Linear Regression & Correlation Test

BODY MEASUREMENTS DATA MODELING
BY CHRISTIAN THEMIN

20 MAY 2020

Introduction

This report will investigate the Body Measurements Dataset (bdims.csv) to check whether there is any statistical significant relationship between a person’s Chest Diameter (inch) and Height (cm) using R programming.

Problem Statement

Investigators want to understand the relationship between the Chest Diameter and Height in order to make accurate predictions. This report will use the measurement of Linear Regression and Correlation.

A hypothesis test will also be implemented to test any assumptions that are in doubt and whether there is any statistically significant between the two variables. It will then be discussed with a conclusion.

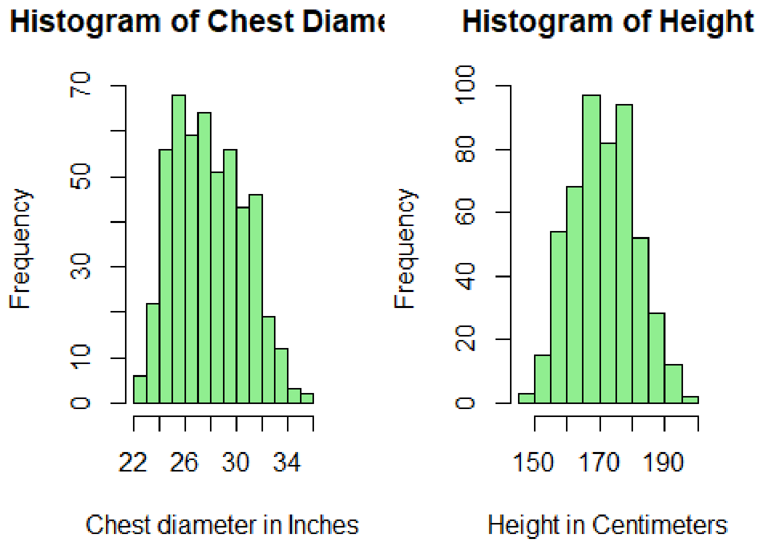
Data

The bdims.csv data is collected from the Data Respiratory folder accessible for RMIT students only. The dataset consists of 507 physically active individuals and 25 variables:

1. Biacromial Diameter (cm)
2. Biiliac Diameter (cm)
3. Bitochanteric Diameter (cm)
4. Chest Depth (cm)
5. Chest Diameter (inch)
6. Elbow Diameter (cm)
7. Wrist Diameter (cm)
8. Knee Diameter (cm)
9. Ankle Diameter (cm)
10. Shoulder Girth (cm)
11. Chest Girth (cm)
12. Waist Girth (cm)
13. Navel Girth (cm)
14. Hip Girth (cm)
15. Thigh Girth (cm)
16. Bicep Girth (cm)
17. Forearm Girth (cm)
18. Knee Diameter (cm)
19. Calf Maximum Girth (cm)
20. Ankle minimum Girth (cm)
21. Wrist Minimum Girth (cm)
22. Age (years)
23. Weight (kg)
24. Height (cm)
25. Sex (male/female)

Check for Outliers

There were no outliers/missing value found in the dataset using Histogram.



```
par(mfrow=c(1,2))
hist(bdims$che.di, col="lightgreen", main = "Histogram of Chest Diameter",
     xlab = "Chest diameter in Inches")
hist(bdims$hgt, col="lightgreen", main = "Histogram of Height", xlab = "Height in
Centimeters")
```

Summary Statistics

The Descriptive Statistics for Chest Diameter is as follows:

```
##    Min    Q1 Median     Mean    Q3      SD  Max   n
## 1 22.2 25.65   27.8 27.97377 29.95 2.74165 35.6 507
```

```
# Descriptive Statistics of Chest Diameter
bdims %>% summarise (
  Min = min(che.di),
  Q1 = quantile(che.di, probs = .25),
  Median = median(che.di),
  Mean = mean(che.di),
  Q3 = quantile(che.di, probs = .75),
  SD = sd(che.di),
  Max = max(che.di),
  n = n()
)
```

Summary Statistics

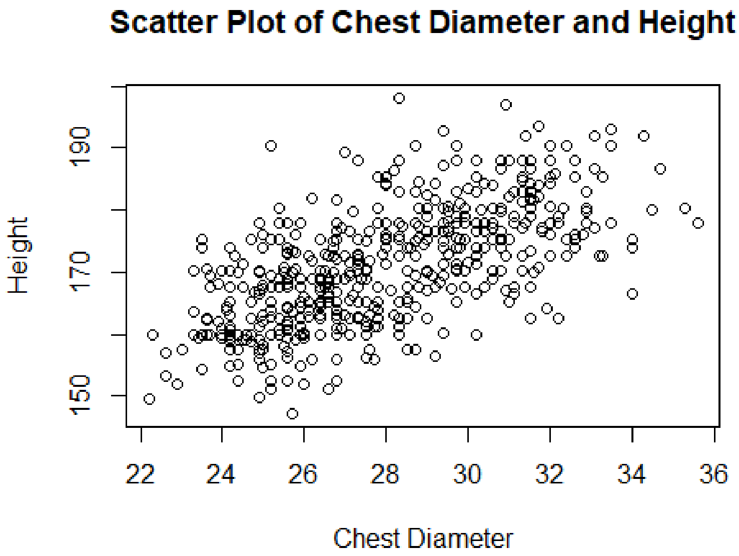
The Descriptive Statistics for Height is as follows:

```
##     Min    Q1 Median     Mean    Q3       SD   Max   n
## 1 147.2 163.8  170.3 171.1438 177.8 9.407205 198.1 507
```

```
# Descriptive Statistics of Height
bdims %>% summarise (
  Min = min(hgt),
  Q1 = quantile(hgt, probs = .25),
  Median = median(hgt),
  Mean = mean(hgt),
  Q3 = quantile(hgt, probs = .75),
  SD = sd(hgt),
  Max = max(hgt),
  n = n()
)
```

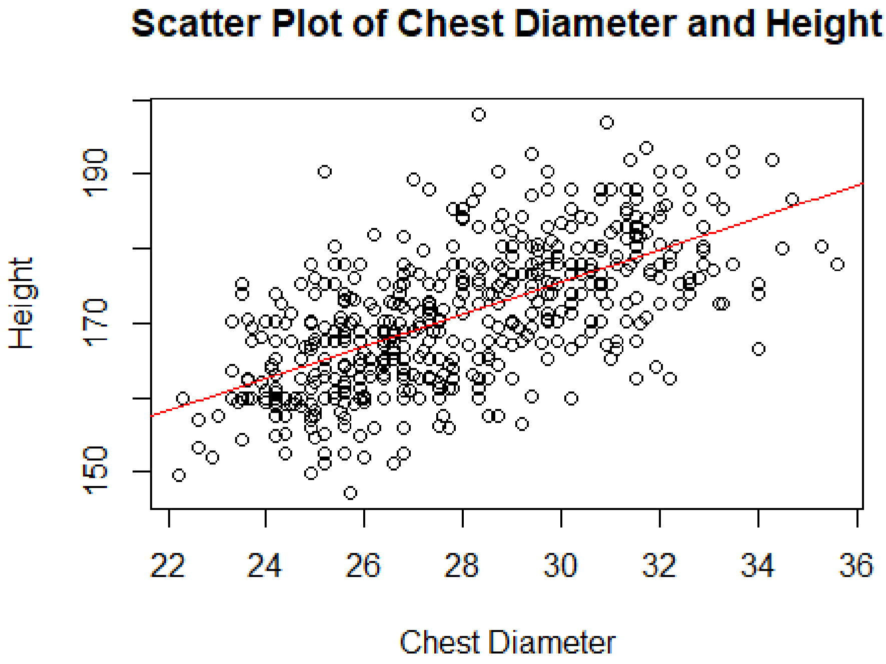
Hypothesis Test for Linear Regression Model

The Scatter Plot is used to check whether there is any *linear relationship* between the two variables having Chest Diameter as the Predictor and Height as the dependent variable.



```
plot(hgt~che.di, data = bdims, xlab="Chest Diameter", ylab="Height", main = "Scatter Plot of Chest Diameter and Height")
```

Fitting a regression line to the sample data



The result shows that there is a *positive linear relationship* between the two variables as can be seen from the red line applied on the Scatter plot.

Statistical Test | F-Statistic

To test the statistical significance of the F-Statistic, we set the following hypothesis test:

H0: The data does not fit the linear regression model

HA: The data fits the linear regression model

In R Studio, we will perform the following code:

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

Statistical Test | F-Statistic

```
## Call:
## lm(formula = hgt ~ che.di, data = bdims)
##
## 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 output shows there is a statistically significant, with F(1, 505) = 327, p < 0.001.
The R-squared = 0.393, the Chest diameter explained 39.3% of the variability in Height.

Statistical Test | Intercept & Slope

To test the statistical significance of the constant/intercept, we set the following statistical hypothesis:

$$H0:\alpha=0$$

$$HA:\alpha\neq0$$

And for the slope, we set the following statistical hypothesis:

$$H0:\beta=0$$

$$HA:\beta\neq0$$

We will check the Hypothesis in R by performing the following code:

```
fitmodel %>% summary() %>% coef()
fitmodel %>% confint()
```

Statistical Test | Intercept & Slope

```
##                Estimate Std. Error  t value      Pr(>|t|)
## (Intercept) 110.971967  3.3436580 33.18879 5.346989e-129
## che.di        2.151009  0.1189595 18.08186  1.017694e-56

##                  2.5 %     97.5 %
## (Intercept) 104.402773 117.541160
## che.di        1.917292   2.384725
```

The intercept of the regression for height was statistically significant,
a = 110.972, p < 0.001, 95% CI (104.403, 117.541).

The slope of the regression for chest diameter was statistically significant,
b = 2.151, p < 0.001, 95%CI (1.917, 2.385)

Making Predictions

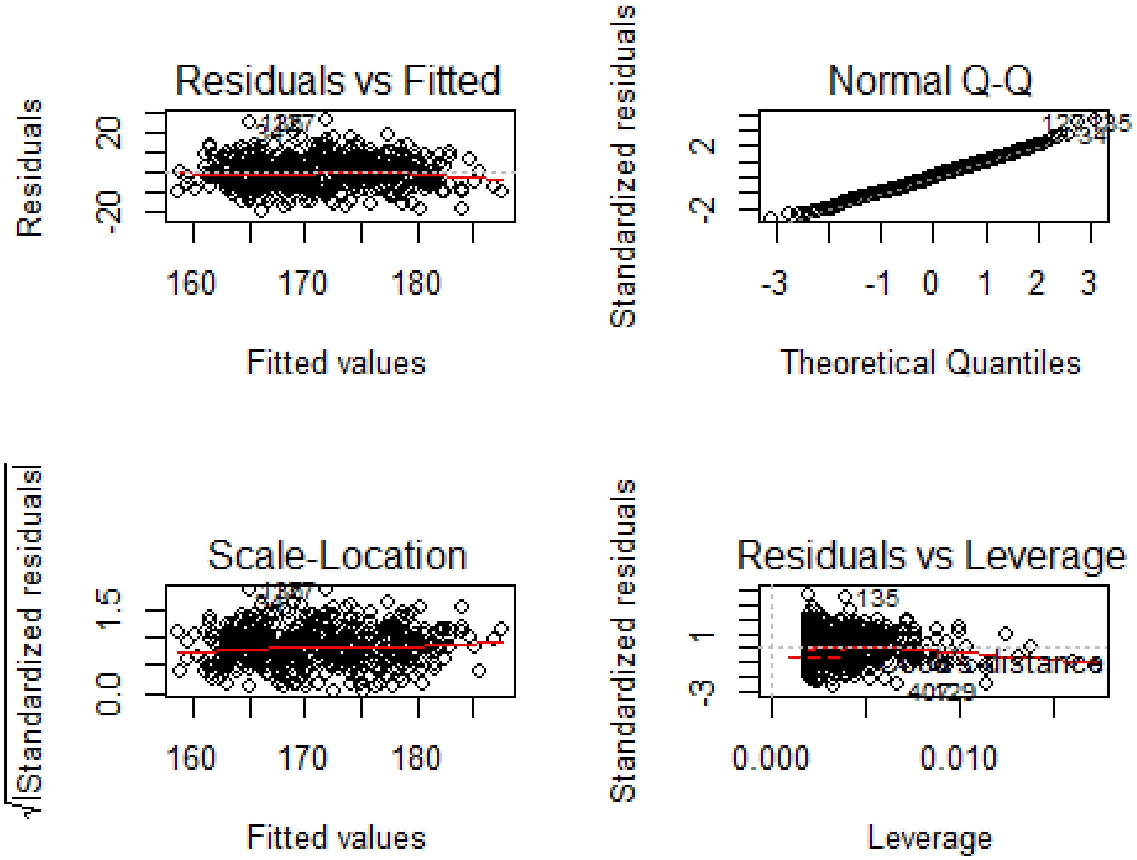
Using the estimated linear regression model, let’s predict the chest diameter of 34.5 inch and height of 165 cm.

We will perform the following code:

```
predict(fitmodel,new=data.frame(che.di=34.5))
            ##          1
            ## 185.1818
```

This result of the prediction is 185.182, it did not predict the value well. The prediction has an error or residual of -20.182mm (165 - 185.182)

Assumptions test for Linear Regression



```
par(mfrow=c(2,2))
plot(fitmodel)
```

Assumptions test for Linear Regression

Independence: Independence was assumed as each height and chest diameter measurement came from different people.

Linearity: The scatter plot suggested a linear relationship. Other non-linear relationships were ruled out. There were no non-linear trends in the Residual vs Fitted Plot.

Normality of Residuals: Normal Q-Q plot didn’t show any obvious departures from normality.

Influential Cases: There appeared to be no influential cases.

Homoscedasticity: Homoscedasticity looks fit to the line according to the scale-location plot and the variance in residuals appeared to be constant.

Correlation Test

To test the statistical significance for r, we can set the hypothesis test as:

H0:r=0 | HA:r≠0

The following code is executed in R Studio:

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

```
##          hgt che.di
## hgt     1.00   0.63
## che.di  0.63   1.00
##
## n= 507
##
##
## P
##         hgt  che.di
## hgt          0
## che.di   0
```

The correlation between Chest diameter and Height to r = 0.63 and the p-value is < 0.001.

The p-value for r can be readily calculated by converting r to a t statistic of the chest diameter which is 18.08.

Correlation Test

Therefore, a two-tailed p-value for r and 95% CI can be calculated as:

```
2*pt(q=18.08, df = 507-2, lower.tail = FALSE)
## [1] 1.038739e-56

r <- cor(bdims$hgt,bdims$che.di, use="complete.obs")
CIr(r, n=507, level=.95)
## [1] 0.5709813 0.6770164
```

The positive correlation was statistically significant,
$r = 0.63$, $p < 0.001$, 95%CI [0.571, 0.677]

Conclusion

Linear Regression Model

A linear regression model was fitted to predict the dependent variable (height) using the independent variable (Chest Diameter). Prior fitting the regression, a scatter plot assessed the bivariate relationship between Height and Chest Diameter. The scatter plot demonstrated evidence of a positive linear relationship. The overall regression model was statistically significant, F(1, 505) = 327, p < 0.001. The R-squared = 0.393, the Chest diameter explained 39.3% of the variability in Height. The positive slope for Chest Diameter was statistically significant, b = 2.151, p < 0.001, 95%CI (1.917, 2.385). Final inspection of the residuals supported normality and homoscedasticity.

Conclusion

Correlation Model

The Pearson correlation was calculated to measure the strength of the linear relationship between Chest Diameter and Height.

The positive correlation was statistically significant, r = 0.63, p < 0.001, 95%CI [0.571, 0.677].

Prediction Discussion

There was a statistically significant and positive linear relationship between the Chest Diameter and Height, when making the prediction using the real-world data, it did not predict the value well.

The prediction result is 185.18 with residual error of -20.182mm when measuring Chest Diameter of 34.5 inch and Height of 165 cm.

References

RMIT Lecture Slides, Module Notes, Worksheets, and Tutorials.

All works are done individually by Christian Themin.