Linear Regression Model-Predicting Forearm length by using Height

ZQ3383

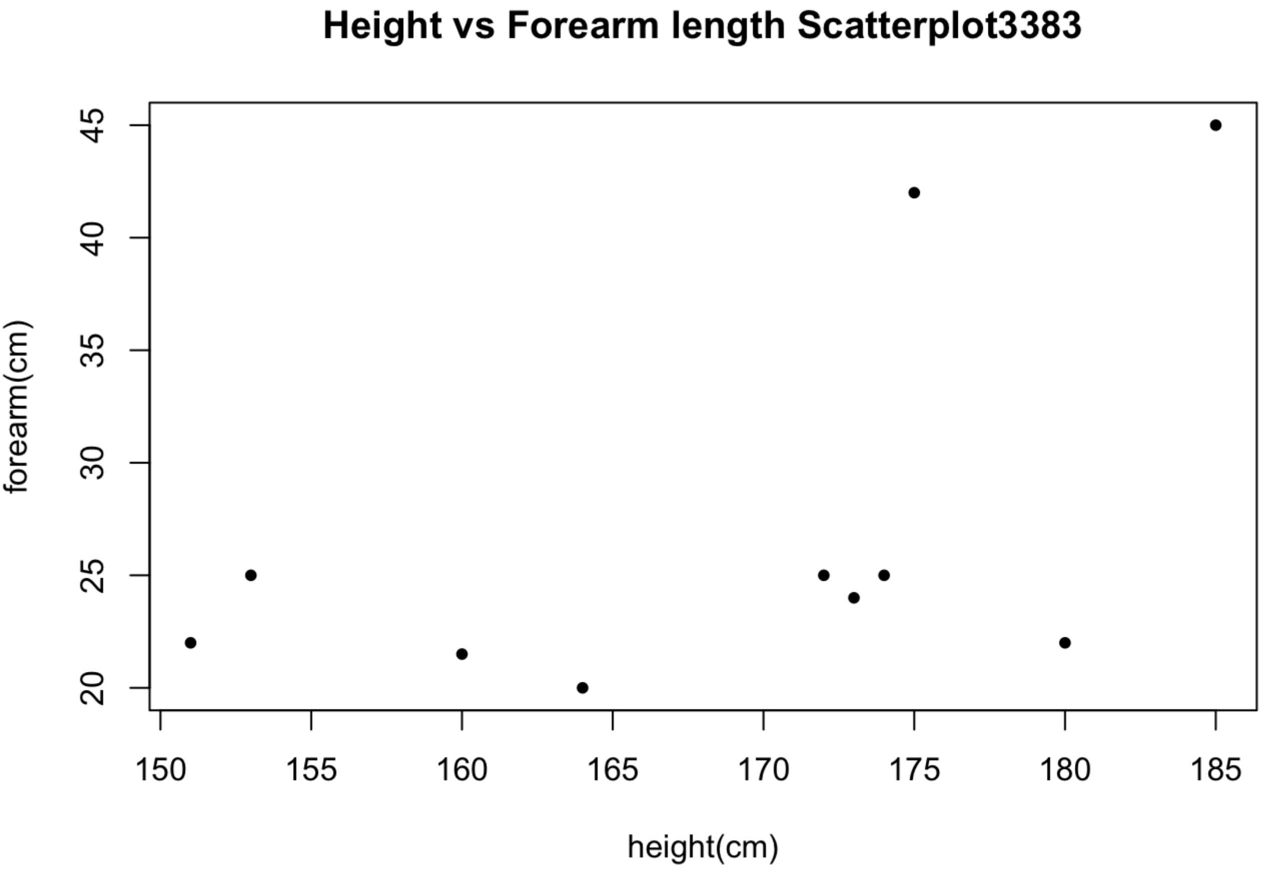
September 22, 2020

I. Introduction

Part 1 This is an analysis for a linear regression model: predicting forearm length by using height. The sample data was randomly collected among students who are taking STA302 this semester. I set the seed of my randomization to be the last four digits of my student number (3383), and randomly sample 10 data points. The result I get is 17 44 75 130 168 169 175 189 242 323, so I choose my data using the result as id in the given .csv file.

```
set.seed(3383)
numbers <- sort(sample.int(366,10))
```

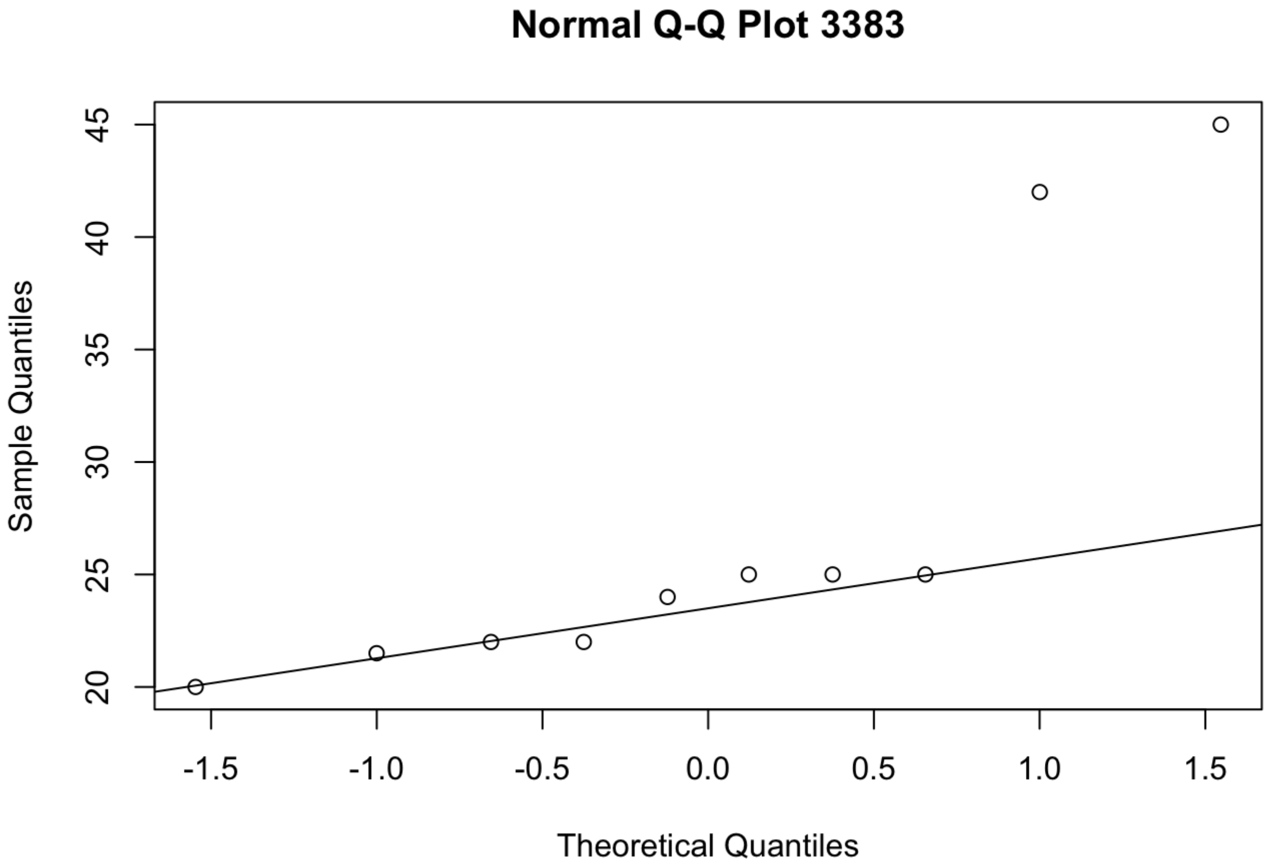
Since the path name may reveal my identity, I choose to hard code my data into a data frame named ZQ3383_data. My response variable is called forearm_y_zq3383 and my explanatory variable is called height_x_zq3383. This is because we need to use height to predict the value of forearm. In order to increase the uniqueness and helps me to understand my code, I add a _x at the end of variable height and _y at the end of variable forearm when I name my variables. This helps me to keep track of which variable is explanatory variable and which one is response variable. I also add my initials to highlight the fact that my code is unique.



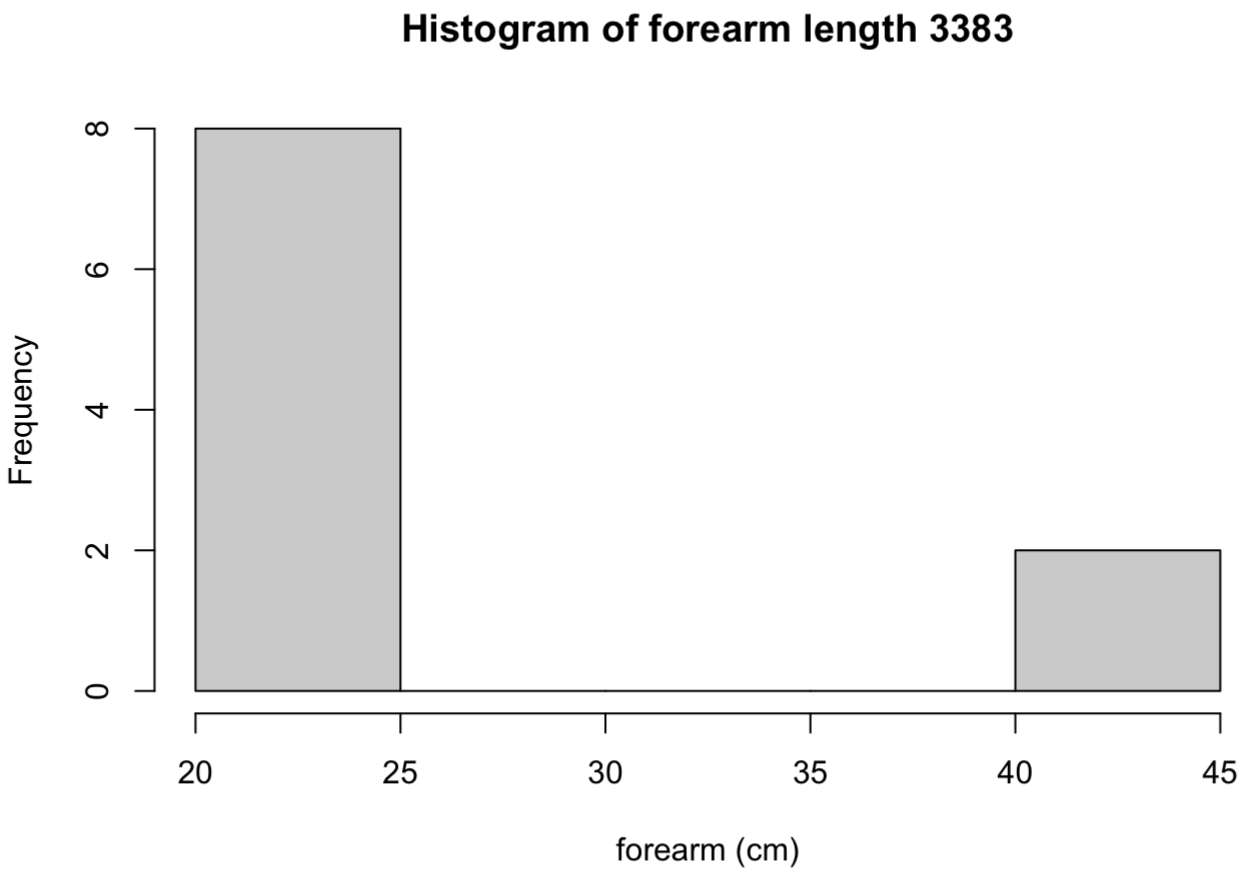
My explanatory variable is height(cm), and my response variable is forearm length(cm). I choose them because my linear regression model is using height to predict forearm length.

II. Exploratory Data Analysis

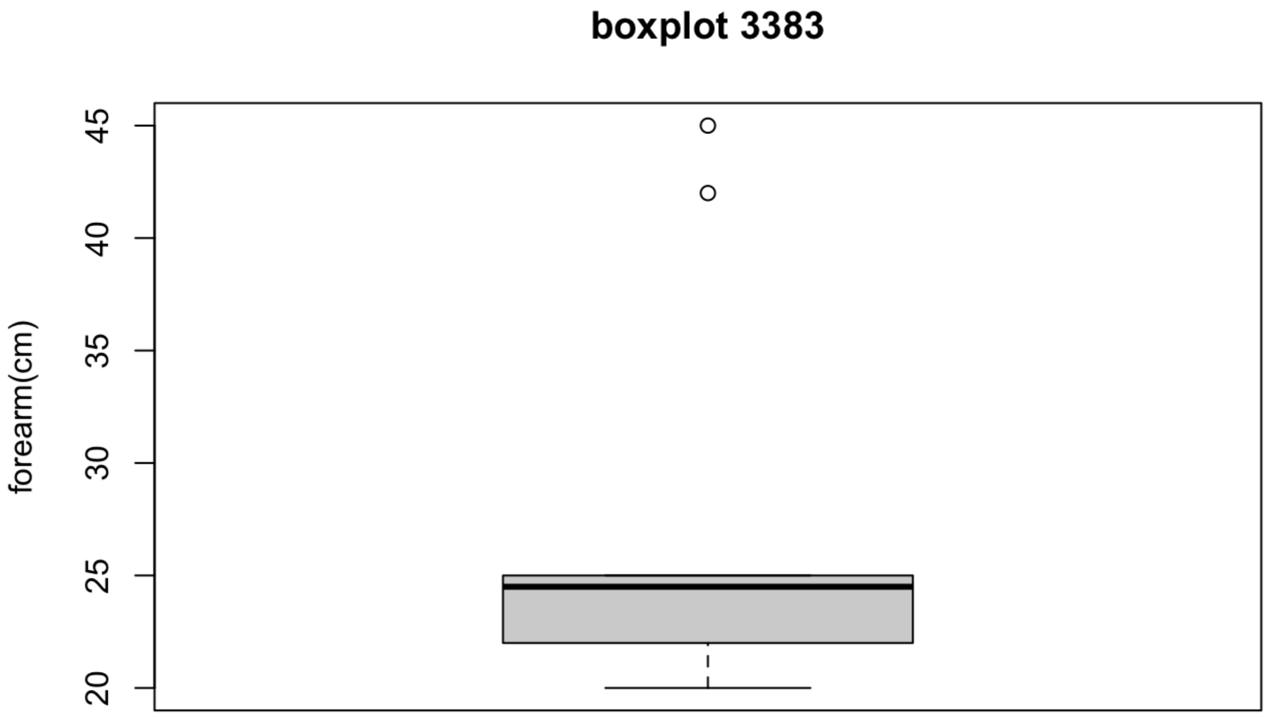
Part 2 Here is an example of the Q-Qplot:



This is an example of histogram, which shows the distribution of the data set.



Here is an example of boxplot, which comparing to the histogram above, shows the distribution of the given data set more significantly. Note that boxplot can still varies even through the data set is small.



Shown from three different plots above, my response variable is approximately normal. Since most of the point falls on the qqline, and the frequency of forearm length between 20 and 25 are much larger than the rest distributions. Also, according to the boxplot, the data are mostly included in the interquartile range except for two outliers.
Summary of the data shown below, all measured in cm:

```
summary(ZQ3383_data)
```

```
##  height_x_zq3383 forearm_y_zq3383
##  Min.   :151.0   Min.   :20.00
##  1st Qu.:161.0   1st Qu.:22.00
##  Median :172.5   Median :24.50
##  Mean   :168.7   Mean   :27.15
##  3rd Qu.:174.8   3rd Qu.:25.00
##  Max.   :185.0   Max.   :45.00
```

From data shown above, we can tell that mean value for predictor, which is height_x_zq3383 is 168.7, and mean value for response, which is forearm_y_zq3383 is 27.2. The median for height_x_zq3383 is 172.5, and the median for forearm_y_zq3383 is 24.5.

From data above, we can tell that the variance of height_x_zq3383 is 127.57 and variance of forearm_y_zq3383 is 77.67. The standard deviation of height_x_zq3383 is 11.29 and standard deviation for forearm_y_zq3383 is 8.81. The IQR for height_x_zq3383 is 13.8, and the IQR for forearm_y_zq3383 is 3.0. The range of height_x_zq3383 is 34, and the range of forearm_y_zq3383 is 25.0.

```
## [1] 127.5667
```

```
## [1] 77.66944
```

```
## [1] 11.29454
```

```
## [1] 8.813027
```

Point(175,42) and point (185,45) are unusual points. This conclusion come from the three data plots above. As shown above, Q-Q plot demonstrated these two points are not on the normal line. Then on Histogram of forearm length, the distribution of forearm length between 40cm and 45cm has the lowest distribution. Also, on the boxplot, which is commonly used as comparing distributions against each other, forearm_y_zq3383 value of 42cm and 45cm are labeled as the outliers. Therefore,to sum it up, point(175,42) and point (185,45) are unusual points. ## III. Methods and Model

Part 3 Here is a linear regression model from the data provided:

```
height_to_forearm_model_3383 <- lm(forearm_y_zq3383~height_x_zq3383, data = ZQ3383_da
ta)
summary(height_to_forearm_model_3383)
```

```
##
## Call:
## lm(formula = forearm_y_zq3383 ~ height_x_zq3383, data = ZQ3383_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.090  -4.889  -2.720   4.183  12.096
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -46.6057    38.6262  -1.207   0.2621
## height_x_zq3383   0.4372     0.2285   1.913   0.0921 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.743 on 8 degrees of freedom
## Multiple R-squared:  0.3139, Adjusted R-squared:  0.2282
## F-statistic: 3.661 on 1 and 8 DF,  p-value: 0.09206
```

Since every linear regression model should follow the format $\hat{y} = \overset{\wedge}{\beta_0} + \overset{\wedge}{\beta_1} * x$ According to the result calculated above, we can tell $\overset{\wedge}{\beta_0} = -46.6057$, and $\overset{\wedge}{\beta_1} = 0.4372$. Therefore, the linear regression model should be $\overset{\wedge}{forearm} = -46.6057 + 0.4372 * height$

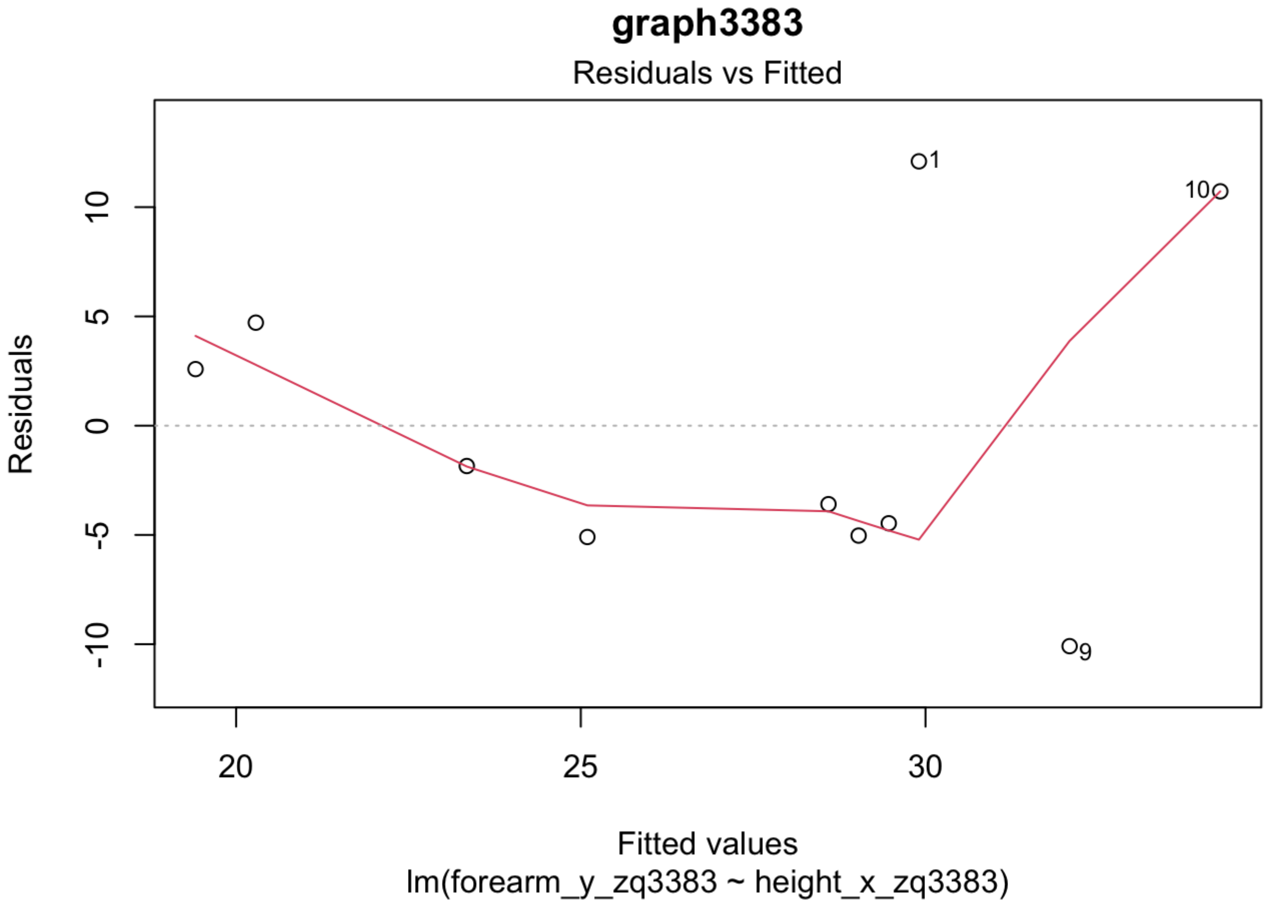
Then we need to do the hypothesis test. Note that H_0 means $\beta_1 = 0$, and H_α means $\beta_1 \neq 0$. From summary shown above, the p-value for height_x_zq3383 is 0.0921, which is larger than 0.05 as the given value of level of significance. Therefore, we don't reject H_0, which means $\beta_1 = 0$. Then re repeat the same for β_0. Note that H_1 means $\beta_0 = 0$, and H_α means $\beta_0 \neq 0$. From summary shown above, the p-value for intercept is 0.2621, which is larger than 0.05 as given. Therefore, we don't reject H_1, which means $\beta_0 = 0$.

To sum it up, according to the summary shown above, both of the regression parameters are zero. Since $\beta_0 = 0$, then we can tell that the line of best fit passes through the origin. $\beta_1 = 0$ shows that there is no correlation between height and forearm length. To sum it up, according to my data, there isn't a significant relationship between forearm length and height.

IV. Discussions and Limitations

Part 4 According to definition, lurking variable is a variable neither response or explanatory variable, but can affect the relationship between them. Thus, since our model is the linear regression model: using height to predict forearm length, a person's weight can be a lurking variable. The reason is that if someone with a height of 150cm, weights 60kg, it's very likely to have a larger forearm length than someone with a height of 180cm, weights 60kg. It's a hidden factor that affects the relationship between height and forearm length directly.

One issue or limitation about this model is the data size is not big enough, therefore the result is not accurate enough since an outlier can be more consequential in small samples. Thus, the limitation of the sample could be point(175,42) and point (185,45).



We can also tell from the residual plots that the line is not flat enough, and that’s because of the outliers.

Another example for a pair of variables to explore a simple linear model can be “the time people spend on a particular assignment (hrs)” and “the mark they received (points)” The explanatory variable is the time people spend, and the response variable is the mark received. The reason I assign them like this is because in real life, it makes more sense to predict the grade from the time we spend on this course; because we always know the amount of time we studied before we take the test, and then get the result mark.