University of Stirling MATPMD0

Computing Science & Mathematics 2022

**MATPMD0 INTRODUCTORY STATISTICS FOR DATA SCIENCE**

**PROJECT : AUTUMN SEMESTER 2022**

**Submission due 19th 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 is the process of analyzing and visualizing data to produce new discoveries and a deeper understanding of the data. When performing EDA, data analysts follow these processes.

1. ****Summarizing**** a dateset using descriptive statistics.
2. ****Visualizing**** a dateset using charts.
3. ****Identifying**** missing values.

Dateset :dateset is the ordered collection of data.The dateset is statistically measured using its mean, median, mode, range, and other properties.

Description of variable

dateset contains 254 observations and 2 variables.Body fat and abdomen are numeric variables ,data is Quantitative .we can use abdomen as dependent variable and body fat as independent variable.

**Head of data**

**The first six rows of the dateset can be viewed.**

|  |  |  |
| --- | --- | --- |
|  | inb | da |
| 1 | 30.946 | 103.269 |
| 2 | 17.063 | 95.481 |
| 3 | 20.770 | 83.558 |
| 4 | 20.803 | 92.365 |
| 5 | 37.883 | 108.844 |
| 6 | 12.058 | 83.653 |

**Dimension of data**

This shows the rows and columns of your loaded dateset.

252 2

We can see that the dateset has **252**rows and ****2**** columns

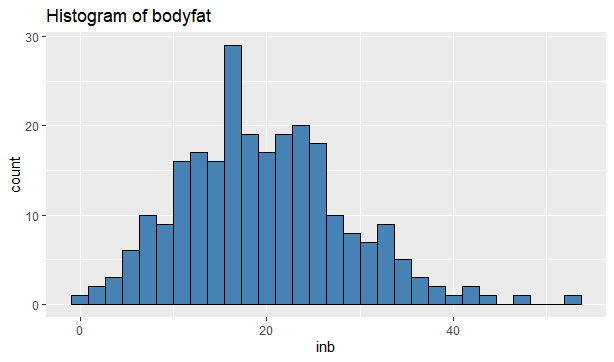
**Summary**

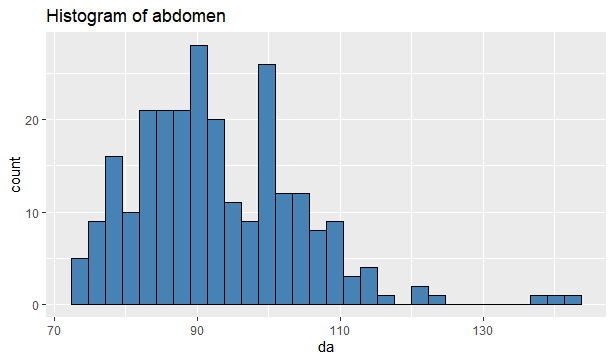
|  |  |  |
| --- | --- | --- |
|  | inb | da |
| Min. | 0.694 | 73.44 |
| 1st Qu.: | 13.586 | 85.08 |
| Median: | 18.720 | 90.86 |
| Mean | :19.772 | 92.97 |
| 3rd Qu.: | 25.171 | 99.84 |
| Max. | 53.491 | 142.33 |

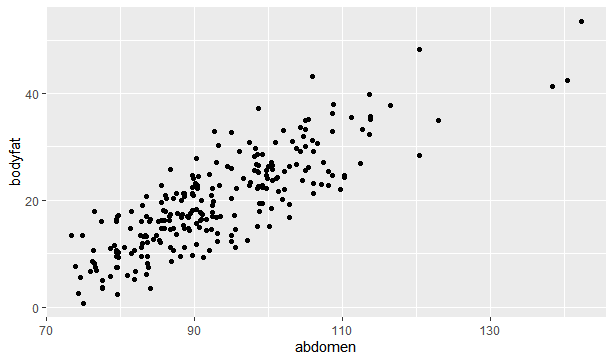
Summary is a very useful function that lists the highlights of each property in your dateset one by one. Each attribute's breakdown is listed in a table that is created by the function for each attribute.The difference between the 75th Q (upper quartile) and the 25th Q( lower quartile) is used to compute the inter quartile range . A variable's value in the middle of a set of data is its median value. The Arithmetic average (mean) is calculated by dividing the total number of observations by the sum of the values.

**Data visualization**

We can also create charts to visualize the values in the dateset.







Charts shows that data is normally distributed .

**Correlation of data**

We create a correlation matrix to view the [correlation coefficient](https://www.statology.org/pearson-correlation-coefficient/" \t "https://www.statology.org/exploratory-data-analysis-in-r/_blank) between each pairwise combination of numeric variables in the dateset

|  |  |  |
| --- | --- | --- |
|  | inb | da |
| inb | 1.0000000 | 0.8254362 |
| da | 0.8254362 | 1.0000000 |

**Identify missing values**

We can count the total number of missing values in each column of the dateset

|  |  |
| --- | --- |
| inb | da |
| 0 | 0 |

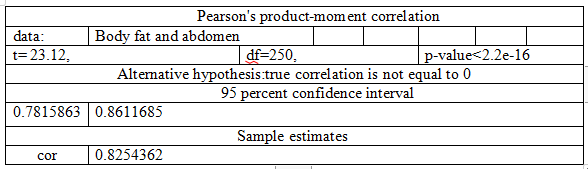
From the output we can see that there are zero missing values in each column.

1. **Calculate the correlation coefficient for the two variables given and comment on the relationship between body fat and abdomen size.**

Pearson correlation coefficient (r)is a most popular method to determine a linear relationship between two variables and the value is lie between -1 to+1.Pearson correlation is frequently used when there is a linear relationship between two numerical continuous variables.The correlation coefficient is computed using Cor() and Cor.test(). Test for association/correlation between matched samples with cor.test(). It provides the correlation's p-value as well as the correlation's significance level (or correlation coefficient).

**Correlation of body fat and abdomen**

0.8254362

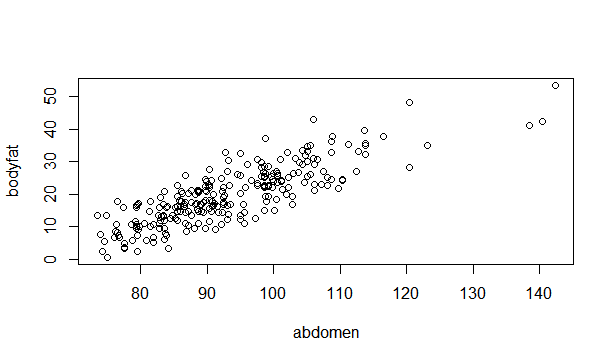


We can say that variables have a positive linear relationship because its value is close to 1.

Body fat and abdomen are correlated at 0.8254362. The test's p-value is 2.2e-16, which is below the threshold of significance set at 0.05. We can draw the conclusion that there is a strong link between abdomen and body fat.

**Visualization of data**

scatter plot is use to check the relationship between variable is linear or not .plot show that it falls almost along a straight line with positive slope



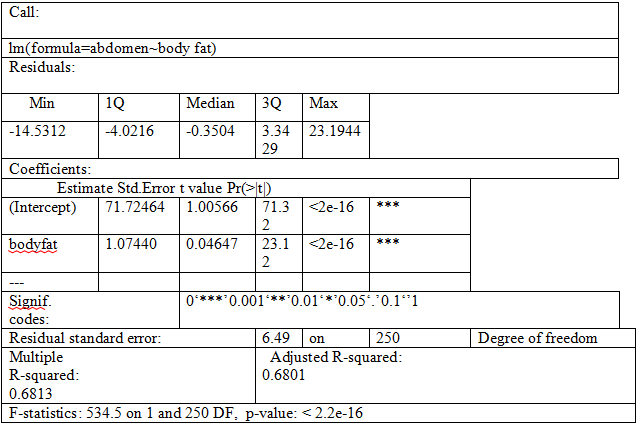
**3. Investigate a model to test the relationship between body fat and abdomen size. You must include output from R to support your findings. Details you should include are:**

* **a description of the model;**

**Regression modelling that uses a straight line to depict the relationship between variables is known as linear regression. It looks for the value of the regression coefficient(s) that minimizes the total model error to choose the line that best matches your data..Results includes median, mean, and maximum values of variables of body fat and abdomen here we have one dependent variable and one independent variable that’s why we use simple linear regression**

|  |  |  |
| --- | --- | --- |
|  | inb | da |
| Min. | 0.694 | 73.44 |
| 1st Qu.: | 13.586 | 85.08 |
| Median: | 18.720 | 90.86 |
| Mean | :19.772 | 92.97 |
| 3rd Qu.: | 25.171 | 99.84 |
| Max. | 53.491 | 142.33 |

* **a summary of the fitted model with interpretation of test statistics and parameter estimates;**

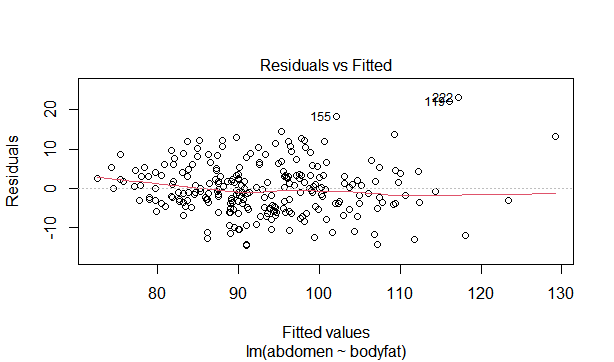


The estimated regression line equation looks like this abdomen = 71.72 + 1.07\*body fat. We calculate the model parameters, value of y- intercept is 71.72464, and the estimated impact of body fat on the abdomen is 1.07440. p value of model is 2.2e-16which is nearly zero shows that how well the model fits the data These findings allow us to reject null hypothesis and accept the alternative, which is that there is a strong positive correlation between body fat and abdomen.. Metrics used to assess how well the model fits our data include residual standard error (RSE), R-squared (R2), and the F-statistic.

* **evidence as to whether assumptions of the model have been met;**

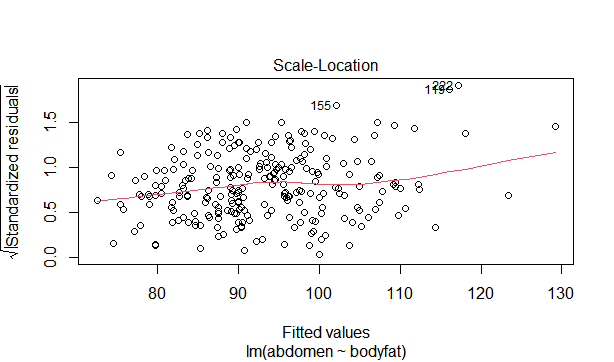
#### **DATA is linear**

we determine whether the data are linear is not by examining the Residual versus Fitted plot . The red line goes smoother should not typically follow a pattern where it is roughly horizontal at zero.



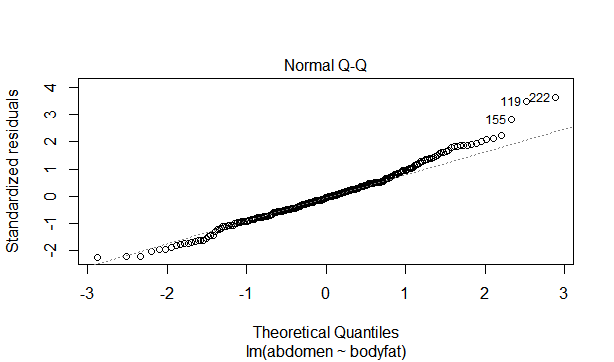
#### CONSTANT VARIANCE of RESIDUAL ERRORS

. The fitted values versus the square root of the standardized residuals are shown in this plot.It’s a scale location plot where red line should be distributed uniformly from the residual points to signify constant variance.



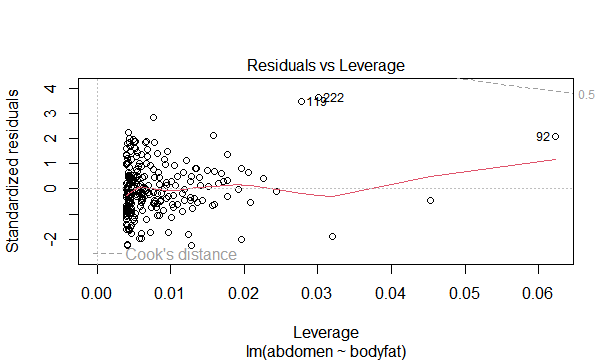
**Normality of residuals**.

It is assumed that the residual errors are normally distributed. For this purpose we use Q-Q plot and straight dashed line should be followed the residuals points.



****Residuals error terms independence****

**Leverage vs. Residuals used to find influential which shows extreme values of data are included or omitted from the study.because values impact on regression results.Look for a data point outside of a dashed line and values in this instance have an impact on the regression findings .Results will be changed in those cases**



* **conduct a formal test to question whether there is a significant linear relationship between body fat and abdomen size.**

**A t-test for the regression slope is used to determine significance.The link between a predictor variable and a response variable is measured using linear regression. Whenever we conduct a linear regression, we are interested in determining whether the link between the predictor variable and the response variable is statistically significant. For this t-test, the following null and alternate assumptions are used**

****H0****: β = 0

****HA****: β ≠ 0

p-value is significantly below 0.05, hence the null hypothesis that = 0 is rejected. As a result, there is a substantial correlation between the variables in the data set's linear regression model.

4. Use the equation to predict the percentage of body fat for a male whose abdomen measures 100cm.

### **Measuring Body Fat Percentage formula for men**

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

The process of developing machine learning models must include evaluating the model accuracy in order to characteristic how well the model is performing in its predictions.

* **MSE** (Mean Squared Error) is a representation of the difference between the original and anticipated values that was calculated by squaring the average difference throughout the data set.

41.7796

* **RMSE** (Root Mean Squared Error) is the error rate divided by the MSE square root..

6.463714

**6. 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.**

**For data import**

data1<-read.csv("data.csv")

data1

bodyfat=data1$inb

abdomen=data1$da

library(tidyverse)

head(data1)

summary(data1)

dim(data1)

ggplot(data=data1, aes(x=inb)) +

geom\_histogram(fill="steelblue", color="black") +

ggtitle("Histogram of bodyfat")

ggplot(data=data1, aes(x=da)) +

geom\_histogram(fill="steelblue", color="black") +

ggtitle("Histogram of abdomen")

ggplot(data,aes(x=abdomen,y=bodyfat))+geom\_point()

cor(data1)

sapply(data1, function(x) sum(is.na(x)))

library(haven)

cor(bodyfat,abdomen)

cor.test(bodyfat,abdomen)

plot(abdomen,bodyfat)

model<-lm(abdomen~bodyfat)

model

summary(model)

esti\_y=fitted.values(model)

esti\_y

sum(esti\_y)

sum(abdomen)

resi=residuals(model)

> resi

sum(resi)

par(mfrow=c(1,1))

plot(model,1)

plot(model,2)

plot(model,3)

plot(model,5)

mean(model\_sum$residuals^2)

sqrt(mean(model\_sum$residuals^2))