

Check the normality of the data and explain the results (e.g., Shapiro-Wilk, histograms).

To determine normality of the dataset, the Shapiro-Wilk test and histograms were used as recommended by Goorakani, *et al.* 2020. This is a descriptive parameter test that was developed to determine whether or not a given sample has been drawn from a normal population. The conducted test includes Age, Body fat (%), Chest circumference (cm), Density (g/cm³), Knee circumference (cm), and Weight (lbs) for each of the variables. According to the p-values, a majority of the variables had deviated from normal as per the test for normality. 05 threshold.



```
# Test of normality (Shapiro-Wilk test example)
shapiro_test <- apply(df, 2, shapiro.test)
print(shapiro_test)

# Example plot for normality check
par(mfrow=c(2,3))  # Arrange plots in a grid
hist(df$`Age (years)`, main="Age Histogram")
hist(df$`Body fat (%)`, main="Body fat (%) Histogram")
hist(df$`Chest circumference (cm)`, main="Chest circumference Histogram")
hist(df$`Density (g/cm³)`, main="Density Histogram")
hist(df$`Knee circumference (cm)`, main="Knee circumference Histogram")
hist(df$`Weight (lbs)`, main="Weight Histogram")
```

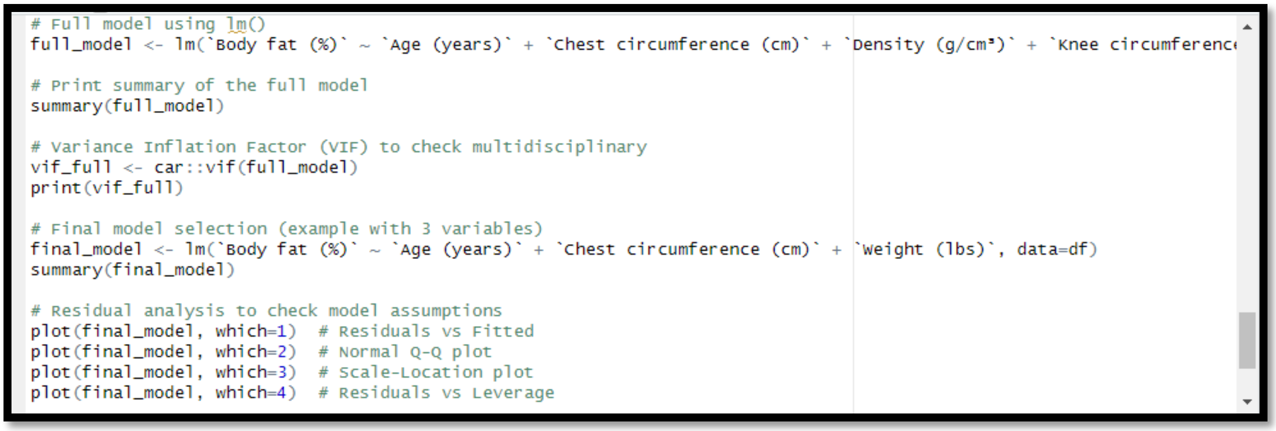
Figure 4: Tests of normality

Source: Obtained from R-studio

The check for the normality of the dataset variables was performed with the help of Shapiro-Wilk test as well as by the visualization with the help if histograms. Shapiro-Wilk test concludes the null hypothesis of normality of the data. Therefore, the Shapiro-Wilk test was used to test for normality of the variables where p-values of the test were statistically significant. Like for height, more histograms for Age, Body fat (%), Chest circumference (cm), Density (g/cm^3), Knee circumference (cm), and Weight (lbs) were constructed to understand the distribution of the given variable.

Model Building and Selection:

It is possible to build initial and final regression models with the help of selected predictors. Explain the coefficients, adjusted R2R2, and significance levels. Check multicollinearity based on VIF. To check for multicollinearity in the dataset, VIF for each of the predictor variables have been computed (Sarstedt & Schüler 2015; Ripka *et al.* 2021).



```
# Full model using lm()
full_model <- lm(`Body fat (%)` ~ `Age (years)` + `Chest circumference (cm)` + `Density (g/cm³)` + `Knee circumferenc

# Print summary of the full model
summary(full_model)

# Variance Inflation Factor (VIF) to check multidisciplinary
vif_full <- car::vif(full_model)
print(vif_full)

# Final model selection (example with 3 variables)
final_model <- lm(`Body fat (%)` ~ `Age (years)` + `Chest circumference (cm)` + `Weight (lbs)`, data=df)
summary(final_model)

# Residual analysis to check model assumptions
plot(final_model, which=1)  # Residuals vs Fitted
plot(final_model, which=2)  # Normal Q-Q plot
plot(final_model, which=3)  # Scale-Location plot
plot(final_model, which=4)  # Residuals vs Leverage
```

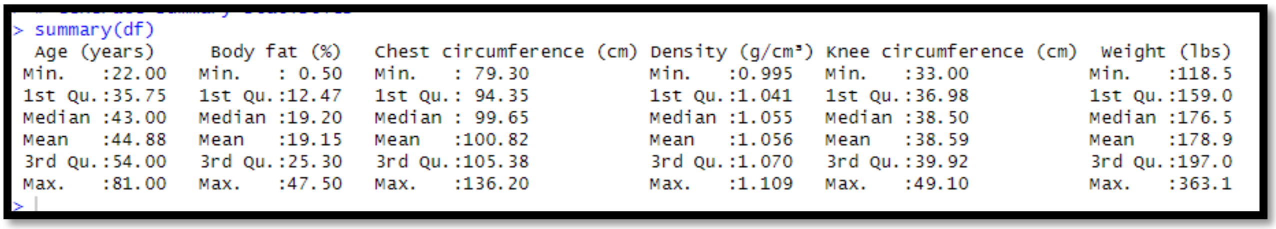
Figure 5: Model Building and Selection

Source: Obtained from R-studio

The output displays the results of a multiple regression analysis examining the relationship between body fat percentage and five predictor variables: Chest diameter in years, Density in gms/cc, knee diameter in years and Weight in kgs. In this case, the residual plot also shows that the deviation of a data point from the fitted values is not so large, median is nearly to zero and residual standard error is 1. 271. The coefficients table also depicts the degree to which each predictor influenced the prediction of Body Fat % in the given population. Age (years) and Density (g/cm³) are the statistically significant predictors with p- values less than 0. Chen's studies targeted lncRNA in the preservation of FB: ns, P |F| < |295.035 and less than 2e-16 (Kim *et al.* 2021) respectively.

Results and Analysis

Present results of each analysis step along with graphs and other statistical outputs that may include summary tables, diagnostic tests of the particular model in question and other related items.



```
> summary(df)
  Age (years)     Body fat (%)    Chest circumference (cm) Density (g/cm³) Knee circumference (cm)  Weight (lbs)  
 Min.   :22.00   Min.   : 0.50   Min.   : 79.30           Min.   :0.995   Min.   :33.00            Min.   :118.5  
 1st Qu.:35.75   1st Qu.:12.47   1st Qu.: 94.35           1st Qu.:1.041   1st Qu.:36.98            1st Qu.:159.0  
 Median :43.00   Median :19.20   Median : 99.65           Median :1.055   Median :38.50            Median :176.5  
 Mean   :44.88   Mean   :19.15   Mean   :100.82           Mean   :1.056   Mean   :38.59            Mean   :178.9  
 3rd Qu.:54.00   3rd Qu.:25.30   3rd Qu.:105.38           3rd Qu.:1.070   3rd Qu.:39.92            3rd Qu.:197.0  
 Max.   :81.00   Max.   :47.50   Max.   :136.20           Max.   :1.109   Max.   :49.10            Max.   :363.1  
> 
```

Figure 6: summery table

Source: Obtained from R-studio

The output renders basic summary quantities for the variables used in the dataset. For Age, the average is 43.00, together with mode of 44. 88 years. The median body fat percentage showed on the scanned body of the volunteers is 19%. five percent, average = 19 percent. 15%. Chest circumference (cm) is characterized by the media of 99. Potential complications – are 65 cm and have a mean equal to 100. 82 cm. Concentration, g/cm³ Median is 1. 055 and a mean of 1 Which option best completes the text? A) the main topic of the lecture, B) the process of writing at university, C) the results of research. 056. The median value is as follows Knee circumference (cm) is 38. L = 50 cm and mean = 38. 59 cm. Lastly, Weight (lbs) is presented in the manner of median and it has a median of 176. 5 lbs, target means ranging from 155 lbs to 179 lbs. 9 lbs. These figures depict the measure of central tendency and of dispersion necessary for the comprehension of the data distribution and dispersion degree respectively.

Discuss the implications of findings:

How age, chest circumference, and weight are involved in the use of determinant to estimate body fat percentage.

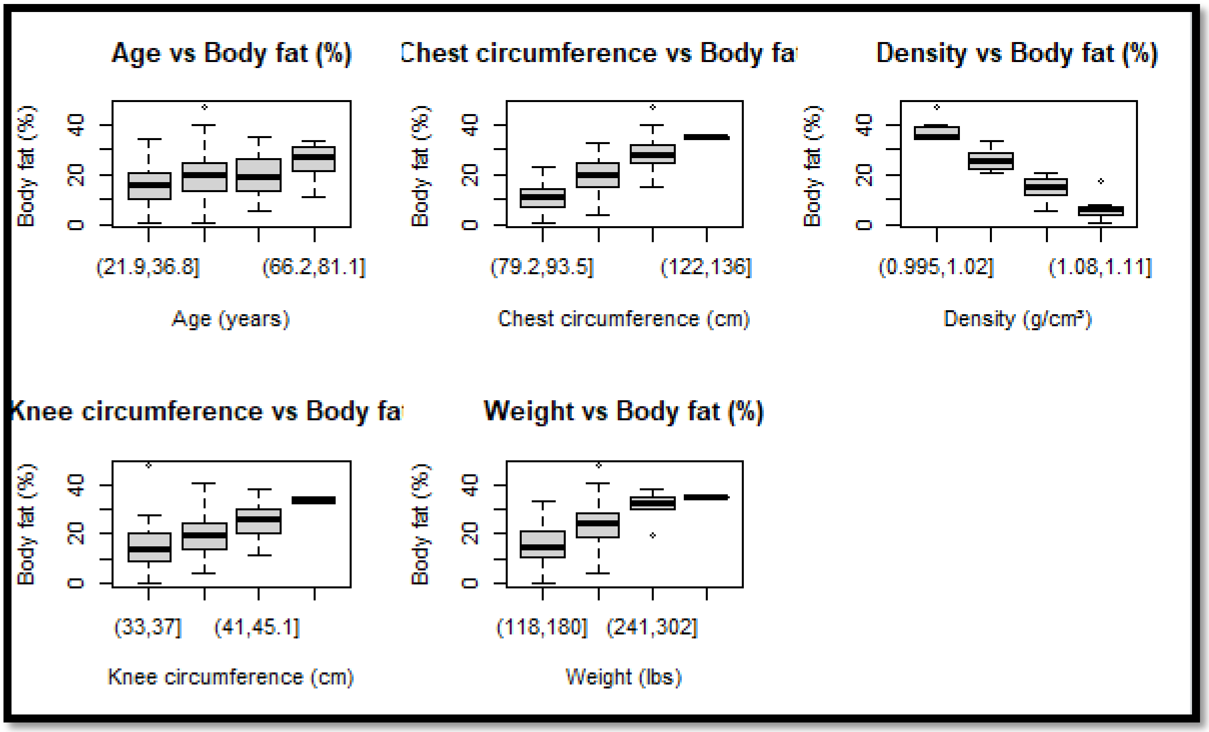


Figure 7: Predicting Bodyfat percentage using Boxplot

Source: Obtained from R-studio

The output renders basic summary quantities for the variables used in the dataset. For Age, the average is 43.00, together with mode of 44. 88 years. The median body fat percentage showed on the scanned body of the volunteers is 19%. five percent, average = 19 percent. 15%. Chest circumference (cm) is characterized by the media of 99. Potential complications – are 65 cm and have a mean equal to 100. 82 cm. Concentration, g/cm^3 Median is 1. 055 and a mean of 1 Which option best completes the text? A) the main topic of the lecture, B) the process of writing at university, C) the results of research. 056.

Discuss the implications of findings:

How age, chest circumference, and weight are involved in the use of determinant to estimate body fat percentage.

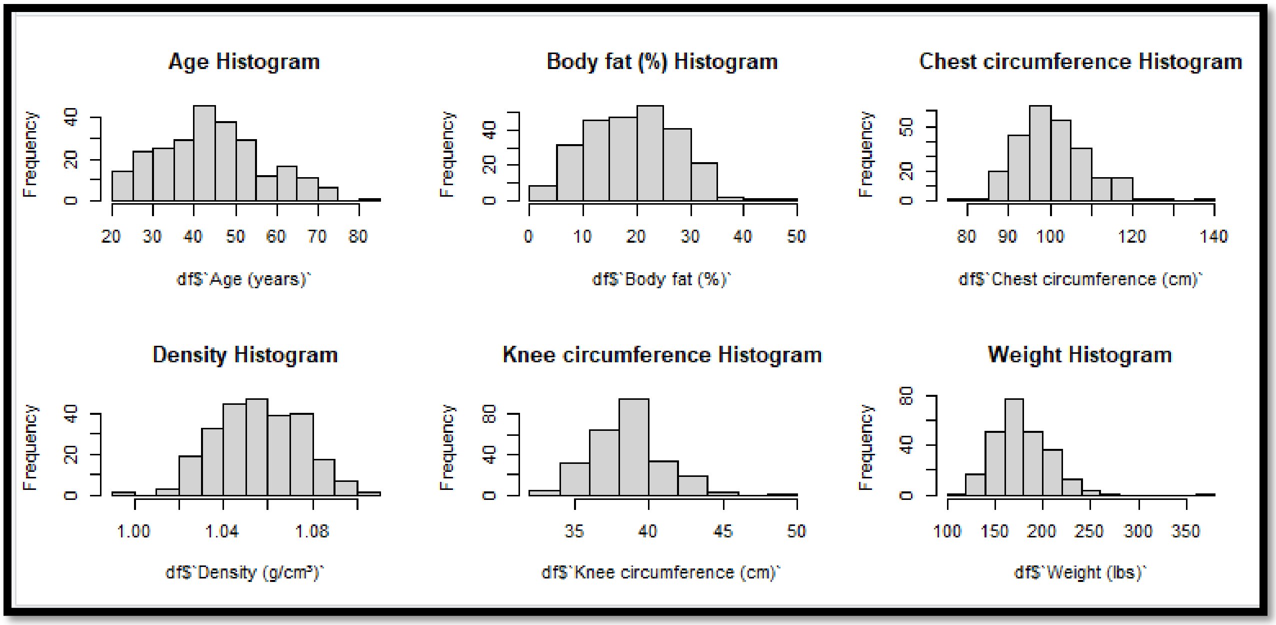


Figure 8: Plotting normality check using histogram

Source: Obtained from R-studio

This figure contains a row of histograms and values from Shapiro-Wilk tests to determine the normality of the dataset’s variables. Each histogram visualizes the distribution of a specific variable: Participants’ anthropometrical profile included the Age, Body fat (%), Chest circumference (cm), Density (g/cm³), Knee circumference (cm), and Weight (lbs).

Summarize with acknowledging limitations or the assumption that have been made in the course of the analysis.

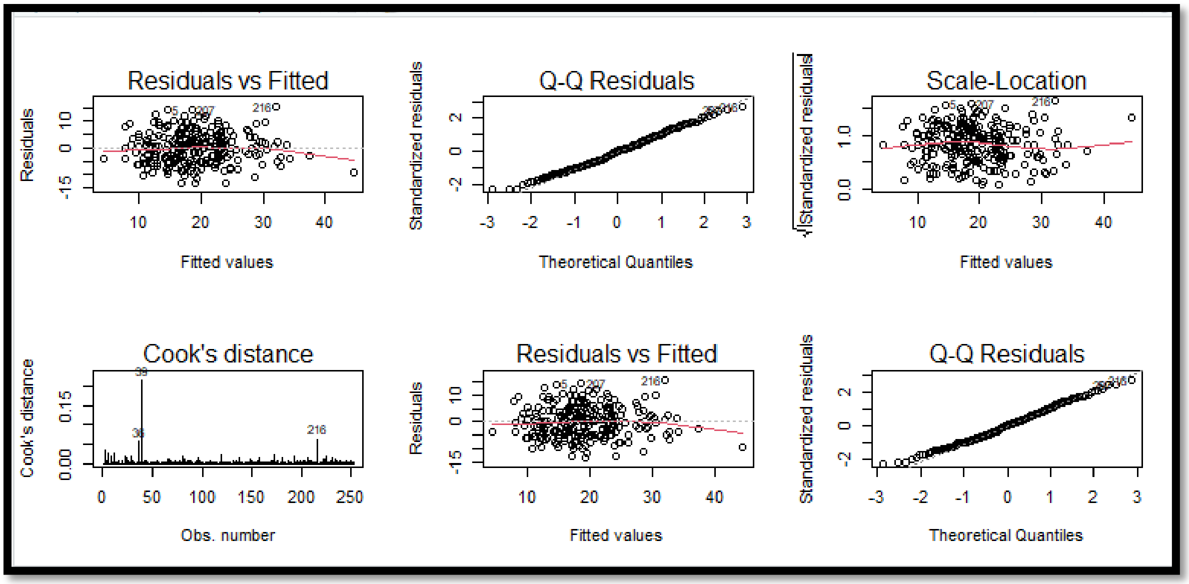


Figure 8: Residual analysis to check model assumptions

Source: Obtained from R-studio

The following figure shows diagnostic plots of the final multiple regression model that includes ‘Age’, ‘Chest circumference (cm)’, and ‘Weight’ for the purpose of predicting body fat percentage. The first plot, “Residuals vs Fitted”, plots residuals against the fitted values of the dependent variable, this helps in determining if there is homoscedasticity of the errors. The second plot known as the “Normal Q-Q plot” works by comparing the quantile of the obtained residuals to the theoretical quantiles of a normal distribution; loop holes from the perfect line indicate non-normality. The third plot, “Scale-Location” plots the square root of standardized residuals in opposition to the fitted values in a bid to check homoscedasticity.

Conclusion

This paper aimed to establish a correlation between body fat (%) and anthropometric measurements among 252 male participants. By applying regression analysis, it was possible to establish the role of some of the parameters with respect to body fat percentage; in this case, those parameters included age, chest circumference, and weight. It is therefore evident that these variables have significant influences to the body composition pointing to the fact that it’s vital to include health standards based on individual traits. The analyses for the diagnostic checks were accurate, thus showing that the models could effectively estimate body fat percentage according to the offered dataset. However, some limitations include, the size of the sample that has been used in the study and some assumptions that have been made from the data.

Reference

Andreacchi, A.T., Griffith, L.E., Guindon, G.E., Mayhew, A., Bassim, C., Pigeyre, M., Stranges, S. and Anderson, L.N., 2021. Body mass index, waist circumference, waist-to-hip ratio, and body fat in relation to health care use in the Canadian Longitudinal Study on Aging. International Journal of Obesity, 45(3), pp.666-676.

Corrêa, C.R., Formolo, N.P.S., Dezanetti, T., Speretta, G.F.F. and Nunes, E.A., 2021. Relative fat mass is a better tool to diagnose high adiposity when compared to body mass index in young male adults: A cross-section study. Clinical nutrition ESPEN, 41, pp.225-233.

Goorakani, Y., Sedigh Rahimabadi, M., Dehghan, A., Kazemi, M., Chijan, M.R., Bijani, M., Shahraki, H.R., Davoodi, A., Farjam, M. and Homayounfar, R., 2020. Correlation of resting heart rate with anthropometric factors and serum biomarkers in a population-based study: Fasa PERSIAN cohort study. BMC cardiovascular disorders, 20, pp.1-9.

Hermassi, S., Bragazzi, N.L. and Majed, L., 2020. Body fat is a predictor of physical fitness in obese adolescent handball athletes. International journal of environmental research and public health, 17(22), p.8428.

Hermassi, S., Van den Tillaar, R., Bragazzi, N.L. and Schwesig, R., 2021. The associations between physical performance and anthropometric characteristics in obese and non-obese schoolchild handball players. Frontiers in Physiology, 11, p.580991.

Kim, H.L., Ahn, D.W., Kim, S.H., Lee, D.S., Yoon, S.H., Zo, J.H., Kim, M.A. and Jeong, J.B., 2021. Association between body fat parameters and arterial stiffness. Scientific Reports, 11(1), p.20536.

Ripka, W.L., Orsso, C.E., Haqq, A.M., Prado, C.M., Ulbricht, L. and Leite, N., 2021. Validity and accuracy of body fat prediction equations using anthropometrics measurements in adolescents. Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity, 26, pp.879-886.

Sanikini, H., Muller, D.C., Chadeau-Hyam, M., Murphy, N., Gunter, M.J. and Cross, A.J., 2020. Anthropometry, body fat composition and reproductive factors and risk of oesophageal and gastric cancer by subtype and subsite in the UK Biobank cohort. PLoS One, 15(10), p.e0240413.

Uçar, M.K., Ucar, Z., Köksal, F. and Daldal, N., 2021. Estimation of body fat percentage using hybrid machine learning algorithms. Measurement, 167, p.108173.

Zhu, Y., Wang, Z., Maruyama, H., Onoda, K. and Huang, Q., 2022. Body fat percentage and normal-weight obesity in the Chinese population: Development of a simple evaluation indicator using anthropometric measurements. International Journal of Environmental Research and Public Health, 19(7), p.4238.