HARVINDER SINGH SETHI

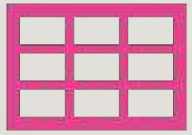
INTRODUCTION

- A good introduction provides a brief background to the problem, defines important terms, and leads to a strong rationale.
- Our dataset is all about different types of body parts, their measurements.
- We will be using linear regression model to test the relationship and see how the model works.

PROBLEM STATEMENT

- We are going to find that whether there is statistical significant relationship between a person's chest diameter (che.di) and height (hgt).
- H0: Dataset will not fit linear regression (Corelation = 0) AND there is NO statistical significant relationship.
- HA: Dataset is suitable for linear regression, AND there is statistical significant relationship.
- We will use linear regression model and see F-statistics results to decide.

DATA



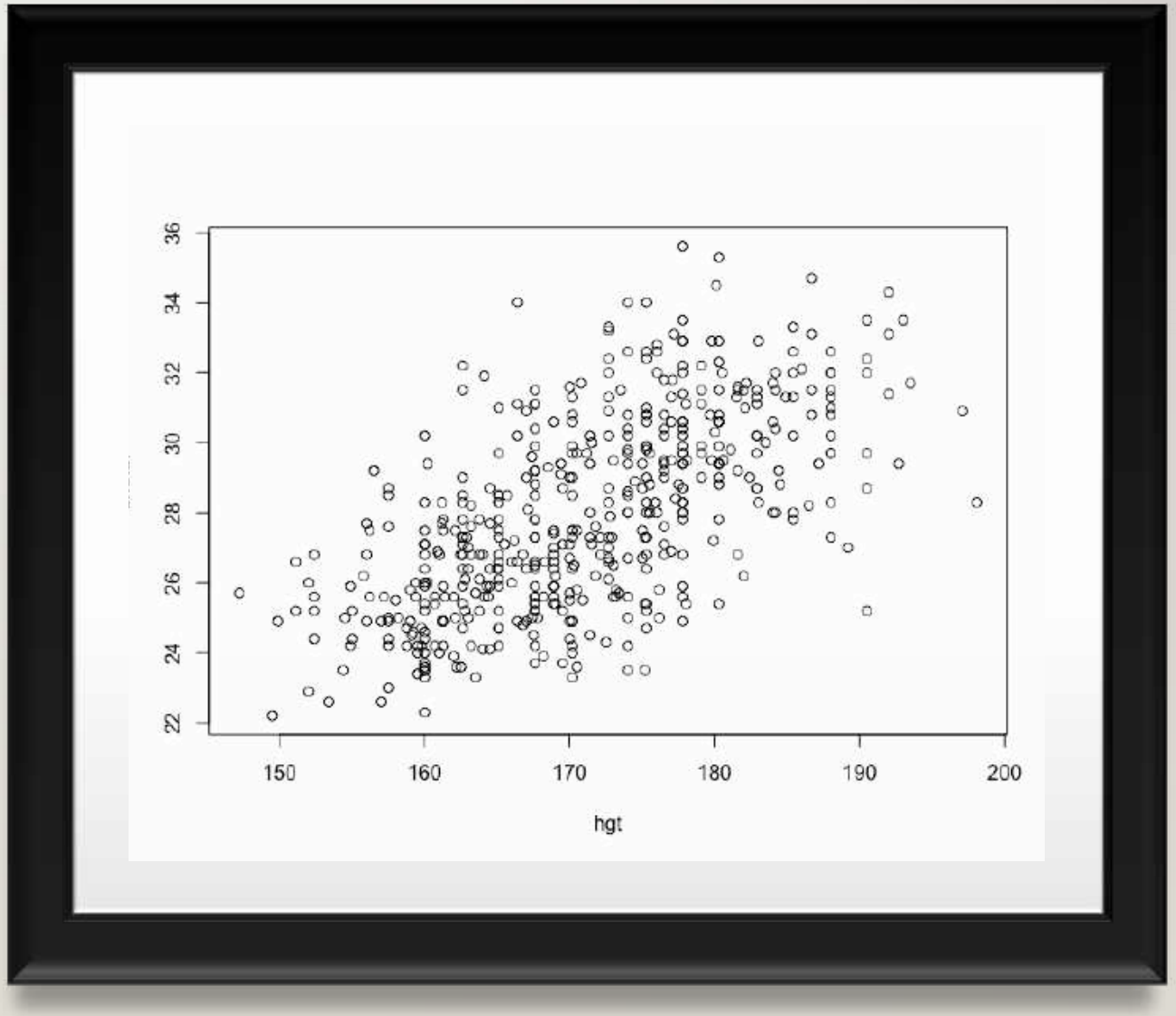
We first load the dataset and check whether both columns are in the same measurements(cms).

Fortunately we have already have the clean data with no NA, missing or whitespaces in the entries. But we check for Linearity and normality of data.

We then can make assumptions based on our problem statement and model used.

SCATTER PLOT

- Plotted the scatter plot using plot(che.di ~ hgt,data= bdims) , to check and see if the data is linear or not.
- We then, Visualized a linear positively increasing relationship between chest diameter and height.
- Our dependent variable(y) is chest diameter and independent variable (x) is height.
- So we can determine a person's chest diameter based on their height.



DESCRIPTIVE STATISTICS

SUMMARY() CODE FOR CHEST DIAMETER, HEIGHT AND OUTPUT.

```
summary(bdims$che.di)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
22.20   25.65   27.80   27.97   29.95   35.60
```

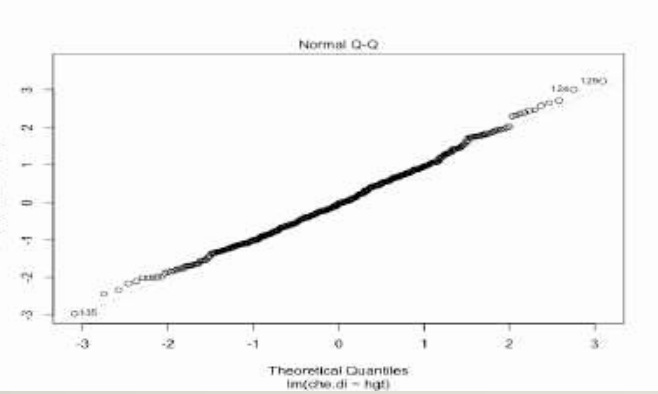
```
summary(bdims$hgt)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
147.2   163.8   170.3   171.1   177.8   198.1
```

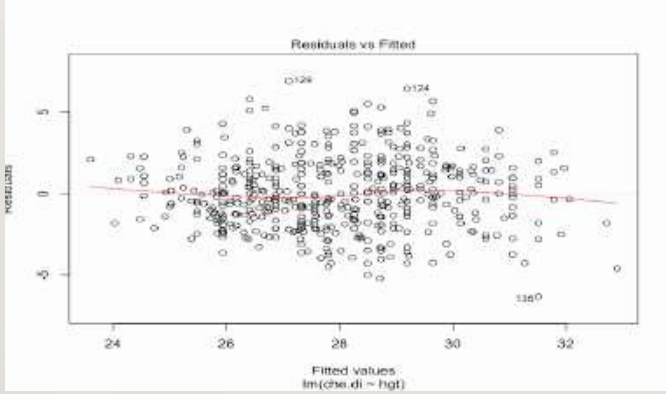
LINEAR REGRESSION MODEL

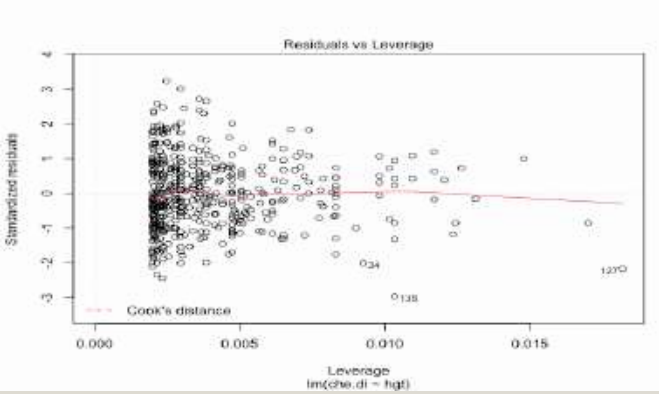
- I used the linear regression model to check the relationship between the chest diameter and height.
- model = lm(che.di ~ hgt,data= bdims)
- Then I checked the summary of my model for Intercept, slope, T-values, P- values.

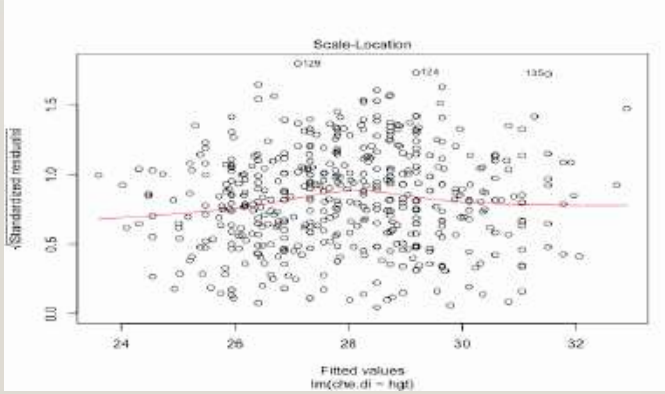
```
> model %>% summary()

Call:
lm(formula = che.di ~ 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 .
hgt           0.1827     0.0101  18.082   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 2.138 on 505 degrees of freedom
Multiple R-squared:  0.393,     Adjusted R-squared:  0.39
F-statistic:   327 on 1 and 505 DF,  p-value: < 2.2e-16
```

VISUALISATION









- **Residual vs. Fitted :-** Check this plot for non-linear trends. If the relationship between fitted values and residuals is flat, this is a good indication that you are modelling a linear relationship.
- **Normal Q-Q :-** We check the normal Q-Q plot to determine if there were any gross deviations from normality.
- **Scale-Location :-** This is another plot used to check homoscedasticity. The red line should be close to flat and the variance in the square root of the standardised residuals should be consistent across predicted.
- **Residual vs Leverage :-** This plot is used to identify cases that might be unduly influencing the fit of the regression model, for example, outliers. What we need to look for are values that fall in the upper and lower right hand side of the plot beyond the red bands. These bands are based on Cook's distances. In the diagnostic plot above, there are no values that fall outside the bands, and therefore, no evidence of influential cases.

- R-code :- plot(model)

CORRELATION

- We find the correlation between both the columns as that's also the other way of showing whether or not there is a linear relationship.
- cor(bdims$che.di , bdims$hgt) = 0.6268931
- Since the value is greater than 0 so we can say that there is strong positive corelation fitting the linear regression model.
- library(psychometric)

  corelation = cor(bdims$che.di , bdims$hgt)

  CIr(corelation, 505, 0.95)

```
> CIr(corelation, 505, 0.95)
[1] 0.5708642 0.6771105
```

Note:- The confidence interval does not capture H0, so H0 is rejected and we can say that there is a statistical significanct positive correlation between person's height and chest diameter.

HYPOTHESIS FOR LINEAR REGRESSION MODEL

INTERCEPT :-

Ho = No statistical significance between chest diameter and height. So the intercept =0.

Ha = There is a statistical significance between chest diameter and height. So the intercept != 0.

SLOPE :-

Ho = There is no increase in chest diameter as the height is varied. So the slope = 0.

Ha = There is variance in chest diameter as the height is varied. So the slope != 0.

SLOPE AND INTERCEPT

- As we can see the **slope** =0.1827 i.e. for every one unit increase in height the chest diameter increases by 0.1827
- And 0.16 and 0.20 are the **lower and upper bound** of the 95% CI of the height slope for the model predicting chest diameter.
- And **Intercept = -3.294** signifies when height =0 chest diameter = **-3.294**
- And -6.697 and 0.107 are the **lower and upper bound** of the 95% CI of the height intercept for the model predicting chest diameter..

```
> model = lm( che.di ~ hgt,data= bdims)
> model

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

Coefficients:
(Intercept)          hgt
    -3.2947       0.1827

> #lm(y,x)
> #The constant, or intercept, is the average value
>
> model %>% confint()
                 2.5 %    97.5 %
(Intercept) -6.6972252 0.1079121
hgt          0.1628512 0.2025541
```

LINEAR REGRESSION MODEL RESULT

The linear model was Statically significant as , $F(1, 505) = 327$, $p < 0.001$

Height explained 39.3% of the variability in Chest diameter

We reject the null hypothesis for linear regression because the P value from f statistics is less then 0.05

Slope is 0.1827

We reject the null hypothesis for slope because p value for slope is less then 0.05

Intercept point -3.947

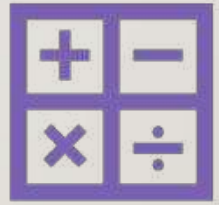
We fail to reject the intercept null hypothesis, because p value for intercept is greater than 0.05

```
> model %>% summary()

Call:
lm(formula = che.di ~ 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 .
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
```

CONCLUSION

We reject the Null Hypothesis

Since, the p-value = < 2.2e-16 i.e. less than 0.05



There was a statistically significance positive linear relationship between chest diameter and height. Height explained 39.3% (R-Squared) variability in chest diameter.

DISCUSSION

STRENGTHS

Easy to apply the model and state the results from it

LIMITATION

Simple linear regression is NOT suitable for categorical data

SUGGESTIONS

Make sure the data is linear and if its not try to make it linear by substituting the values by its mean or log(). Else go for another model.

REFERENCES

- https://astral-theory-157510.appspot.com/secured/MATH1324_Module_09.html#correlation
- https://towardsdatascience.com/everything-you-need-to-know-about-linear-regression-b791e8f4bd7a