Are forearm length and height linearly related (.)

JW0935 (.)

September 26, 2020 (.)

I. Introduction

Part 1 (.)

Data collection

I use sample() function to randomly select 11 sample data from the dataset that surveyed 346 adult students of their height and forearm length.

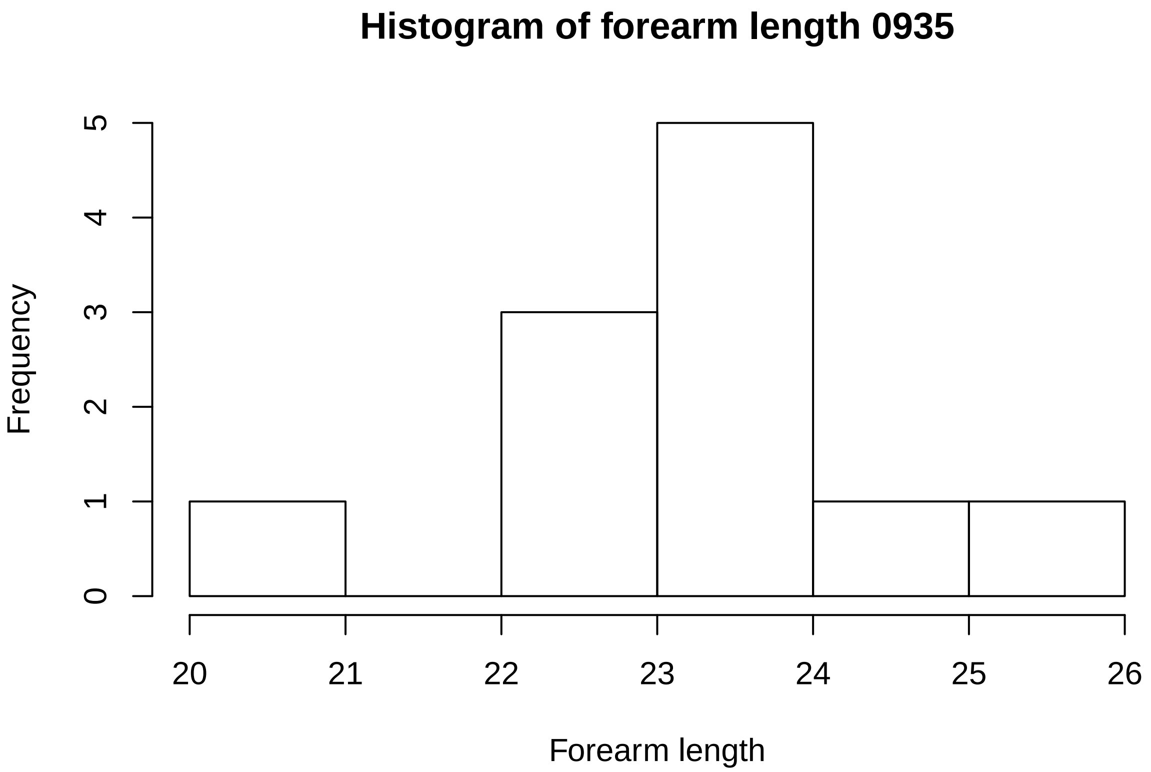
Explanatory variable & Response variable

My explanatory variable is height and response variable is forearm length because height is relatively easier to obtain compared to forearm length, and through this study, we can use height to predict forearm length when needed. For instance, if someone wants to buy clothes in-store and not sure about size, they can use their height to estimate forearm length to choose the right size.

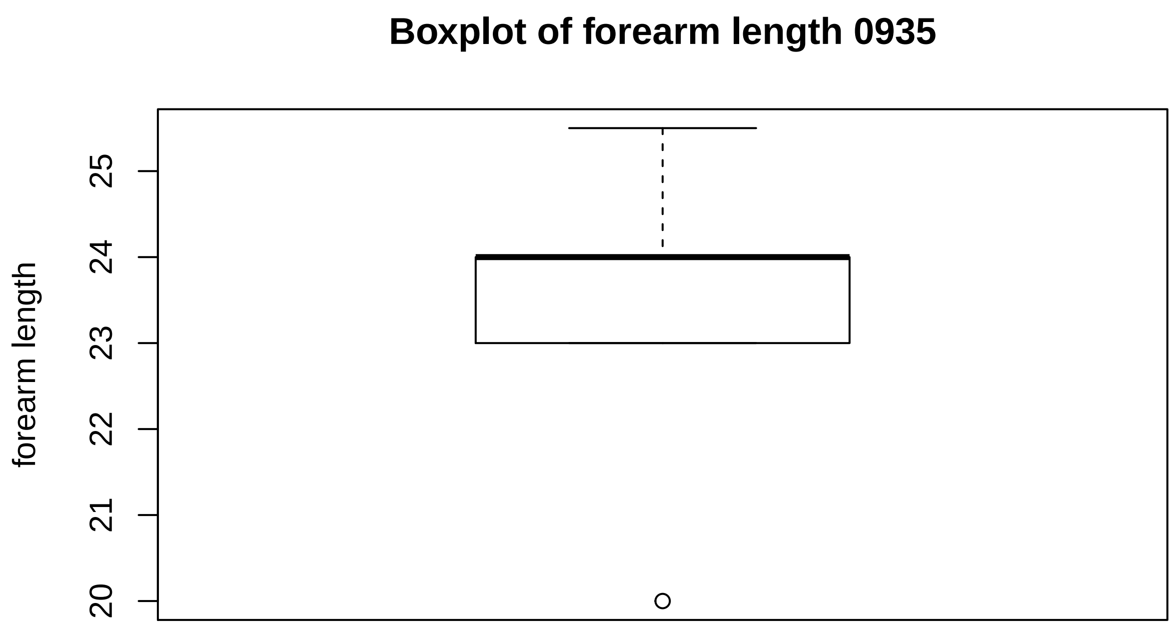
II. Exploratory Data Analysis

Part 2 (.)

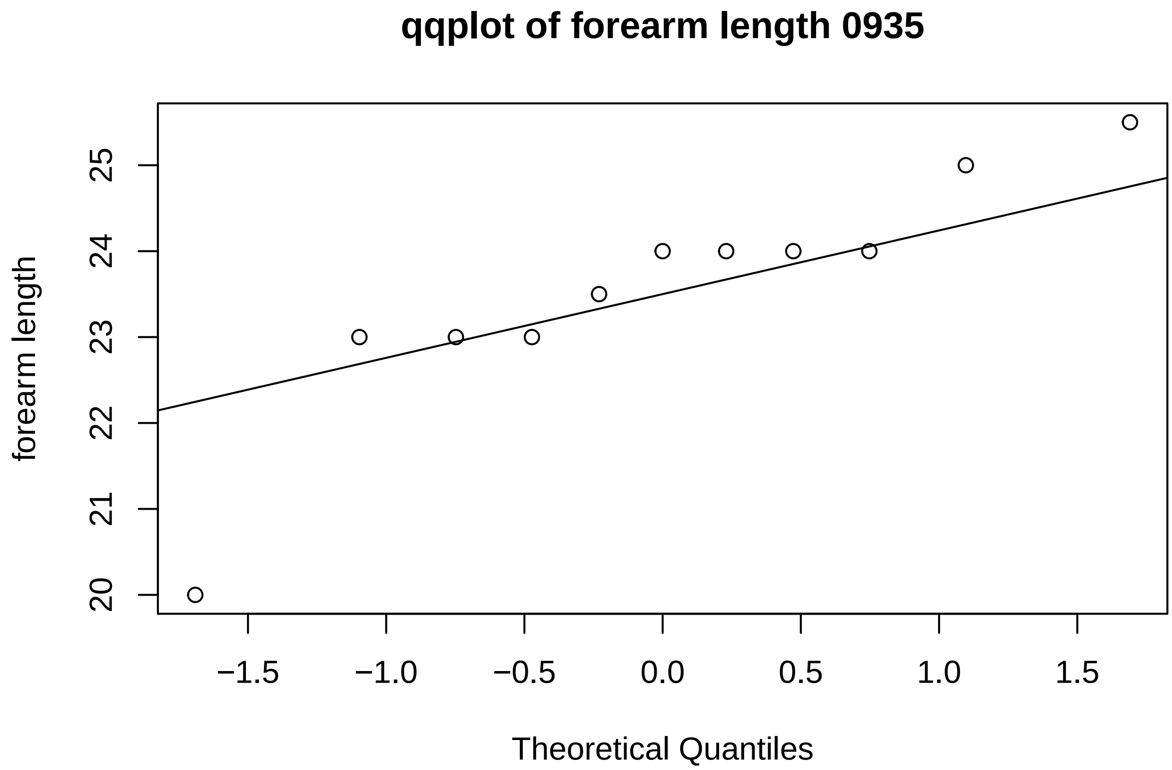
Graph & Interpretation



From the histogram, we can see that it is roughly left skewed and not symmetrical distributed, and it seems not to follow a normal distribution.



From the boxplot, we can see it is left skewed with a median of 24. Besides, there is an outlier whose forearm length is 20cm.



From the qqplot, we can see the points are not evenly distributed along the line, so we are more confident it not follows a normal distribution. Besides, we can see there is an outlier, proved our findings in the boxplot.

Calculation

Centre

```
mean(dataset2$forearm)
```

```
## [1] 23.54545
```

```
median(dataset2$forearm)
```

```
## [1] 24
```

```
mean(dataset2$height)
```

```
## [1] 171.2727
```

```
median(dataset2$height)
```

```
## [1] 168
```

```
# calculate mean and median for forearm length and height
```

From all three plots and calculations, the mean forearm length of the sample is 23.5454545 and the median is 24, which are the centres of forearm length. The mean height is 171.2727273 and the median height is 168, which are the centres of height.

Spread

```r
sd(dataset2$forearm)
```

```
## [1] 1.422226
```

```r
IQR_forearm <- IQR(dataset2$forearm)
IQR_forearm
```

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

```r
sd(dataset2$height)
```

```
## [1] 7.836975
```

```r
IQR_height <- IQR(dataset2$height)
IQR_height
```

```
## [1] 9.5
```

```r
#calculate sd and IQR for forearm length and height
```

The standard deviation of forearm length is 1.4222262 and the IQR of forearm length is 1, which are the spreads of forearm length. The standard deviation of height is 7.8369753 and the IQR of height is 9.5, which are the spreads of height.

Outlier

```r
lowerq_forearm <- quantile(dataset2$forearm,0.25) - 1.5*IQR_forearm
dataset2$forearm[which(dataset2$forearm < lowerq_forearm)]
```

```
## [1] 20
```

```r
# find points < 1st quartile - 1.5IQR of forearm length
```

Calculating 1st quartile - 1.5IQR, we find an outlier with forearm length 20, proved our findings in boxplot and qqplot.

III. Methods and Model

Part 3 (.)

Fitting SLR model

```r
linear_mod <- lm(forearm ~ height, data = dataset2)
summary(linear_mod)
```

```
##
## Call:
## lm(formula = forearm ~ height, data = dataset2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.12900 -0.16508 -0.00622  0.43872  1.77353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.85169    8.74779   0.783   0.4536
```

```
## height       0.09747    0.05103   1.910   0.0884 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.265 on 9 degrees of freedom
## Multiple R-squared:  0.2885, Adjusted R-squared:  0.2094
## F-statistic: 3.649 on 1 and 9 DF,  p-value: 0.08844
```

```
# fitting into SLR model and summary
```

Fitting the sample data into a SLR model, we have $\hat{\beta}_0 = 6.85169$, $\hat{\beta}_1 = 0.09747$, and the estimated model : $\hat{Y}_i = 6.85169 + 0.09747x_i$, where $\hat{Y}_i$ is expected forearm length and x_i is height.

Testing

When testing whether $\beta_0 = 0$, p-value is 0.4536 which is greater than significance level of 5%, so we do not reject $H_0 : \beta_0 = 0$, which means β_0 is not statistically different from 0.

And when testing whether $\beta_1 = 0$, p-value is 0.0884 which is also greater than significance level of 5%, so we do not reject $H_0 : \beta_1 = 0$, which means β_1 is not statistically different from 0.

Interpretation

$\hat{\beta}_0$: when height is 0cm, the expected forearm length is 6.85169cm.

$\hat{\beta}_1$: when height increase by 1cm, the expected forearm length increase by 0.09747cm.

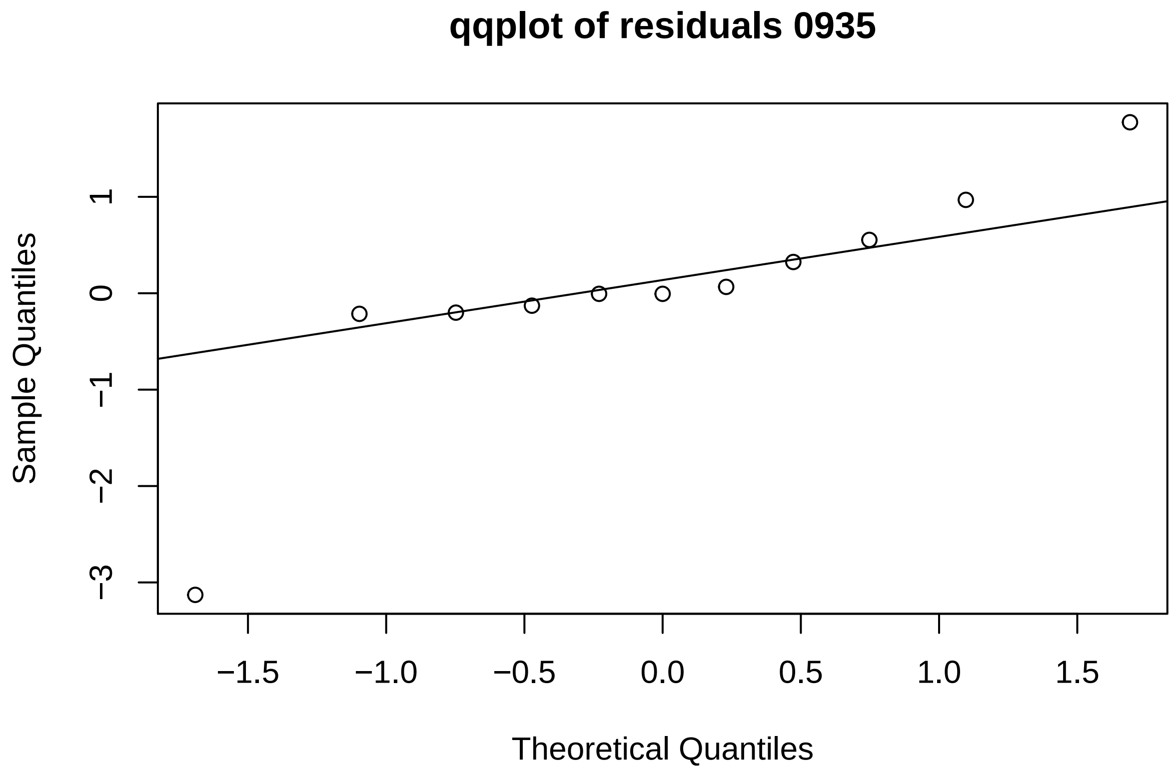
IV. Discussions and Limitations

Part 4 (.)

Limitations

Since we only use height as the explanatory variable to predict forearm length, there might be some potential lurking variables. Except for height, some other factors may potentially affect forearm length. For example, personal exercise habits may be also related to their forearm length. If the sampled objects like to play basketball, they might have a longer forearm length than others who do not like playing with the same height.

Besides, due to our small sample size, the accuracy is low and the error can be large.



By checking the qqplot of residuals, we can see residuals are not evenly distributed along the line, so it is not normally distributed, violating the model assumption of normality, which is also a limitation of fitting the SLR model.

Extension

Except for our study between height and forearm length, we can explore other linear relationships. For instance, we can study the relationship between disposable income and housing area, fitting them into an SLR model, where disposable income is the explanatory variable and housing area is the response variable.