MATPMD0 - INTRODUCTORY STATISTICS FOR DATA SCIENCE

PROJECT: AUTUMN SEMESTER 2022

Student Number: 3122831

Declaration: In submitting this project, I declare that this is all my own work, and I did not seek help to complete it.

Abstract

The report presents a statistical analysis of body fat and abdomen measures and develops a predictive linear regression model to estimate a male's body fat percentage across significant abdomen measurements. The dataset includes two continuous quantitative variables, body fat and an abdomen measurement, collected from 252 males. As the initial step, the statistical summary of the data and, in addition, variance, standard deviation, and outliers of the whole dataset have been analysed. The data were also examined based on the skewness of the variables. It was ensured that the outliers in the data were mild and that the prediction wouldn't be adversely affected by any extreme anomalies. The data was split into the training data(70%) and the test data(30%), and the model was built on the training data, performing the predictions on test data to determine whether the model's assumptions had been met while also analysing the slope of the regression line and the relationship between the variable strength. Additionally, the data revealed a positive linear relationship, and the regression model appears to be a satisfactory fit for the data interpreting few findings and outcomes, taking into consideration that the dataset contains human data, has fewer variables, a variability of 59.76%, and other statistics & error rates.

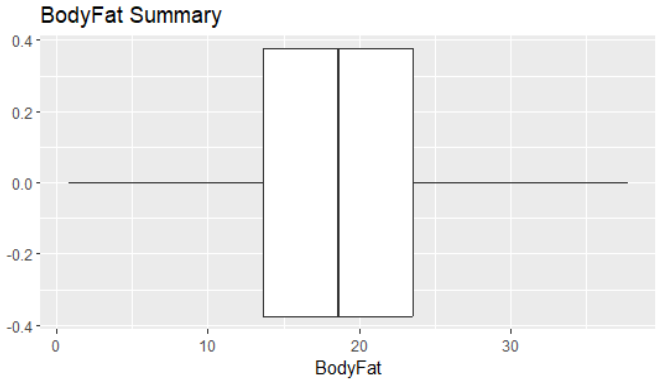
1. Perform an exploratory data analysis, describing the type of variables in the data set.

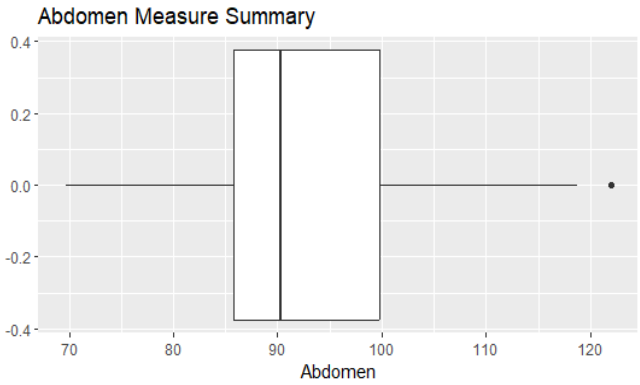
The following summary statistics of the dataset and plots have been used to acquire and analyse the basic characteristics of the dataset, such as mean, median, standard deviation, QR and so on.

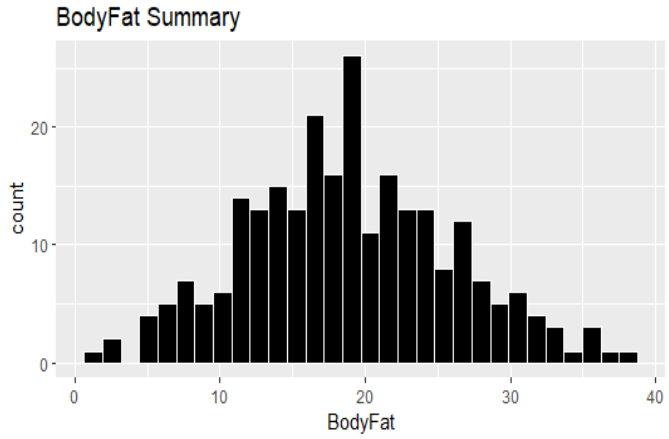
Quartile of variables									
BodyFat					*Abdomen*				
0%	*25%*	*50%*	*75%*	*100%*	*0%*	*25%*	*50%*	*75%*	*100%*
0.832	*13.681*	*18.637*	*23.555*	*37.705*	*69.701*	*85.838*	*90.300*	*99.800*	*121.946*

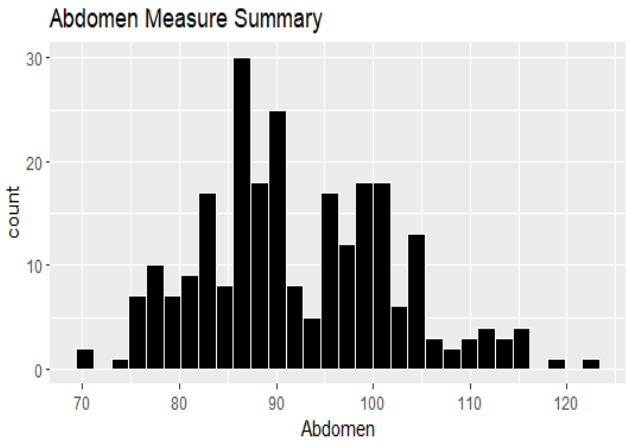
Summary Statistics	
BodyFat	*Abdomen*
Min. : 0.832	*Min. : 69.70*
1st Qu.:13.688	*1st Qu. : 85.84*
Median:18.637	*Median: 90.30*
Mean :18.865	*Mean : 92.31*
3rd Qu.:23.547	*3rd Qu. : 99.80*
Max. :37.705	*Max. :121.95*

Continued overleaf…









Skewness	
BodyFat	*Abdomen*
0.167	*0.369*

The Abdomen variable appears to be slightly positively skewed, with a mild outlier and gap on the upper fence, having a skewness of 0.3691 and a mean of 92.31, higher than the median of 90.30. On the other hand, the bodyfat variable has a skewness of 0.167 (near zero), with a relatively closer mean and median and supports our test by indicating the variable is normally distributed. Attempts have been made to clean the data and remove outliers; however, while testing with the cleaned data without outliers, the correlation coefficient and the coefficient of determination tend to be smaller than the outlier-containing data; additionally, since it is merely a mild outlier with no significant differences in the correlation and model fit, however, the outliers will be removed, and the data will be split into train and test sets for further analysis. (R code has been attached in section 6, and the outcome has been provided in the upcoming section).

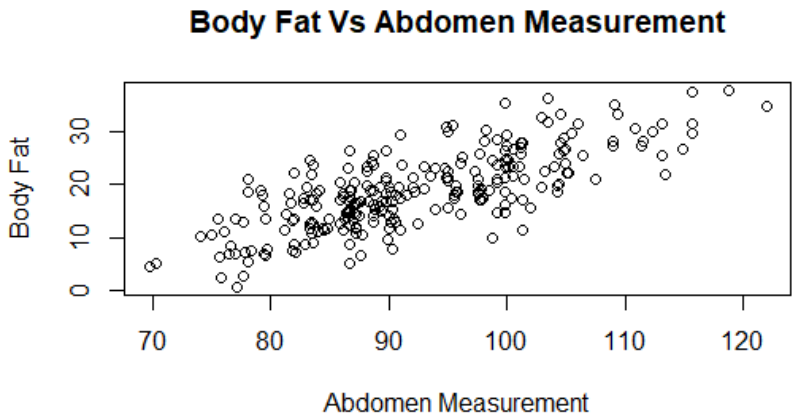
Furthermore, we consider the BodyFat as Y (response or dependent variable) and the Abdomen measure as X (independent or predictor variable) to proceed with the tests and build the model. Since errors in X are insignificant and the Y variable is normally distributed with no outliers according to our assessment, normalisation transformation is not carried out.

Dispersion Parameters:

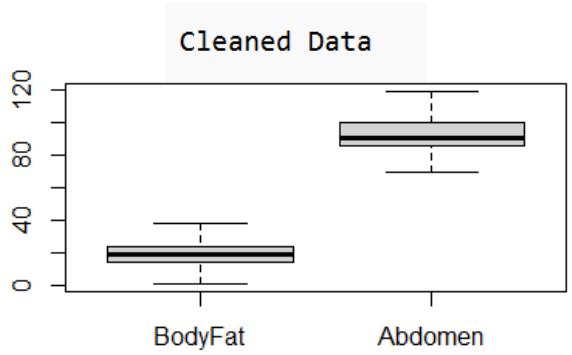
	BodyFat	*Abdomen*
Range	*36.873*	*52.245*
Variance	*51.381*	*98.347*
Standard Deviation	*7.168*	*9.917*

In comparison, the Abdomen measure's standard deviation is relatively small compared to its mean of 92.31, showing that the data points are more clustered around the mean value. The standard deviation of BodyFat is rather significant compared with its mean of 18.865, indicating a broad spread of data points and more variability.

Visualising The Variables:



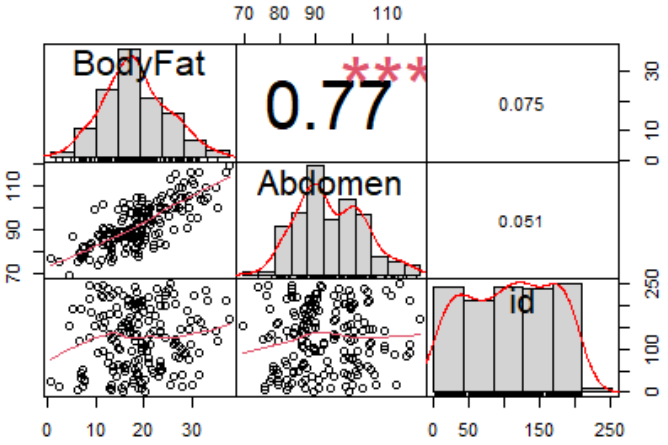
In order to determine the relationship, we took into account the independent variable Abdomen on the X-axis and the response variable BodyFat on the Y-axis of the scatter plot. The slope of the relationship between the continuous variables is assumed to be positively linear at first glance, suggesting that one variable increases as the other increases and shares a positive linear relationship.



The outliers have been removed from the dataset, and it has been divided into training data (70%) and test data (30%) for the subsequent stages in the analysis, using randomly chosen observations. Therefore, the following conclusion will be based on the linear model of train data and train data. (R code is included below.)

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

	BodyFat	*Abdomen*	*id*
BodyFat	*1.000*	*0.773*	*0.075*
Abdomen	*0.773*	*1.000*	*0.051*
id	*0.075*	*0.051*	*1.000*



The training data has been considered for evaluating the correlation coefficient. The correlation coefficient of the two variables, body fat and abdomen measure, in the training dataset is 0.773, which is close to 1, indicating a strong positive linear relationship between the variables. Thus, when the value of the abdomen measurement increases, the likelihood of an increase in body fat will also increase. However, a correlation alone does not imply causation; additional factors, such as the coefficient of determination that influence the relationship, may also be considered, and the r-squared value is 0.598, which is equivalent to the coefficient of determination in our data, indicating that the model fits the data satisfactorily. Moreover, a hypothesis test has been conducted to evaluate if the correlation coefficient is significantly different from zero, indicating a variable linear relationship. (The variable "ID" has been added while splitting the data and can be ignored in the chart.)

The hypothesis of the correlation coefficient:

As the F-statistic (from the F-test) measures the strength of the variables in the model, and the p-value is the probability that the F-statistics indicates the strength of the variable relationship, we assume that the null hypothesis of our test as the relationship or coefficient being equal to zero, where the alternative hypothesis is the other way around since we want to determine if the two variables are correlated.

H0 : ρ = 0 Relationship of coefficient is zero
H1 : ρ ≠ 0 Relationship of coefficient is different from zero
The hypothesis was analysed using the linear model summary, and with the findings,

F-statistic: 258.4 on 1 and 174 DF, p-value: < 2.2e-16

The **F statistic** of *258.4* and the p-value of <2.2e-16 (around zero), which falls below the significance level of 0.05, indicate sufficient evidence to support the **alternate hypothesis ρ≠0** that the correlation is significantly different from zero and the predictor significantly influences the response variable. Consequently, a robust linear relationship between the abdomen measure and body fat is suggested from both hypothesis tests, and the trained model seems to fit the data satisfactorily.

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

(a) A description of the model:

A linear regression model has been employed to predict the body fat given the measurements of the abdomen since the response, or dependent variable, body fat increases, or decreases based on the continuous change

in abdomen measurement, indicating that the relationship between the variables is linear. The relationship between the explanatory and response variables is depicted by a straight line referred to as the line of best fit, which is used to anticipate the relationship between the variables. Despite the fact that linear regression is robust to outliers, one outlier was discovered in the data on abdomen measures and has been eliminated, allowing us to forecast the values and apply them to actual data. The dataset is divided into training and test data, with training data comprising a random 70% of the total data and the rest employed as test data to build the linear model. The performance and accuracy of the training model will next be assessed using the test data.

(b) A summary of the fitted model with the interpretation of test statistics and parameter estimates:

The summary of the fitted linear model from the training data is estimated below,

```
Residuals:
    Min      1Q  Median    3Q     Max
-12.706  -2.964  -0.367   3.110  12.319

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -32.217    3.202      -10.06   <2e-16 ***
Abdomen       0.554     0.034      16.07    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.589 on 174 degrees of freedom
Multiple R-squared:  0.5976,      Adjusted R-squared:  0.5953
F-statistic: 258.4 on 1 and 174 DF,  p-value: < 2.2e-16
```

The **residuals** depict the disparity between the observed value and the value predicted by the model, which provides the measure of the fitted model's error of deviation. Our results show a surprisingly wide range of residual values, from -12.706 to a maximum of 12.319, with the median -0.367, showing that most residuals are negative. On the other hand, the **first quartile, the median, and the third quartile** are relatively near to one another, indicating that the residuals are symmetrically distributed around the median. The **intercept** (the expected value of the body fat when the abdomen is zero) is significantly different from zero and has a negative effect on the response variable, according to the estimate of -32.217, representing the relationship between both variables of our model. This estimate includes a standard error of 3.20161 and a t-value of -10.06. On the other hand, the **predictor abdomen** (the anticipated change in the bodyfat when the abdomen varies) positively affects the response variable with an estimate of 0.554, a t-value of 16.09, and a standard error of 0.034.

The **residual standard error** to the degree of freedom of 250 indicates that the model can predict body fat with an average error of 4.589 and that 250 observations can be varied in the fitting procedure. By the **R-squared** value of 0.598 (equal to the small r-squared), 59.80% of the total variation in the body fat percentage is explained by the linear relationship with the abdomen measure, and the remaining 40.20% explains that the relationship is not strictly linear. Given that the **R-squared** value is in the range of 50% to 70%, the model appears to be satisfactorily fitted and neither over nor under-fits the data. In addition, the **F-statistic** provides the relationship between two variables, and a lower p- of < 2.2e-16 suggests a strong relationship. Altogether, the f-statistic, residual standard error and R-squared values imply that the model provides a better fit, which is backed by the model's small residuals for the coefficients, high t-values, and low p-values. A hypothesis test for the slope of the regression line has been provided below, analysing the coefficients from the linear model.

The hypothesis test for the slope of the regression line:

The regression line depicts the expected change in response variable Body Fat concerning the predictor Abdomen measure. Hence, we assume the null hypothesis that the slope of the regression line is equal to zero; however, the alternate hypothesis, as shown below, states that the opposite is the case.

H0 : β = 0 Slope of regression line equals zero
H1 : β ≠ 0 Slope of regression line different from zero
This hypothesis can be interpreted from the outcome,

Coefficients:				
	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
Abdomen	*0.55432*	*0.03449*	*16.07*	*<2e-16 ****

The hypothesis test has been performed based on the trained data results. From the summary, the predictor "Abdomen" has a p-value of <2.2e-16, which falls below the significance level of 0.05, indicating sufficient evidence to support the **alternate hypothesis β≠0** that the variables are linearly associated.

Confidence Interval of Model Parameters:

	2.5 %	*97.5 %*
(Intercept)	*-38.536*	*-25.898*
Abdomen	*0.486*	*0.622*

The confidence interval of the intercept and predictor slope is generated based on the presumptions. The confidence interval parameter, which states that the likelihood of the Intercept true value falling within the range (38.536, -25.898) (below the slope) while repeating the study and obtaining estimates numerous times, is 95%, whereas the true value of abdomen predictor's likelihood of falling in the range (0.486, 0.622), explaining the model's prediction uncertainty.

(c) Evidence as to whether assumptions of the model have been met:

Since the calculation of the prediction intervals relies on the residual's normality, predictions may not be accurate if the residuals or errors are not normal; therefore, evaluating the residuals' normality is a significant stage. The Q-Q plot, Anderson-Darling Normality test, and residual plots are interpreted to check whether the model's assumptions have been met.

Anderson-Darling Normality Test to verify the normality of errors:

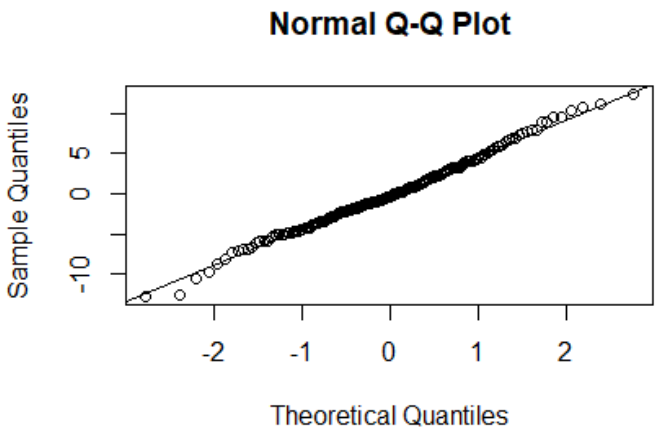
In addition to the Q-Q normal plot, the Anderson-Darling test has been performed to evaluate the residual's normality. The hypothesis of the AD test is,

H0: Residuals follow the normal distribution
H1: Residuals do not follow the normal distribution
The AD normality test findings are,

data: model_nooutlier$residuals
A = 0.379, p-value = 0.4021

The p-value of 0.379, being higher than the significance level of 0.05 to the test statistic of 0.42051, indicates insufficient evidence to reject the null hypothesis; thus, the residuals are normally distributed. As a result, it implies that the error has a normal distribution and that the model's assumptions and inferences are met.

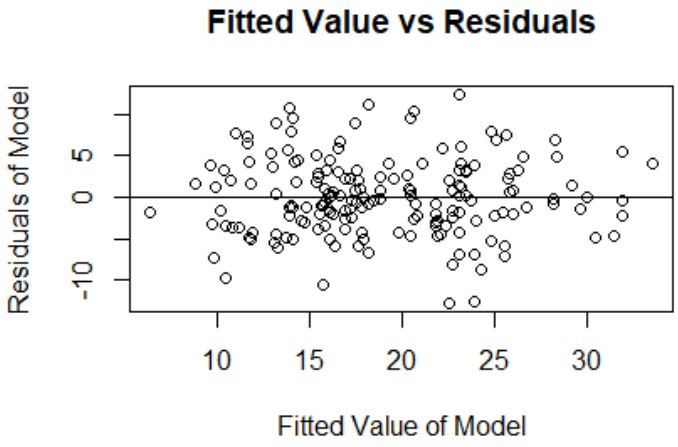
The Normal Q-Q plot and the Q-Q Line:



A Quantile-Quantile plot has been adopted, which depicts the difference between the observed response variable and the model's residuals or fitted values to evaluate the normality of the residuals. The normal Q-Q plot demonstrates that the residuals are normally distributed since the points form a straight line and fall on the Q-Q line, which acts as a reference line to assess if the points are on a straight line to be normally distributed.

Residual Plot:

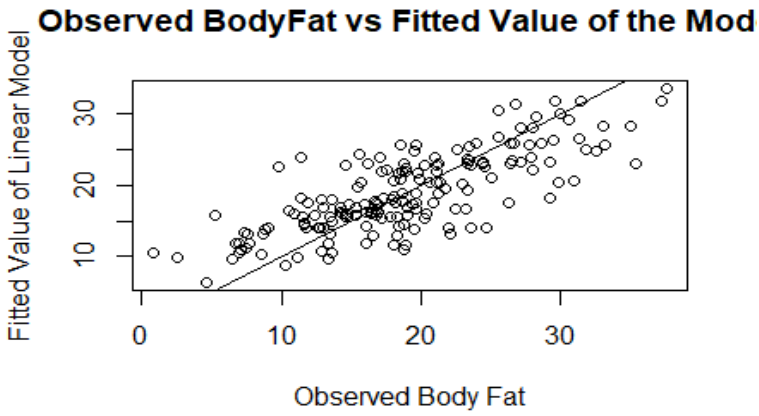
(a) Fitted value vs Residuals



The fitted values versus residuals scatter plot provides insight into the constant variance, which has fitted values on the X-axis and the residual on the Y-axis. The graph shows that the assumption of constant variance is met and the model fits the data well, given that there are no apparent patterns in the distribution of the points around the horizontal line Y=0, and the residuals are roughly equal at each level of the fitted value.

The **independence of error** assumption has also been validated from the plot since there is no correlation or trend between the fitted values and the errors.

(b) Observed Response Value vs Fitted Value:



The points are close to the line rather than dispersed, which demonstrates the adequacy of the model, according to the scatter plot of the observed body fat values and the model's fitted values with a slope of 1 and intercept of 0.

(d) Conduct a formal test to question whether there is a significant linear relationship between body fat and abdomen size.

The two-sided t-test has been conducted on the entire dataset to analyse the linearity of the relationship between the independent variables since the observations are independent and the residuals have a normal distribution.

Hypothesis test to identify the linear relationship of variables:

Research question: The abdominal circumference and the body fat percentage are linearly related.

The hypothesis of the t-test is,

H0: Two variables are not linearly related $\beta=0$
H1: Two variables are linearly related $\beta\neq0$
Significance Level (α): 0.05
$\alpha/2 = 0.025$

Test Results:

SXX	SYY	SXY	Beta hat	Sigma-squared	t-obs	p-value
24685.1	12896.72	13432.6	0.5441582	22.34902	18.08479	2.596887e-47

The observed (calculated) value of the test statistic is 18.08479, and since it is a two-sided t-test, the p-values for both the fences are calculated, and the sum of both the values has been obtained as the p-value, which is 2.597×10^{-47} (close to zero). The test has provided evidence that the p-value is less than the ($\alpha/2=0.025$, taken two-sided), suggesting insufficient evidence to prove the null hypothesis; hence, the variables are linearly related. Even though the t-test is performed on the entire dataset, the results of "**the hypothesis test for the slope of the regression line**" performed on the trained data in the preceding section support the conclusions.

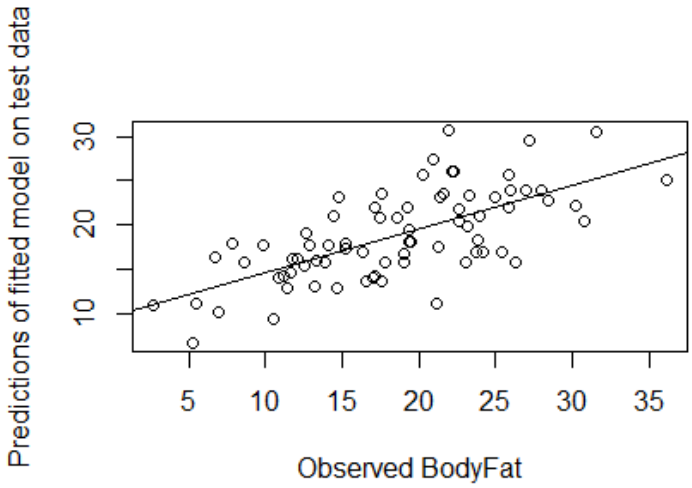
4. Use the equation to predict the percentage of body fat for a male whose abdomen measures 100cm.

Predicted Body Fat against the abdomen measure 100cm - 23.215%

Using the linear model, the body fat percentage was estimated as 23.215% across the abdomen measurement of 100cm. In addition, the body fat against the abdomen measure of 90cm has been predicted as **17.607**%; however, the actual percentage may vary due to other influencing factors that are not accounted for in the model and error standards.

5. Assess the predictive performance of the model.

(a) Observed Response Variable Vs Predictions on Test Data with Best Fitted Line:



The fitted line in the graph shows that the observed and expected values are positively correlated. The graph's interpretation indicates that the points are neither close to the fitted line nor dispersed, suggesting that it makes a reasonable prediction but is not always accurate. The plot will assist us in identifying trends and patterns in the data; however, other evaluation metrics must be taken into account to analyse the model's performance.

The r-squared, mean squared error (MSE), and root mean squared error (RMSE) values have been chosen to evaluate the linear regression model's performance rather than the F1 score or precision(since the F1 score and precision are generally used for classification models (for categorical data)).

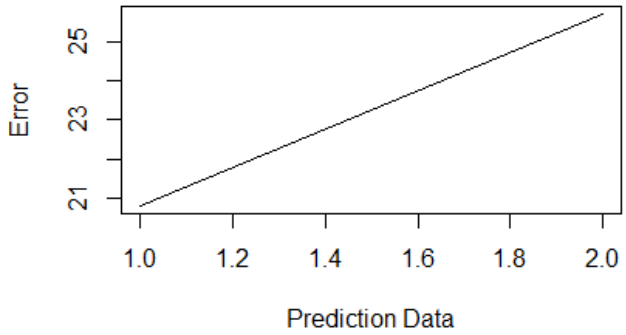
(b) RSE, MSE, & R-Squared of the trained linear model:

MSE	RMSE	R-Squared
25.699	5.069	0.598

A relatively high MSE and RMSE values suggest that the prediction is somewhat different from the actual values, and the R-Squared value indicates that the predictor explains approximately 60% of the variance, taking into consideration that the human data generally has an R-squared ranging from 50-70%, altogether suggests that the model needs to fit better with the data. To further check the fitting of the model, the error in test and training data has been analysed in the subsequent step.

(c) Evaluating the error in test and training data:

Training Data Error	Test Data Error
20.81727	25.69878



The mean squared error, which represents the discrepancy between the predicted and actual values of training and test data, has been obtained from the evaluation of test data using the model fit on the training data(cross-validation). When deployed to a new data or test dataset, the probability of the model generalising the data and generating the correct predictions is less since the error in the test set is larger than the error in the training set (25.699 vs 20.817), suggesting that the model is overfitting (performing well) to training data, given that the size of the dataset is small and the fact that the model has learned the characteristics peculiar to the training data, the likelihood of the model overfitting is greater.

The evaluation metrics for the model trained on the entire dataset for testing purposes have been analysed, and it has been confirmed that the errors seem to be nearly equivalent to the trained model. Hence, more data can be added to the dataset to improve the r-squared value and performance, and algorithms like ensemble bagging and boosting can increase model accuracy. In the ensemble methods, bagging creates multiple datasets of original data and combines the individual predictions to make a final prediction, which lowers the prediction variance; in contrast, boosting adjusts the predictions by learning the outcomes from the prior model.

6. In this final section, include all R code that you have used for this project verbatim.

```
# Installed packages
> install.packages("tidyverse")
> install.packages("readxl")
> install.packages("GPL2025")
> install.packages("dataset")
> install.packages("survival")
> install.packages("car")
> install.packages("nortest")
> install.packages("PerformanceAnalytics")
> install.packages("corrplot")
> install.packages("e1071")

# To import and read the excel file to the R console readxl library has been used
> library("readxl")
# The excel file has been read to a data frame named main_data.df, which will be used
> main_data.df <- read_excel("3122831BodyFatData.xlsx",range="A1:B253")
```

```
> attach(main_data.df)
The following objects are masked from bodyfat_data.df:
   Abdomen, BodyFat
> View(main_data.df)
> dim(main_data.df)
[1] 252   2
# The data has 252 items in 2 columns
# is.na(main_data.df)
> sum(is.na(main_data.df))
[1] 0
# CHECKED MISSING VALUE IN THE DATASET, and the dataset seems clean with no missing values as the
output is zero.
> str(main_data.df)
tibble [252 × 2] (S3: tbl_df/tbl/data.frame)
 $ BodyFat: num [1:252] 15.92 23.2 20.31 6.47 12.57 ...
 $ Abdomen: num [1:252] 89.2 93.9 99.1 75.7 85.8 ...
> typeof(Abdomen)
[1] "double"
> typeof(BodyFat)
[1] "double"

# FINDING QUARTILES
> quantile(main_data.df$BodyFat, type=6)
    0%     25%     50%     75%    100%
 0.8320 13.6805 18.6365 23.5545 37.7050
> quantile(main_data.df$Abdomen, type=6)
    0%     25%     50%     75%    100%
 69.701  85.838  90.300  99.800 121.946

# SUMMARY STATISTICS OF THE DATA
> summary(main_data.df)
   BodyFat         Abdomen
 Min.   : 0.832   Min.   : 69.70
 1st Qu.:13.688   1st Qu.: 85.84
 Median :18.637   Median : 90.30
 Mean   :18.865   Mean   : 92.31
 3rd Qu.:23.547   3rd Qu.: 99.80
 Max.   :37.705   Max.   :121.95

#STANDARD DEVIATION
> sd(main_data.df$BodyFat)
[1] 7.168078
> sd(main_data.df$Abdomen)
[1] 9.917006

# RANGE
> max(main_data.df$BodyFat) - min(main_data.df$BodyFat)
[1] 36.873
> max(main_data.df$Abdomen) - min(main_data.df$Abdomen)
[1] 52.245
```

```
# VARIANCE
> var(main_data.df$BodyFat)
[1] 51.38134
> var(main_data.df$Abdomen)
[1] 98.34701

# SKEWNESS CHECK
> library(e1071)
> skewness(main_data.df$Abdomen)
[1] 0.3691489
> skewness(main_data.df$BodyFat)
[1] 0.1665765

# BOX PLOT OF THE DATA
> library(ggplot2)
> ggplot(main_data.df, aes(x=Abdomen)) + geom_boxplot() + ggtitle("Abdomen Measure Summary")
> ggplot(main_data.df, aes(x=BodyFat)) + geom_boxplot() + ggtitle("BodyFat Summary")

# HISTOGRAM OF THE DATA
> ggplot(main_data.df, aes(x=BodyFat)) + geom_histogram(color="white", fill="black") + ggtitle("BodyFat
Summary")
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> ggplot(main_data.df, aes(x=Abdomen)) + geom_histogram(color="white", fill="black") + ggtitle("Abdomen
Measure Summary")
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
# The boxplot and histogram have been attached to the report.
# Referred both hist and box plot from: http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-
start-guide-r-software-and-data-visualization

# PLOTTING GRAPH OF VARIABLES TO CHECK THE RELATIONSHIP
> plot(BodyFat~Abdomen, data=main_data.df, main="Body Fat Vs Abdomen Measurement",
xlab="Abdomen Measurement", ylab="Body Fat")

# HYPOTHESIS TEST TO PROVE LINEAR RELATIONSHIP OF VARIABLES
# n has been defined as the total number of observations in one variable
> n = length(bodyfat_data.df$BodyFat)
> df = n-2
# As we are doing a two-sided t-test, alpha has been considered as 0.025
> alpha = 0.025
> sum_x <- sum(bodyfat_data.df$Abdomen)
> sum_x
[1] 23262.26
> sum_y <- sum(bodyfat_data.df$BodyFat)
> sum_y
[1] 4753.905
> sum_xy <- sum((bodyfat_data.df$Abdomen)*(bodyfat_data.df$BodyFat))
> sum_xy
[1] 452268.2
> sum_x_sqr <- sum((bodyfat_data.df$Abdomen)^2)
> sum_x_sqr
[1] 2172037
```

```
> sum_y_sqr <- sum((bodyfat_data.df$BodyFat)^2)
> sum_y_sqr
[1] 102577.7
# SXX, SYY, and SXY is calculated below to find the beta hat and sigma squared
> s_xy<- sum_xy-((sum_x*sum_y)/n)
> s_xy
[1] 13432.6
> s_xx <- sum_x_sqr-((sum_x^2)/n)
> s_xx
[1] 24685.1
> s_yy<-sum_y_sqr-((sum_y^2)/n)
> s_yy
[1] 12896.72
> beta_hat<-s_xy/s_xx
> beta_hat
[1] 0.5441582
> sigma_sqr<-(s_yy-((s_xy^2)/s_xx))/(n-2)
> sigma_sqr
[1] 22.34902
# The observed t-value has been calculated below
> t_obs <- beta_hat/(sqrt(sigma_sqr/s_xx))
> t_obs
[1] 18.08479
# For the two-sided t-test, we are calculating p-values on both sides of the distribution and adding them up
to get the P_VALUE
> p_right <- pt(abs(t_obs), df = n-2, lower.tail = FALSE)
> p_left <- pt(-abs(t_obs), df = n-2, lower.tail = TRUE)
> p_value <- p_right+p_left
> p_value
[1] 2.596887e-47
# The command below will provide an output if p_value ≤ alpha = TRUE, and prove there is insufficient
evidence to support the null hypothesis
> p_value <= alpha
[1] TRUE
```

 # LINEAR MODEL OF THE WHOLE DATASET – This was created just as a test and to compare the results with the trained data model. This model has not been used to interpret the results.

```
# FINDING CORRELATION
> cor(main_data.df$Abdomen, main_data.df$BodyFat)
[1] 0.7528407
> cor(main_data.df$Abdomen, main_data.df$BodyFat)^2
[1] 0.5667691
# DEVELOPING A LINEAR MODEL
> lin_model<-lm(BodyFat~Abdomen, data=main_data.df)
> summary(lin_model)

Call:
lm(formula = BodyFat ~ Abdomen, data = main_data.df)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-12.5520  -3.2580  -0.2494   3.0950  12.4839

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -31.36685    2.79348  -11.23   <2e-16 ***
Abdomen       0.54416    0.03009   18.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.727 on 250 degrees of freedom
Multiple R-squared:  0.5668,      Adjusted R-squared:  0.565
F-statistic: 327.1 on 1 and 250 DF,  p-value: < 2.2e-16
```

 # REMOVING THE OUTLIERS FROM THE ABDOMEN AND STORING THE DATA IN A NEW DATAFRAME

```
# Interquartile range has been calculated and tried to remove the outliers from the upper and lower fences.
> dim(main_data.df)
[1] 252   2
> quartiles <- quantile(main_data.df$Abdomen, probs=c(.25, .75), na.rm = FALSE)
> iqr <- IQR(main_data.df$Abdomen)
> lower_quart <- quartiles[1] - 1.5*iqr
> upper_quart <- quartiles[2] + 1.5*iqr
> newdata_no_outlier <- subset(main_data.df, main_data.df$Abdomen > lower_quart &
main_data.df$Abdomen < upper_quart)
> boxplot(newdata_no_outlier)$stats
      [,1]    [,2]
[1,]  0.8320  69.701
[2,] 13.6840  85.838
[3,] 18.6060  90.300
[4,] 23.4825  99.800
[5,] 37.7050 118.717
# outlier and iqr have been referred from notes and https://absentdata.com/how-to-find-outliers-in-excel/#:~:text=Lower%20range%20limit%20%3D%20Q1%20%E2%80%93%20(,will%20be%20considered%20an%20outlier

# FINDING THE CORRELATION OF THE DATA WITH NO OUTLIERS
> cor(newdata_no_outlier$Abdomen, newdata_no_outlier$BodyFat)
[1] 0.7471266

# SPLITTING DATA INTO TRAIN(0.7) AND TEST(0.3)
> library(dplyr)
> set.seed(123)
> newdata_no_outlier$id <- 1:nrow(newdata_no_outlier)
> traindata_nooutlier <- newdata_no_outlier %>% dplyr::sample_frac(0.70)
> View(traindata_nooutlier)
> View(newdata_no_outlier)
> testdata_nooutlier  <- dplyr::anti_join(newdata_no_outlier, traindata_nooutlier, by = 'id')
# Referred for splitting data: https://www.statology.org/train-test-split-r/
```

 # CREATED A LINEAR MODEL ON THE TRAIN DATA – WHICH IS USED THROUGHOUT THE REPORT

```
> model_nooutlier<-lm(BodyFat~Abdomen, data=traindata_nooutlier)
> summary(model_nooutlier)
```

```
Call:
lm(formula = BodyFat ~ Abdomen, data = traindata_nooutlier)

Residuals:
    Min      1Q  Median      3Q     Max
-12.7062  -2.9639  -0.3668   3.1099  12.3195

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.21712    3.20161  -10.06   <2e-16 ***
Abdomen       0.55432    0.03449   16.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.589 on 174 degrees of freedom
Multiple R-squared:  0.5976,      Adjusted R-squared:  0.5953
F-statistic: 258.4 on 1 and 174 DF,  p-value: < 2.2e-16

# FINDING CORRELATION OF THE TRAIN DATA
> cor(traindata_nooutlier$Abdomen, traindata_nooutlier$BodyFat)
[1] 0.7730309

# CORRELATION GRAPH
> result = cor(traindata_nooutlier, method = "pearson", use = "complete.obs")
> round(result,3)
         BodyFat Abdomen    id
BodyFat   1.000   0.773 0.075
Abdomen   0.773   1.000 0.051
id        0.075   0.051 1.000
> library("PerformanceAnalytics")
> chart.Correlation(traindata_nooutlier, histogram=TRUE, pch=19)
Warning messages:
1: In par(usr) : argument 1 does not name a graphical parameter
2: In par(usr) : argument 1 does not name a graphical parameter
3: In par(usr) : argument 1 does not name a graphical parameter

# Q-Q PLOT AND Q-Q LINE
> library(nortest)
> qqnorm(model_nooutlier$residuals)
> qqline(model_nooutlier$residuals)

# ANDERSON DARLING TEST TO FIND NORMALITY OF RESIDUALS
> ad.test(model_nooutlier$residuals)

        Anderson-Darling normality test

data:  model_nooutlier$residuals
A = 0.379, p-value = 0.4021
```

```
# PLOTTING THE MODEL'S FITTED VALUE AGAINST THE RESIDUALS
> plot (model_nooutlier$fitted.values, model_nooutlier$residuals, main="Fitted Value vs Residuals",
xlab="Fitted Value of Model",  ylab="Residuals of Model")
> abline(h=0)

# CONFIDENCE INTERVAL OF THE MODEL
> confint(model_nooutlier)
                2.5 %     97.5 %
(Intercept) -38.5361016 -25.8981313
Abdomen       0.4862614   0.6223882

# PLOTTING THE FITTED LINE
> plot (traindata_nooutlier$BodyFat, model_nooutlier$fitted.values, main="Observed BodyFat vs Fitted
Value of the Model",  xlab="Observed Body Fat",  ylab="Fitted Value of Linear Model")
> abline(a=0,b=1)

# PLOTTING HORIZONTAL LINE, WHICH IS A MEAN OF Y-VALUE
> rline <- lm(BodyFat~Abdomen, data=traindata_nooutlier)
> with(traindata_nooutlier, segments(Abdomen, fitted(rline), Abdomen, BodyFat, col="red"))
> with(traindata_nooutlier, plot(Abdomen,BodyFat, main="BodyFat vs Abdomen Measure",
xlab="Abdomen", ylab="BodyFat"))
> abline(h = 18.8647, lwd=2, col="blue")
> with(traindata_nooutlier, segments(Abdomen,18.8647, Abdomen,BodyFat, col="red"))

# PLOTTING THE REGRESSION LINE
> with(traindata_nooutlier, plot(Abdomen,BodyFat, main="Regression Line", xlab="Abdomen Measure",
ylab="Body Fat"))
> abline(rline,lwd=2,col="blue")
> with(traindata_nooutlier, segments(Abdomen,fitted(rline),Abdomen,BodyFat,col="red"))

# PREDICTING BODYFAT FOR 100CM ABDOMEN USING THE LINEAR MODEL
# To predict the body fat while the abdomen measure is 100CM using the model
# Referred: https://www.digitalocean.com/community/tutorials/predict-function-in-r
> predict_100 <- as.data.frame(100)
> colnames(predict_100) <- "Abdomen"
> predicted_value_nooutlier <- predict(model_nooutlier, newdata = predict_100)
> predicted_value_nooutlier
       1
23.21536

# PREDICTING BODY FAT FOR 90CM ABDOMEN MEASURE
> predict_90 <- as.data.frame(90)
> colnames(predict_90) <- "Abdomen"
> predicted_bodyfat_90 <- predict(lin_model, newdata = predict_90)
> print(predicted_bodyfat_90)
       1
17.60739
> library(caret)
> library(tidyverse)
```

The "full data" values below represent the error standard computed for the model built using the entire dataset. This has been done as a part of testing to evaluate and assess the model's performance. Similarly, various training models were developed, and tests were run to obtain the same output.
PREDICTIONS ON TEST DATA
> predictions <- predict(model_nooutlier, newdata = testdata_nooutlier)
Predictions - commented out as the output is too large
> predictions_fulldata <- predict(lin_model, newdata = main_data.df)

MEAN SQUARED ERROR
> mse <- mean((testdata_nooutlier$BodyFat - predictions) ^ 2)
> mse
[1] 25.69878
> mse_fulldata <- mean((main_data.df$BodyFat - predictions_fulldata) ^ 2)
> mse_fulldata
[1] 22.17165

ROOT MEAN SQUARED ERROR
> rmse <- sqrt(mean((testdata_nooutlier$BodyFat - predictions) ^ 2))
> rmse
[1] 5.069396
> rmse_fulldata <- sqrt(mean((main_data.df$BodyFat - predictions_fulldata) ^ 2))
> rmse_fulldata
[1] 4.708678

R-SQUARED
> r_squared <- summary(model_nooutlier)$r.squared
> r_squared
[1] 0.5975768
> r_squared_fulldata <- summary(lin_model)$r.squared
> r_squared_fulldata
[1] 0.5667691

TRAIN ERROR AND TEST ERROR
> train_predictions <- predict(model_nooutlier, newdata = traindata_nooutlier)
> test_predictions <- predict(model_nooutlier, newdata = testdata_nooutlier)
> test_error <- mean((testdata_nooutlier$BodyFat - test_predictions) ^ 2)
> train_error <- mean((traindata_nooutlier$BodyFat - train_predictions) ^ 2)
> fulldata_error <- mean((main_data.df$BodyFat - predictions_fulldata) ^ 2)
> fulldata_error
[1] 22.17165
> train_error
[1] 20.81727
> test_error
[1] 25.69878
the formulas for mse, rmse, precision has been referred from: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e
#PLOTTING TRAIN AND TEST ERROR
> plot(c(train_error, test_error), type = "l", xlab = "Prediction Data", ylab = "Error")

PLOTTING PREDICTIONS OF TRAIN DATA MODEL ON TEST DATA

```
> plot(testdata_nooutlier$BodyFat, predictions, xlab = "Observed BodyFat", ylab = "Predictions of fitted
model on test data")
> abline(lm(predictions ~ testdata_nooutlier$BodyFat))

# PLOTTING PREDICTIONS ON THE WHOLE DATASET
> plot(main_data.df$BodyFat, predictions_fulldata)
> # abline(lm(predictions ~ main_data.df$BodyFat))

# PRECISION OF THE MODEL
> sse <- sum((testdata_nooutlier$BodyFat - predictions) ^ 2)
> tss <- sum((testdata_nooutlier$BodyFat - mean(testdata_nooutlier$BodyFat)) ^ 2)
> precision <- 1 - (sse / tss)
> precision
[1] 0.4539505
```