Correlation between forearm length and height

as5312

September 26, 2020

## I. Introduction

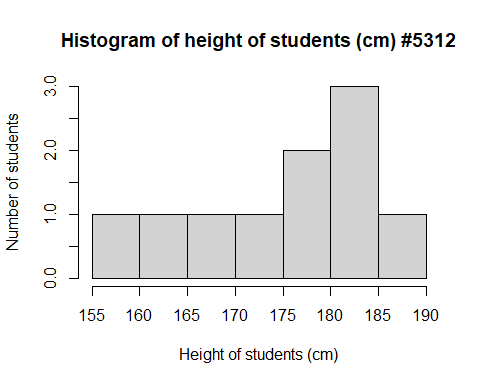
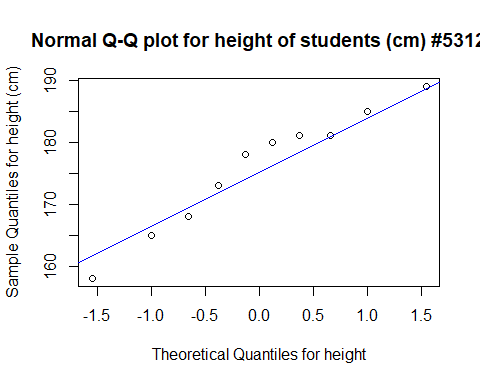
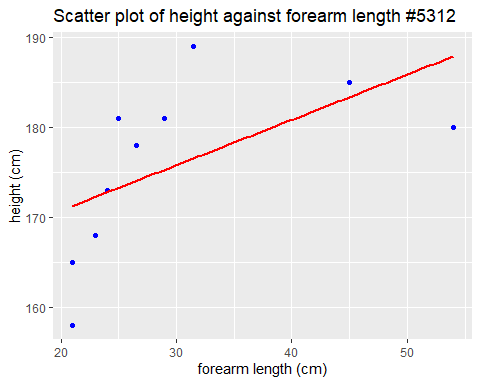
Part 1

The title to summarize assignment 1 is “Correlation between forearm length and height”. The sample data was collected through a survey taken by the current class of STA302 students on quercus where they were asked to enter their height and forearm length in cm. I then randomly took a sample of this data and used it as my data set as instructed. The explanatory variable is forearm length and the response variable is height. I chose forearm length as the explanatory variable to show that it is a good predictor of the height of a person. So the objective of this assignment will be to show that there is some correlation between the length of a person’s forearm and their height.

## II. Exploratory Data Analysis

Part 2

## `geom\_smooth()` using formula 'y ~ x'



## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 158.0 169.2 179.0 175.8 181.0 189.0

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 21.00 23.25 25.75 30.00 30.88 54.00

The response variable, height, is approximately normal as can be seen by the qqplot. In the qqplot we can see that the majority of the points lie either on or very close to the line of best fit (blue) except on the ends, indicating that the data is approximately normally distributed.

The histogram however does not indicate normal distribution since the sample data of 10 students collected was random and so it does not properly describe the entire data set. Hence, we see that the histogram does not have a bell-shaped curve and is not symmetric about the mean.

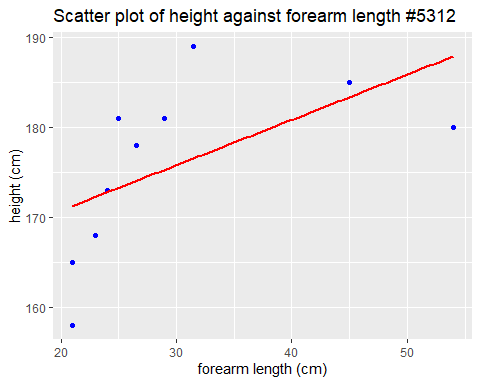
From the scatter plot above we can see that there is a weak, positive correlation between forearm length and height since some points lie close to the line but none of them are directly on the line. Perhaps with a better data set we would see a relatively stronger positive correlation.

To describe the spread of data I used the summary(height) and summary(forearm) functions. From the summary of height we can see that the center/median of the data is 179. The first quartile is 169.2 and the third quartile 181 giving us 11.8 as the IQR which tells us that the data points are not too spread out. We can also see that the data has a symmetrical distribution since the mean and median are relatively close together. For the data points of forearm length we can see that the IQR is 21.13 which tells us that the data points are more spread out. The mean and median of forearm length are also not as close indicating that the data is not very normally distributed. In addition to that, the data also has some unusual points like the point (30, 184) which are quite far away from the line of best fit and do not fit the trend.

## III. Methods and Model

Part 3

##   
## Call:  
## lm(formula = height ~ forearm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.2649 -5.7668 0.9324 5.2688 12.4441   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 160.6829 8.0634 19.928 4.19e-08 \*\*\*  
## forearm 0.5039 0.2539 1.985 0.0824 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.376 on 8 degrees of freedom  
## Multiple R-squared: 0.33, Adjusted R-squared: 0.2462   
## F-statistic: 3.94 on 1 and 8 DF, p-value: 0.08242

 Using the lmod() function we can see that the linear regression model that we get is:

Looking at the linear regression model above we can see that the regression parameters are not zero. Our first regression parameter intercepts the y-axis at 163.2684 (as shown in the summary of the linear model) so it does not pass through the origin. Our second regression parameter is 0.2739 which tells us that there is a positive correlation between forearm length and height, so when forearm length increases, so does height. Interpreting further we see that 163.2684 is the height of a student whose forearm length is 0 cm. However, since x cannot possibly take on the value 0 (since it is not within the range of the predictor values) in this study, the interpretation of the constant is not useful. Interpreting further we see that since the regression coefficient is 0.2739. This means that in this model, a person’s height increases by 0.2739 cm with each additional cm of forearm length. So using this linear model we can conclude that there is a positive correlation between a person’s forearm length and their height, i.e. as forearm length increases, height increases.

## IV. Discussions and Limitations

Part 4

Ethnicity of the students is a potential lurking variable since forearm length and height are also dependent on a students ethnicity.This lurking variable is not controlled in this study since we took data from students of different ethnic backgrounds. Another limitation of our fit is that the data that students submitted is prone to measurement error which also explains why there are quite a few unusual data points in our scatter plot shown in Part I.

Another pair of variables to explore a simple linear regression model is the student’s score on the statistics aptitude test before beginning the course and their final statistics grade. Here, the independent or predictor variable is the score on the aptitude test and it will be used to predict the dependent or response variable which is the final statistics grade for the student.