Exploring effects on the length of forearm by height

TONGFEI ZHOU

9/25/2020

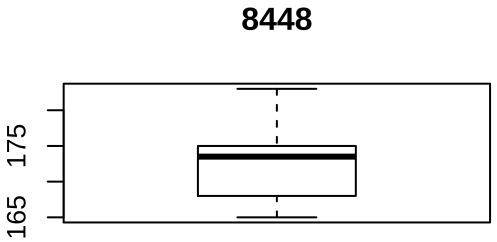
I. Introduction

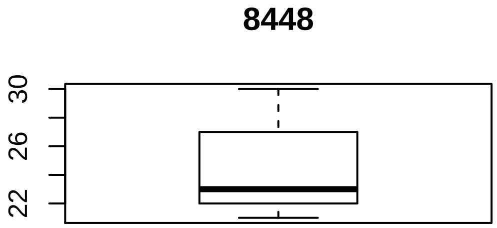
The title of this assignment is: **Exploring effects on the length of forearm by height**, and this assignment is aimed to use simple linear regression model to explore the potential effects on the length of forearm by the height a person owns.

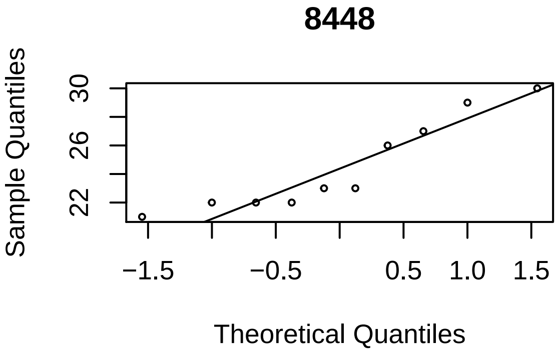
Since I did not get the data from the live class, I used the data collected from the STA302 Quercus Page of A1, provided by Professor Shivon, and set the seed with my last 4 digit and sample 10 data points randomly from the data set.

Since often people say the taller one is, the longer one's forearm's length should be, and due to the reason that using height to predict the length of forearms seems to be more senisible to myself, and I have great interest in studying with knowing a person's height, can I know his or her length of forearm. Therefore, I personally choose height as independent, or rather, explanatory variable and forearms as response variable. It is also why my title is called the effects on length of forearms by height.

II. Exploratory Data Analysis







From the boxplots we can see that the min of height is 165cm and max is 183cm, with median is around 173-174cm and is higher than the mean, 173cm. The first quantile of height is around 167cm and the thrid quantile is around 175cm with standard deviation of 6.253888.

The min of forearm is 21cm and the max is 30cm, with median 23cm which is lower than the mean, 24.5cm. The first quantile of forearm is around 22-23cm and the third quantile is around 27cm with standard deviation of 3.24037

From the qqplot, we can see that even though y_i are plotted around the line and some of them are quite far from the line, we can still say that they are approximately following the normal distribution. The first quantile of the length of forearm is approximately 23cm and the third quantile is around 27cm.

Notice that there are two pairs in the data where their height number is the same and there is no outliers in this data set.

III. Methods and Model

```
##
## Call:
## lm(formula = forearm ~ height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2045 -1.7216 -0.0625  1.8125  3.7955
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.4432    23.2538  -1.567   0.1557
## height        0.3523     0.1343   2.622   0.0305 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.52 on 8 degrees of freedom
## Multiple R-squared:  0.4622, Adjusted R-squared:  0.395
## F-statistic: 6.877 on 1 and 8 DF,  p-value: 0.03054
```

From the summary we can see that the estimated regression line is

$$\hat{y} = -36.4432 + 0.3523 * height$$

, with standard error being 23.2538 and 0.1343 respectively. Since we choose $\alpha = 0.05$ significance level and the p-value of the slope is 0.0305, and therefore we can reject the null hypothesis of

$$\beta_1 = 0$$

and the p-value of my intercept is 0.1557, which is relatively large(compared to the significance level $\alpha = 0.05$) so that there is no statistically significant evidence to reject the null hypothesis of

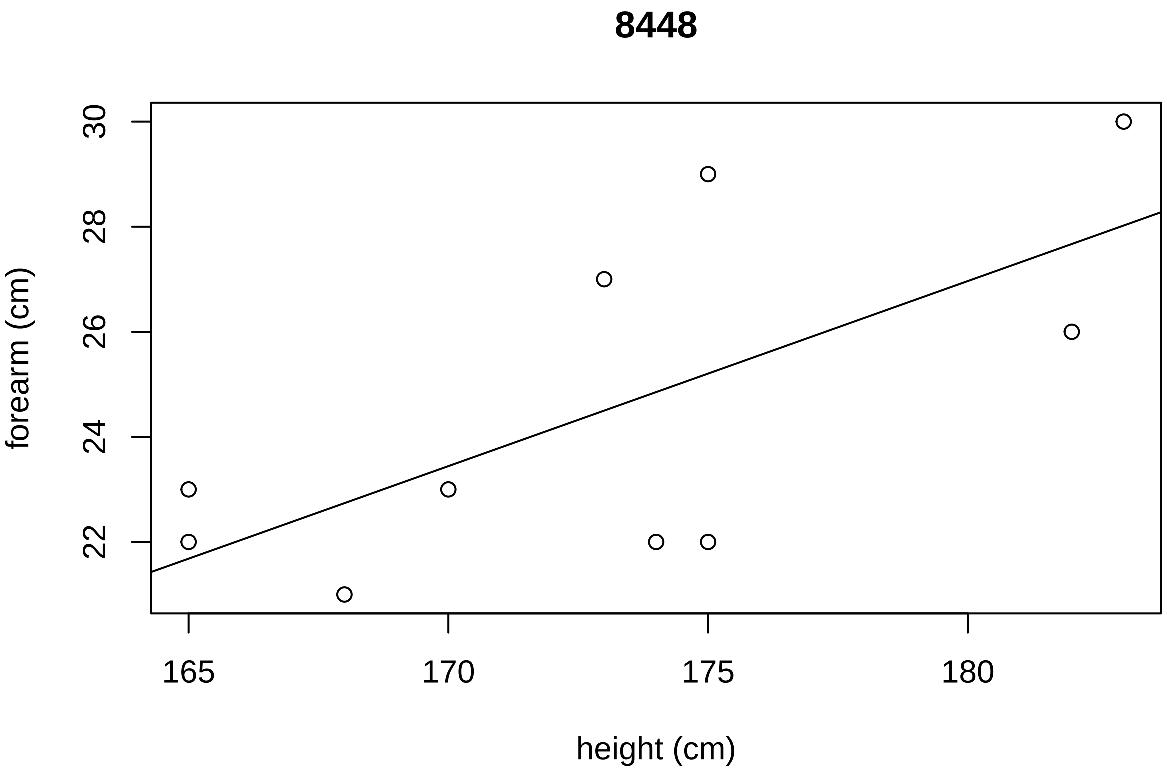
$$\beta_0 = 0$$

.

The interpretation of the slope is that with extra one unit change of the height, the average change of the length of forearm will be 0.3523. The interpretation of the intercept in this context is meaningless because there is no possible situation for height to be 0 and certainly the length of the forearm cannot be negative.(it should always stay as positive)

IV. Discussions and Limitations

```r
plot(height, forearm, main="8448", xlab="height (cm)", ylab="forearm (cm)")
abline(lmod)
```



Speaking of the limitation and restriction of our model, from the picture above, one can see that the points are scattered quite randomly, and since we only have 10 data points in this assignment, it is quite possible that the variance of the error is not constant and so the model might be wrong. Also, when testing the intercept in this model, we fail to reject the null hypothesis of $\beta_0 = 0$, and so it certainly has some troubles here.

Furthermore, it is worth pointing out that due to the restriction of the linear model we construct and the complexity of the real world, perhaps the relationship between the height and the length of the forearm is not linear but quadratic or so.

What is more, our model is only a **simple** linear one. However, it is quite certain that there are something that the model cannot account for, which, perhaps is a lurking variable rather than random error. For example, gene is perhaps a potential lurking variable since some families have longer forearm than others under the same height, or shorter one, which may have some impacts on the data we collected(e.g. height with 165cm but one with 22 and the other is 23 or so.)

Last, the relationship between the **hours spent on studied for one course (per day)** and **the grades earned in the final of the course** may be explained by a simple linear model. I will choose **the grades earned in the final of the course** as the response variable since one should spend hours studying the course first and then get the grades.