Simulate Two Correlated Random Variables

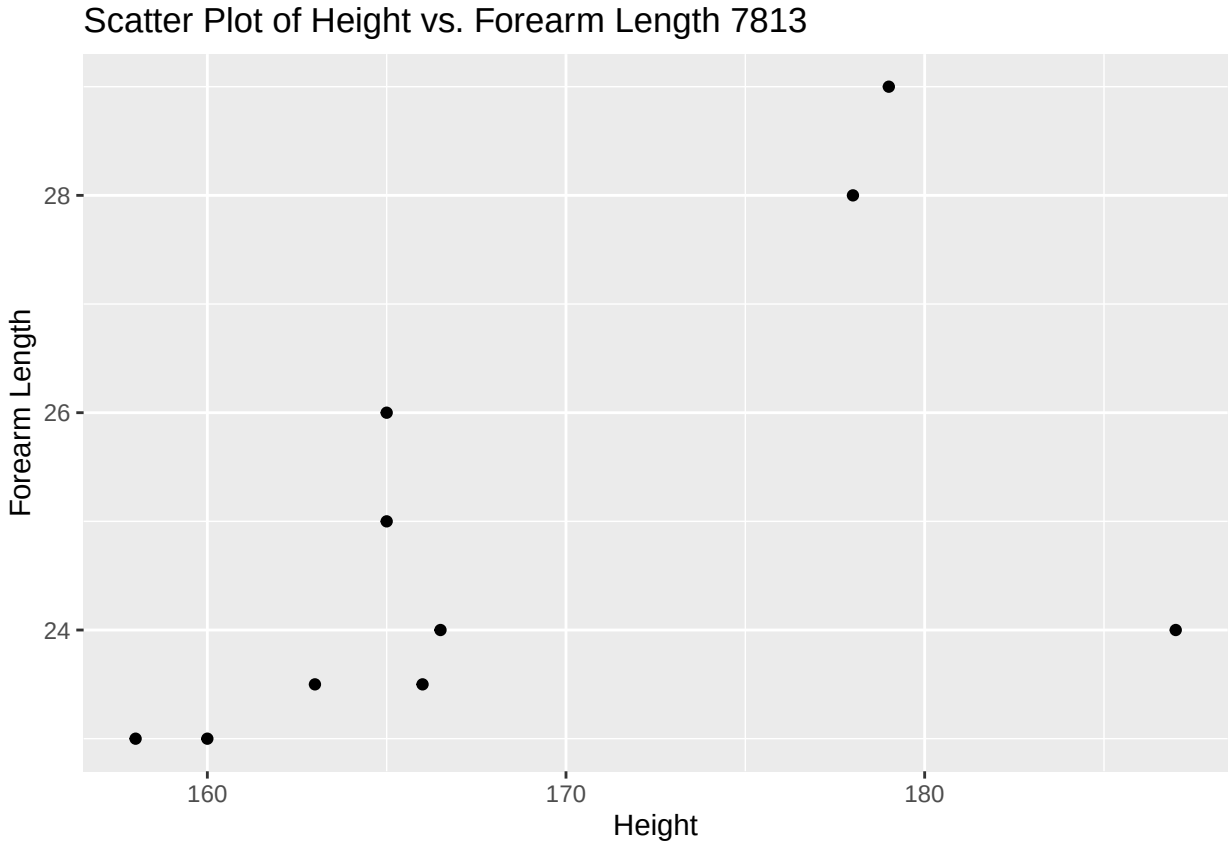
Zishu Zhu

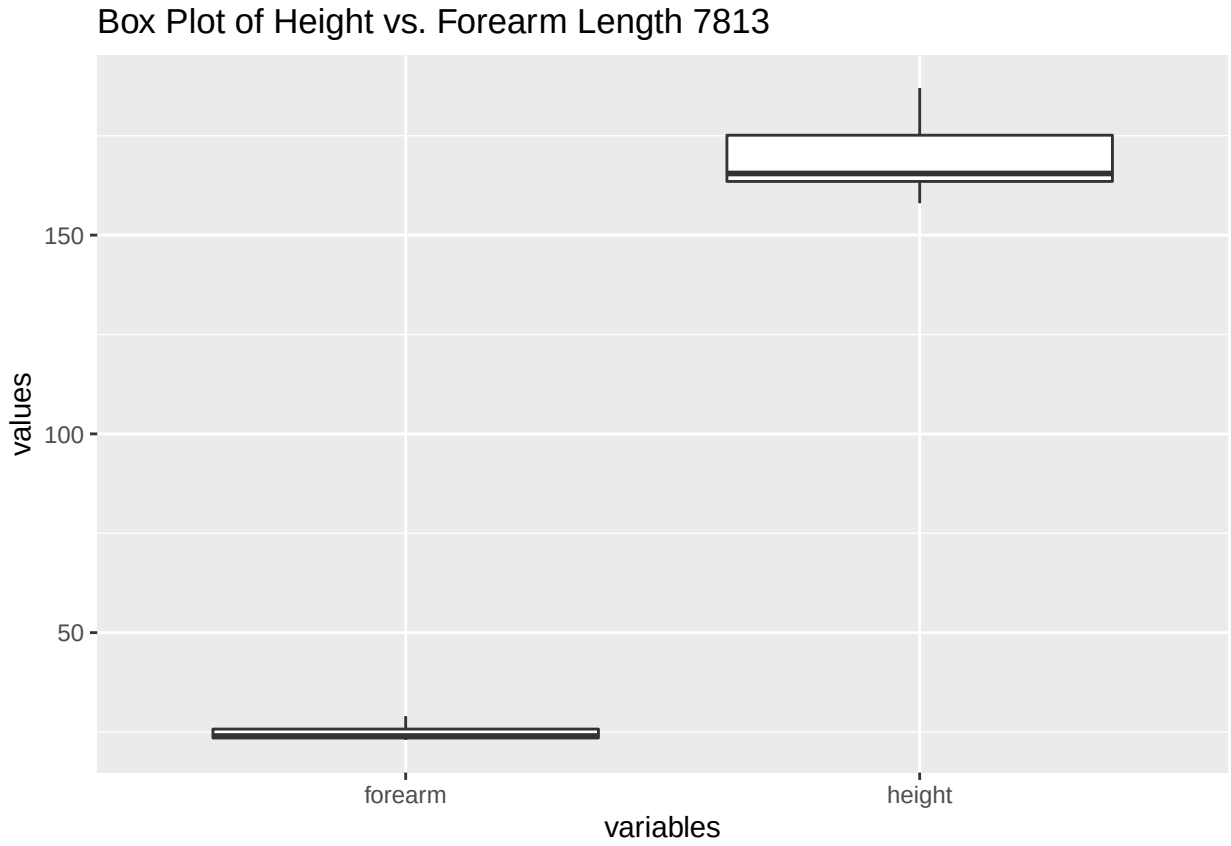
September 26, 2020

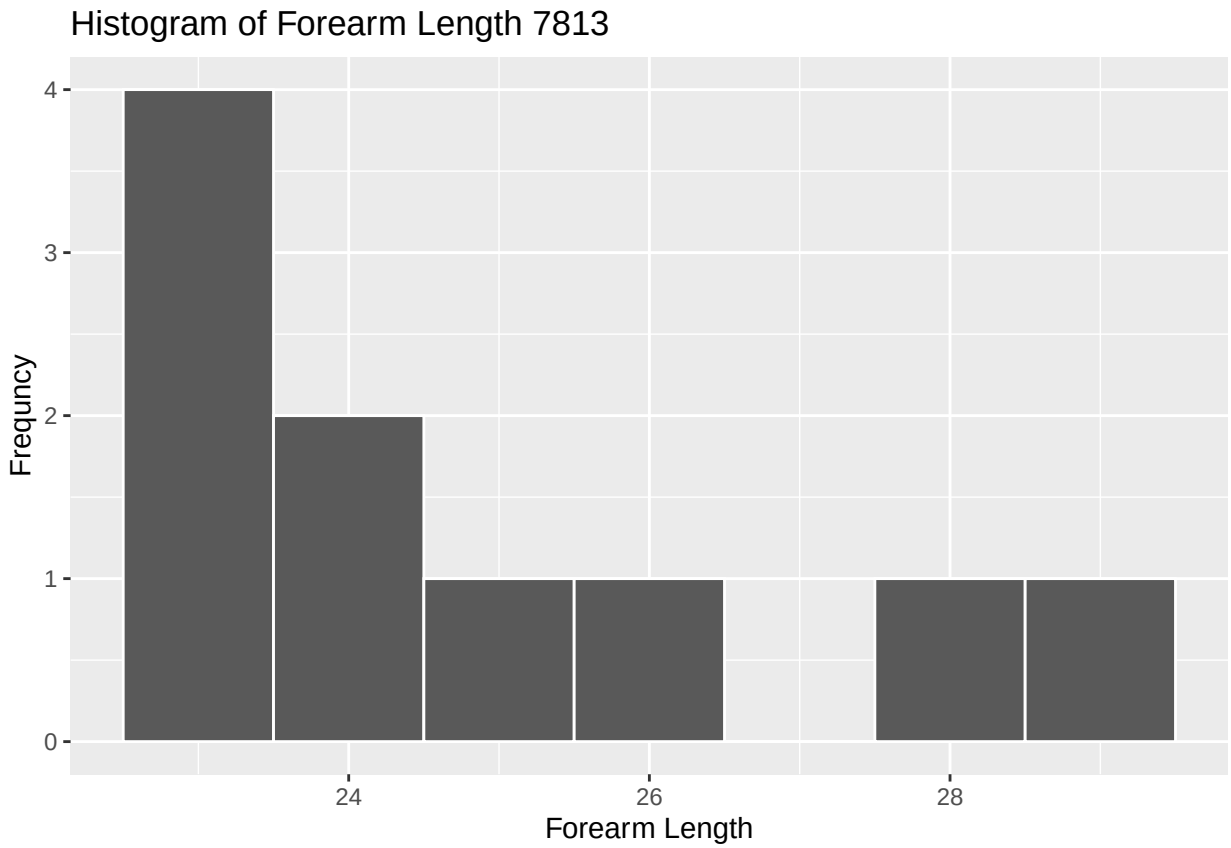
I. Introduction

I have surveyed all of the 347 students in sta 032 class and collected 10 adults' forearm length and height. My response variable is forearm length and explanatory variable is height.

II. Exploratory Data Analysis




Box Plot of Height vs. Forearm Length 7813
150
100
50
values
forearm
height
variables


Histogram of Forearm Length 7813
4
3
2
1
0
Frequncy
24
26
28
Forearm Length

I have drawed three plots above. The first one is a scatter plot of Height vs. Forearm Length. There is a weak linear association between height and forearm length. Because is only ten sample points on the plot, I cannot say that these two variables can form a linear association.

The second graph is a boxplot which is used to analyze the variability of the sample data. The middle 50% of values for height is from 163.5cm to 175.125cm. The middle 50% of values for forearm is from 23.5cm to 25.75cm. And from the box plot we can see that both of the reponse variable "forearm", and predictor variable "height" are right skewed. There is no outlier for both variables.

The histogram is a right skewed histogram and it is asymmetrical. The histogram is not bell-shaped. From the histogram, we can see that my response variable "Forearm Length" is not approximately normal distributed.

Centre

The centre of the sample data is as follow. The mean forearm length of the sample data is 24.9cm and the median forearm length of the sample data is 24cm. The mean height of the sample data is 168.75cm and the median height of the sample data is 165.5cm.

Spread

The spread of the sample data is as follow. The standard deviation of forearm length of the sample data is 2.1186998cm and the IQR of forearm length of the sample data is 2.25cm. The standard deviation of height of the sample data is 9.3667556cm and the IQR of height of the sample data is 11.625cm.

Unusual points

There is no outlier or unusual point for either response variable height of the sample data or predictor variable forearm of the sample data.

III. Methods and Model

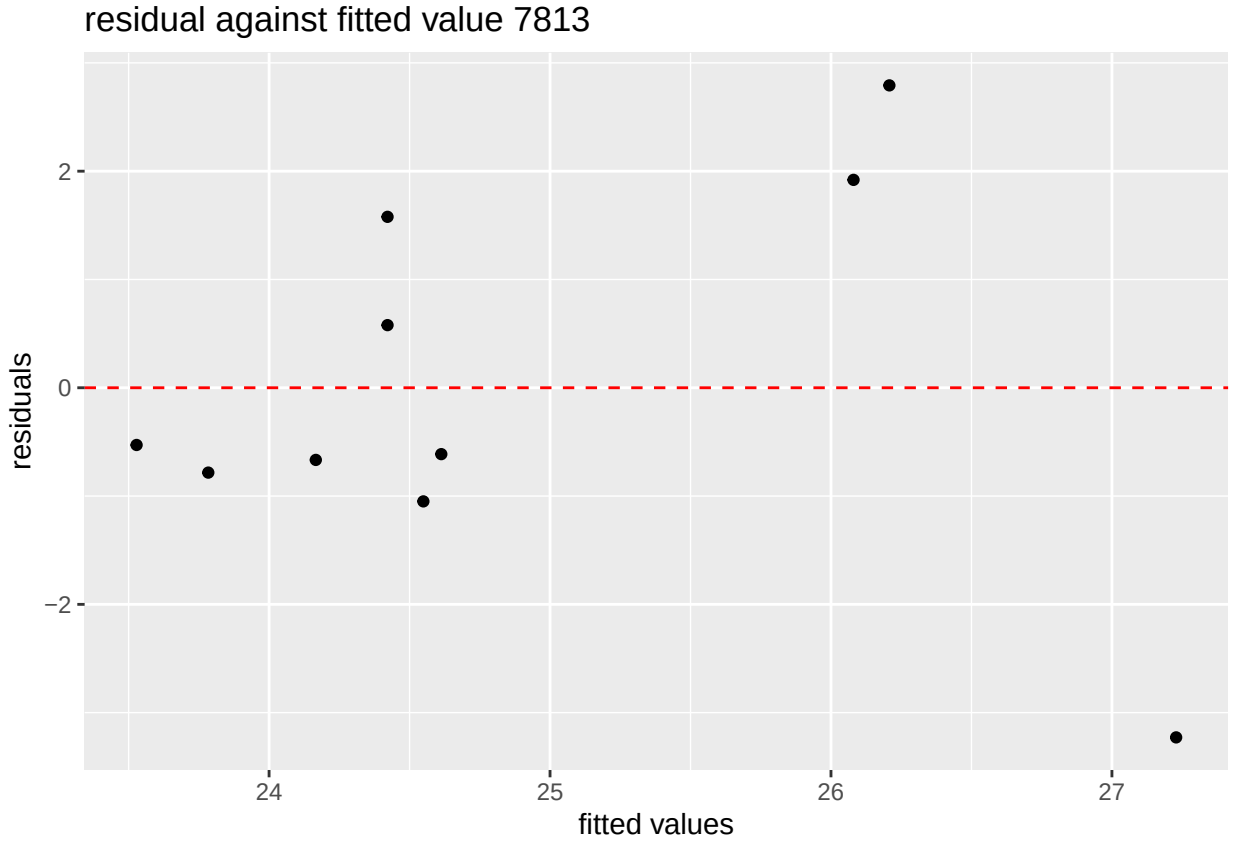
```
##
## Call:
## lm(formula = forearm ~ height, data = the_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2286 -0.7543 -0.5707  1.3285  2.7922
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.36881   11.15865   0.302   0.7704
## height       0.12759    0.06603   1.932   0.0894 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.856 on 8 degrees of freedom
## Multiple R-squared:  0.3182, Adjusted R-squared:  0.233
## F-statistic: 3.733 on 1 and 8 DF,  p-value: 0.08941
```

• From the summary of the linear model. $b_0 = 3.36881$ and $b_1 = 0.12759$. The fitted linear model is $\hat{y}_i = 3.36881 + 0.12759 \cdot x_i$. $\hat{y}_i$ stands for the expected forearm length and x_i stands for the height.

• By looking at the p-value for intercept β_0 is 0.7704 and it is larger than 0.05. We should not reject the null hypothesis that $\beta_0 = 0$ at 5% significance level. Therefore, β_0 is not statistically different from zero. And the p-value for slope β_1 is 0.0894 and it is also larger than 0.05. Thus, we should not reject the null hypothesis that $\beta_1 = 0$ at 5% significance level. So β_1 is also not statistically different from zero.

• $b_0 = 3.36881$ means that the expected forearm length is 3.36881 cm when the height is 0 cm. $b_1 = 0.12759$ means that when the height is increased by 1 cm, the expected forearm length will increase by 0.12759 cm.

IV. Discussions and Limitations



• One of the potential lurking variable is hand length. The limitation of my fit is that from the graph of residual against fitted value, we can find that the sample data has non-constant variance(increasing variance).
• The distance and speed of a car is driving is another pair of variables. The response variable is distance.