Statistic relationship between Height and Chest diameter

AKSHAT VIJAYVARGIA (s3826627)

INTRODUCTION

As a human we evolve from time to time, we can see how with the flow of life and work, we are gradually loosing some parts of our body. For instance, there was a time, when ancient people used to have big earlobes, but with its no major use, earlobes are getting small and getting evolved with time. But only our body structure and things and cavity inside, keeps our body very well shaped without evolving. The data, which we will analyze today, is very similar to above stated lines and it is none other than measurements of our body girth and skeletal diameter. Firstly, lets understand the meaning, girth is the distance around something or you can say circumference and width. And the diameter is the length of a straight line through the center of an object or space, the synonyms are caliber, breadth and thickness. This data majorly has every body part's measurements as well as the gender, the height and the weight. The data consists of 507 physically active individuals - 247 men and 260 women.

PROBLEM STATEMENT

The investigation aims to compare the relation and understand if there is any statistical significant relationship between a person's chest diameter (che.di) and height (hgt). Our investigation will involve the summary of each variable and simple linear regression model to understand the relationship between two variables and to make a regression model. Moreover, we will use F-statistics for measuring the regression model.

DATA

The data involves 507 observations and 25 variables of measurements of wrist, ankle etc., sex, height and weight. Firstly, we will import the dataset through read_excel and will convert any NP or “-” to NA.

DESCRIPTIVE STATISTICS AND VISUALISATION

- Summary Statistics of Height

```
# SUMMARY STATISTICS OF HGT(HEIGHT)
summary(BODY$hgt, na.rm=TRUE)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
147.2   163.8   170.3   171.1   177.8   198.1 
```

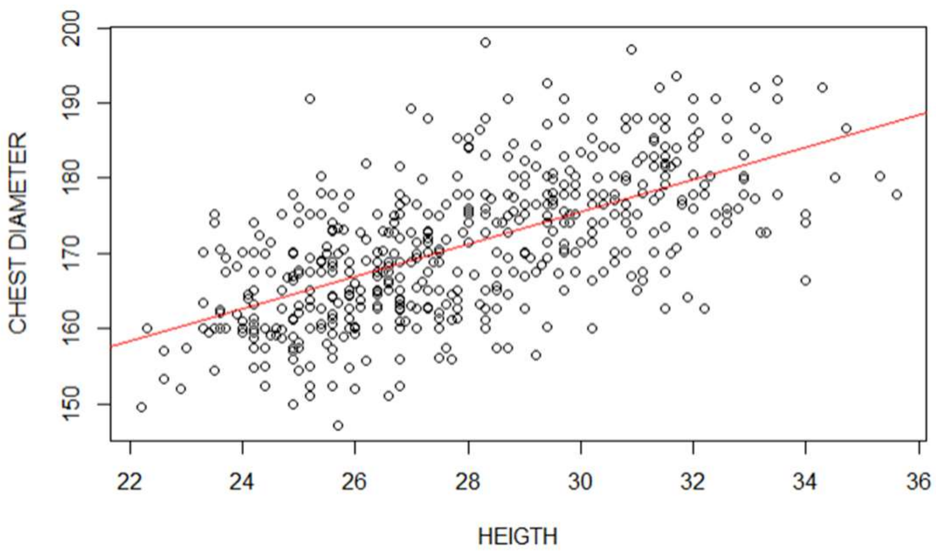
- Summary Statistics of Chest Diameter

```
# SUMMARY STATISTICS OF CHE.DI(CHEST DIAMETER)
summary(BODY$che.di, na.rm=TRUE)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
22.20   25.65   27.80   27.97   29.95   35.60 
```

CONT…

Let’s make a scatterplot to know the relationship

```
# SCATTERPLOT
plot(BODY$hgt~BODY$che.di,data = BODY,xlab="HEIGTH",ylab="CHEST DIAMETER")
abline(lm(BODY$hgt~BODY$che.di,data = BODY), col= "red")
```



From the scatter plot, we can say that these two variable has a positive relationship between them because the values of both the variables sort to have position near the red line, which shows the strength of the relationship. This also indicates that when the height increased the chest diameter will be increased as well.

HYPOTHESIS TESTING

The data exhibited a positive linear trend. A positive linear relationship occurs when as the predictor variable increases in value, so too do the values for the dependent variable. In this situation, higher height values are associated with higher chest diameter. This makes scientific sense. However, if height had decreased with increasing values for chest diameter, the relationship would be negative.

As the data exhibit signs of a positive linear relationship, we can proceed with fitting the linear regression model using R. We will work through the important code and output. Let's run the regression using the lm() function.

CONT……

```
# TESTING LINEAR REGRESSION MODEL

BODY_model <- lm(hgt~che.di,data = BODY)
BODY_model %>% summary()
```
```

Call:
lm(formula = hgt ~ che.di, data = BODY)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.0529  -5.2298   0.0753   4.8582  26.2545 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  110.972      3.344   33.19   <2e-16 ***
che.di         2.151      0.119   18.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.336 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
```

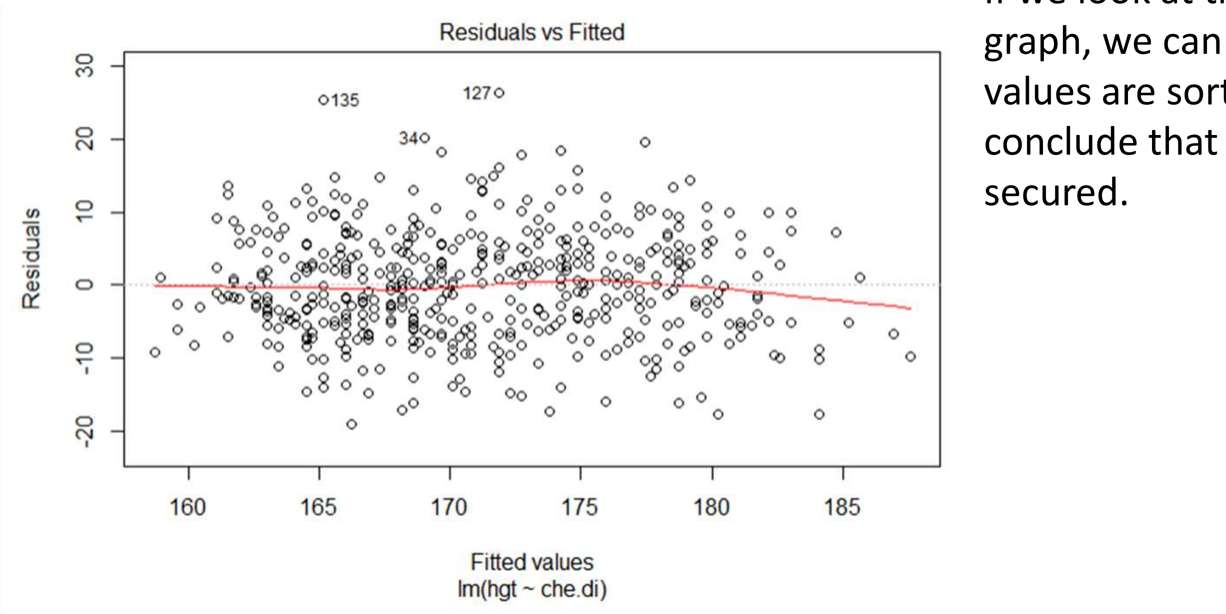
The F-test for the linear regression has the following statistical hypotheses:
H0:The data do not fit the linear regression model
HA:The data fit the linear regression model
Let's calculate the p-value for the F statistic.

```{r}
# TESTING P-VALUE

pf(q = 327,1,505,lower.tail = FALSE)

# The above p-value is 1.00343e-56 or it is p<0.01, that means it is very small compared to the level of significance 0.05. We reject the null hypothesis (H0).
```
```
[1] 1.00343e-56
```

It is statistically significant evidence that the data fit a linear regression model.

TESTING ASSUMPTIONS

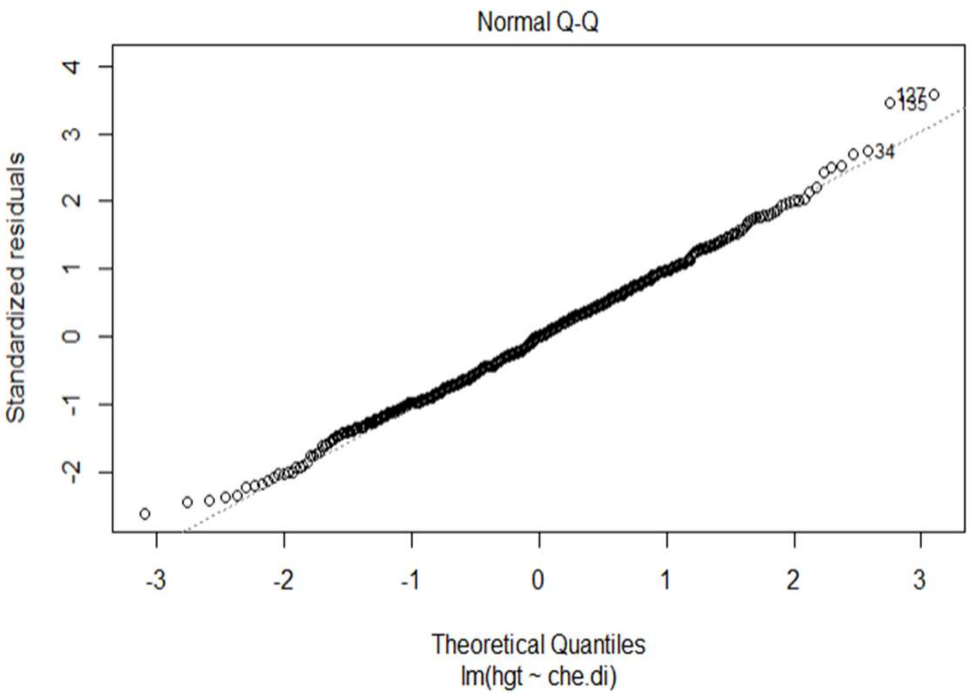
```{r}
# WE CAN ALSO CHECK THE SERIES OF PLOT

plot(BODY_model,which=1)
```



If we look at the this first graph, which is a scatter-graph, we can see our values for residuals and fitted values are sort of flat(the redline/abline). We can conclude that assumption of homoscedasticity is fairly secured.

TESTING ASSUMPTIONS

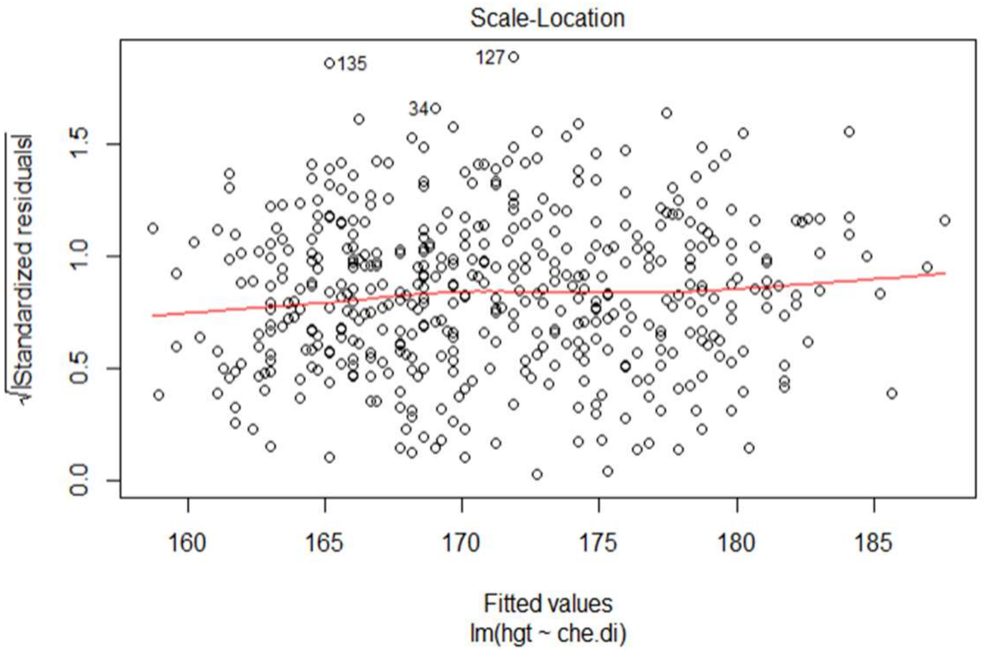
```{r}
# WE CAN ALSO CHECK THE SERIES OF PLOT

plot(BODY_model,which=2)
```



If we look at this Normal Q-Q plot, we can easily deduce that there is no major deviations in the values. So, we are free from issues from normality and residuals fit the normal distribution.

TESTING ASSUMPTIONS

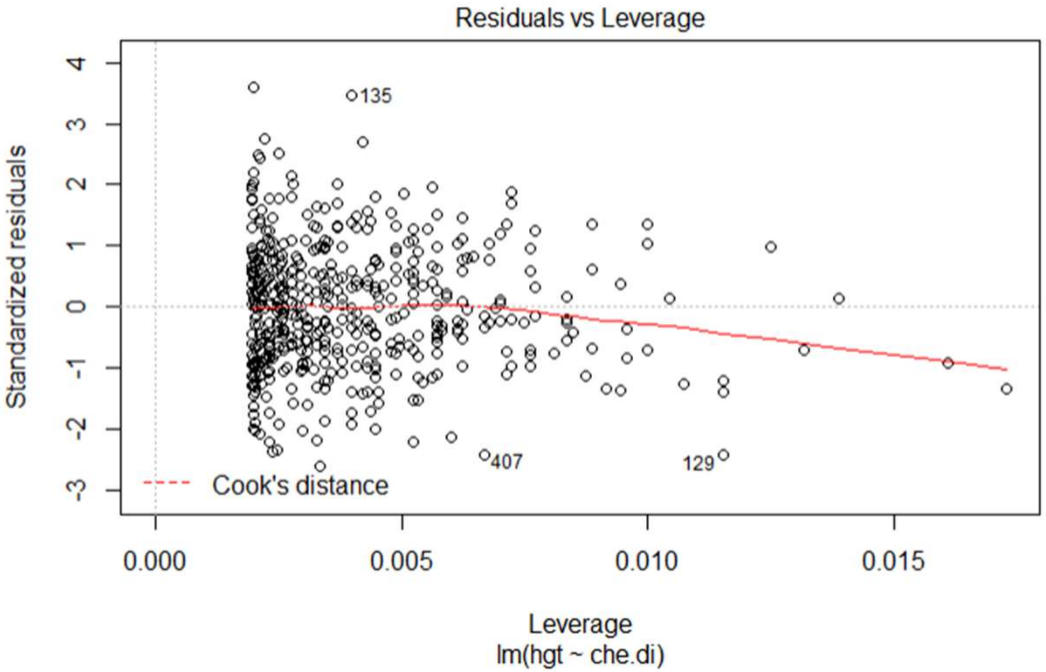
```{r}
# WE CAN ALSO CHECK THE SERIES OF PLOT

plot(BODY_model,which=3)
```



Let’s understand the Scale-Location graph, we can see that our abline(red line) is nearly straight and flat. This means variance in the square root of the standardized residuals is consistent. So, our assumption is free from risk.

TESTING ASSUMPTIONS

```{r}
# WE CAN ALSO CHECK THE SERIES OF PLOT

plot(BODY_model,which=5)
```



In this plot, we can look that our centres of the small dots are under the red line. So, we can finally say that there is no factor to disturb our regression model.

DISCUSSION

According to our scatterplot, we can say that the two variable share a positive relationship between them. Furthermore, we look at the F-test, the result also shows that variables-height and chest diameter meet the linear regression model.

“In this investigation, we found that there is a positive relationship between a person’s chest diameter (che.di) and height (hgt).”

The strength of my investigation is that I got the data from RMIT University, which is one of the best Government university in Australia. Moreover, the data can also be seen on Kaggle(https://www.kaggle.com/vikrant4/bdims), which is a platform for all authentic open source data.

The limitation of the investigation is that, I am restricted some factors i.e. Height and Chest Diameter to show the dependency and relationships. Therefore, my investigation may not fully true.

The direction for future investigation is collecting and considering more data and more variables or factors to know the truth or at least get closer to it.

REFERENCES

- Definition Of Girth-https://www.thefreedictionary.com/girth
- Definition of Diameter-https://www.merriam-webster.com/dictionary/diameter
- Understanding the Data-https://www.rdocumentation.org/packages/openintro/versions/1.7.1/topics/bdims
- Module 10-file:///C:/Users/61422/AppData/Local/Packages/Microsoft.MicrosoftEdge_8wekyb3d8bbwe/TempState/Downloads/Week_10_Demo%20(2).pdf
- Reference for the assignment-https://astral-theory-157510.appspot.com/secured/MATH1324_Module_09.html