Exercise No -6

Title of Experiment: Statistical Analysis Using R.

College: V.P.Institute of Management Studies & Research, Sangli

Solution Provider Name: Pranita Anil Kamble

#R Version: 3.6.3

#R Studio Version: 1.3.1093

Abstract:

Statistical analysis is implemented by using:

1) mean and mode

2) covariance

3) standard deviation

4) variance

5) linear regression

6) prediction

Install Necessary Packages:

1. install.packages("reader")

2. install.packages("modeest")

3. install.packages("ggpubr")

4. install.packages("ggplot2")

Q1) Calculate mean and mode of StudyHours from given table.

Introduction:

1. The **mean** is the average of a data set.

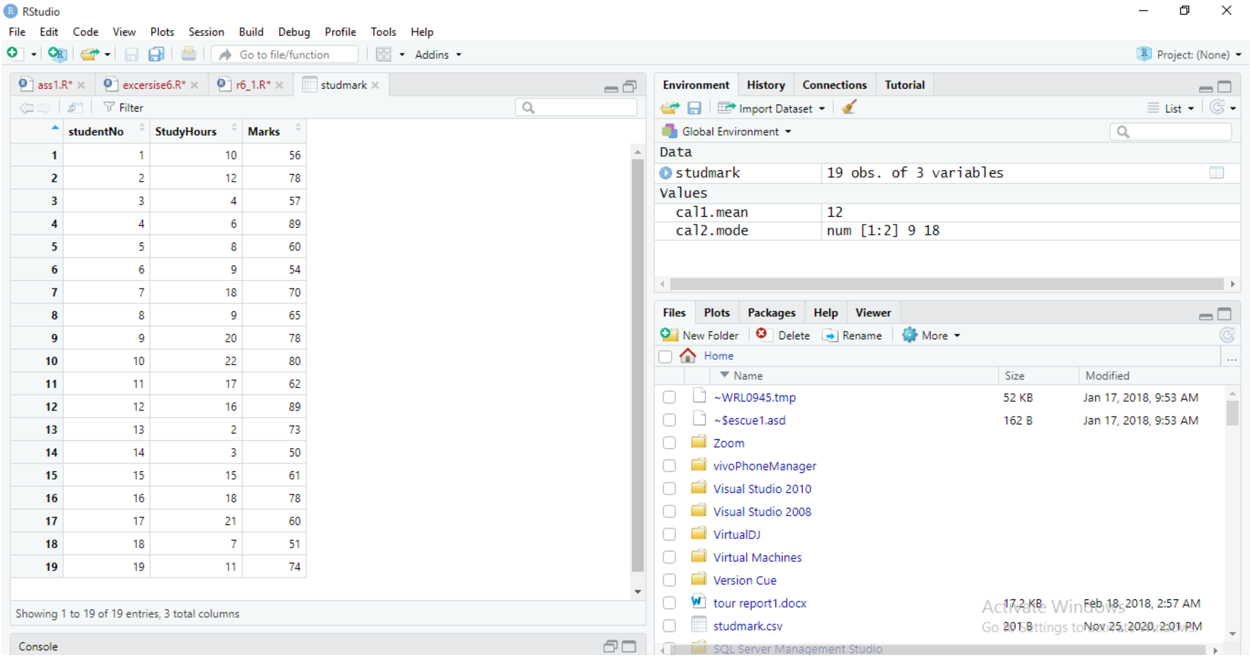
$$\bar{x} = \frac{\sum x}{N}$$

2. The **mode** is the most common number in a data set.

Solution:

```
> #R Version 3.6.3
> #R Studio Version 1.3.1093
> library(readr)
> library(modeest)
> library (ggpubr)
> library(ggplot2)
> rm(list=ls())
> studmark <- read_csv("R/studmark.csv")#to load the data

-- ----------------Column specification ----------------------------------------------
cols(
  studentNo = col_double(),
  StudyHours = col_double(),
  Marks = col_double()
)

> cal1.mean <- mean(studmark$StudyHours)#mean function for Studyhours
> print(cal1.mean)
[1] 12
> cal2.mode = mfv(studmark$StudyHours) #to calculate mode
> print(cal2.mode)
[1]  9 18
```

Screen:

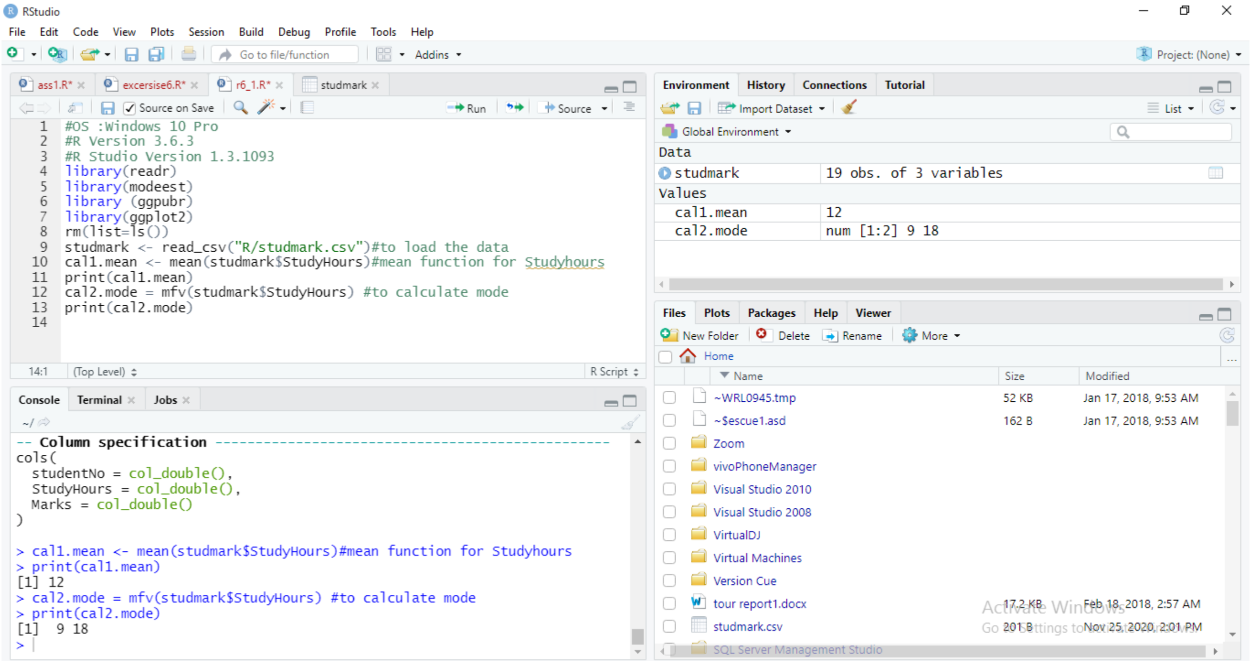


	studentNo	StudyHours	Marks
1	1	10	56
2	2	12	78
3	3	4	57
4	4	6	89
5	5	8	60
6	6	9	54
7	7	18	70
8	8	9	65
9	9	20	78
10	10	22	80
11	11	17	62
12	12	16	89
13	13	2	73
14	14	3	50
15	15	15	61
16	16	18	78
17	17	21	60
18	18	7	51
19	19	11	74

Showing 1 to 19 of 19 entries, 3 total columns

```
Data
studmark        19 obs. of 3 variables
Values
  cal1.mean     12
  cal2.mode     num [1:2] 9 18
```

Name	Size	Modified
~WRL0945.tmp	52 KB	Jan 17, 2018, 9:53 AM
~$escue1.asd	162 B	Jan 17, 2018, 9:53 AM
Zoom		
vivoPhoneManager		
Visual Studio 2010		
Visual Studio 2008		
VirtualDJ		
Virtual Machines		
Version Cue		
tour report1.docx	17.2 KB	Feb 18, 2018, 2:57 AM
studmark.csv	201 B	Nov 25, 2020, 2:01 PM
SQL Server Management Studio		



```r
#OS :Windows 10 Pro
#R Version 3.6.3
#R Studio Version 1.3.1093
library(readr)
library(modeest)
library (ggpubr)
library(ggplot2)
rm(list=ls())
studmark <- read_csv("R/studmark.csv")#to load the data
cal1.mean <- mean(studmark$StudyHours)#mean function for Studyhours
print(cal1.mean)
cal2.mode = mfv(studmark$StudyHours) #to calculate mode
print(cal2.mode)
```

```
-- Column specification --------------------------------------------------------
cols(
  studentNo = col_double(),
  StudyHours = col_double(),
  Marks = col_double()
)

> cal1.mean <- mean(studmark$StudyHours)#mean function for Studyhours
> print(cal1.mean)
[1] 12
> cal2.mode = mfv(studmark$StudyHours) #to calculate mode
> print(cal2.mode)
[1]  9 18
>
```

```
Data
studmark       19 obs. of 3 variables
Values
  cal1.mean    12
  cal2.mode    num [1:2] 9 18
```

Name	Size	Modified
~WRL0945.tmp	52 KB	Jan 17, 2018, 9:53 AM
~$escue1.asd	162 B	Jan 17, 2018, 9:53 AM
Zoom		
vivoPhoneManager		
Visual Studio 2010		
Visual Studio 2008		
VirtualDJ		
Virtual Machines		
Version Cue		
tour report1.docx	17.2 KB	Feb 18, 2018, 2:57 AM
studmark.csv	201 B	Nov 25, 2020, 2:01 PM
SQL Server Management Studio		

Q2) Calculate covariance between StudyHours and Marks.

Explanation:

Covariance is measure of relationship between two variables.

1) Positive Covariance: Tend to move in same direction.
2) Negative Covariance: Tend to move in inverse direction.

The formula is:

Cov(X,Y) = Σ E((X-μ)E(Y-ν)) / n-1 where:

X is a random variable

E(X) = μ is the expected value (the mean) of the random variable X and

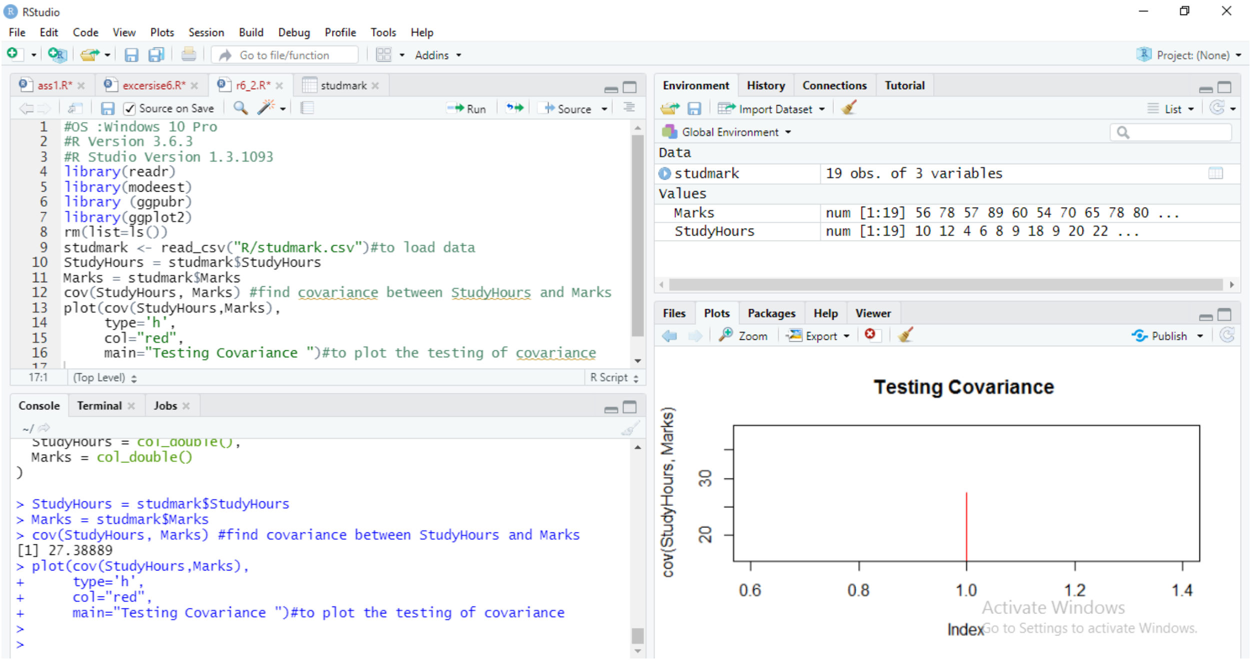
E(Y) = ν is the expected value (the mean) of the random variable Y

n = the number of items in the data set

Solution:

```
> #OS :Windows 10 Pro
> #R Version 3.6.3
> #R Studio Version 1.3.1093
> library(readr)
> library(modeest)
> library (ggpubr)
> library(ggplot2)
> rm(list=ls())
> studmark <- read_csv("R/studmark.csv") #to load data

-- ---------------------Column specification ------------------------------------------------
cols(
  studentNo = col_double(),
  StudyHours = col_double(),
  Marks = col_double()
)

> StudyHours = studmark$StudyHours
> Marks = studmark$Marks
> cov(StudyHours, Marks)  #find covariance between StudyHours and Marks
[1] 27.38889
> plot (cov(StudyHours,Marks),
+       type='h',
+       col="red",
+       main="Testing Covariance ") #to plot the testing of covariance
```

Screen:



```
#OS :Windows 10 Pro
#R Version 3.6.3
#R Studio Version 1.3.1093
library(readr)
library(modeest)
library (ggpubr)
library(ggplot2)
rm(list=ls())
studmark <- read_csv("R/studmark.csv")#to load data
StudyHours = studmark$StudyHours
Marks = studmark$Marks
cov(StudyHours, Marks) #find covariance between StudyHours and Marks
plot(cov(StudyHours,Marks),
      type='h',
      col="red",
      main="Testing Covariance ")#to plot the testing of covariance
```

```
Data
studmark      19 obs. of 3 variables
Values
Marks         num [1:19] 56 78 57 89 60 54 70 65 78 80 ...
StudyHours    num [1:19] 10 12 4 6 8 9 18 9 20 22 ...
```

```
  StudyHours = col_double(),
  Marks = col_double()
)

> StudyHours = studmark$StudyHours
> Marks = studmark$Marks
> cov(StudyHours, Marks) #find covariance between StudyHours and Marks
[1] 27.38889
> plot(cov(StudyHours,Marks),
+       type='h',
+       col="red",
+       main="Testing Covariance ")#to plot the testing of covariance
> 
> 
```

Q3)Calculate Standard Deviation of marks obtained by students.

Explanation: The standard deviation measures the spread of the data about the mean value. It is useful in comparing sets of data which may have the same mean but a different range.

Formula:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ = population standard deviation

N = the size of the population

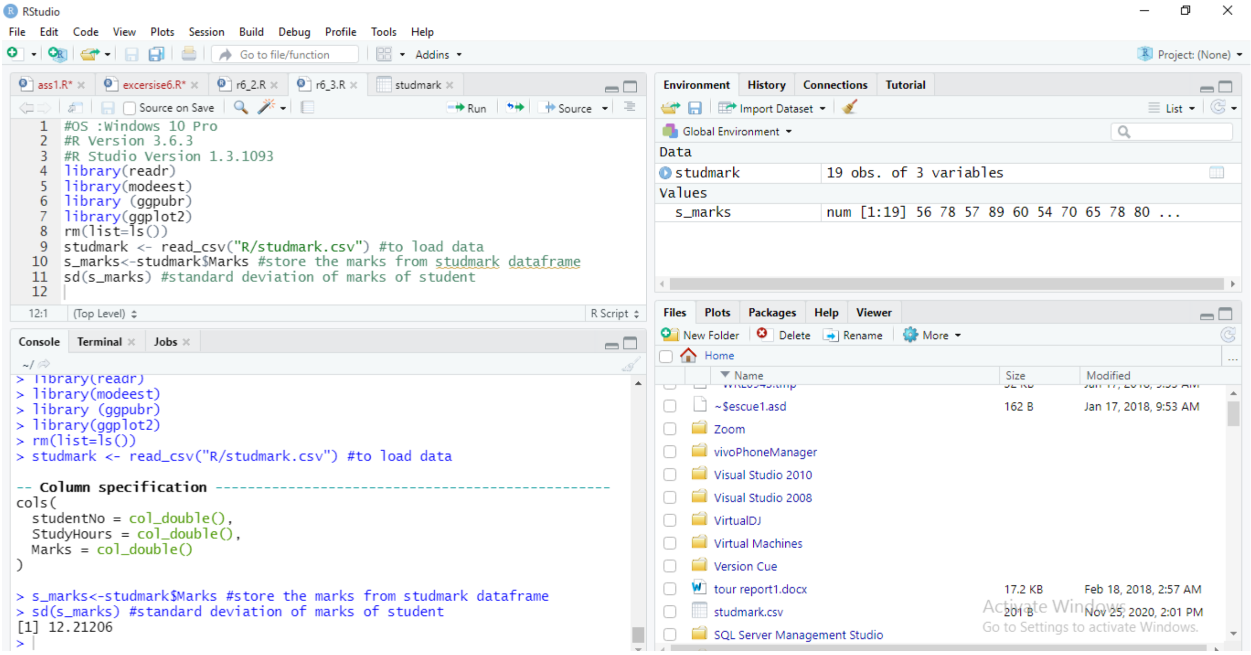
x_i = each value from the population

μ = the population mean

Solution:

```
#OS :Windows 10 Pro
> #R Version 3.6.3
> #R Studio Version 1.3.1093
> library(readr)
> library(modeest)
> library (ggpubr)
> library(ggplot2)
> rm(list=ls())
> studmark <- read_csv("R/studmark.csv")             #to load data

-- Column specification ------------------------------------------------
cols(
  studentNo = col_double(),
  StudyHours = col_double(),
  Marks = col_double()
)

> s_marks<-studmark$Marks                              #store the marks from studmark dataframe
> sd(s_marks)                                          #standard deviation of marks of student
[1] 12.21206
```

Screen:



```
#OS :Windows 10 Pro
#R Version 3.6.3
#R Studio Version 1.3.1093
library(readr)
library(modeest)
library (ggpubr)
library(ggplot2)
rm(list=ls())
studmark <- read_csv("R/studmark.csv") #to load data
s_marks<-studmark$Marks #store the marks from studmark dataframe
sd(s_marks) #standard deviation of marks of student
```

```
> library(readr)
> library(modeest)
> library (ggpubr)
> library(ggplot2)
> rm(list=ls())
> studmark <- read_csv("R/studmark.csv") #to load data

-- Column specification --------------------------------------------
cols(
  studentNo = col_double(),
  StudyHours = col_double(),
  Marks = col_double()
)

> s_marks<-studmark$Marks #store the marks from studmark dataframe
> sd(s_marks) #standard deviation of marks of student
[1] 12.21206
> 
```

```
Data
studmark      19 obs. of 3 variables
Values
  s_marks     num [1:19] 56 78 57 89 60 54 70 65 78 80 ...
```

Name	Size	Modified
~$escue1.asd	162 B	Jan 17, 2018, 9:53 AM
Zoom		
vivoPhoneManager		
Visual Studio 2010		
Visual Studio 2008		
VirtualDJ		
Virtual Machines		
Version Cue		
tour report1.docx	17.2 KB	Feb 18, 2018, 2:57 AM
studmark.csv	201 B	Nov 25, 2020, 2:01 PM
SQL Server Management Studio		

Q4) Calculate variance of StudyHours from above data.

Explanation: Variance is used to measure of how data is spread in the dataset. It is calculated as the average squared deviation of each number from the mean of a data set.

Formula:

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

S^2 = sample variance

x_i = the value of the one observation

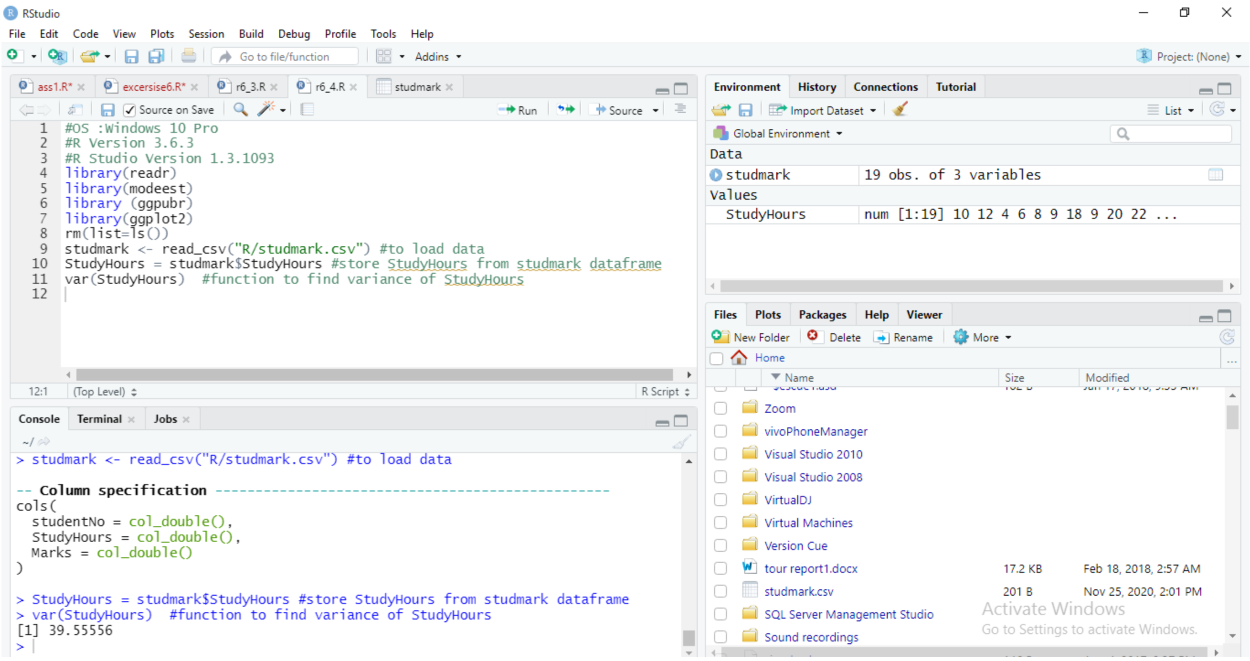
$\bar{x}$ = the mean value of all observations

n = the number of observations

Solution:

```
> #OS :Windows 10 Pro
> #R Version 3.6.3
> #R Studio Version 1.3.1093
> library(readr)
> library(modeest)
> library (ggpubr)
> library(ggplot2)
> rm(list=ls())
> studmark <- read_csv("R/studmark.csv") #to load data

----------- Column specification ------------------------------------------------
cols(
  studentNo = col_double(),
  StudyHours = col_double(),
  Marks = col_double()
)
> StudyHours = studmark$StudyHours                    #store StudyHours from studmark dataframe
> var(StudyHours)                                      #function to find variance of StudyHours
[1] 39.55556
```

Screen:



```
#OS :Windows 10 Pro
#R Version 3.6.3
#R Studio Version 1.3.1093
library(readr)
library(modeest)
library (ggpubr)
library(ggplot2)
rm(list=ls())
studmark <- read_csv("R/studmark.csv") #to load data
StudyHours = studmark$StudyHours #store StudyHours from studmark dataframe
var(StudyHours)  #function to find variance of StudyHours
```

```
Data
studmark      19 obs. of 3 variables
Values
StudyHours    num [1:19] 10 12 4 6 8 9 18 9 20 22 ...
```

Name	Size	Modified
Zoom		
vivoPhoneManager		
Visual Studio 2010		
Visual Studio 2008		
VirtualDJ		
Virtual Machines		
Version Cue		
tour report1.docx	17.2 KB	Feb 18, 2018, 2:57 AM
studmark.csv	201 B	Nov 25, 2020, 2:01 PM
SQL Server Management Studio		
Sound recordings		

```
> studmark <- read_csv("R/studmark.csv") #to load data

-- Column specification --------------------------------------------------------
cols(
  studentNo = col_double(),
  StudyHours = col_double(),
  Marks = col_double()
)

> StudyHours = studmark$StudyHours #store StudyHours from studmark dataframe
> var(StudyHours)  #function to find variance of StudyHours
[1] 39.55556
> 
```

Q5) Determine the relation between StudyHours and marks obtained by students using linear regression.

Explanation: Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

The formula for a regression line is

Y' = bX + A

where Y' is the predicted score, b is the slope of the line, and A is the Y intercept.

A non-parametric procedure, due to Spearman, is to replace the observations by their ranks in the calculation of the correlation coefficient.

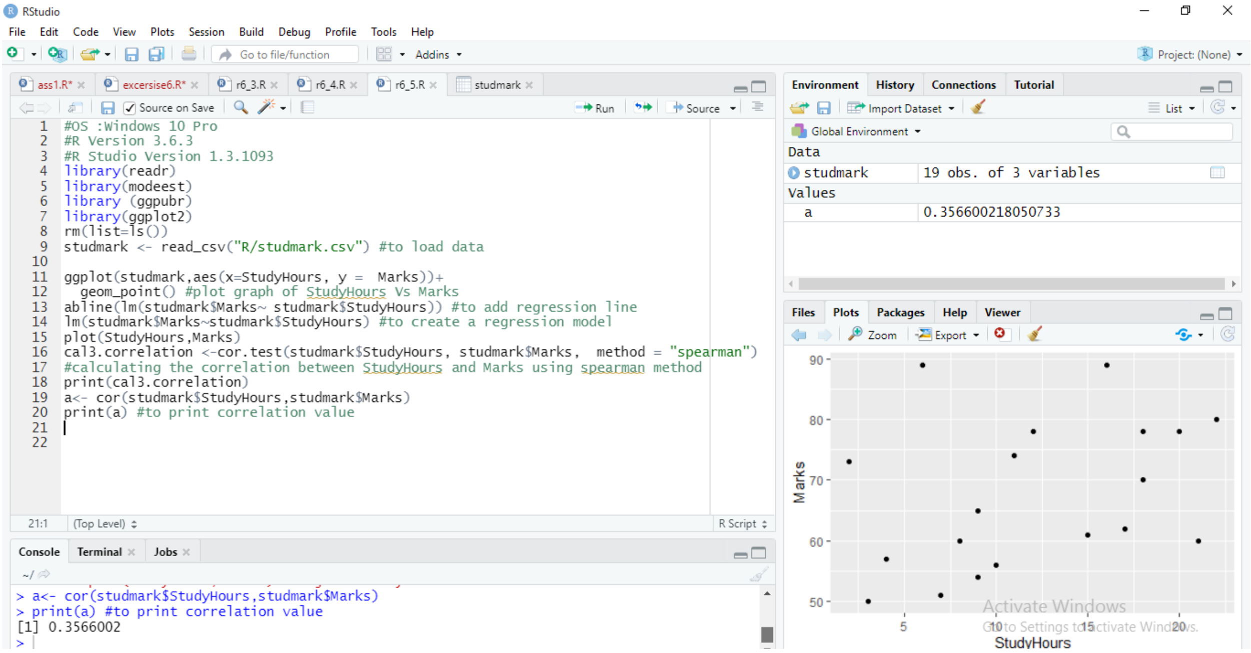
$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

Solution:

#OS :Windows 10 Pro
#R Version 3.6.3
#R Studio Version 1.3.1093
>library(readr)
>library(modeest)
>library (ggpubr)
>library(ggplot2)

```
>rm(list=ls())
>studmark <- read_csv("R/studmark.csv") #to load data

>ggplot(studmark,aes(x=StudyHours, y =  Marks))+
  >geom_point() #plot graph of StudyHours Vs Marks
>abline(lm(studmark$Marks~ studmark$StudyHours)) #to add regression line
>lm(studmark$Marks~studmark$StudyHours) #to create a regression model
>plot(StudyHours,Marks)
>cal3.correlation <-cor.test(studmark$StudyHours, studmark$Marks,  method = "spearman")
#calculating the correlation between StudyHours and Marks using spearman method
>print(cal3.correlation)
>a<- cor(studmark$StudyHours,studmark$Marks)
>print(a) #to print correlation value
>[1] 0.3566002
```

Screen:



```
#OS :Windows 10 Pro
#R Version 3.6.3
#R Studio Version 1.3.1093
library(readr)
library(modeest)
library (ggpubr)
library(ggplot2)
rm(list=ls())
studmark <- read_csv("R/studmark.csv") #to load data

ggplot(studmark,aes(x=StudyHours, y =  Marks))+
  geom_point() #plot graph of StudyHours Vs Marks
abline(lm(studmark$Marks~ studmark$StudyHours)) #to add regression line
lm(studmark$Marks~studmark$StudyHours) #to create a regression model
plot(StudyHours,Marks)
cal3.correlation <-cor.test(studmark$StudyHours, studmark$Marks,  method = "spearman")
#calculating the correlation between StudyHours and Marks using spearman method
print(cal3.correlation)
a<- cor(studmark$StudyHours,studmark$Marks)
print(a) #to print correlation value
```

```
> a<- cor(studmark$StudyHours,studmark$Marks)
> print(a) #to print correlation value
[1] 0.3566002
>
```

```
Data
studmark    19 obs. of 3 variables
Values
  a         0.356600218050733
```

Q6) Predict the marks of students who study for 14 hours in a week.

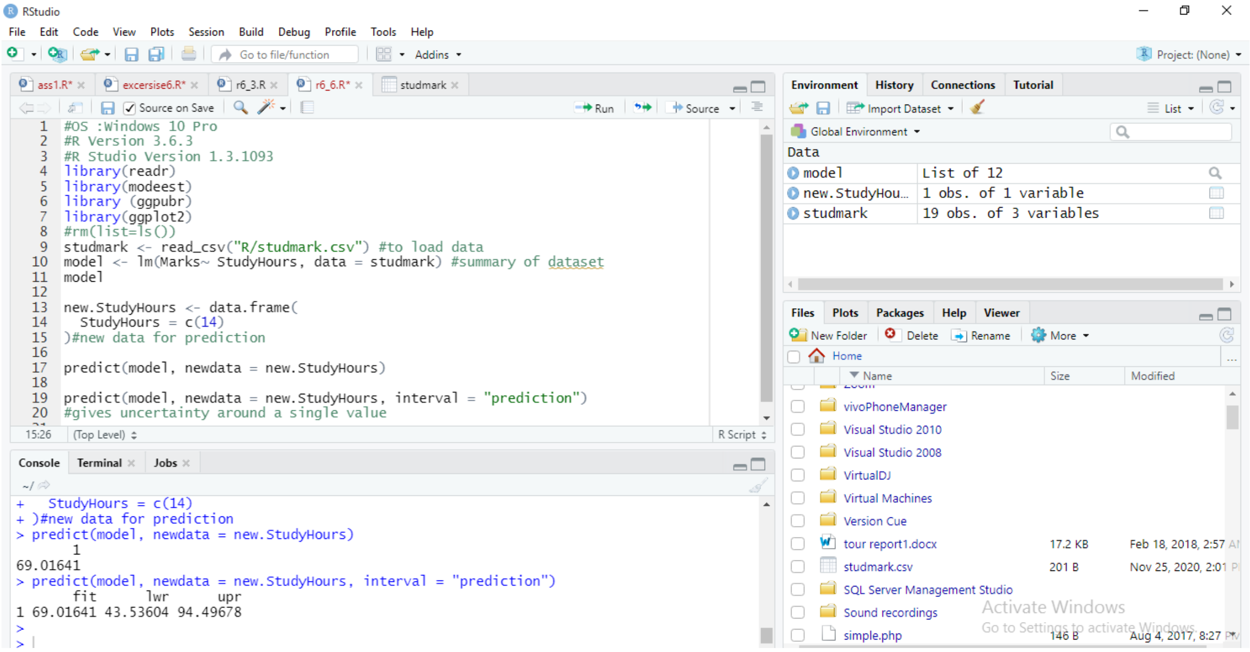
Explanation:

Predictions are precise when the observed values cluster close to the predicted values. Regression predictions are for the mean of the dependent variable. If you think of any mean, you know that there is variation around that mean. The same applies to the predicted mean of the dependent variable.

Solution:

```
> #OS :Windows 10 Pro
> #R Version 3.6.3
> #R Studio Version 1.3.1093
> library(readr)
> library(modeest)
> library (ggpubr)
> library(ggplot2)
> #rm(list=ls())
> studmark <- read_csv("R/studmark.csv") #to load data

-- Column specification --------------------------------------------------------
cols(
  studentNo = col_double(),
  StudyHours = col_double(),
  Marks = col_double()
)

> # Load the data
> #data("studmark", package = "dataframes")
> # Build the model
> model <- lm(Marks~ StudyHours, data = studmark)#summary of dataset
> model

Call:
lm(formula = Marks ~ StudyHours, data = studmark)

Coefficients:
(Intercept)   StudyHours
    59.3226       0.6924

> new.StudyHours <- data.frame(
+   StudyHours = c(14)
+ )#new data for prediction
> predict(model, newdata = new.StudyHours)
       1
69.01641
> predict(model, newdata = new.StudyHours, interval = "prediction")
       fit      lwr      upr
1 69.01641 43.53604 94.49678
>
```

Screen:


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
ass1.R* excersise6.R* r6_3.R r6_6.R* studmark
Source on Save
Run
Source
15:26
(Top Level)
R Script
Console Terminal Jobs
~/
Environment History Connections Tutorial
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home
Activate Windows
Go to Settings to activate Windows.

```r
#OS :Windows 10 Pro
#R Version 3.6.3
#R Studio Version 1.3.1093
library(readr)
library(modeest)
library (ggpubr)
library(ggplot2)
#rm(list=ls())
studmark <- read_csv("R/studmark.csv") #to load data
model <- lm(Marks~ StudyHours, data = studmark) #summary of dataset
model

new.StudyHours <- data.frame(
  StudyHours = c(14)
)#new data for prediction

predict(model, newdata = new.StudyHours)

predict(model, newdata = new.StudyHours, interval = "prediction")
#gives uncertainty around a single value
```

```
+   StudyHours = c(14)
+ )#new data for prediction
> predict(model, newdata = new.StudyHours)
       1
69.01641
> predict(model, newdata = new.StudyHours, interval = "prediction")
       fit      lwr      upr
1 69.01641 43.53604 94.49678
>
> |
```

Data	
model	List of 12
new.StudyHou…	1 obs. of 1 variable
studmark	19 obs. of 3 variables

Name	Size	Modified
vivoPhoneManager		
Visual Studio 2010		
Visual Studio 2008		
VirtualDJ		
Virtual Machines		
Version Cue		
tour report1.docx	17.2 KB	Feb 18, 2018, 2:57 AM
studmark.csv	201 B	Nov 25, 2020, 2:01 PM
SQL Server Management Studio		
Sound recordings		
simple.php	146 B	Aug 4, 2017, 8:27 PM

XXXXXXXXXXXXXXXXXX