

Linear Regression Model Analysis on Forearm Length and Height

J.W 3519

September 26, 2020

I. Introduction

The sample data was collected through a introduction survey that was conducted at the beginning of the Fall 2020 STA302 course. It was given to all the students in the course where they recorded their own height and forearm length. The survey was given to more than 300 students and a sample of size 10 was drawn.

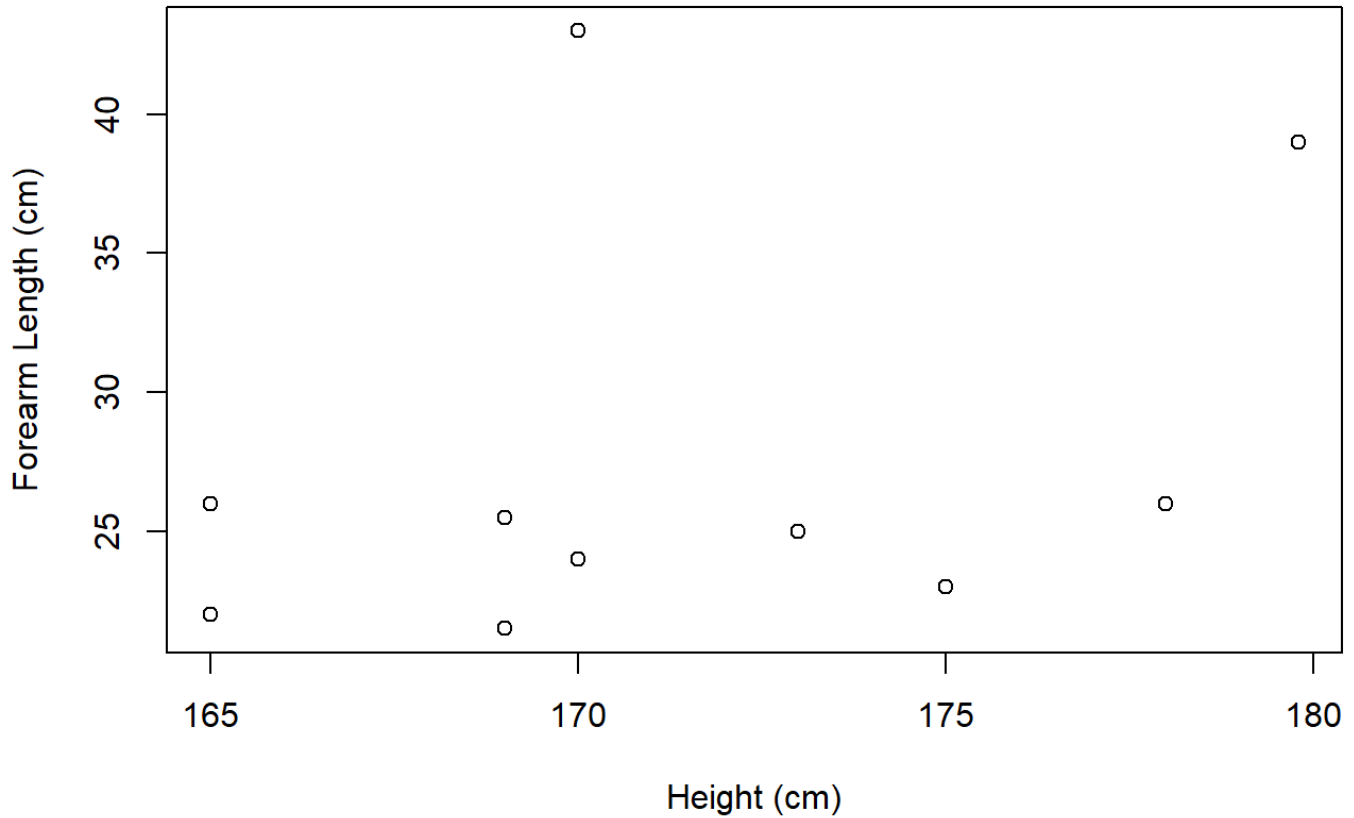
For this analysis, the explanatory variable is the height and the response variable is the forearm length. The reason for these two choices is because of our purpose of this analysis. As mentioned in the title, we are creating a linear regression model on the height and forearm length to determine if there is a relationship between the two. More specifically, we want to see if height has a direct effect on the forearm length. In our daily lives, we normally associate taller people with having longer arms and so we want to see if this is true by having the height as our explanatory variable.

II. Exploratory Data Analysis

Since we want our code to be reproducible on any machine, the following data is from the csv file that contains all the data achieved from the introduction survey and a sample of size 10 was drawn with `set.seed(3519)`. Then, we created a new data frame to store the data separately. By doing so, the code is now reproducible. The following table shows the sample data we got.

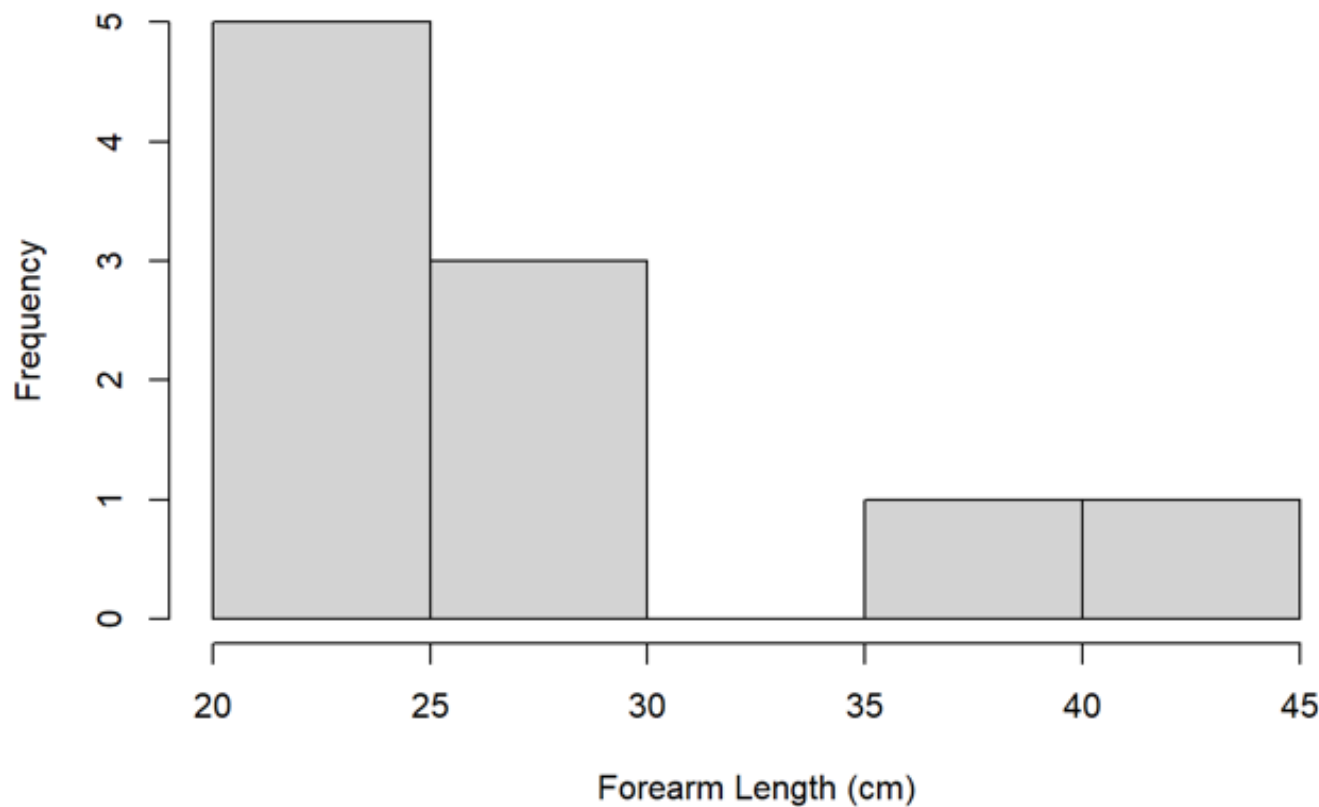
##	height	forearm
## 1	169.0	21.5
## 2	175.0	23.0
## 3	173.0	25.0
## 4	165.0	22.0
## 5	170.0	43.0
## 6	170.0	24.0
## 7	178.0	26.0
## 8	179.8	39.0
## 9	165.0	26.0
## 10	169.0	25.5

Scatterplot of Forearm Length vs Height

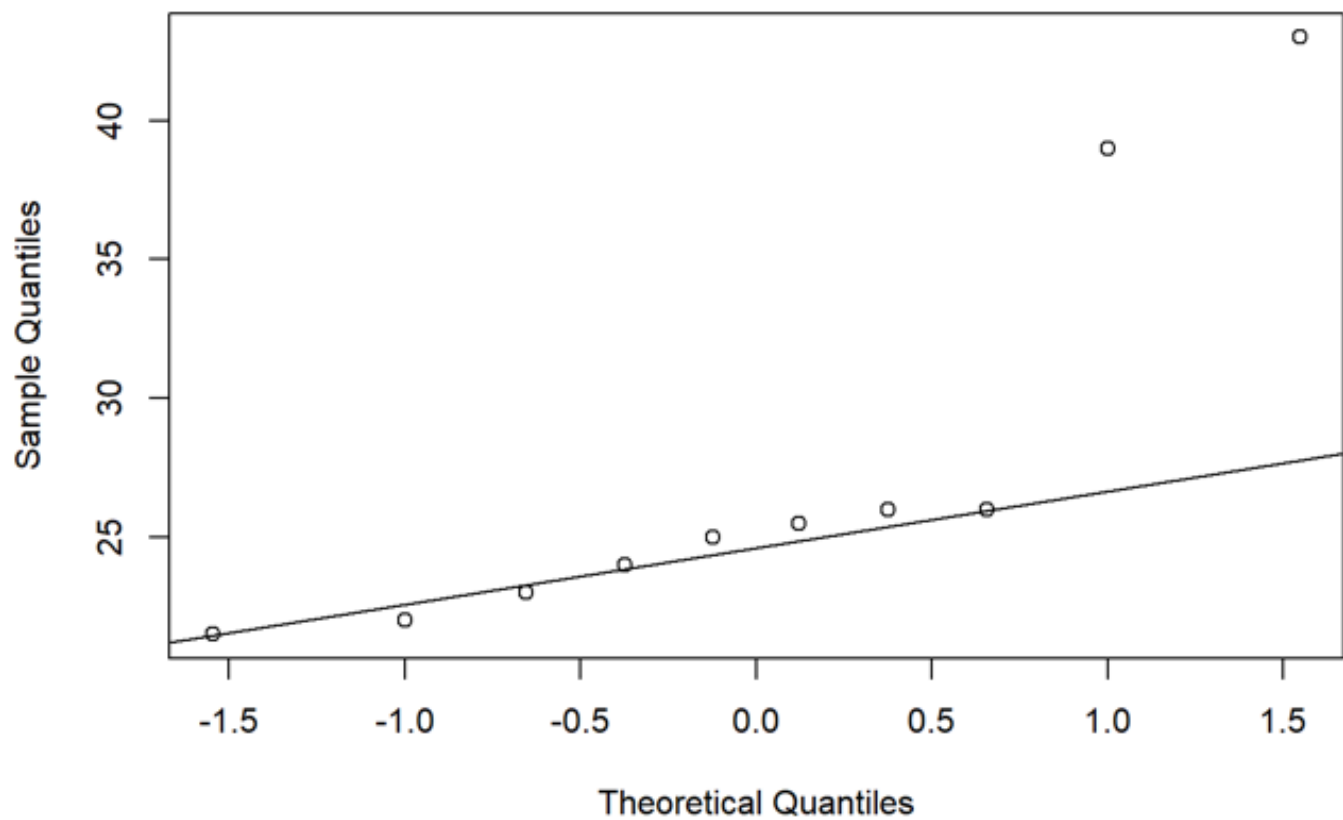


The scatterplot above shows the relationship between height and forearm length and in this case, the plot does not give a clear picture if the relationship between the two variables is linear or not. We can also see there are two unusual points that are away from the majority of the points. As a result, we can conclude those points as outliers.

Histogram of Forearm Length



Normal QQ Plot: Forearm Length



Is the response variable normal?:

In order to check for normality in the response variable, we can look at two plots. First, we look at the histogram for forearm length. This plot shows us the frequency of people who have each certain forearm length. If the histogram indicates a symmetric distribution, then the variable in question is likely to be normal. However, the histogram of the forearm length we have above, is right-skewed thus the response variable is not likely to be normal. Next, we can look at the normal probability plot or QQ plot. We can conclude the variable is normal if the points fall on the linear line. In our plot, we can see that there are two points that are clearly not on the line and they are not even close. Thus, by observing these two plots, we can conclude that the response variable in forearm length is not normal.

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      165.0   169.0   170.0   171.4   174.5   179.8
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      21.50   23.25   25.25   27.50   26.00   43.00
```

Two of the main numerical measures for the centre is the mean and median. The mean represents the average of the data set while median represents the middle of the set of data. Based on the summary of the variables, we can see that the sample mean for forearm length is 27.5cm and median is 25.25cm. For height, the sample mean is 171.4cm and the median is 170cm.

```
## [1] 5.022129
```

```
## [1] 7.348469
```

One of the numerical measures of spread is the standard deviation. The standard deviation represents the spread of the data and shows the distance between each data and the mean. Above we can see that the standard deviation of the height and forearm length is 5.022 and 7.348 respectively.

Another measure of spread are quartiles and from the summary table we can see that the 1st quartile and 3rd quartile for height is 169cm and 174.5cm respectively. For forearm length, we have 23.25cm and 26cm. Quartiles represent the number of data points that are divided into equal parts. In this case, the 1st quartile represents the middle of the smallest data to the median and the 3rd quartile represents the middle of the median to the largest data.

III. Methods and Model

```
##
## Call:
## lm(formula = forearm ~ height)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.313  -4.434  -2.557   1.070  16.191
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -58.3425     83.3349  -0.70    0.504
## height         0.5009      0.4861   1.03    0.333
##
## Residual standard error: 7.323 on 8 degrees of freedom
## Multiple R-squared:  0.1172, Adjusted R-squared:  0.006831
## F-statistic: 1.062 on 1 and 8 DF,  p-value: 0.3329
```

The following summary describes the linear regression model that we created. Based on the summary we get that the fitted line is:

$$\hat{y} = -58.3425 + 0.5009x$$

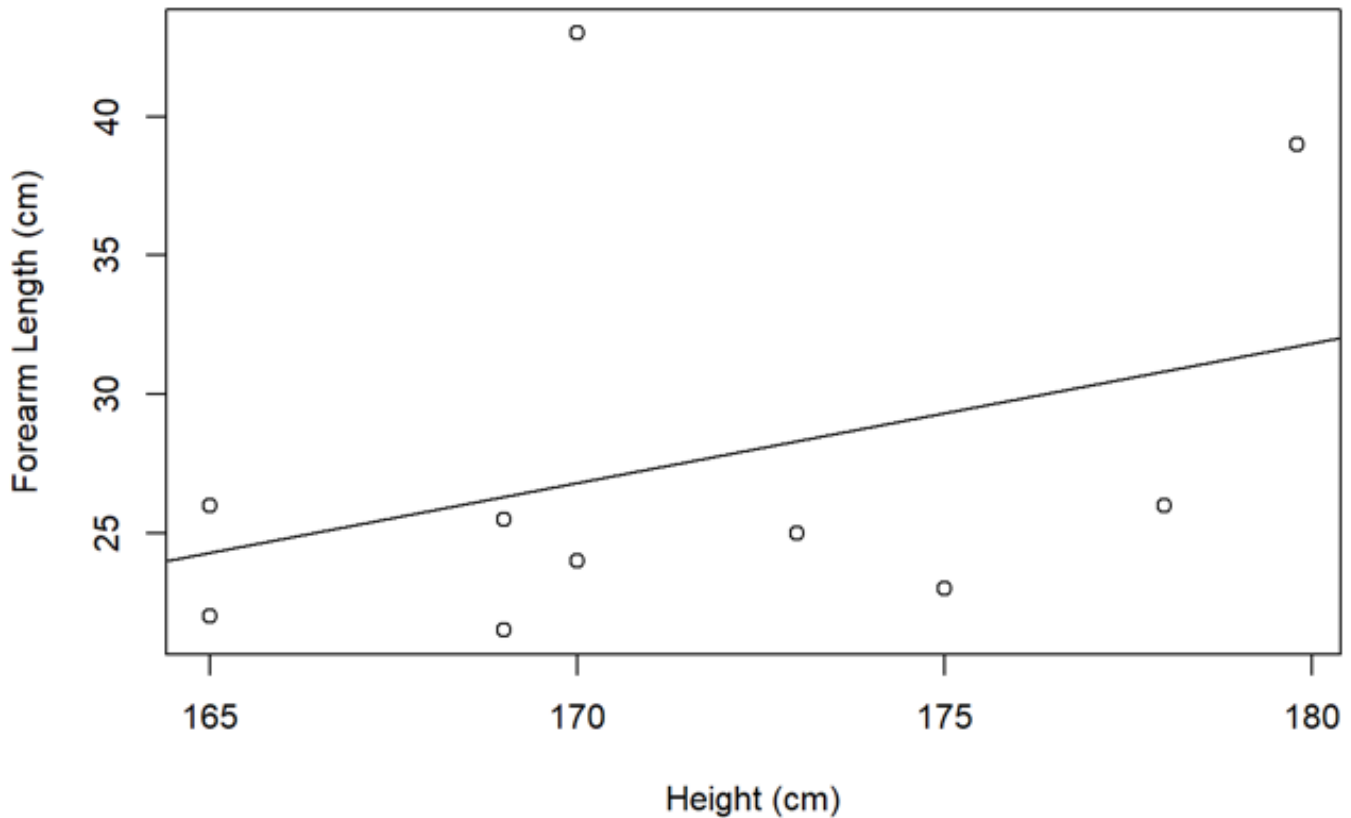
As we can see from the summary table, none of the regression parameters are 0. The interpretation of the slope estimate is that if we are to increase the height by 1 cm, then the forearm length increases on average by 0.5009 cm. For the intercept estimate, we have -58.3425 which represents the forearm length when the height is 0. Obviously, this interpretation is useless as having a height of 0 does not make any sense in reality and also having a negative forearm length makes no sense. Thus, there is no practical interpretation for the intercept estimate.

The linear line in the plot above is the linear regression model between height and forearm length. First we should check if our model is a good fit. If we look at our p-value for our two coefficients, we see that they are much larger than the significance level of 5%. As a result, we do not consider the linear model to be statistically significant meaning it is not a good fit.

IV. Discussions and Limitations

A lurking variable in this analysis can be the age of the person. Age can be associated to a person's growth and thus it can affect the relationship between the height and forearm length. One of the issue we mentioned above is that the linear model we created is not a good fit. A possible reason for this is due to the size of our sample. Thus, the limitation of our fit is that the sample size is quite small. If we were to have a bigger sample size, it can lead to a better fit as there are more data points to evaluate thus the accuracy should increase.

Scatterplot of Forearm Length vs Height



As we can see from the plot, there are a low amount of data points and thus the linear model does not look accurate.

If we were to conduct another data analysis using a simple linear regression model, another pair variable we can use is number of hours study and the grade percentage on exam. We would want to response variable to be the grade percentage on exam because we want to see if studying more would increase the performance on an exam.