# MTH522 Homework 1

Anubhav Shankar

2022-10-03

```
library(UsingR)
```

```
## Loading required package: MASS
```

```
## Loading required package: HistData
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
##
## Attaching package: 'UsingR'
```

```
## The following object is masked from 'package:survival':
##
##     cancer
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:Hmisc':
##
##     src, summarize
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ggpubr)
library(rmarkdown)
library(knitr)
```

We’ll now load the **Pearson’s Father-Son Height** data for further evaluation

```
fason <- father.son # Load the data and save it into an object
glimpse(fason) # Get a quick overview of the data
```

```
## Rows: 1,078
## Columns: 2
## $ fheight <dbl> 65.04851, 63.25094, 64.95532, 65.75250, 61.13723, 63.02254, 65…
## $ sheight <dbl> 59.77827, 63.21404, 63.34242, 62.79238, 64.28113, 64.24221, 64…
```

```
str(fason) # Same as glimpse()
```

```
## 'data.frame':    1078 obs. of  2 variables:
##  $ fheight: num  65 63.3 65 65.8 61.1 ...
##  $ sheight: num  59.8 63.2 63.3 62.8 64.3 ...
```

```
head(fason) # Preliminary exploration
```

```
##    fheight  sheight
## 1 65.04851 59.77827
## 2 63.25094 63.21404
## 3 64.95532 63.34242
## 4 65.75250 62.79238
## 5 61.13723 64.28113
## 6 63.02254 64.24221
```

So, the data contains of two numerical features - **Father’s Height (fheight)** and **Son’s Height (sheight)**. There are a total of 1078 entries.

Let’s create a simple linear regression model by making the son’s height the dependent variable and father’s height as the independent variable.

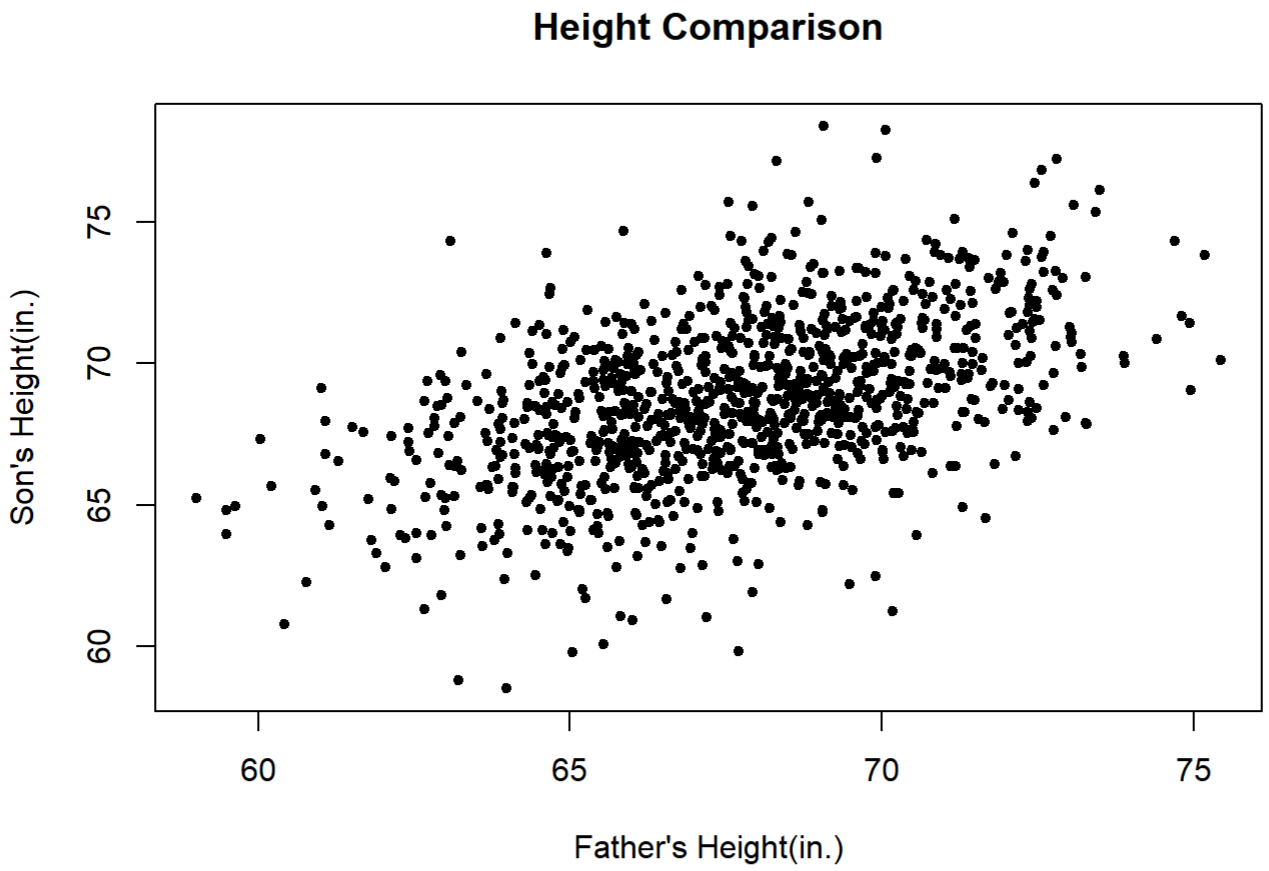
```
model_fit <- lm(fason$sheight ~ fason$fheight) # Create a simple Linear Regression Model
summary(model_fit) # Get the model statistics
```

```
## 
## Call:
## lm(formula = fason$sheight ~ fason$fheight)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.8772 -1.5144 -0.0079  1.6285  8.9685 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   33.88660    1.83235   18.49   <2e-16 ***
## fason$fheight  0.51409    0.02705   19.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.437 on 1076 degrees of freedom
## Multiple R-squared:  0.2513, Adjusted R-squared:  0.2506 
## F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

**The Goodness of Fit (R^2) for this data is only 0.2506 i.e. only 25% of the data variance is explained by the independent variable, father’s height in this case. Also, we see that father’s height is a significant variable having a p-value << 0.05**

Let us now create a simple scatter plot to see the relationship between the two variables.

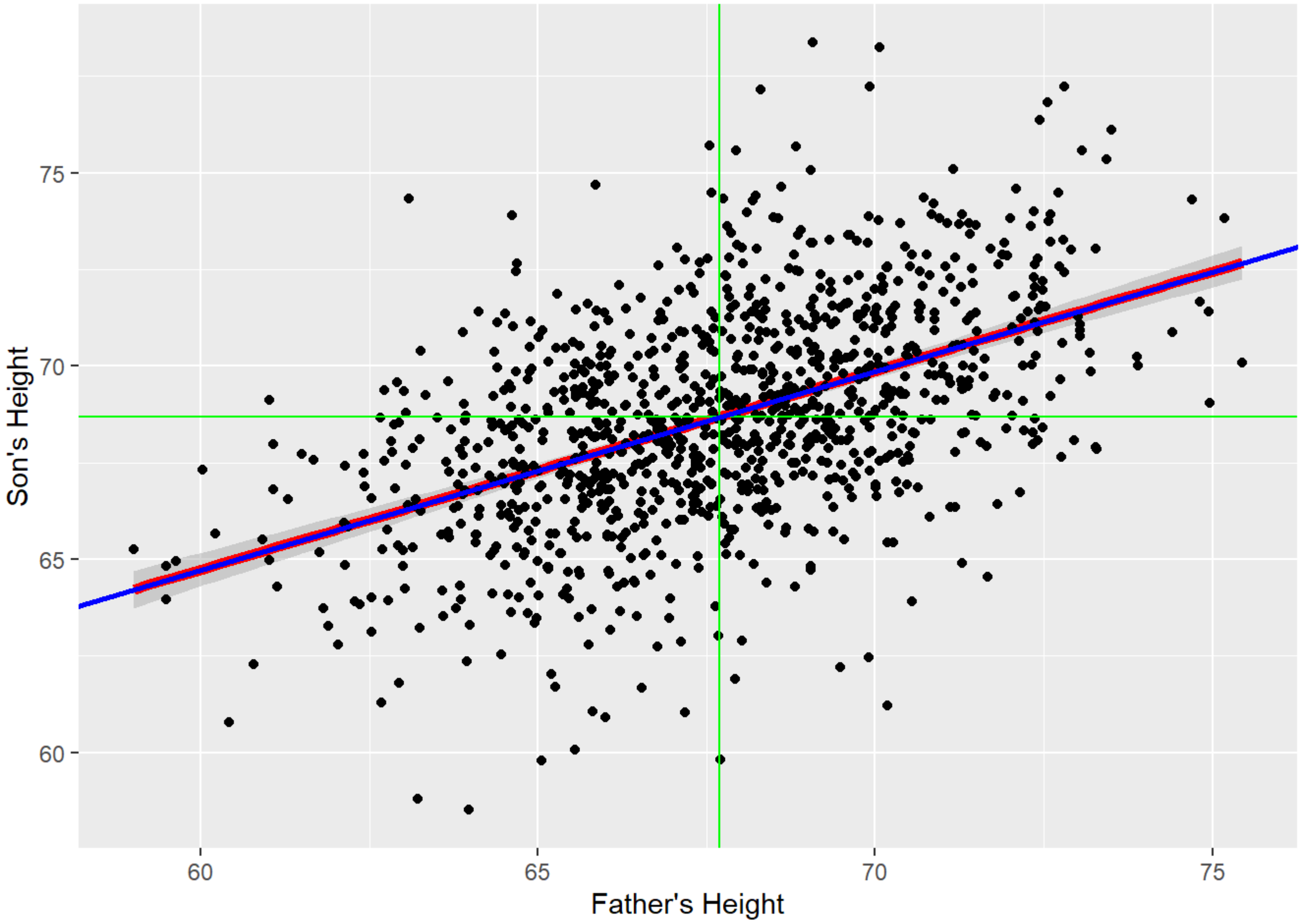
```
plot(fason$fheight, fason$sheight, xlab = "Father's Height(in.)", ylab = "Son's Height(in.)",
     pch = 20) + title("Height Comparison") # Create a simple Scatter Plot
```



We see that there is a strong concentration of the observations. This might indicate some *correlation*. However, let’s create a more detailed plot by adding a regression line and a SD line along with plotting the respective means of the heights.

```
plot.new() # Use this when creating a .Rmd to prevent errors if you have multiple plots in a single file
ggplot(fason,aes(x = fheight, y = sheight)) + # This line initiates the plotting function
geom_smooth(method = "lm", color = "red", size = 2) + # This line creates the regression line
geom_point() + # This line generates the points of the scatter plot
points(mean(fason$fheight),mean(fason$sheight)) + # This line basically sets the coordinates
geom_vline(xintercept = mean(fason$fheight),color = "green") + # Adds a vertical line passing through the mean of father's height
geom_hline(yintercept = mean(fason$sheight),color = "green") + # Adds a vertical line passing through the mean of son's height
geom_abline(slope = model_fit$coefficients[2],intercept = model_fit$coefficients[1],color = "blue",size = 1) + # This line creates the SD line by getting the intercept and slope from the regression model
xlab("Father's Height") + # Adds a x-axis label
ylab("Son's Height") # Adds a y-axis label
```

```
## `geom_smooth()` using formula 'y ~ x'
```



We see that the **SD line and the regression line overlap perfectly**. Concurrently, the mean of father's height and the son's height are pretty close by as well.