ASSIGNMENT 3

APPLIED ANALYTICS

NAME : YOGESH HARESH BOJJA

STUDENT ID : S3789918

INTRODUCTION

We have been provided with ‘bdims.csv’. This dataset has body girth measurements and skeletal diameter of 507 physically active individuals - 247 men and 260 women. Nine skeletal measurements (diameter measurements) and twelve girth (or circumference) measurements, as well as age, weight, height, and gender, are available in this dataset.

PROBLEM STATEMENT

In this investigation we will seek to understand if there is any statistically significant relationship between a person's chest diameter (che.di) and height (hgt).

We will find the relationship between person's chest diameter and height using correlation and based on that we will build the linear regression model for both the attributes. We will be doing various tests through visualizations to check whether the data fits good for building a linear model. We will be testing different hypothesis and based on that we will decide how good our model is.

DATA

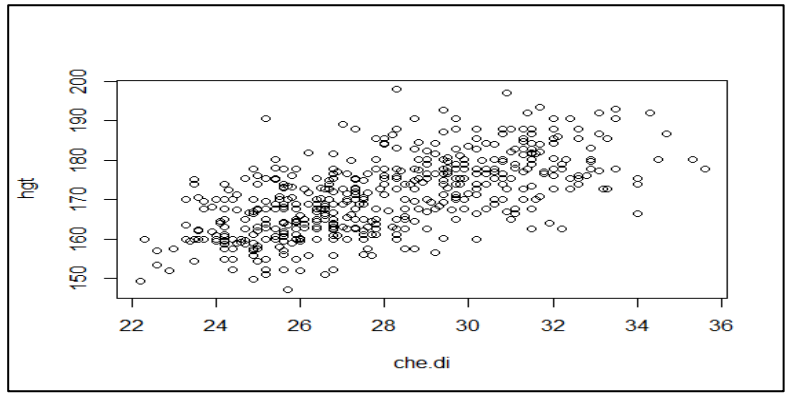
1) We are reading the data in df variable using read.csv() function. As we are dealing with person's chest diameter and height, we will subset the required columns from entire dataset as shown in the code below. We will check whether there are any null values present in the column using is.na() function. We can see from the output that there are no null values present hence there is no need to handle them.

```
> df <- read.csv("bdims.csv")
> df <- df[, c("che.di", "hgt")]
>
> sum(is.na(df$che.di))
[1] 0
> sum(is.na(df$hgt))
[1] 0
```

2) We want to check how does person's height depend on person's chest diameter, hence 'hgt' will be our dependent variable and 'che.di' will be our independent variable. We will plot the points from the table to check how are the point distributed.

```
> plot(hgt ~ che.di, data = df, xlab = "che.di", ylab = "hgt")
>
```

In the output of plot() we can see that there is moderate positive linear relationship between 'hgt' and 'che.di'. Let us approve this visuals by testing different hypothesis further.



BUILDING LINEAR MODEL

In R we build linear model using lm() function. Its first parameter is dependent variable and independent variable separated with ~, and second parameter is variable name of the data. With the help of summary() we can check the computed values of slope, intercept, residuals etc.

```
> bdims_model <- lm(hgt~che.di, data = df)
> summary(bdims_model)

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

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
```

Value of intercept is 110.972 while slope of the model is calculated 2.151. R-squared reflects the proportion of variability in the dependent variable that can be explained by a linear relationship with the predictor variable. R-squared = 0.393 says that 39.3% of dependent variables variability is depended on the independent variable. F-statistics which is required further for testing hypothesis is also computed by lm() function.

Here we are going to see how are the values of intercept and slope calculated without lm() function.

1) **Intercept :** Value of independent variable when dependent variable is equal to 0 is said to be intercept.

```
> a = mean(df$hgt) - b*mean(df$che.di)
> print(a)
[1] 110.972
```

2) Slope : Slope is the amount of increase or decrease in the dependent variable on one-unit change of independent variable.

```
> Lxx <- sum(df$che.di^2) - ((sum(df$che.di)^2)/n)
> print(Lxx)
[1] 3803.421
> Lyy <- sum(df$hgt^2) - ((sum(df$hgt)^2)/n)
> print(Lyy)
[1] 44778.73
> Lxy <- sum(df$che.di*df$hgt) - ((sum(df$che.di)*sum(df$hgt))/n)
> print(Lxy)
[1] 8181.192
> b = Lxy/Lxx
> print(b)
[1] 2.151009
```

HYPOTHESIS TESTING (Testing overall regression model)

With the help of F-statistics computed by lm() we will check how well does the regression model fit our data. Hence the null and alternative hypothesis are given below.

H_0 : The data do not fit the linear regression model.
H_A : The data fit the linear regression model.

F-statistics can be computed with the help of anova() function too as shown below.

```
> anova(bdims_model)
Analysis of Variance Table

Response: hgt
           Df Sum Sq Mean Sq F value    Pr(>F)
che.di      1  17598 17597.8  326.95 < 2.2e-16 ***
Residuals 505  27181    53.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F- statistics calculated is 326.95. We will calculate p-value of the F-statistics and compare with the area of significance i.e. 0.05. Degree of freedom is 1 and n-2(507-2 = 505).

```
> pf(q=326.95, 1, 505, lower.tail = FALSE)
[1] 1.01882e-56
>
```

As p-value is less than the area of significance we can reject the null hypothesis. Hence, we can conclude that the data fits the linear regression model well.

HYPOTHESIS TESTING (Testing statistical significance of constant α)

The constant or intercept is the value of dependent variable when the value of independent variable is 0. In our example hgt is 110.972 when che.di is 0 which does not make sense. 'a' is the intercept which is calculated through our sample dataset hence intercept **a** = 110.972 can not be generalized to the intercept of the entire population which is '**α**'. Hence, we check whether α holds same as **a**.

$H_0 : \alpha = 0$
$H_A : \alpha \neq 0$

95% confidence interval(CI) can be found by using confint() function as below.

```
> confint(bdims_model)
                2.5 %     97.5 %
(Intercept) 104.402773 117.541160
che.di        1.917292   2.384725
```

95% CI calculated for a is [104.40, 117.54]. $H_0 : \alpha = 0$ is not captured by this interval. Therefore null hypothesis is been rejected.

HYPOTHESIS TESTING (Testing statistical significance of slope β)

Slope is the average increase in dependent variable following one-unit increase in the independent variable. Slope of the regression line was calculated to be 2.151 from the lm() function before.

$H_0 : \beta = 0$
$H_A : \beta \neq 0$

1) Testing hypothesis by CI : With the help of confint() function we get the 95% confidence interval for the slope.

```
> confint(bdims_model)
                2.5 %     97.5 %
(Intercept) 104.402773 117.541160
che.di        1.917292   2.384725
```

From the above code and output we can see that 95% CI for slope is [1.917, 2.384]. This interval does not capture $H_0: \beta=0$. Therefore we can reject the null hypothesis.

2) Testing hypothesis by t-statistics :
Formula to calculate t-statistics is $slope \Big/ \sqrt{\frac{residuals\ mean\ sq}{Lxx}}$

Residuals mean sq can be calculated by anova() function.

Mean sq of residuals is found to be 53.8.

```
> anova(bdims_model)
Analysis of Variance Table

Response: hgt
           Df Sum Sq Mean Sq F value    Pr(>F)
che.di      1  17598 17597.8  326.95 < 2.2e-16 ***
Residuals 505  27181    53.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Lxx can be calculated by formula : $\sum_{i=1}^{i=n} x_i^2 - \frac{\left(\sum_{i=1}^{i=n} x_i\right)^2}{n}$

```
> sum(df$che.di^2) - (sum(df$che.di)^2/507)
[1] 3803.421
>
```

Lxx is 3803.421

As we now know slope, Lxx and residuals mean sq we can calculate t-statistics for the slope as shown below.

```
> t <- 2.151/sqrt(53.8/3803.421)
> 2*pt(q=t,df=505,lower.tail = FALSE)
[1] 9.750446e-57
```

As p-value is less than the area of significance i.e. 0.05 we can reject the null hypothesis. Hence there is statistically significant evidence that che.di was positively related to hgt.

DESCRIPTIVE STATISTICS AND VISUALIZATION (Residual vs Fitted)

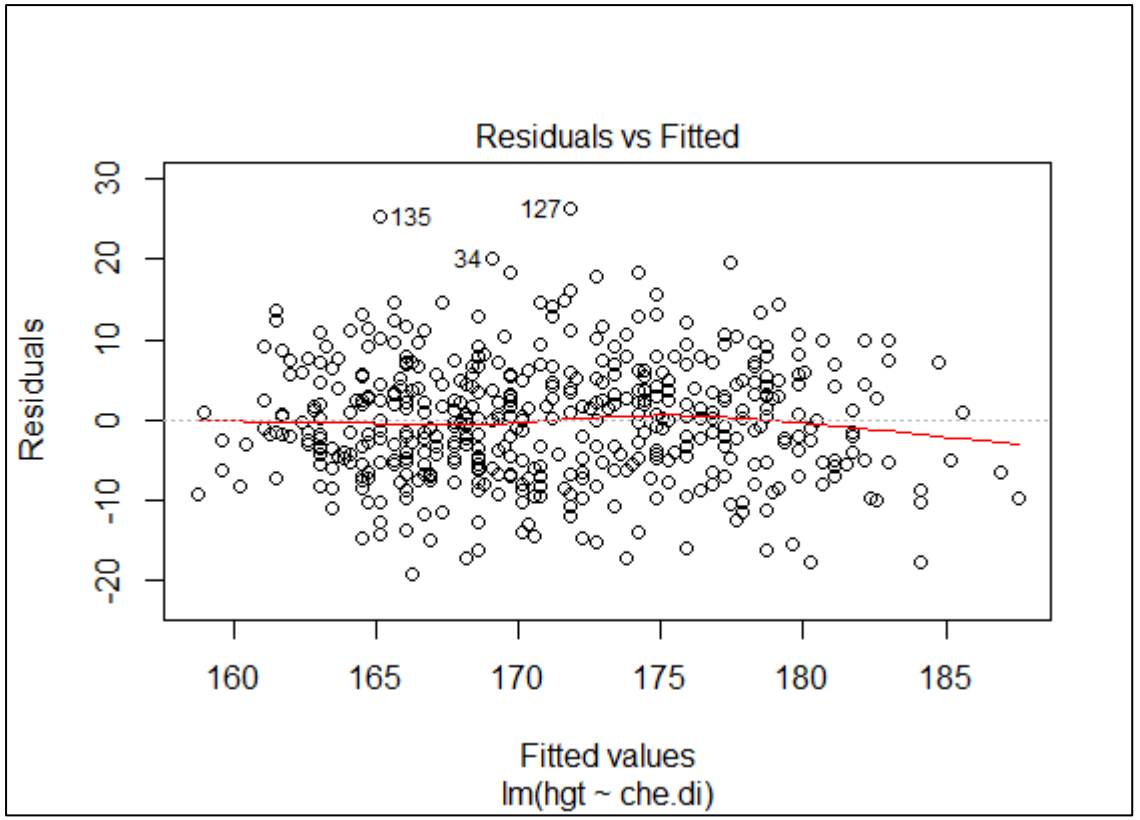
Before we report the final regression model, we need to test the following assumption.

- Independence
- Linearity
- Normality of residuals
- Homoscedasticity

Independence checks how independent are the measurements between the participants. **Linearity** determines is the relationship between the attributes positive, negative or is there no relation at all. Distribution of residuals or error is determined by **normality of residuals**. **Homoscedasticity** is related to the assumption of homogeneity of variance for the two-sample t-test.

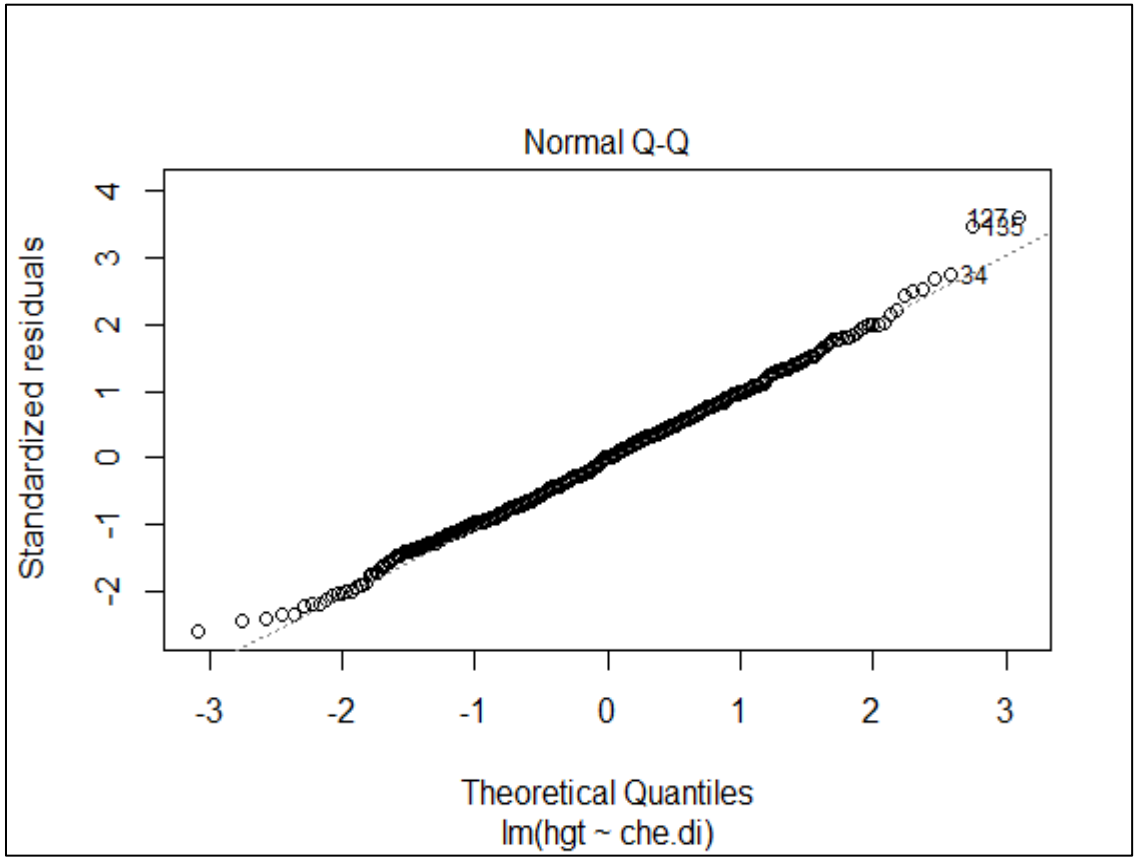
Plot() function is used to plot series of graphs to check the diagnostics of a regression model. Red line is almost horizontal which tells us that relationship between residuals and fitted value is flat hence the relationship is linear. As the points are distributed evenly around the red line it confirms the assumption of homoscedasticity. Points are not skewed towards the end its concentrated around the red line evenly.

```
> plot(bdims_model)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
```



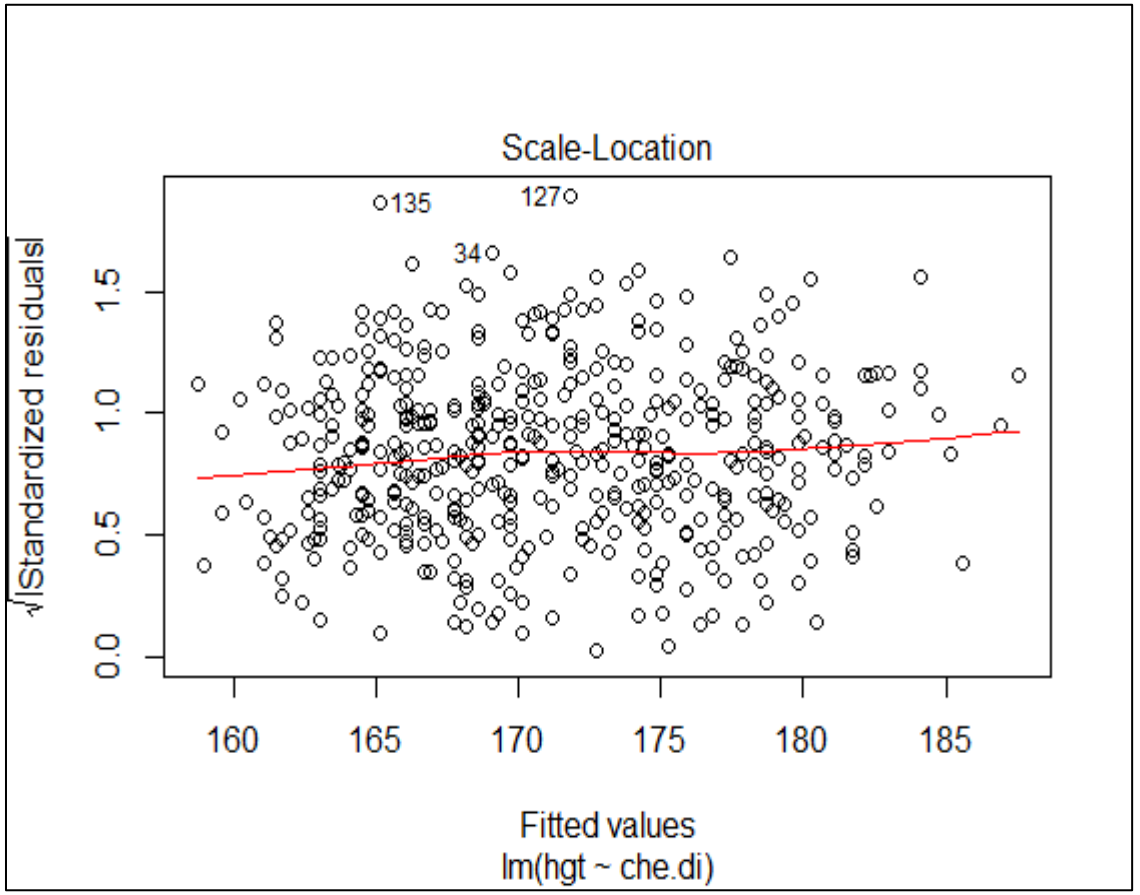
DESCRIPTIVE STATISTICS AND VISUALIZATION (Normal Q-Q)

Gross deviation from normality is observed by Q-Q plot. If the Q-Q plot deviates from the ideal line and makes S shape, then we can say that there is gross deviation from normality. In the plot we can see that all the points lie almost in the straight line hence there is no gross deviation observed. Hence, we can say that residuals are normally distributed.



DESCRIPTIVE STATISTICS AND VISUALIZATION (Scale location)

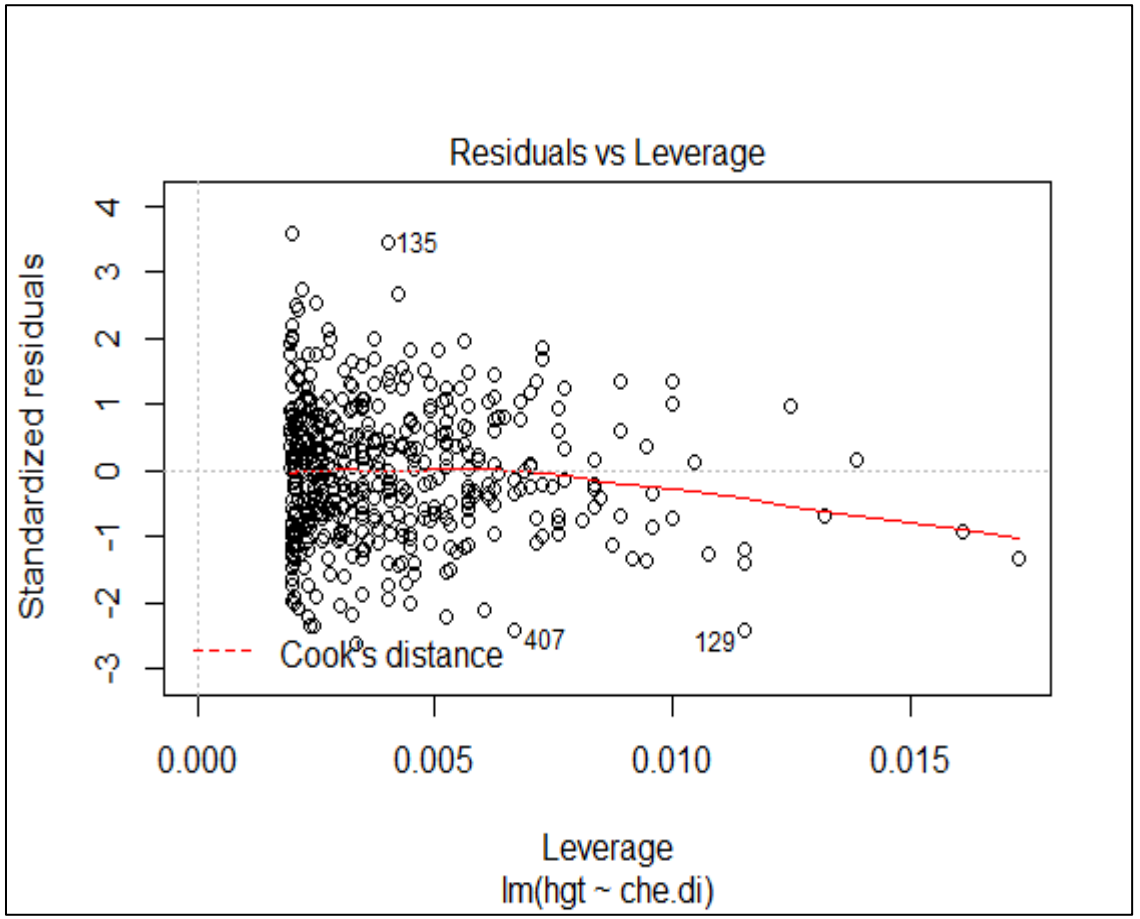
Scale location is used to check homoscedasticity. The red line is close to flat and the variance in the square root of the standardized residuals is consistent across predicted values. Therefore plot confirms the assumption of homoscedasticity safe.



DESCRIPTIVE STATISTICS AND VISUALIZATION (Residual vs Leverage)

This plot determines the cases that might be unduly influencing the fit of the regression model. These cases are called as outliers. Hence, we remove the outliers which are considered influential.

All the outliers are numbered with their case number, but we eliminate those outliers which falls beyond the upper and lower red band. These bands are located at the cook's distance. Cases that fall beyond these bands are considered influential hence we handle them.

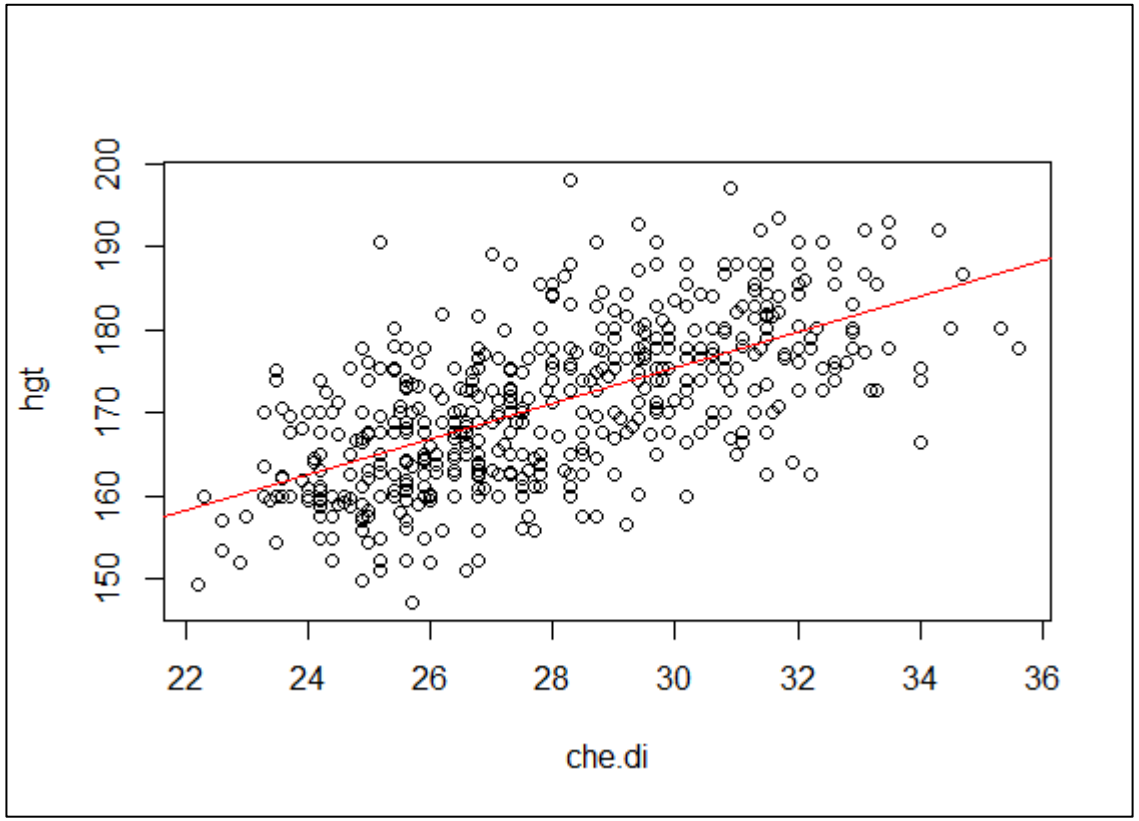


DESCRIPTIVE STATISTICS AND VISUALIZATION (Plotting regression model)

We have plotted the points using plot() function where hgt is dependent variable and che.di is independent variable. Dependent and independent variable is separated by ~. Bdims_model is the linear regression model produced by lm() function. We pass this model to abline() function. Abline() draws the linear regression model. Red line shown in the graph is the linear regression model. Red line is positive, and it increases up gradually as the independent variable increases. From the previous slides we have computed following values necessary for building linear model:

a : 110.972
b : 2.151

```
> plot(hgt ~ che.di, data = df, xlab = "che.di", ylab = "hgt")
> abline(bdims_model, col= "red")
>
```



DISCUSSION ON REGRESSION MODEL

A linear regression model was fitted to predict the dependent variable, hgt, using che.di as a single predictor. Prior to fitting the regression, a scatter plot assessing the bivariate relationship between hgt and che.di was inspected. The scatter plot demonstrated evidence of a positive linear relationship. Other non-linear trends were ruled out. The overall regression model was statistically significant, $F(1,505) = 326.95$, $p < 0.001$, and explained 39.3% of the variability in hgt, $R^2 = 0.393$. The estimated regression equation was hgt = 110.972 + 2.151*che.di. The positive slope for che.di was statistically significant, b= 2.151, $t(505) = 18.085$, $p < 0.001$, 95% CI [1.917, 2.384]. Final inspection of the residuals supported normality and homoscedasticity.

CORRELATION

```
> r <- cor(df$hgt, df$che.di)
> print(r)
[1] 0.6268931
```

Pearson's correlation coefficient(r) is used to check the strength of relationship between two variables. Correlation coefficient ranges from -1 to 1. If the relationship is positive it tends to be close to 1. If the relationship is negative then r tends to be close to -1. r = 0 implies there is absolutely no relationship between the two variables.

R reports the correlation between the two variable is r = 0.626831. Hence from the r value it can be said that both the variable holds positive relationship.

HYPOTHESIS TESTING OF CORRELATION

Let us test the hypothesis for correlation between independent and dependent variable. Let the null hypothesis be that there is no relationship between the two variables.

H_0 : r = 0
H_A : r ≠ 0

1) Testing by test statistics :

Test statistics is calculated by the formula, $t = r\sqrt{\frac{n-2}{1-r^2}}$

```
> t <- r*sqrt((n-2)/(1-r^2))
> print(t)
[1] 18.08186
```

Hence t statistics is 18.08 and we calculate the p-value of it as follows :

```
> 2*pt(q = t,df = 50 - 2,lower.tail=FALSE)
[1] 4.590554e-23
>
```

As p-value is less than the area of significance i.e. 0.05 we reject the null hypothesis and we can say that there is relationship between the two variables.

2) Testing hypothesis by CI :

```
> library(psychometric)
> CIr(r=r, n=n, level = 0.95)
[1] 0.5709813 0.6770164
```

95% CI of the correlation is found by CIr() function. CIr() calculates it to be [0.571, 0.677]. As this range of CI does not capture the H_0 we can reject the null hypothesis. Hence by CI method too we come to the conclusion that there is significant relationship between two variables.

❖ **DISCUSSION ON CORRELATION :**

A Pearson's correlation was calculated to measure the strength of the linear relationship between che.di and hgt. The positive correlation was statistically significant, r =0.6268931 , $p < 0.001$, 95% CI [0.571, 0.677].

REFERENCES

- Applied Analytics Module 9 notes
- (R Documentation and manuals | R Documentation, 2020)
- (Journal of Statistics Education, V11N2: Heinz, 2020)