linear regression between forearm length and height

ST4232

2020-09-26

I. Introduction

In the assignment #1, we are going to find the linear regression between forearm length and height.

We take a survay in the lecture time and select the data of our height and forearm length in groups, then we collect the data from 346 students in to a csv file.

Since we usually know more about our height than our forearm length, I set the height as explanatory variable and the forearm length as response variable so that we can use the height to predict the forearm length.

II. Exploratory Data Analysis

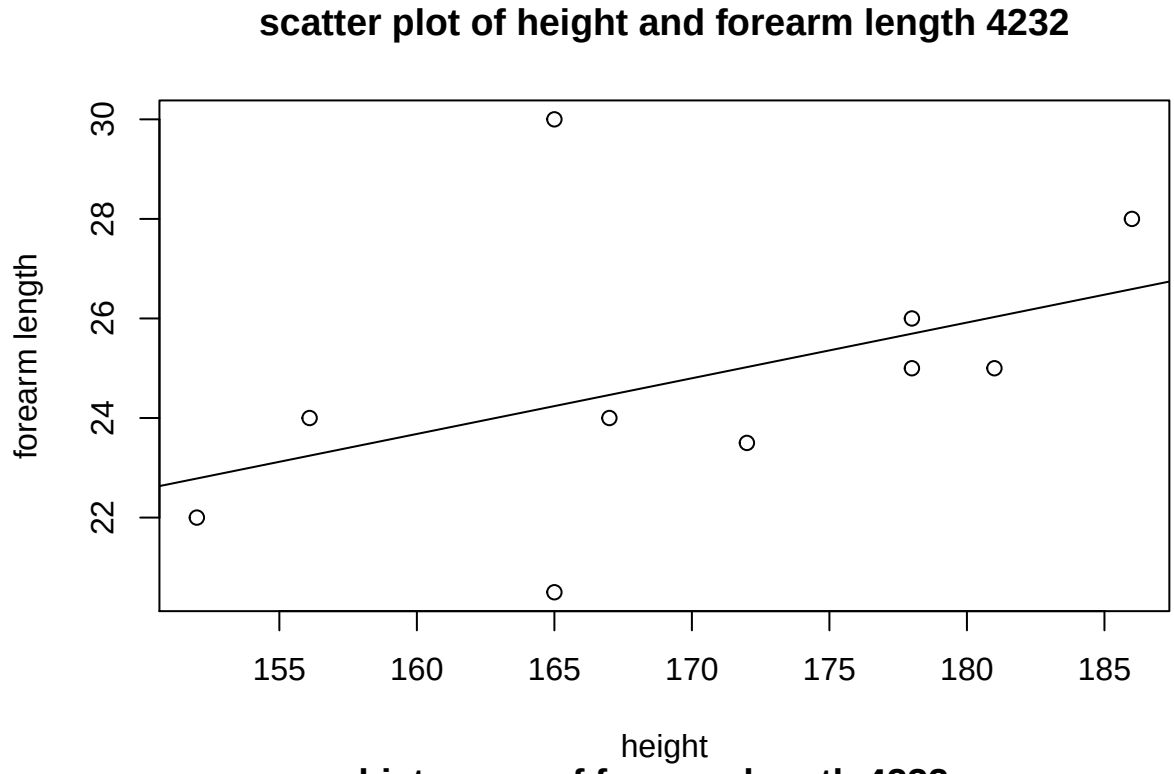
First, we use a scatter plot to see the relationship between the height and the forearm length, we can see when the height increase, most of the forearm length increase, the dots are near from the line which best fits the data, thus we can say there might be a moderate linear correlation

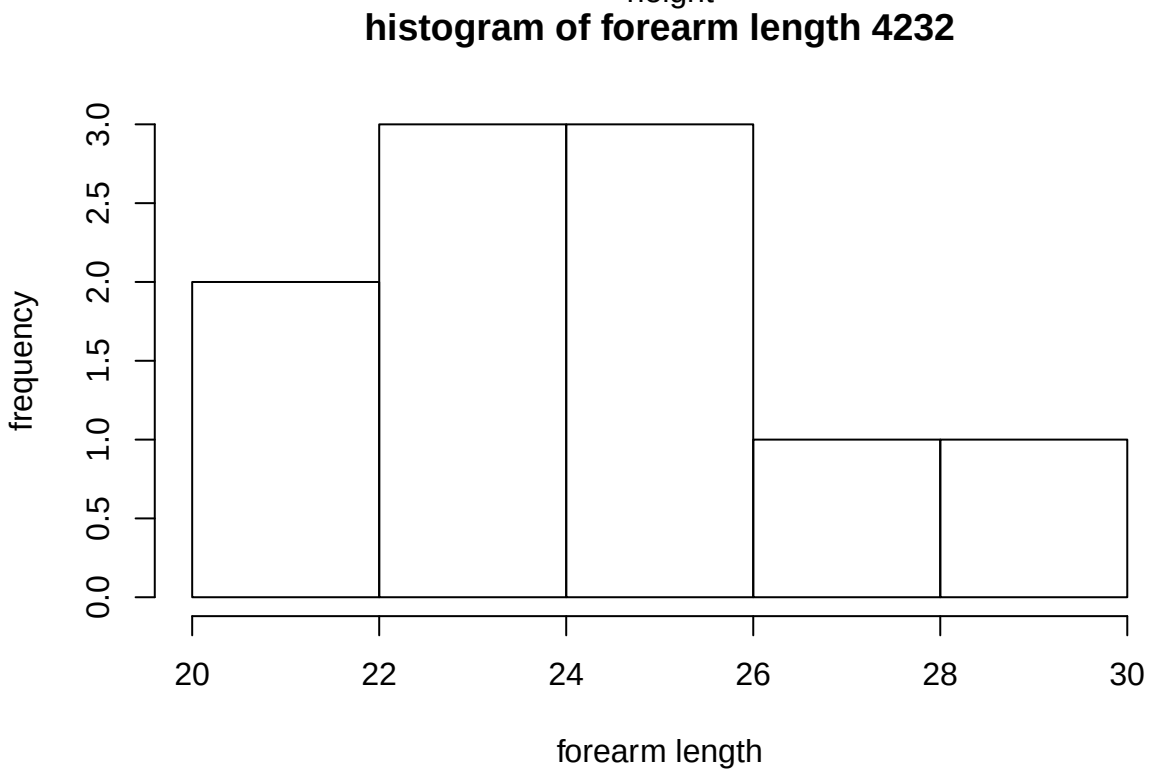
To check if the response variable is approcimately normal, we use the histogram and the qqplot. From the histogram, the graph does not represent the bell-shaped curve. From the qqplot, the dots are deviate from the black line, so the response variable doesn't seems normal.

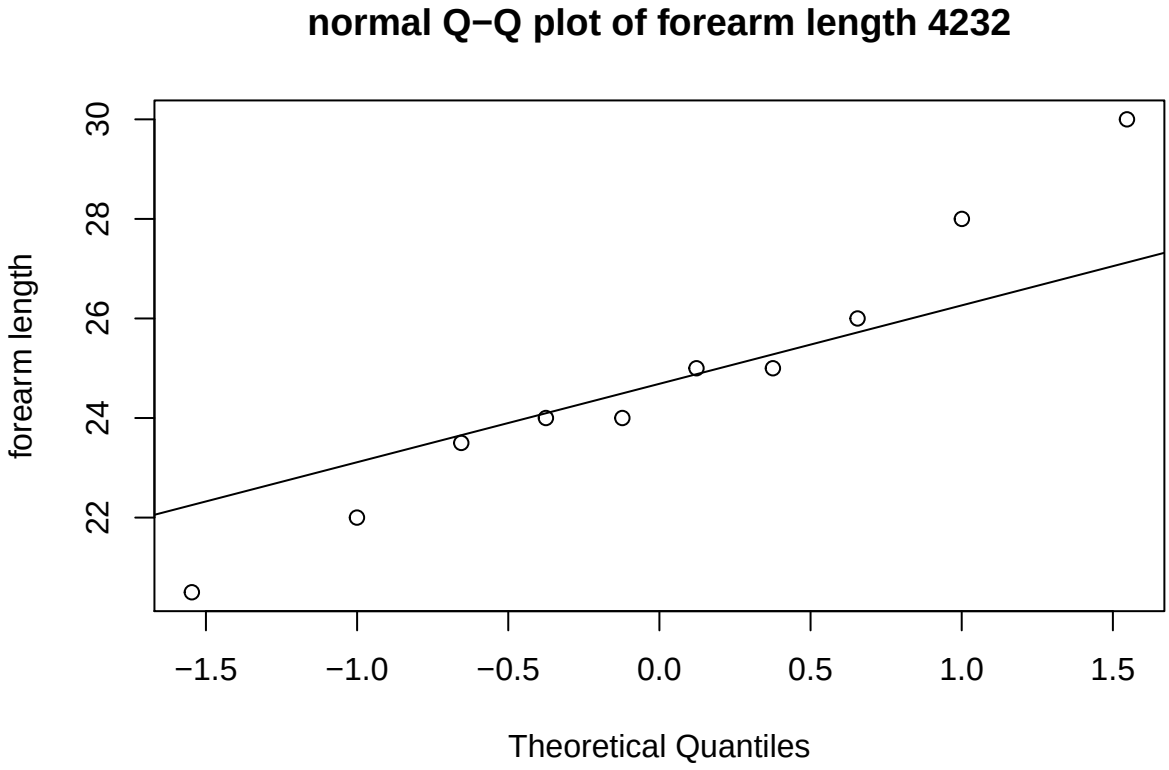
The median height of the sample data is 169.5 cm and the mean height of the sample data is 170.0 cm. The median forearm length of the sample data is 24.5 cm and the mean height of the sample data is 24.8 cm.

The standard deviation of the height of the sample data is 10.980. The standard deviation of the forearm length of the sample data is 2.750.

Thhe unusual point of the forearm length is 30. For the data of height, we do not have an unusual point.


scatter plot of height and forearm length 4232
forearm length
22
24
26
28
30
155
160
165
170
175
180
185
height


histogram of forearm length 4232
frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
20
22
24
26
28
30
forearm length



III. Methods and Model

From the summary we know that $\hat{\beta}_0$ equals to 5.77821, $\hat{\beta}_1$ equals to 0.11189. The expected forearm length equals to the sum of $\hat{\beta}_0$ and $\hat{\beta}_1$ times the height. We set the expected forearm length to $\hat{y}_i$ and the height to x_i, the formula for the linear regression model will be $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$ which is $\hat{y}_i = 5.77821 + 0.11189 * x_i$.

Since the p-value for the hypothesis test (if β_0 equals to 0) is 0.6800, which is larger than the significance level 0.05, we fail to reject the hypothesis test which means β_0 is not statistically different from 0. Since the p-value for the hypothesis test (if β_1 equals to 0) is 0.1960, which is larger than the significance level 0.05, we alse fail to reject the hypothesis test which means β_1 is not statistically different from 0.

$\hat{\beta}_0$ shows if height equals to 0 cm, the expected forearm length will be 5.77821 cm. $\hat{\beta}_1$ shows if height increase by 1 cm, the expected forearm length will be increase by 0.11189 cm.

```
##
## Call:
## lm(formula = forearm ~ height, data = sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7394 -0.9685 -0.5786  0.6438  5.7606
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.77821   13.49720   0.428    0.680
## height       0.11189    0.07924   1.412    0.196
##
## Residual standard error: 2.61 on 8 degrees of freedom
## Multiple R-squared:  0.1995, Adjusted R-squared:  0.09943
## F-statistic: 1.994 on 1 and 8 DF,  p-value: 0.1957
```

IV. Discussions and Limitations

The lurking variable to predict the forearm length might be the foot length. The reason why I choose foot length as the lurking variable is that they are all limb and it's more relevant between limb length than height.

We check the normality of the residual to see if height works for the linear regression model. Since it doesn't seems the varience is constant, we can say the residual might not nonrmal.Thus the model is not appropriate for the height.

There's another pair of variables to explore a simple linear regression model. I set the time we use for study as explanatory variable and GPA as response variable since I think more time on study will get higher GPA.