Analysis on person's Chest diameter and Height

Shonil Dabreo

Analysis on Admitted patients

- Introduction
- Problem Statement
- Data
- Descriptive Statistics and Visualization
- Hypothesis Testing
- Discussion
- References

Introduction

- Body measurements of different person such as chest, hip, shoulder, wrist, knee, etc are recorded.
- The problem is to understand the statistical significant relationship between a person's Chest Diameter (che.di) and Height (hgt).
- Statistics is undertaken to find the relationship between a person's Chest Diameter (che.di) and Height (hgt). Linear regression model is used to explain the relationship between variables. The results of the statistics between an independent and dependant variable are interpreted.

Problem Statement

- The goal is to understand if there is any statistical significant relationship between a person's Chest Diameter (che.di) and Height (hgt).
- The problem is established by assuming a null hypothesis that the data doesn't fit the regression model.
- Linear regression model is used to explain any statistical significant relationship between an independent and dependent variable.

Data

Data preprocessing is extremely important because it allows improving the quality of the raw experimental data. The primary aim of preprocessing is to eliminate those small data contributions associated with the experimental error.
Data Cleaning is one of the Data preprocessing step for collecting and preparing the data.

Steps of Data Cleaning:-

- Identify and remove outliers
- Feature Engineering

Data

Identify and remove outliers

- The bdims (i.e. Body dimensions) dataset and necessary packages are imported into the studio for analysis. The dataset has 25 columns and 507 entries.
- The is.na function was used to find the null values and then the sum of those values were calculated.
- The dataset didn’t had any missing values present in it.
- Moreover, the data was already cleaned with appropriate datatype for each feature.

Data

Feature Engineering

- The attributes of interest are selected to calculate the statistical information.
- The ‘che.di’ and ‘hgt’ are the attributes which are selected to perform the statistics.
- ‘che.di ‘is the measurement of the person’s Chest Diameter (cm).
- ‘hgt’ is the measurement of the person’s Height (cm).

Descriptive Statistics and Visualization

Descriptive Statistics

- Descriptive statistics of Chest Diameter is calculated using summarize function.

```
# Descriptive statistics
bdims %>% summarise(
  Min = min(che.di, na.rm = TRUE),
  Q1 = quantile(che.di, probs = .25, na.rm = TRUE),
  Median = median(che.di, na.rm = TRUE),
  Q3 = quantile(che.di, probs = .75, na.rm = TRUE),
  Max = max(che.di, na.rm = TRUE),
  Mean = mean(che.di, na.rm = TRUE),
  SD = sd(che.di, na.rm = TRUE),
  n = n(),
  Missing = sum(is.na(che.di))
)
```

Code

```
# 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
```

Output

Descriptive Statistics and Visualization

Descriptive Statistics

- Descriptive statistics of Height is calculated using summarize function.

```
bdims %>% summarise(
  Min = min(hgt, na.rm = TRUE),
  Q1 = quantile(hgt, probs = .25, na.rm = TRUE),
  Median = median(hgt, na.rm = TRUE),
  Q3 = quantile(hgt, probs = .75, na.rm = TRUE),
  Max = max(hgt, na.rm = TRUE),
  Mean = mean(hgt, na.rm = TRUE),
  SD = sd(hgt, na.rm = TRUE),
  n = n(),
  Missing = sum(is.na(hgt))
)
```

Code

```
# 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
```

Output

Descriptive Statistics and Visualization

Descriptive Statistics

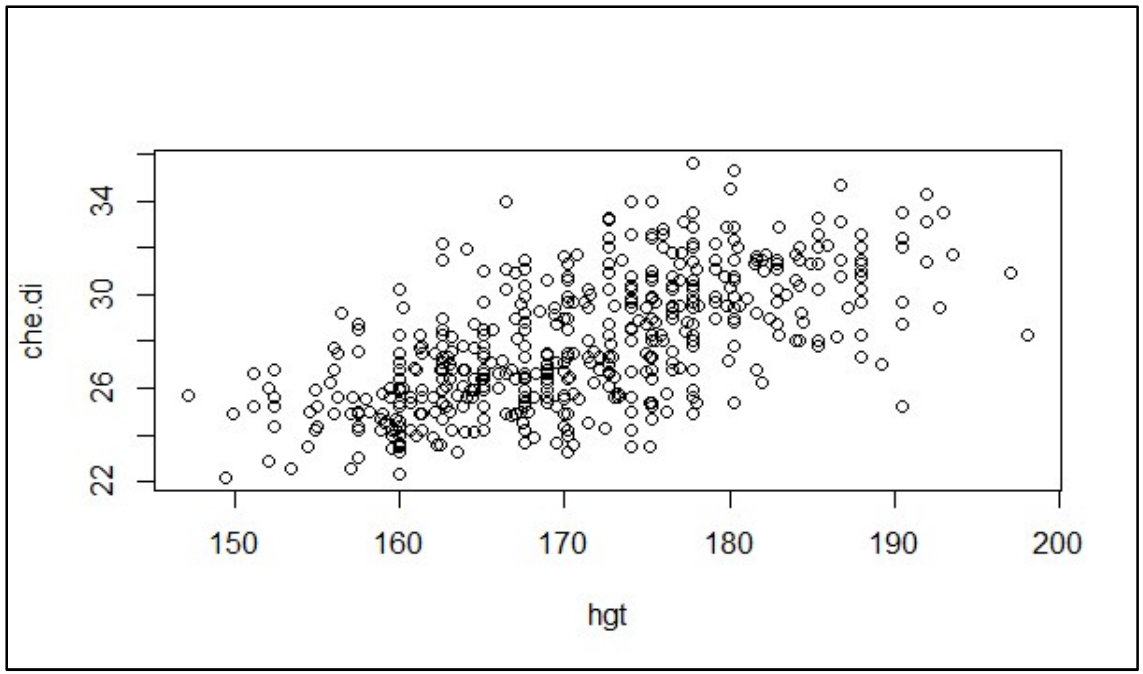
- The mean of Chest Diameter and Height is 28 cm and 171cm respectively. Whereas, median of Chest Diameter and Height is 27.8 cm and 170 cm respectively.
- The mean and median of both Chest Diameter and Height are close to each other. So we can say that the data is fairly balanced, or symmetric, on each side.
- The standard deviation of Chest Diameter is 2.74 and the standard deviation of Height is 9.41.

Descriptive Statistics and Visualization

Visualization

```
# scatter plot
plot(che.di ~ hgt, data = bdims, xlab = "hgt", ylab = "che.di")
```

Code



Output

Descriptive Statistics and Visualization

Visualization

- Scatter plot of the Chest Diameter (che.di) with respect to the Height (hgt) is plotted for visual analysis.
- Looking at the output, we can say that, as the value of Height increases, the value of the Chest Diameter also increases.
- This is a positive linear relationship.
- A positive linear relationship occurs when as the predictor variable (hgt) increases in value, so too do the values for the dependent variable (che.di).
- As the data exhibit signs of a positive linear relationship, we can fit the data into the linear regression model.

Hypothesis Testing

- Assuming null hypothesis as the data doesn't fit the regression model and alternate hypothesis as the data fit the regression model.
- Null hypothesis was rejected as $p < .001$. Therefore, the linear model was statistically significant with $F(1, 505) = 327$ and $p < .001$.
- The Height explained 39.3% of variability in the Chest Diameter.

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

Code

```
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 .
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
```

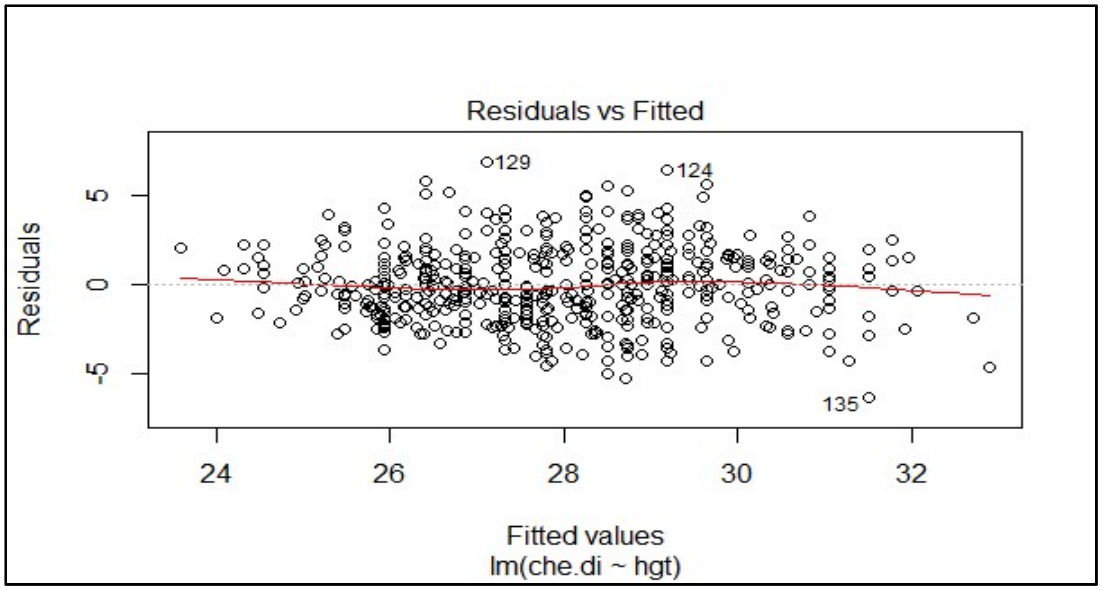
Output

Hypothesis Testing

- The estimated average Chest Diameter when Height= 0 was -3.29 cm.
- A Height with 0 is impossible. Therefore, the intercept usually don't have a meaningful interpretation.
- As **p >= .05** for the intercept, the intercept of the regression was close to being statistically significant with a = -3.29**.**
- For every one unit increase in Height, the mean Chest Diameter was estimated to increase on average by 0.18 cm. This is a positive change.
- As **p < .05** for the slope, the slope of the regression was statistically significant with b = 0.18.
- The Chest Diameter (intercept) isn't significant which means that its appearance is based on the Height (slope) of the person and not the other way round.

Hypothesis Testing

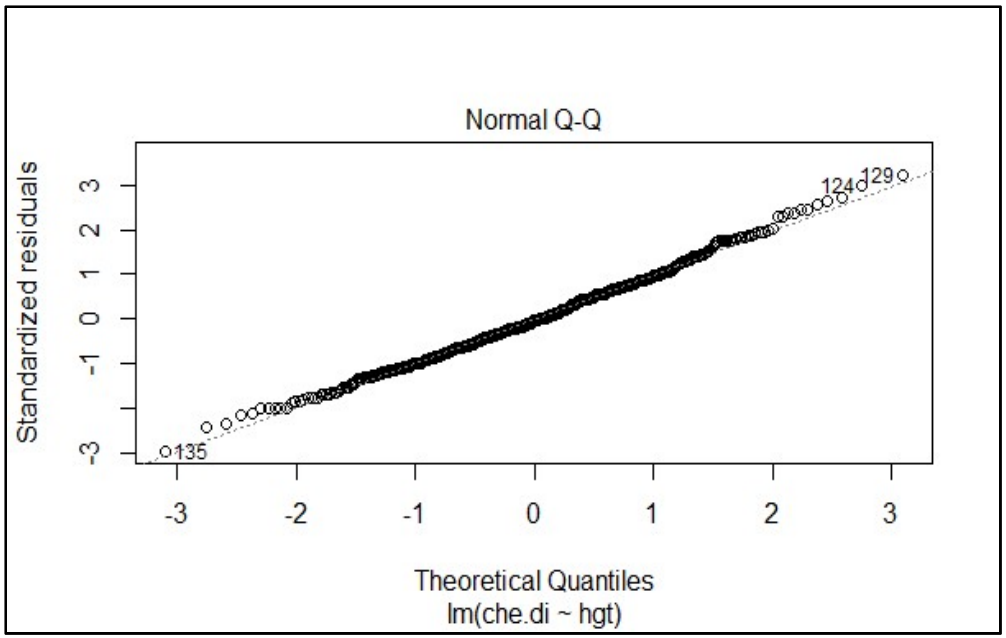
- To report the final regression model, assumptions must be validated for linear regression.
- **Independence** was assumed as each Chest Diameter and Height measurement came from different people.
- **Linearity:** The scatter plot suggested a linear relationship (i.e. flat red line). There were no non-linear trends in the Residual vs. fitted plot.



Output

Hypothesis Testing

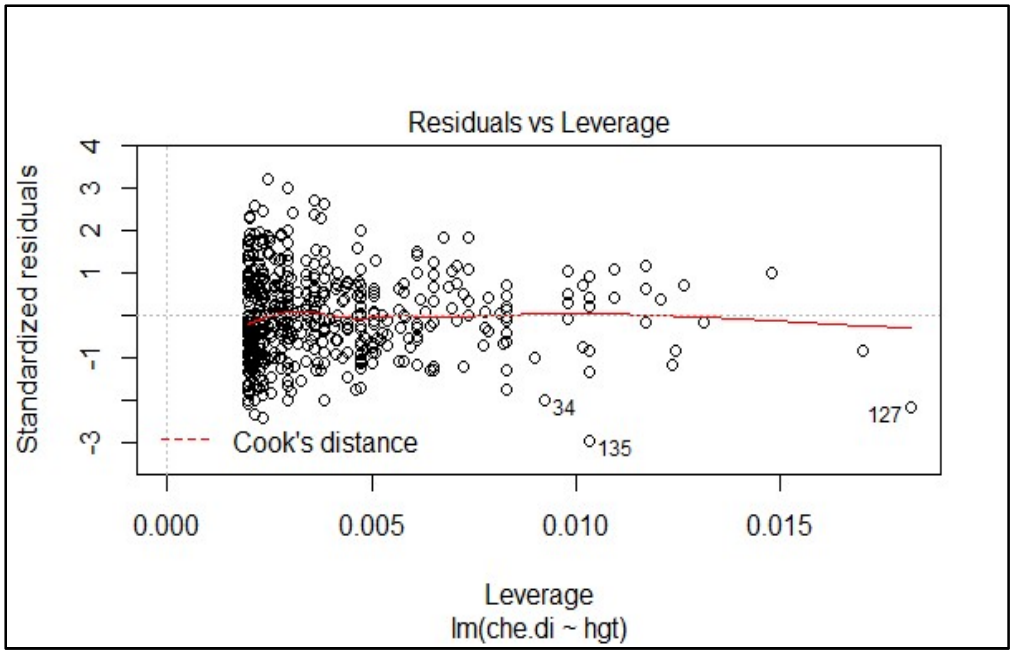
- **Normality of residuals:** The normal Q-Q plot was used to determine if there were any gross deviations from normality (e.g. obvious S shapes or non-linear trends).
- The plot suggested that there were none major deviations from normality.



Output

Hypothesis Testing

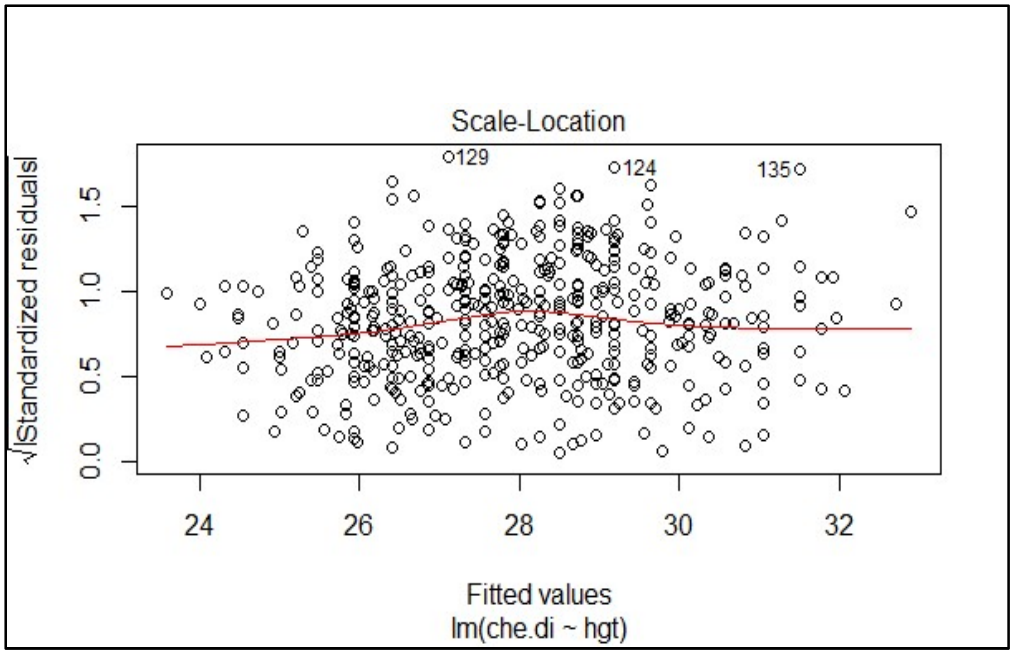
- **Influential cases:** The residual vs. leverage plot is used to identify cases that might be unduly influencing the fit of the regression model, for example, outliers. However, all the outliers are not influential.
- There are no values that fall outside the cook's distance bands, and therefore, there was no evidence of influential cases.



Output

Hypothesis Testing

- **Homoscedasticity:** In the scale location plot, the variance in the square root of standardized residuals appeared to be consistent across fitted values (i.e. flat red line).



Output

Discussion

- The scatter plot indicated a positive linear relationship before fitting the data into the linear regression model and other non-linear relationships were ruled out.
- Final inspection of the residuals supported normality and homoscedasticity.
- A person's Height was estimated to explain up to 39.3% of the variability in Chest Diameter.
- Overall, there was a statistically significant positive linear relationship between Chest Diameter and Height.

References

- Science direct. Data Pre-processing [online]. Available at < https://www.sciencedirect.com/topics/engineering/data-preprocessing > [Accessed 10 May 2020]

- Astral theory. Appspot. MATH1324_Module 09 [online]. Available at < https://astral-theory-157510.appspot.com/secured/MATH1324_Module09.html#example_-_oxygen_uptake_efficiency_slope > [Accessed 23 May 2020]