Name: Nikhil H

INTRODUCTION

- The Body Measurements Dataset consists of Body girth measurements and skeletal diameter measurements, as well as age, weight, height, and gender, are given for 507 physically active individuals - 247 men and 260 women.
- **Variables in the dataset:** Nine skeletal measurements (diameter measurements) and twelve girth (or circumference) measurements, as well as age, weight, height, and gender, are available in this dataset.
- The source of the data is Heinz G, Peterson LJ, Johnson RW, Kerk CJ. 2003. Exploring Relationships in Body Dimensions. Journal of Statistics Education 11(2).

PROBLEM STATEMENT

The aim is to investigate and understand if there is any statistical significant linear relationship between a person’s chest diameter (che.di) and height (hgt).

DATA PREPARATION

- Import the dataset in R-studio using the read_csv function.
- The important variables for the analysis are che.di(chest diameter) and hgt(height).
- The height is the independent variable(predictor) and the chest diameter is the dependent variable.
- Sort the height and chest diameter to check for unexpected values and outliers.

Example: There might be an input where in the height is entered in meters instead of centimeters. Filter these values as it might affect our data analysis.

- There are no missing values in the dataset.

- There are no outliers in the dataset. Even if there were outliers we won’t filter those values as we need to consider the chest diameter of tallest person as well.

DESCRIPTIVE STATISTICS AND VISULAISATION

- Missing values :

- According to the summary statistics there are no missing values for che.di (chest diameter).

```
# A tibble: 1 x 9
    Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
1  22.2  25.6   27.8  30.0  35.6  28.0  2.74   507       0
```

- According to the summary statistics there are no missing values for hgt(height).

```
# A tibble: 1 x 9
    Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
1  147.  164.   170.  178.  198.  171.  9.41   507       0
```

- Sort function to check for unexpected values and outliers:

Use the code to sort the hgt (height) values: **sort(bdims$hgt)**

From the sort function output we can conclude that there are no unexpected values and outliers in the height variable.

DESCRIPTIVE STATISTICS AND VISULAISATION

```
> sort(bdims$hgt)
  [1] 147.2 149.5 149.9 151.1 151.1 152.0 152.0 152.4 152.4 152.4 152.4 153.4 154.4 154.5 154.9 154.9
 [17] 155.0 155.0 155.8 156.0 156.0 156.0 156.2 156.2 156.5 157.0 157.0 157.2 157.5 157.5 157.5 157.5
 [33] 157.5 157.5 157.5 157.5 158.0 158.2 158.8 158.8 159.0 159.0 159.1 159.2 159.4 159.4 159.5 159.5
 [49] 159.5 159.8 159.8 160.0 160.0 160.0 160.0 160.0 160.0 160.0 160.0 160.0 160.0 160.0 160.0 160.0
 [65] 160.0 160.0 160.0 160.0 160.0 160.0 160.0 160.0 160.2 160.2 160.7 160.7 160.7 160.9 161.0 161.0
 [81] 161.2 161.2 161.2 161.3 161.3 161.3 161.3 161.3 161.3 161.3 161.4 162.0 162.0 162.1 162.1 162.2
 [97] 162.5 162.5 162.6 162.6 162.6 162.6 162.6 162.6 162.6 162.6 162.6 162.6 162.6 162.6 162.6 162.6
[113] 162.8 162.8 162.9 163.0 163.0 163.0 163.2 163.2 163.2 163.2 163.5 163.5 163.8 163.8 163.8 163.8
[129] 164.0 164.0 164.1 164.1 164.3 164.4 164.5 164.5 164.5 164.5 164.5 165.0 165.1 165.1 165.1 165.1
[145] 165.1 165.1 165.1 165.1 165.1 165.1 165.1 165.1 165.1 165.1 165.1 165.1 165.1 165.1 165.5 165.7
[161] 166.0 166.0 166.2 166.4 166.4 166.4 166.4 166.4 166.8 166.8 167.0 167.0 167.0 167.0 167.1 167.4
[177] 167.4 167.5 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6
[193] 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.6 167.8 168.2 168.2
[209] 168.3 168.5 168.9 168.9 168.9 168.9 168.9 168.9 168.9 168.9 168.9 168.9 168.9 169.0 169.0 169.4
[225] 169.4 169.5 169.5 169.5 169.5 170.0 170.0 170.0 170.0 170.0 170.0 170.0 170.0 170.0 170.2 170.2
[241] 170.2 170.2 170.2 170.2 170.2 170.2 170.2 170.2 170.2 170.2 170.2 170.2 170.2 170.3 170.5 170.5
[257] 170.5 170.5 170.8 170.9 171.2 171.4 171.4 171.4 171.4 171.4 171.5 171.5 171.8 171.8 172.1 172.1
[273] 172.5 172.7 172.7 172.7 172.7 172.7 172.7 172.7 172.7 172.7 172.7 172.7 172.7 172.7 172.7 172.7
[289] 172.8 172.9 173.0 173.0 173.0 173.2 173.2 173.4 173.5 174.0 174.0 174.0 174.0 174.0 174.0 174.0
[305] 174.0 174.0 174.0 174.0 174.0 174.0 174.0 174.0 174.0 174.0 174.5 175.0 175.0 175.0 175.0 175.2
[321] 175.2 175.2 175.3 175.3 175.3 175.3 175.3 175.3 175.3 175.3 175.3 175.3 175.3 175.3 175.3 175.3
[337] 175.3 175.3 175.3 175.3 175.3 175.3 175.5 175.5 175.5 175.5 175.9 176.0 176.0 176.0 176.0 176.2
[353] 176.2 176.5 176.5 176.5 176.5 176.5 176.5 176.5 176.5 176.5 176.5 176.5 177.0 177.0 177.0 177.0
[369] 177.1 177.2 177.3 177.5 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8
[385] 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 177.8 178.0
[401] 178.0 178.1 179.1 179.1 179.1 179.1 179.1 179.1 179.7 179.8 179.8 179.9 180.0 180.1 180.3 180.3
[417] 180.3 180.3 180.3 180.3 180.3 180.3 180.3 180.3 180.3 180.3 180.3 180.3 180.3 180.3 180.5 180.6
[433] 180.6 181.1 181.5 181.6 181.6 181.6 181.6 182.0 182.0 182.1 182.2 182.4 182.9 182.9 182.9 182.9
[449] 182.9 182.9 182.9 182.9 183.0 183.0 183.5 184.0 184.0 184.0 184.2 184.2 184.2 184.2 184.4 184.5
[465] 184.9 185.4 185.4 185.4 185.4 185.4 185.4 185.4 186.0 186.5 186.7 186.7 186.7 186.7 187.2 188.0
[481] 188.0 188.0 188.0 188.0 188.0 188.0 188.0 188.0 188.0 188.0 188.0 188.0 189.2 190.5 190.5 190.5
[497] 190.5 190.5 190.5 192.0 192.0 192.0 192.7 193.0 193.5 197.1 198.1
```

- Sort function to check for unexpected values and outliers:

Use the below code to sort the che.di(chest diameter) values: **sort(bdims$che.di)**

From the sort function output we can conclude that there are no unexpected values and outliers in the che.di variable.

DESCRIPTIVE STATISTICS AND VISULAISATION

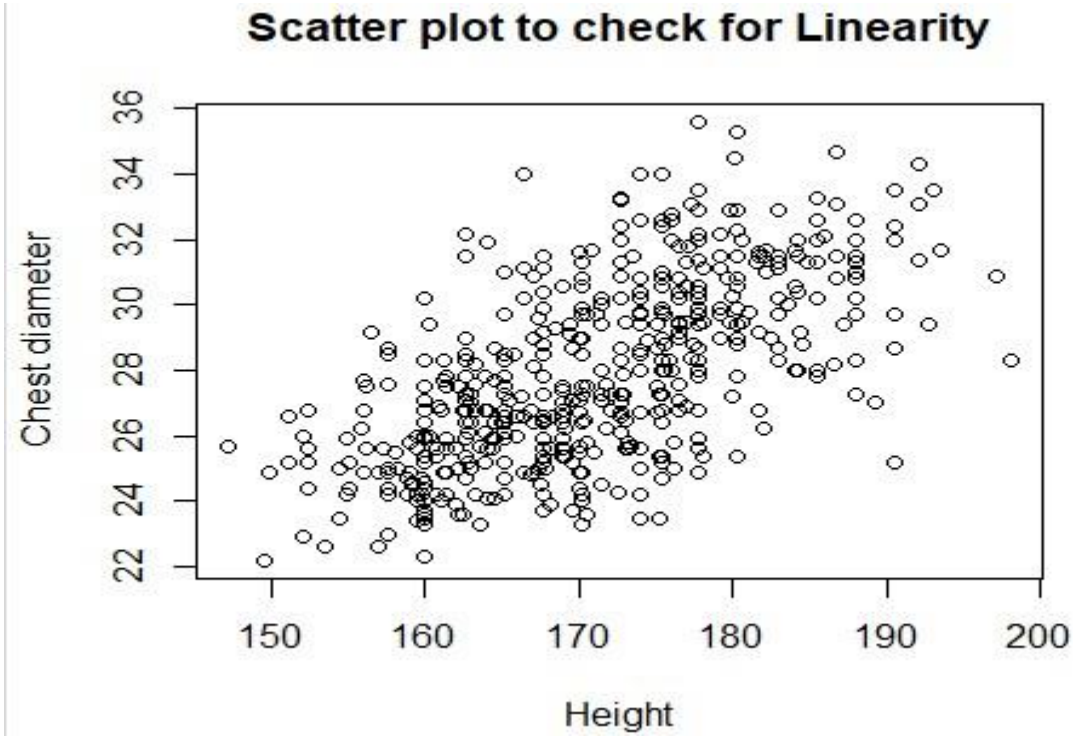
```
> sort(bdims$che.di)
  [1] 22.2 22.3 22.6 22.6 22.9 23.0 23.3 23.3 23.3 23.4 23.5 23.5 23.5 23.5 23.5 23.6 23.6 23.6 23.6
 [20] 23.7 23.7 23.7 23.9 23.9 24.0 24.0 24.0 24.0 24.1 24.1 24.2 24.2 24.2 24.2 24.2 24.2 24.2 24.2
 [39] 24.2 24.2 24.2 24.2 24.3 24.4 24.4 24.4 24.4 24.4 24.5 24.5 24.5 24.6 24.6 24.7 24.7 24.7 24.7
 [58] 24.7 24.7 24.8 24.9 24.9 24.9 24.9 24.9 24.9 24.9 24.9 24.9 24.9 24.9 24.9 24.9 24.9 24.9 25.0
 [77] 25.0 25.0 25.0 25.0 25.0 25.0 25.0 25.0 25.2 25.2 25.2 25.2 25.2 25.2 25.2 25.2 25.2 25.2 25.2
 [96] 25.2 25.4 25.4 25.4 25.4 25.4 25.4 25.4 25.4 25.4 25.4 25.4 25.5 25.5 25.5 25.6 25.6 25.6 25.6
[115] 25.6 25.6 25.6 25.6 25.6 25.6 25.6 25.6 25.6 25.6 25.6 25.6 25.6 25.7 25.7 25.7 25.7 25.8 25.8
[134] 25.8 25.8 25.9 25.9 25.9 25.9 25.9 25.9 25.9 25.9 25.9 25.9 25.9 26.0 26.0 26.0 26.0 26.0 26.0
[153] 26.1 26.1 26.1 26.1 26.2 26.2 26.2 26.2 26.4 26.4 26.4 26.4 26.4 26.4 26.4 26.4 26.4 26.4 26.4
[172] 26.5 26.5 26.5 26.6 26.6 26.6 26.6 26.6 26.6 26.6 26.6 26.6 26.6 26.7 26.7 26.7 26.7 26.8 26.8
[191] 26.8 26.8 26.8 26.8 26.8 26.8 26.8 26.8 26.8 26.8 26.8 26.8 26.8 26.8 26.8 26.9 26.9 27.0 27.0
[210] 27.0 27.0 27.1 27.1 27.1 27.1 27.1 27.1 27.1 27.1 27.2 27.2 27.3 27.3 27.3 27.3 27.3 27.3 27.3
[229] 27.3 27.3 27.3 27.3 27.3 27.3 27.3 27.4 27.5 27.5 27.5 27.5 27.5 27.5 27.5 27.5 27.5 27.5 27.6
[248] 27.6 27.6 27.6 27.7 27.7 27.7 27.8 27.8 27.8 27.8 27.8 27.8 27.8 27.8 27.8 27.9 28.0 28.0 28.0
[267] 28.0 28.0 28.0 28.0 28.0 28.0 28.0 28.0 28.0 28.1 28.2 28.2 28.3 28.3 28.3 28.3 28.3 28.3 28.3
[286] 28.3 28.3 28.3 28.3 28.3 28.3 28.4 28.5 28.5 28.5 28.5 28.5 28.5 28.5 28.5 28.6 28.7 28.7 28.7
[305] 28.7 28.7 28.7 28.7 28.7 28.7 28.7 28.8 28.8 28.8 28.8 28.8 28.9 29.0 29.0 29.0 29.0 29.0 29.0
[324] 29.0 29.0 29.0 29.1 29.2 29.2 29.2 29.2 29.2 29.2 29.2 29.3 29.4 29.4 29.4 29.4 29.4 29.4 29.4
[343] 29.4 29.4 29.4 29.4 29.4 29.4 29.5 29.5 29.5 29.5 29.5 29.5 29.5 29.6 29.7 29.7 29.7 29.7 29.7
[362] 29.7 29.7 29.7 29.7 29.7 29.7 29.7 29.7 29.8 29.8 29.8 29.9 29.9 29.9 29.9 29.9 29.9 29.9 29.9
[381] 30.0 30.0 30.2 30.2 30.2 30.2 30.2 30.2 30.2 30.2 30.2 30.2 30.2 30.3 30.4 30.4 30.4 30.4 30.4
[400] 30.6 30.6 30.6 30.6 30.6 30.6 30.6 30.6 30.6 30.6 30.8 30.8 30.8 30.8 30.8 30.8 30.8 30.8 30.8
[419] 30.9 30.9 30.9 31.0 31.0 31.0 31.0 31.1 31.1 31.1 31.1 31.1 31.1 31.3 31.3 31.3 31.3 31.3 31.3
[438] 31.3 31.3 31.4 31.4 31.5 31.5 31.5 31.5 31.5 31.5 31.5 31.5 31.5 31.5 31.5 31.6 31.6 31.7 31.7
[457] 31.7 31.7 31.8 31.8 31.9 32.0 32.0 32.0 32.0 32.0 32.0 32.0 32.0 32.0 32.0 32.1 32.2 32.2 32.2
[476] 32.3 32.4 32.4 32.4 32.6 32.6 32.6 32.6 32.6 32.8 32.9 32.9 32.9 32.9 32.9 33.1 33.1 33.1 33.2
[495] 33.3 33.3 33.5 33.5 33.5 34.0 34.0 34.0 34.3 34.5 34.7 35.3 35.6
```

- Scatter Plot:

Plot the scatter plot to check for linearity using the below code:

DESCRIPTIVE STATISTICS AND VISULAISATION

plot(bdims$che.di ~ bdims$hgt, data = bdims, xlab = 'Height', ylab = 'Chest diameter', main = 'Scatter plot to check for Linearity')



Description of Scatter plot:

The scatter plot shows the dependent variable (chest diameter) on Y axis and predictor variable (height) on the X-axis. From the scatter plot we can see that as the value of height increases, the value of chest diameter also increases. So the data follows a positive linear relationship.

- Linear regression model

Summary:

Fit the linear regression to the data and check the summary using the below code:

DESCRIPTIVE STATISTICS AND VISULAISATION

bdims_lm <- lm(bdims$che.di ~ bdims$hgt, data = bdims)

bdims_lm %>% summary()

```
Call:
lm(formula = bdims$che.di ~ bdims$hgt, data = bdims)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3102 -1.4326 -0.0696  1.4168  6.8929

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.2947     1.7319  -1.902   0.0577 .
bdims$hgt     0.1827     0.0101  18.082   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.138 on 505 degrees of freedom
Multiple R-squared:  0.393,     Adjusted R-squared:  0.3918
F-statistic:   327 on 1 and 505 DF,  p-value: < 2.2e-16
```

HYPOTHESIS TESTING

- **Description of the summary:**

- **R-squared:** It reflects the proportion of variability in the dependent variable that can be explained by a linear relationship with the predictor variable. Therefore, Person's height measured in centimetres, explained 39.3% of variability in the measurement of chest diameter.

- **Intercept:** The constant or Intercept is the average value of y when x=0. The value represents the average chest diameter(-3.2947) measurement in centimetres when a Person's height is equal to 0.

- **Slope:** The slope of the regression line was reported as b=0.1827 . The slope represents the average increase in y following a one unit increase in x. Hence one unit increase in Person's height was related to an average increase in chest diameter measurement of 0.1827 units.

- **F-statistic:** The F-statistic is used to test the overall regression model. It has the following Hypothesis:

H0 : The data do not fit the linear regression model

HA : The data fit the linear regression model

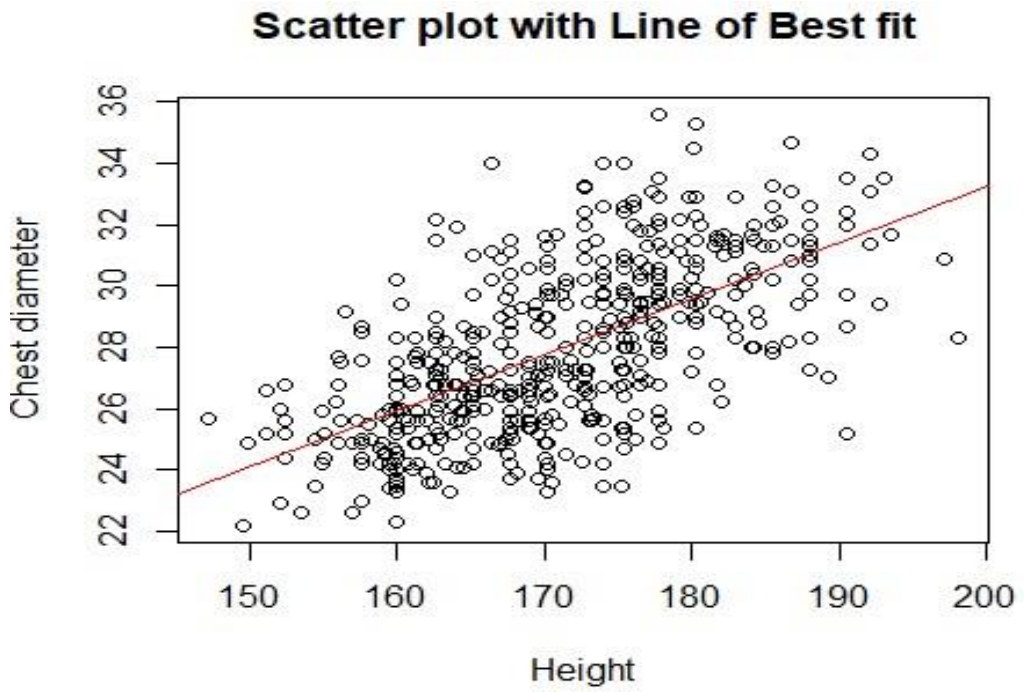
The F distribution is positively skewed, so to calculate the p-value of the observed F statistic, we need to find $Pr(F_{1,n-2}>F)$ for our data it will be Pr(F (1,505) > F). So the p-value for this as mentioned in the summary is pvalue < 2.2e-16 which is less than 0.05. So as p-value < 0.05 we reject H0. There was statistically significant evidence that the data fit a linear regression model.

HYPOTHESIS TESTING

➢ Plot the Line of best fit:

Plot the line of best fit on the scatter plot using the below code:

plot(bdims$che.di ~ bdims$hgt, data = bdims, xlab = 'Height', ylab = 'Chest diameter', main = "Scatter plot with Line of Best fit") abline(bdims_lm, col = "red")



Description:

The line of best fit is highlighted as red color in the plot. The line of best fit has the least sum of squared distances. The equation for line of best fit is: che.di = (-3.2947) + (0.183 *(hgt)) .

HYPOTHESIS TESTING

➢Assumptions:

We must validate the following assumptions for linear regression:

- **Independence**
- **Linearity**
- **Normality of residuals**
- **Homoscedasticity**

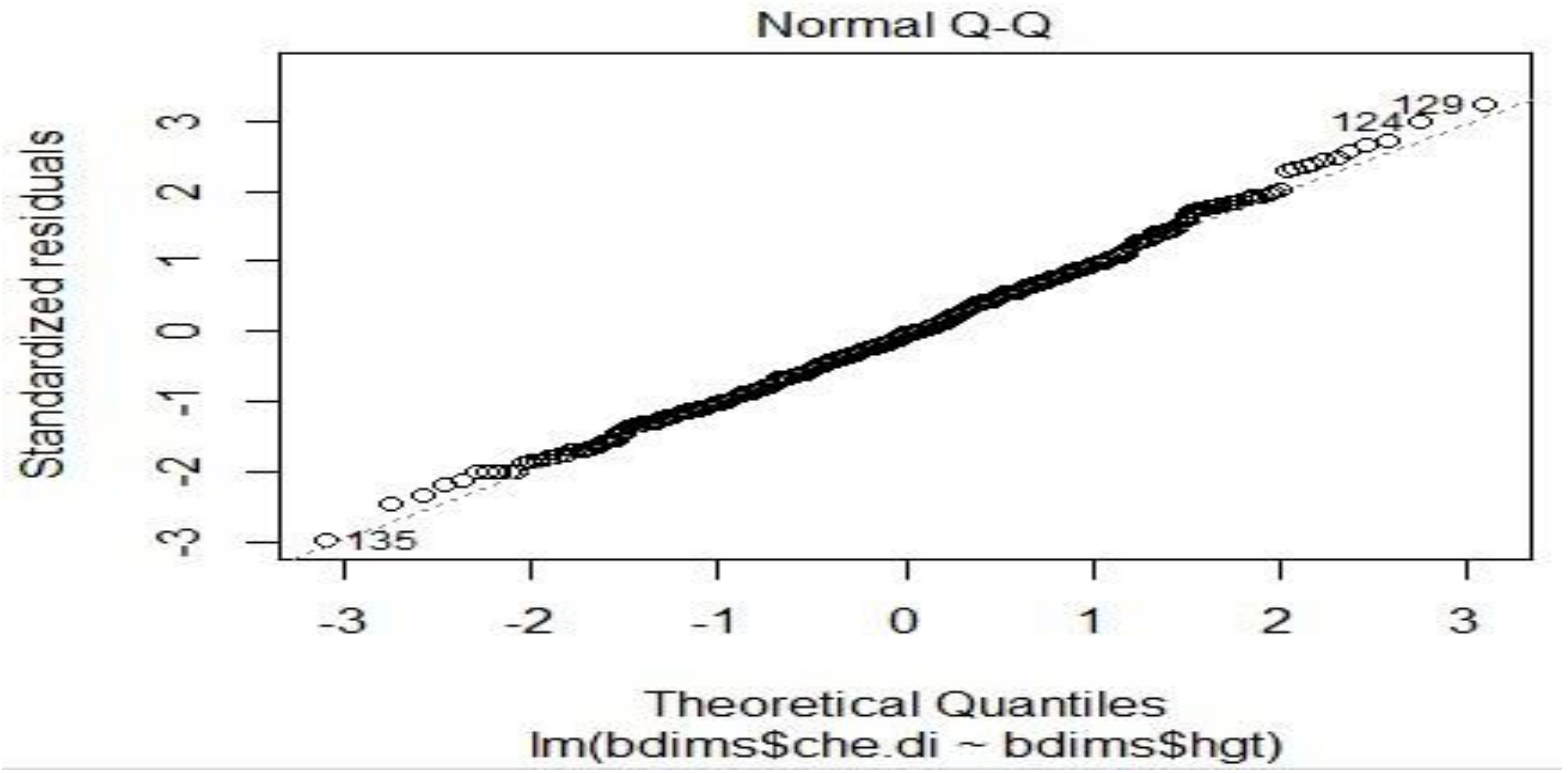
➢ **Independence**:

Independence is checked through the research design. We must ensure that all measurements between participants or observations are independent, for example, we have not included multiple measurements of height and chest diameter from the same people or knowing the measurements of one person do not share a relationship with other peoples' measurements.

HYPOTHESIS TESTING

➢ Normality of residuals – (Normal QQ):

- We check the normal Q-Q plot to determine if there were any gross deviations from normality (e.g obvious S shapes or non-linear trends). The below plot suggests there are no major deviations from normality. It would be safe to assume the residuals are approximately normally distributed.

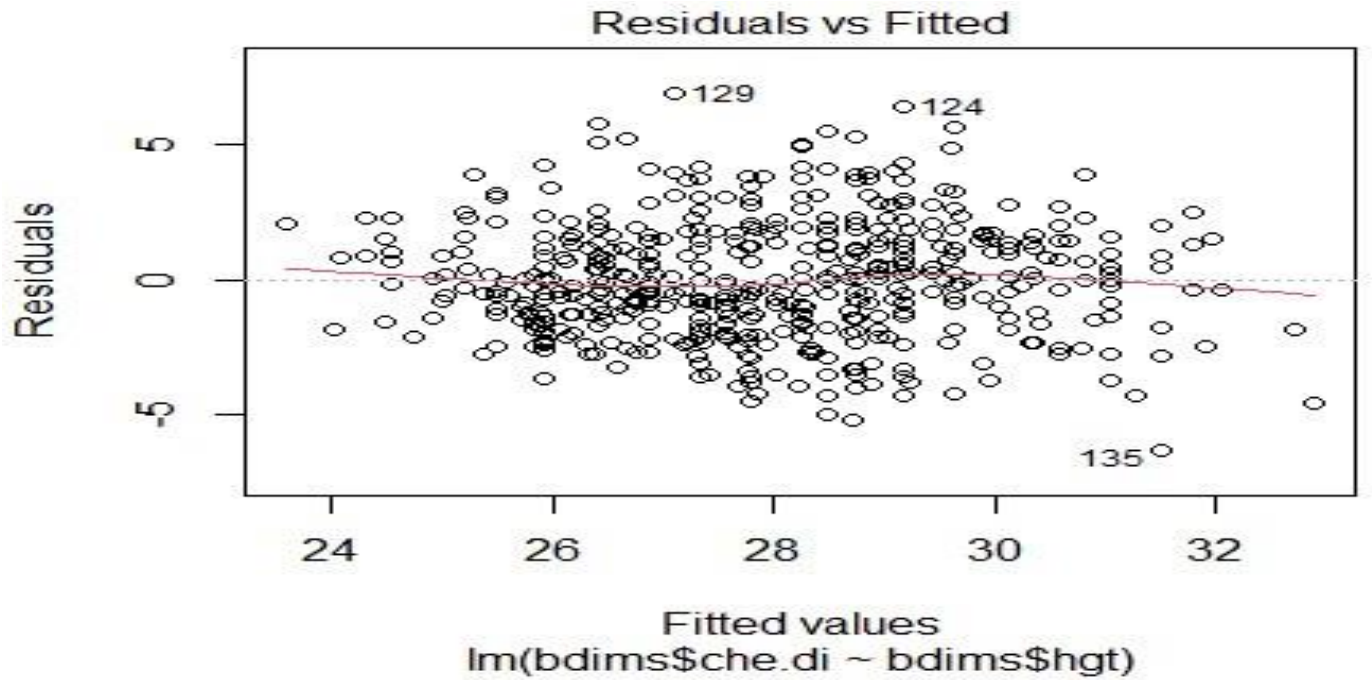


HYPOTHESIS TESTING

- **Linearity – (Residual vs Fitted):**

- Trend line should be flat. Curvature can be a sign of non-linearity
- Variability on axis should be constant across the range of values on the axis. If there is a pattern in variability, this can be a sign of heteroscedasticity.

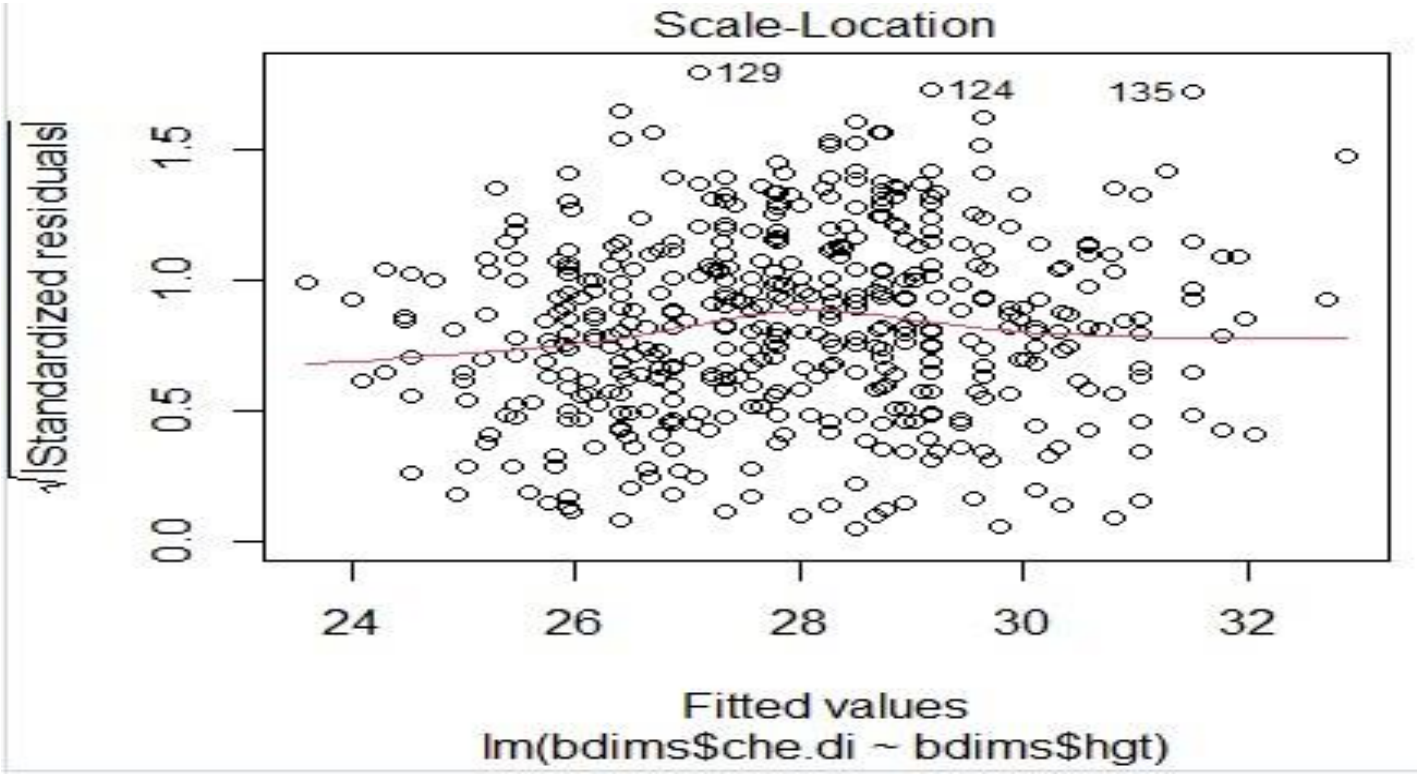
- In the plot below, the variance appears to remain the same. The straighter the line, the safer the assumption of homoscedasticity.



HYPOTHESIS TESTING

➢ **Homoscedasticity – (Scale Location) :**

- Used to check the assumption of Homoscedasticity. The red line should be close to flat.
- Variance in the square root of the standardised residuals should be consistent across predicted (fitted values). In the below plot the red line is almost flat so it is safe to assume Homoscedasticity.

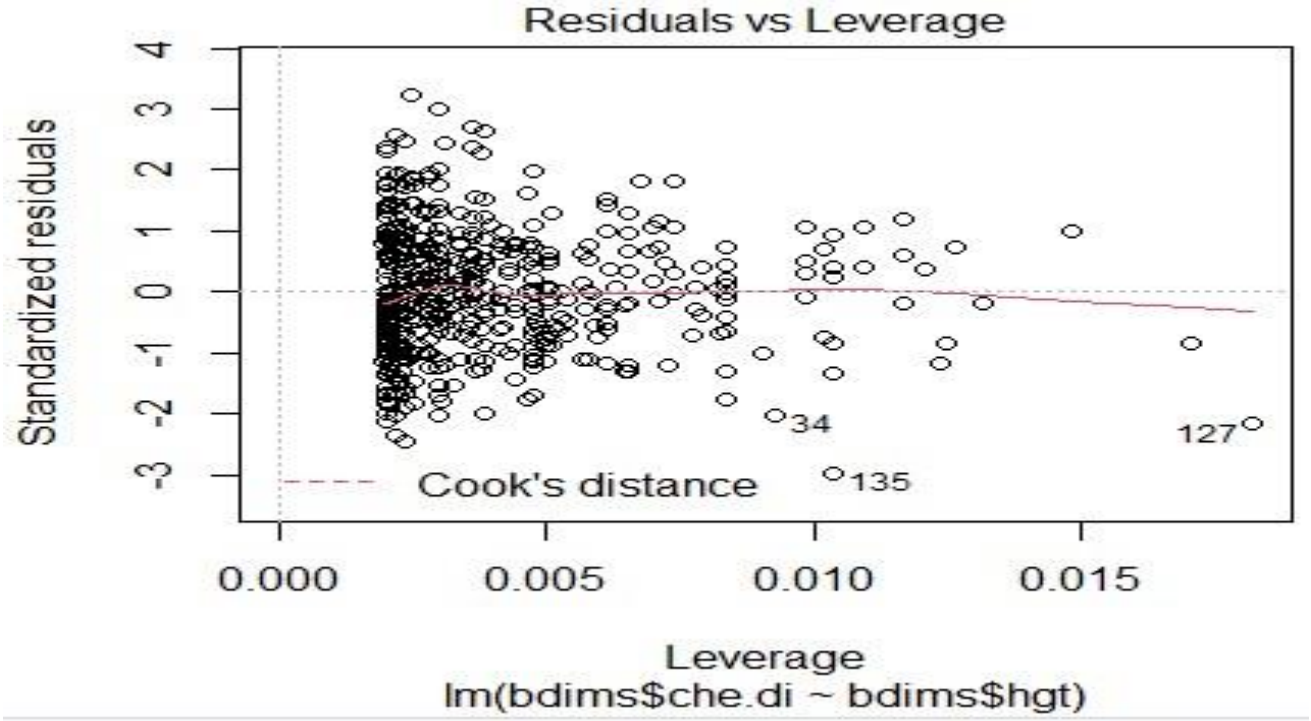


HYPOTHESIS TESTING

➢ **Residual vs Leverage(Influential Cases):**

- This plot is used to locate influential cases. Influential cases are observations that have a disproportional impact on the fit of the regression model. Outliers are at a high risk of being influential, however, not ALL outliers are influential.

- We need to look for are values that fall in the upper and lower right hand side of the plot beyond the red bands. In the plot below there are no values that fall outside the bands, and therefore, no evidence of influential cases. Infact the bands are not even visible.



HYPOTHESIS TESTING

➢**Correlation Coefficient r:**

- Measures the strength and direction of a linear relationship between two variables.

- Ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation). 0 means no correlation.
- We can perform a full correlation analysis by using the below code:

```
library(Hmisc)
bivariate<-as.matrix(dplyr::select(bdims, che.di,hgt))
rcorr(bivariate, type = "pearson")
```

```
        che.di  hgt
che.di    1.00 0.63
hgt       0.63 1.00

n= 507
```

Description of the output:

R reports the correlation between height(hgt) and che.di (chest diameter) to be r=.63 and the p-value = 0, which we write as p<.001.

The hypothesis test for r :

H0 : r = 0

HA : r != 0

```
P
        che.di hgt
che.di          0
hgt      0
```

As the p-value < 0.05 we conclude that we can reject H0. So there is a an evidence of statistically significant correlation between person's height and chest diameter.

Conclusion

➢Simple Linear regression summary:

- Linearity was assumed, normality of residuals OK, homoscedasticity OK, no influential cases.
- $r = 0.63$, $r2 = 0.393$ • F-statistic = 327, $p < 0.001$

➢Decision:

- Overall model: Reject H0

➢Conclusion:

There was a statistically significant positive linear relationship between a person's height and chest diameter.

Discussion

- A linear regression model was fitted to predict the dependent variable, che.di using the hgt(height) measured in centimetres as a single predictor. The scatter plot, demonstrated evidence of a positive linear

relationship and other non-linear trends were ruled out. The overall regression model was statistically significant, **F(1,505)= 327, p<0.001** and explained 39.3% of the variability in che.di measurement, R2 = 0.393. The estimated regression equation was **che.di=** (-**3.2947**) **+ 0.183 *(hgt**). The positive slope for hgt was statistically significant, **b =0.1827, t(505) = 18.082, p<0.001, 95%CI [0.163, 0.203]**. Final inspection of the residuals supported normality and homoscedasticity.

- A Pearson's correlation was calculated to measure the strength of the linear relationship between hgt and che.di. The positive correlation was statistically significant , **r=0.63, p<0.001.**
- From the overall regression model we can state that the height(hgt) and che.di(chest diameter) have positive linear relationship.
- Considering the first record height and chest diameter from the bdims dataset which are respectively 174cm and 28cm respectively. Using the estimated regression equation:

  Predicted che.di = (-**3.2947**) **+ 0.183 *(174**)= **28.5473.**

The prediction has an error or residual of 28 – 28.5473 = -0.547. So the Linear regression model predicts well as there was very less error.