# MTH - 522

## Homework – 1

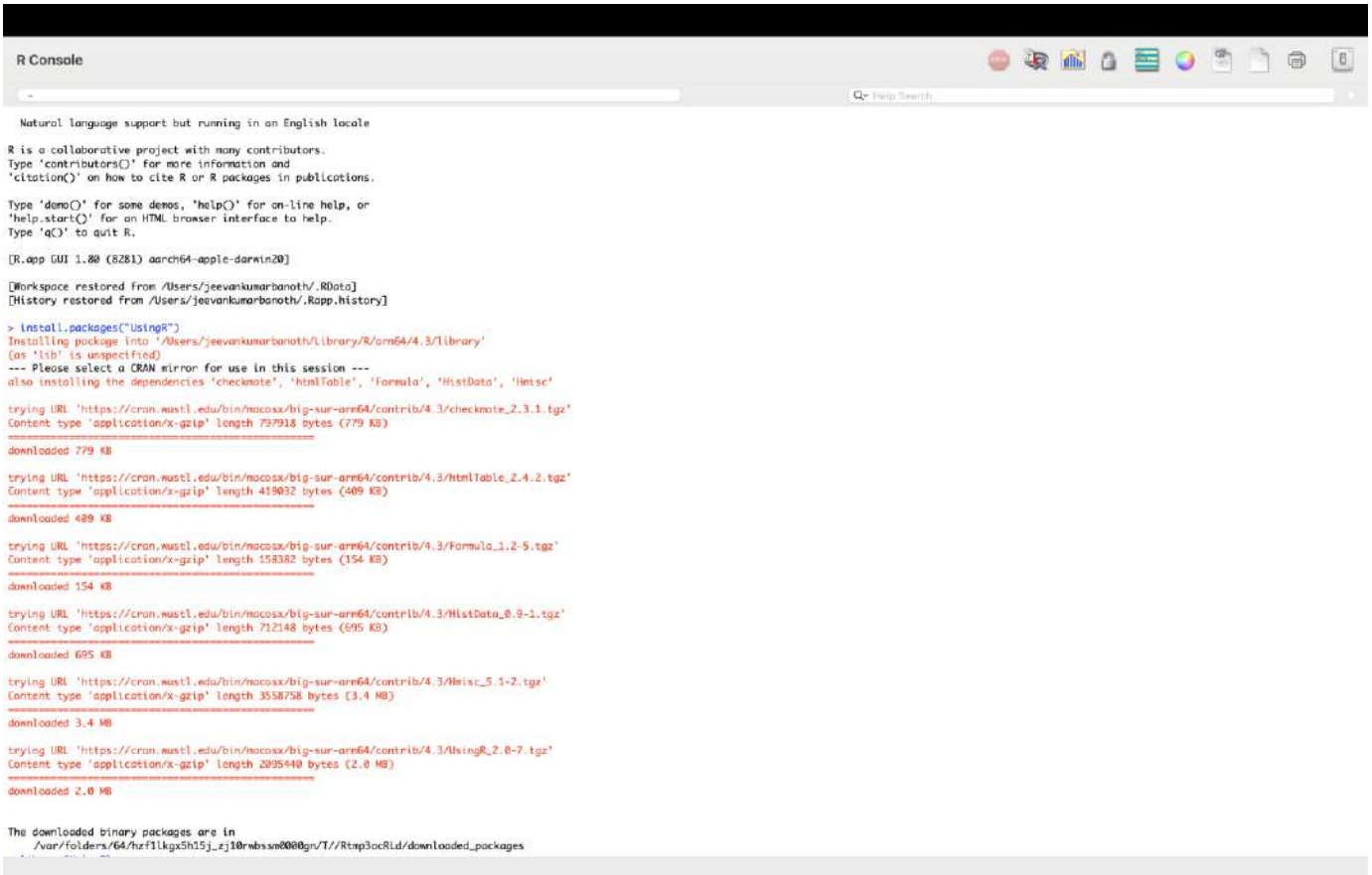
**Jeevan Kumar Banoth**

## 1) Regression on Pearson's father-son data.

a) Get the classical Pearson's father-son data by the following R commands

→ Install the UsingR package

**install.packages("UsingR")**



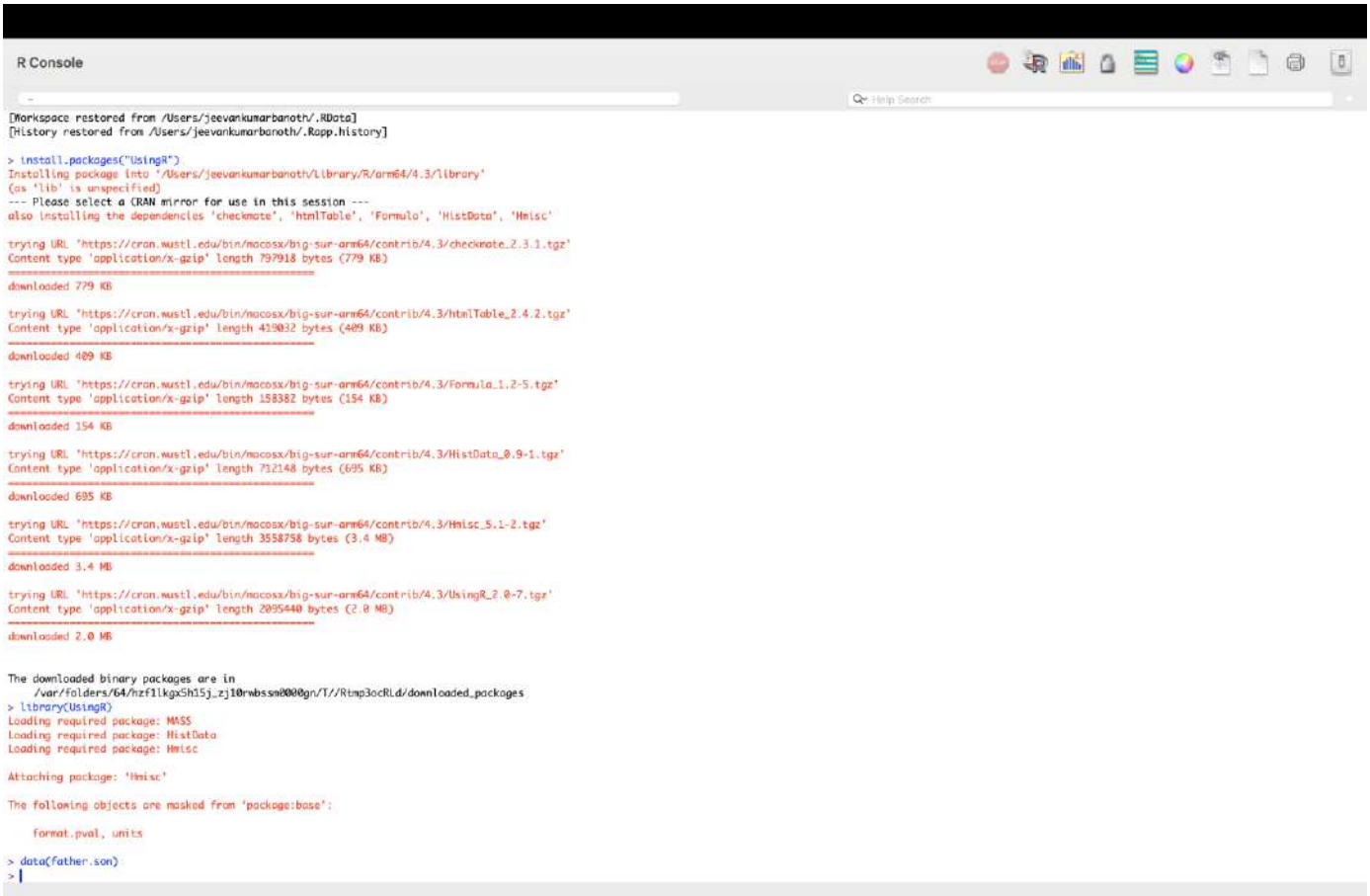
```
R Console

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' for how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' for quit R.

[R.app GUI 1.80 (8281) aarch64-apple-darwin20]

[Workspace restored from /Users/jeevankumarbanoth/.RData]
[History restored from /Users/jeevankumarbanoth/.Rapp.history]

> install.packages("UsingR")
Installing package into '/Users/jeevankumarbanoth/Library/R/arm64/4.3/library'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
also installing the dependencies 'checkmate', 'htmlTable', 'Formula', 'HistData', 'Hmisc'

trying URL 'https://cran.mustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/checkmate_2.3.1.tgz'
Content type 'application/x-gzip' length 797918 bytes (779 KB)
==================================================
downloaded 779 KB

trying URL 'https://cran.mustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/htmlTable_2.4.2.tgz'
Content type 'application/x-gzip' length 419032 bytes (409 KB)
==================================================
downloaded 409 KB

trying URL 'https://cran.mustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/Formula_1.2-5.tgz'
Content type 'application/x-gzip' length 158382 bytes (154 KB)
==================================================
downloaded 154 KB

trying URL 'https://cran.mustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/HistData_0.9-1.tgz'
Content type 'application/x-gzip' length 712148 bytes (695 KB)
==================================================
downloaded 695 KB

trying URL 'https://cran.mustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/Hmisc_5.1-2.tgz'
Content type 'application/x-gzip' length 3558758 bytes (3.4 MB)
==================================================
downloaded 3.4 MB

trying URL 'https://cran.mustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/UsingR_2.0-7.tgz'
Content type 'application/x-gzip' length 2095440 bytes (2.0 MB)
==================================================
downloaded 2.0 MB

The downloaded binary packages are in
    /var/folders/64/hzf1lkgx5h15j_zj10rwbssm0000gn/T//Rtmp3ocRLd/downloaded_packages
```

→ Load the UsingR library

**library(UsingR)**

And loading the father.son dataset from the installed packages by using the command below.

**data(father.son)**



```
R Console
[Workspace restored from /Users/jeevankumarbanoth/.RData]
[History restored from /Users/jeevankumarbanoth/.Rapp.history]

> install.packages("UsingR")
Installing package into '/Users/jeevankumarbanoth/Library/R/arm64/4.3/library'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
also installing the dependencies 'checkmate', 'htmlTable', 'Formula', 'HistData', 'Hmisc'

trying URL 'https://cran.wustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/checkmate_2.3.1.tgz'
Content type 'application/x-gzip' length 797918 bytes (779 KB)
==================================================
downloaded 779 KB

trying URL 'https://cran.wustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/htmlTable_2.4.2.tgz'
Content type 'application/x-gzip' length 419032 bytes (409 KB)
==================================================
downloaded 409 KB

trying URL 'https://cran.wustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/Formula_1.2-5.tgz'
Content type 'application/x-gzip' length 158382 bytes (154 KB)
==================================================
downloaded 154 KB

trying URL 'https://cran.wustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/HistData_0.9-1.tgz'
Content type 'application/x-gzip' length 712148 bytes (695 KB)
==================================================
downloaded 695 KB

trying URL 'https://cran.wustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/Hmisc_5.1-2.tgz'
Content type 'application/x-gzip' length 3558758 bytes (3.4 MB)
==================================================
downloaded 3.4 MB

trying URL 'https://cran.wustl.edu/bin/macosx/big-sur-arm64/contrib/4.3/UsingR_2.0-7.tgz'
Content type 'application/x-gzip' length 2095440 bytes (2.0 MB)
==================================================
downloaded 2.0 MB


The downloaded binary packages are in
    /var/folders/64/hzf1lkgx5h15j_zj10rwbssm0000gn/T//Rtmp3ocRLd/downloaded_packages
> library(UsingR)
Loading required package: MASS
Loading required package: HistData
Loading required package: Hmisc

Attaching package: 'Hmisc'

The following objects are masked from 'package:base':

    format.pval, units

> data(father.son)
>
```

→ **Attaching dplyr library:**

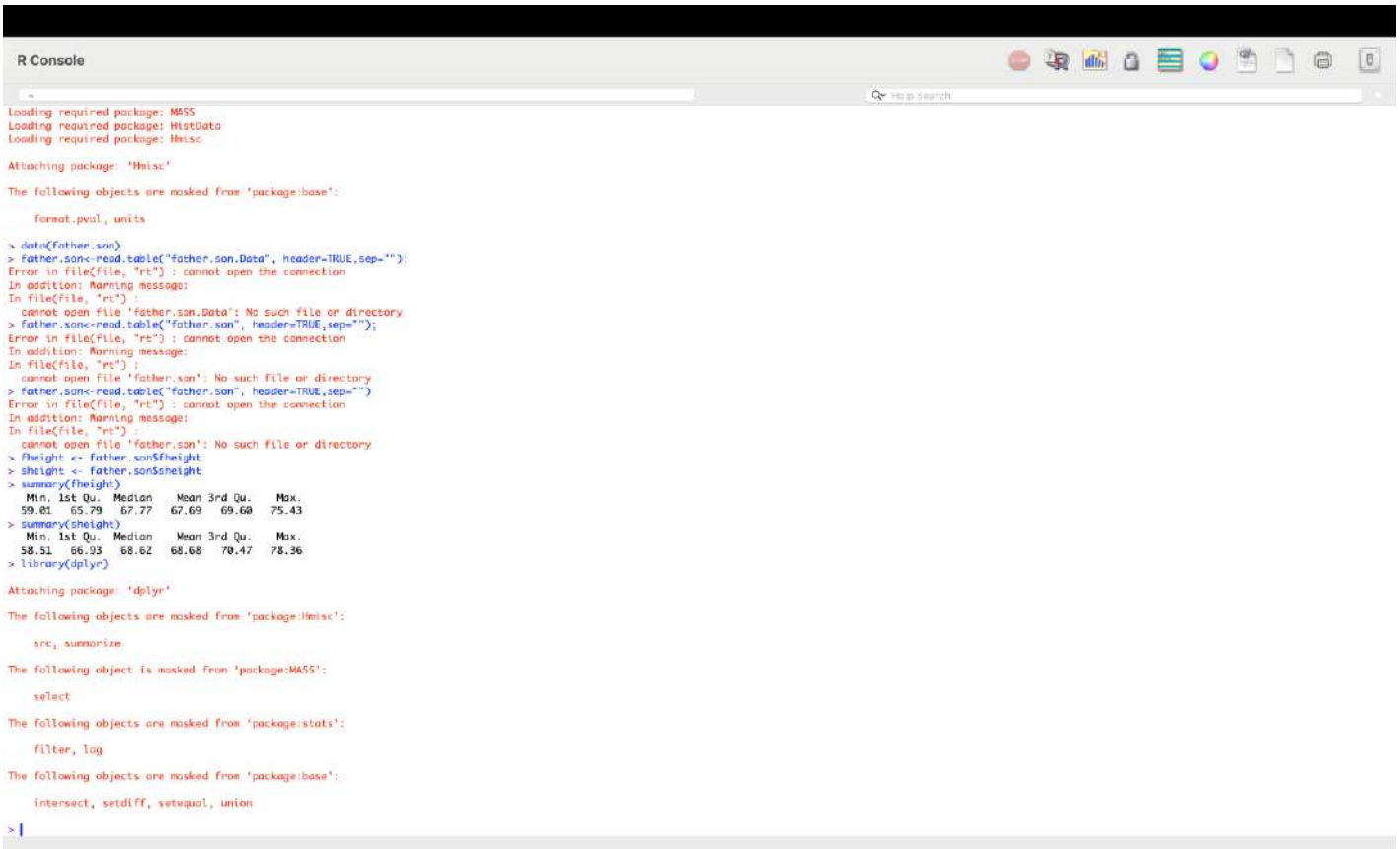
Summary statistics for father's and son’s height.

Using below codes:

→ summary(fheight)

→ summary(sheight)

→ Library(dplyr)



```
R Console
Loading required package: MASS
Loading required package: Hmisc
Loading required package: HistData

Attaching package: 'Hmisc'

The following objects are masked from 'package:base':

    format.pval, units

> data(father.son)
> father.son<-read.table("father.son.Data", header=TRUE,sep=""); 
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'father.son.Data': No such file or directory
> father.son<-read.table("father.son", header=TRUE,sep="");
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'father.son': No such file or directory
> father.son<-read.table("father.son", header=TRUE,sep="")
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'father.son': No such file or directory
> fheight <- father.son$fheight
> sheight <- father.son$sheight
> summary(fheight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  59.01   65.79   67.77   67.69   69.60   75.43 
> summary(sheight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  58.51   66.93   68.62   68.68   70.47   78.36 
> 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

>
```

→ To ensure necessary libraries installed and Loaded:

Using the **glimpse ()** function from the **dplyr** package is a great way to quickly inspect your data frame's structure, including column types and a few initial rows of data. We'll now load the

**Pearson's Father-Son Height** data for further evaluation.

```
R Console
> data(father.son)
> father.son<-read.table("father.son.Data", header=TRUE,sep=""");
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'father.son.Data': No such file or directory
> father.son<-read.table("father.son", header=TRUE,sep="");
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'father.son': No such file or directory
> father.son<-read.table("father.son", header=TRUE,sep="")
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'father.son': No such file or directory
> fheight <- father.son$fheight
> sheight <- father.son$sheight
> summary(fheight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  59.01   65.79   67.77   67.69   69.68   75.43
> summary(sheight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  58.51   66.93   68.62   68.68   70.47   78.36
> 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)
> fason <- father.son
> glimpse(fason)
Rows: 1,078
Columns: 2
$ fheight <dbl> 65.04851, 63.25094, 64.95532, 65.75250, 61.13723, 63.02254, 65.37053, 64.72398, 66.06509, 66.96738, 59.00800, 62.93203, 63.67063, 64.07386, 64.68851, 65.15466, 66.37353, 65.57704, 67.36765, 66.75929, …
$ sheight <dbl> 59.77827, 63.21404, 63.34242, 62.79238, 64.28113, 64.24221, 64.08231, 63.99574, 64.61338, 63.97944, 65.24451, 65.35102, 65.67992, 65.43664, 65.29391, 64.79017, 65.01881, 65.54640, 65.08145, 65.49008, …
>
```

To explore the structure of your **fason** dataset in more detail, you can use the **str ()** function, which displays the structure of an R object. Additionally, the **head ()** function will show the first few rows of your dataset, giving you a glimpse of the actual data values. Here's how you can use these functions:

Both **str ()** and **head ()** are useful functions for getting a sense of the data before you dive deeper into analysis or visualization.

```
R Console
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'father.son': No such file or directory
> fheight <- father.son$fheight
> sheight <- father.son$sheight
> summary(fheight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  59.01   65.79   67.77   67.69   69.60   75.43
> summary(sheight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  58.51   66.93   68.62   68.68   70.47   78.36
> 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)
> fason <- father.son
> glimpse(fason)
Rows: 1,078
Columns: 2
$ fheight <dbl> 65.04851, 63.25094, 64.95532, 65.75250, 61.13723, 63.02254, 65.37053, 64.72398, 66.06509, 66.96738, 59.00800, 62.93203, 63.67063, 64.07386, 64.68851, 65.15466, 66.37353, 65.57704, 67.36765, 66.75929, …
$ sheight <dbl> 59.77827, 63.21404, 63.34242, 62.79238, 64.28113, 64.24221, 64.08231, 63.99574, 64.61338, 63.97944, 65.24451, 65.35102, 65.67992, 65.43664, 65.29391, 64.79017, 65.01881, 65.54640, 65.88145, 65.49008, …
> str(fason)
'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)
   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
>
```

The dataset consists of two numeric columns: Father's Height ("fheight") and Son's Height ("sheight"). It contains 1078 data points.

We will create a straightforward linear regression model where Son's Height is the variable being predicted ("dependent variable") based on Father's Height ("independent variable").

**model_fit <- lm(fason$sheight ~ fason$fheight) # Create a simple Linear Regression Model**
**summary(model_fit)**

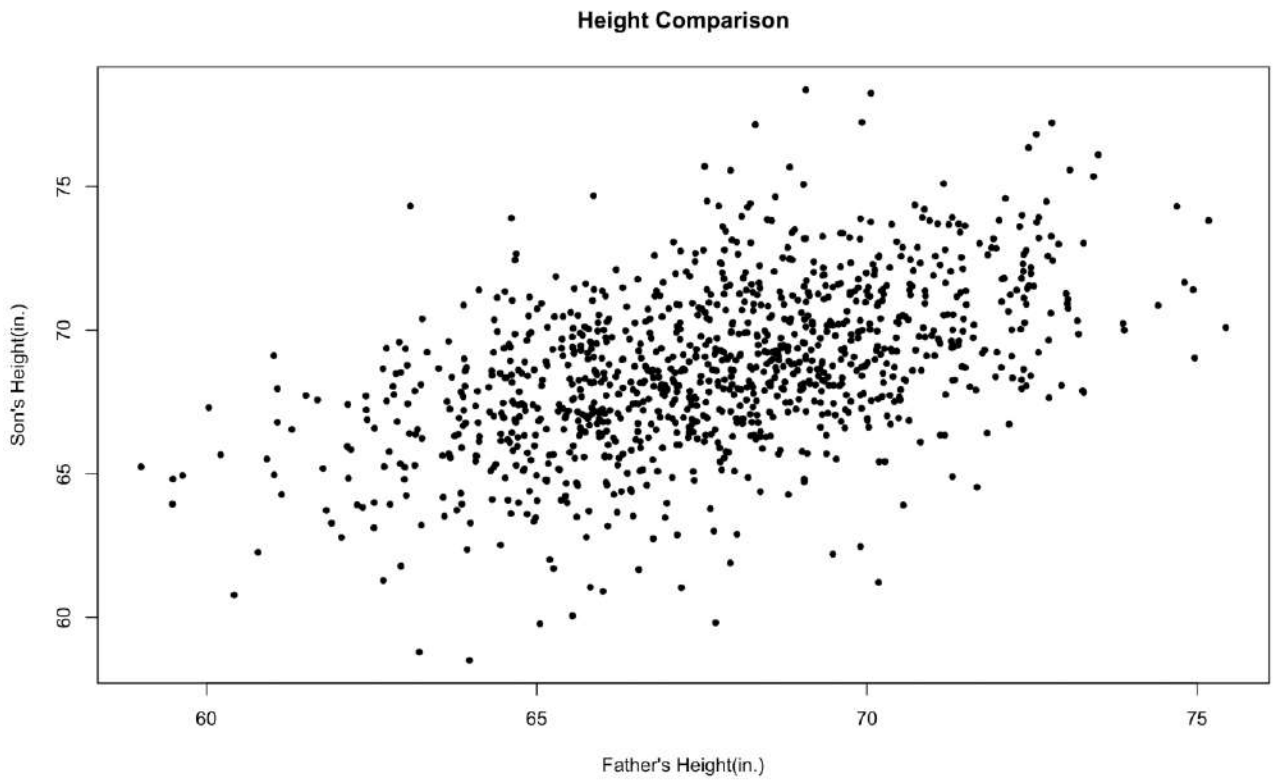
```
R Console

    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)
> fason <- father.son
> glimpse(fason)
Rows: 1,078
Columns: 2
$ fheight <dbl> 65.04851, 63.25094, 64.95532, 65.75250, 61.13723, 63.02254, 65.37053, 64.72398, 66.06509, 66.96738, 59.00800, 62.93203, 63.67063, 64.07386, 64.68851, 65.15466, 66.37353, 65.57704, 67.36765, 66.75929, …
$ sheight <dbl> 59.77827, 63.21404, 63.34242, 62.79238, 64.28113, 64.24221, 64.08231, 63.99574, 64.61338, 63.97944, 65.24451, 65.35102, 65.67992, 65.43664, 65.29391, 64.79017, 65.01881, 65.54640, 65.88145, 65.49008, …
> str(fason)
'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)
  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
> model_fit <- lm(fason$sheight ~ fason$fheight)
> summary(model_fit)

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.

**For creating the scatter plot:**

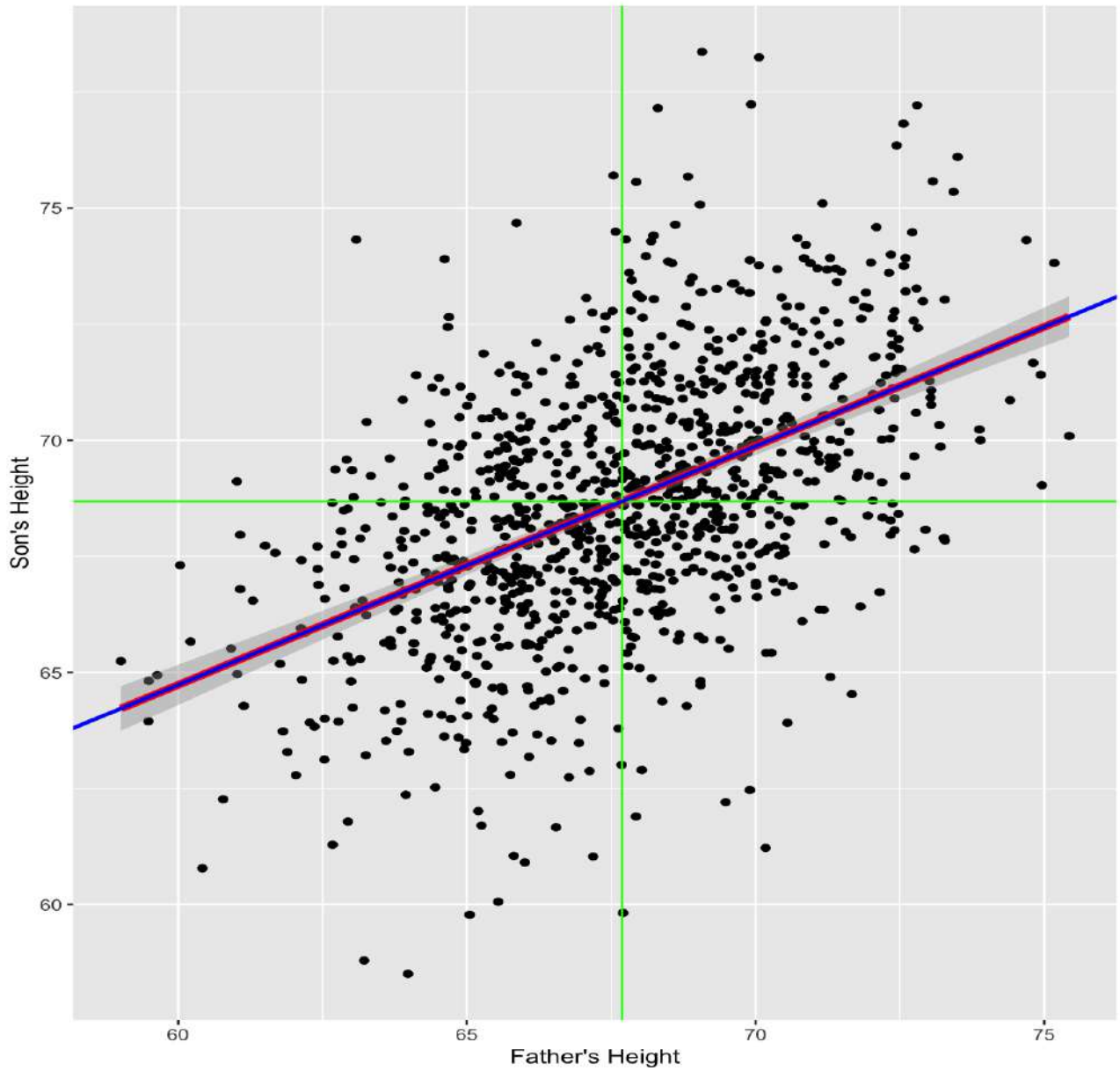
**plot(fason$fheight, fason$sheight, xlab = "Father's Height(in.)", ylab = "Son's Height(in.)", pch = 20) + title("Height Comparison")**




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

```
model_fit <- lm(sheight ~ fheight, data = fason)

> library(ggplot2)

> ggplot(fason, aes(x = fheight, y = sheight)) +

+     geom_point() +  # This line generates the points of the scatter plot

+     geom_smooth(method = "lm", color = "red", size = 2) +  # This line creates the regression line

+      geom_vline(xintercept = mean(fason$fheight), color = "green") +  # Adds a vertical line through the mean of father's height

+     geom_hline(yintercept = mean(fason$sheight), color = "green") +  # Adds a horizontal line through the mean of son's height

+     geom_abline(slope = coef(model_fit)[2], intercept = coef(model_fit)[1], color = "blue", size = 1) +  # Adds the regression line using model coefficients

+     xlab("Father's Height") +  # Adds an x-axis label
```

```
+     ylab("Son's Height")  # Adds a y-axis label
```




The standard deviation line and the regression line are identical, indicating a strong linear relationship. Furthermore, the average heights of fathers and sons are very similar.

---

## Below is the markdown file of my work.

## MTH-522

Jeevan

2024-03-18

## R Markdown

```
library(UsingR)

## Loading required package: MASS

## Loading required package: HistData

## Loading required package: Hmisc

##
## Attaching package: 'Hmisc'

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

data(father.son)
 fheight                                    <-                                    father.son$fheight
 sheight                                    <-                                    father.son$sheight
 summary(fheight)

##               Min.    1st    Qu.        Median               Mean    3rd    Qu.               Max.
##   59.01   65.79   67.77   67.69   69.60   75.43

summary(sheight)

##               Min.    1st    Qu.        Median               Mean    3rd    Qu.               Max.
##   58.51   66.93   68.62   68.68   70.47   78.36

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)
 fason                                          <-                                          father.son
 glimpse (fason)

##                                            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)
```

```
##      '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 ...
```

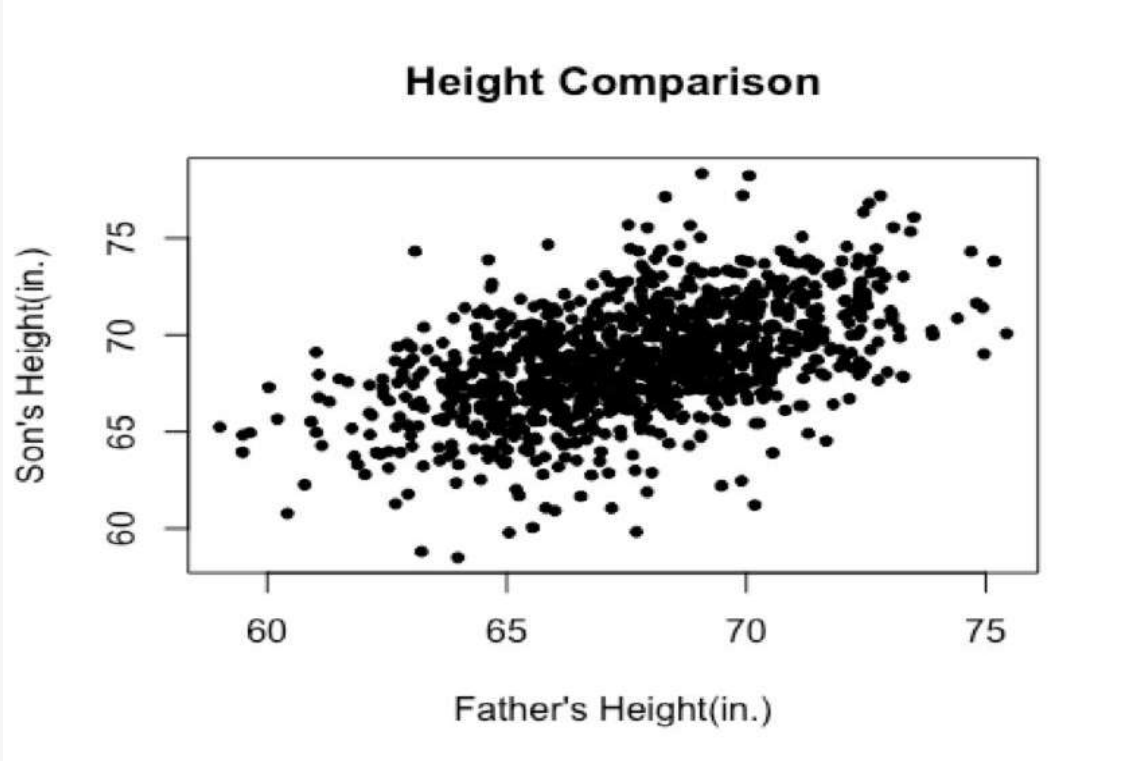
```
head(fason)
```

```
##                                                      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
```

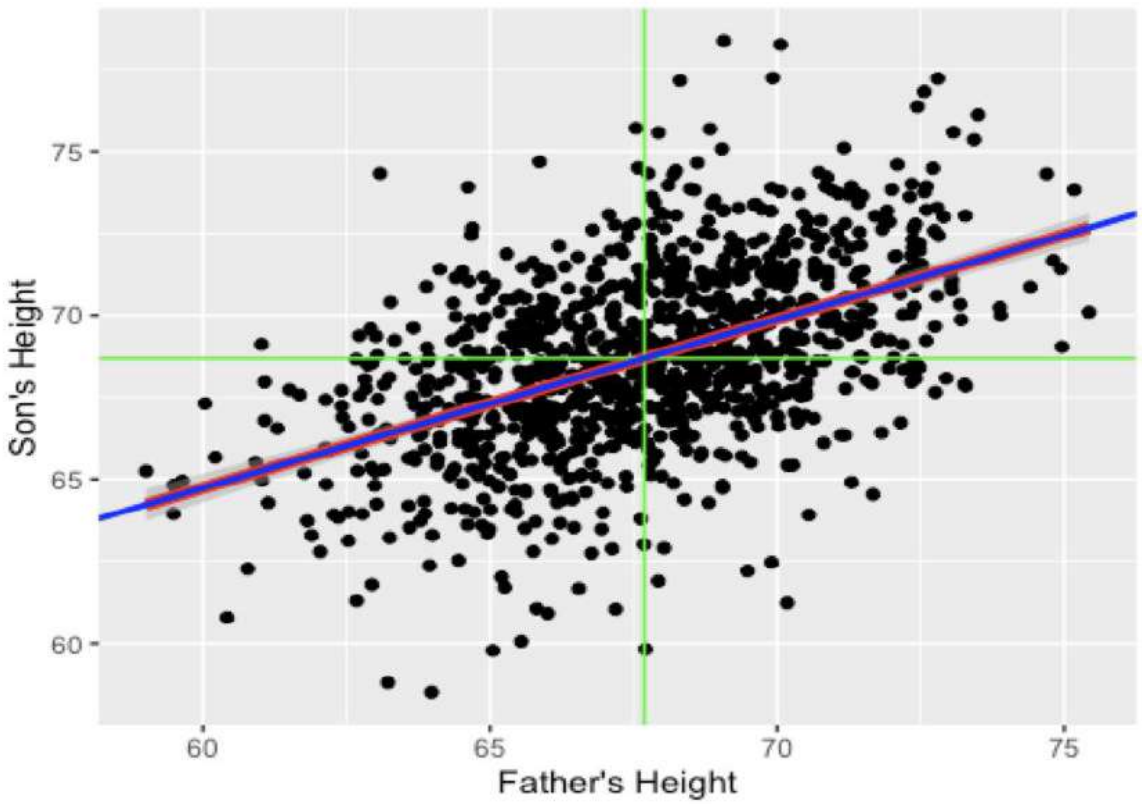
```
model_fit              <-              lm(fason$sheight              ~              fason$fheight)
 summary(model_fit)
```

```
##
##                                                                                          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
```

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




```
## integer(0)

model_fit <- lm(sheight ~ fheight, data = fason)
library(ggplot2)


ggplot(fason, aes(x = fheight, y = sheight)) +
  geom_point() + # This line generates the points of the scatter plot
  geom_smooth(method = "lm", color = "red", size = 2) + # This line creates the regression line
  geom_vline(xintercept = mean(fason$fheight), color = "green") + # Adds a vertical line through the mean of father's height
  geom_hline(yintercept = mean(fason$sheight), color = "green") + # Adds a horizontal line through the mean of son's height
  geom_abline(slope = coef(model_fit)[2], intercept = coef(model_fit)[1], color = "blue", size = 1) + # Adds the regression line using model coefficients
  xlab("Father's Height") + # Adds a x-axis label
  ylab("Son's Height") # Adds a y-axis label

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

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


Son's Height
75
70
65
60
60
65
70
75
Father's Height