

## Homework 3

Your Name: TA sample solutions

Student ID:

Using Case Study 1 data, Case1\_Student Grades\_Large.csv

Using student demographic info and learning behaviors (weekly hrs in different categories) to predict “Grade”

Note: exclude nominal variables, and student performance variables (e.g., Exam) from the list of x variables

Using hold-out evaluation only, 80% as training

Use feature selection to build multiple models, and compare the models based on RMSE

- Backward method using p-value in t-test as metric
- Backward method using AIC as metric
- Forward method using AIC as metric
- Stepwise method using ACI as metric
- Best subset method using Adj-R2 as metric

Q1 Show the R coding, outputs, and your explanations for each step in linear regression.

### ANSWERS:

#### Importing the data into R- studio:

- First, we need to import the dataset into R-studio. Here we used “Case1\_Student Grades\_Large.csv”.

Here is the screenshot of R coding: (loading dataset into R-studio)

The screenshot shows the RStudio interface. In the top-left, the title bar says "RStudio" and "File Edit Code View Plots Session Build Debug Profile Tools Help". Below it is a toolbar with icons for file operations like Open, Save, and Run. The main workspace is titled "HW\_2\_RStudio.R" and contains the following R code:

```
> data <- read.table(file = "C:/Users/satya/OneDrive/Desktop/Case1_Student_Grades_Large.csv", header = T, sep = ",")  
> str(data)  
#> data frame': 10000 obs. of 12 variables:  
#> $ ID : int 1 2 3 4 5 6 7 8 9 10 ...  
#> $ Nationality : chr "India" "India" "India" "India" ...  
#> $ Gender : int 0 0 0 1 1 1 1 1 1 ...  
#> $ Age : int 25 24 26 23 23 18 22 19 25 18 ...  
#> $ Degree : chr "BS" "BS" "BS" "BS" ...  
#> $ Hours.on.Readings : int 14 14 14 14 12 13 13 13 ...  
#> $ Hours.on.Assignments : int 2 2 2 2 2 1 0 0 0 ...  
#> $ Hours.on.Games : int 14 14 14 14 2 7 13 13 13 13 ...  
#> $ Hours.on.Internet : int 6 6 6 7 4 3 3 3 3 ...  
#> $ Exam : num 43.7 62.45 48.9 80.4 ...  
#> $ Grade : num 51.7 72.2 54.4 57.7 88.4 ...  
#> $ GradeLetter : chr "F" "C" "E" "F" ...
```

The "Environment" tab in the top-right pane shows a variable named "data" with the description "10000 obs. of 12 variables". The bottom-right pane has tabs for "Files", "Plots", "Packages", "Help", "Viewer", and "Presentation".

## Step 1: Understand the data & well define independent variables & dependent variables:

- Here, X variables represents the predictor variables, and Y variable represents the target variable.
  - Y variable (target variable/ Dependent variable): Grade
  - X variables (Predictors/ Factors/Independent variables): ID, Nationality, Gender, Age, Degree, Hours on Readings, Hours on Assignments, Hours on Games, Hours on Internet, Exam, Grade Letter

## Excluding nominal & student performance variables (Exam) from the list of X-variables to perform linear regression:

- As per the requirement in the question, we need to exclude nominal variables, and student performance variables (e.g., Exam) from the list of x variables to perform linear regression.
- So, the variables need to be excluded from the list of x-variables are:
  - ID
  - Nationality
  - Gender
  - Degree
  - Exam
  - GradeLetter
- So, the final X-variable list contains the following variables:
  - Age
  - Hours on Readings
  - Hours on Assignments
  - Hours on Games
  - Hours on Internet

Here is the screenshot of R coding: (excluding nominal & stud perform var's)

The screenshot shows the RStudio interface with the following details:

- Console:** Displays R code and its output. The code includes reading data, printing its structure, and filtering it to remove columns for ID, Nationality, Gender, Degree, and Exam. The last printed object is a data frame with 6 variables: Age, Hours.on.Readings, Hours.on.Assignments, Hours.on.Games, Hours.on.Internet, and Grade.
- Environment:** Shows the global environment with a data object containing 10000 observations and 6 variables.
- Data View:** Shows the data frame with the filtered 6 variables.

```
R 4.3.0 -- "/~/.R/4.3.0/bin/R" --slave
> str(data)
'data.frame': 10000 obs. of 12 variables:
$ ID          : int  1 2 3 4 5 6 7 8 9 10 ...
$ Nationality : chr "India" "India" "India" "India" ...
$ Gender      : int  0 0 0 1 1 1 1 1 1 ...
$ Age         : int  25 24 26 21 23 18 22 19 25 18 ...
$ Degree      : chr "B5" "B5" "B5" "B5" ...
$ Hours.on.Readings : int 14 14 14 14 14 12 13 13 13 13 ...
$ Hours.on.Assignments: int 2 2 2 2 2 1 0 0 0 0 ...
$ Hours.on.Games   : int 14 14 14 14 2 7 13 13 13 13 ...
$ Hours.on.Internet : int 6 6 6 6 7 4 3 3 3 3 ...
$ Exam        : num 43.7 62.45 48.9 80.4 ...
$ Grade       : num 51.7 72.2 54.4 57.7 88.4 ...
$ GradeLetter : chr "P" "E" "E" "P" ...
> data = data[, !(names(data) %in% c("ID","Nationality","Gender","Degree", "Exam","GradeLetter"))]
> str(data)
'data.frame': 10000 obs. of 6 variables:
$ Age         : int 25 24 26 21 23 18 22 19 25 18 ...
$ Hours.on.Readings : int 14 14 14 14 14 12 13 13 13 13 ...
$ Hours.on.Assignments: int 2 2 2 2 2 1 0 0 0 0 ...
$ Hours.on.Games   : int 14 14 14 14 2 7 13 13 13 13 ...
$ Hours.on.Internet : int 6 6 6 6 7 4 3 3 3 3 ...
$ Grade       : num 51.7 72.2 54.4 57.7 88.4 ...
```

## Storing X and Y variables:

Here is the screenshot of R coding

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Addins. The main area has tabs for Untitled1 (Source on Save), Run, Source, and Environment. The Environment tab is active, showing the Global Environment and Data pane. The Data pane displays a table of variables:

	data	10000 obs. of 6 variables
Age	int [1:10000]	25 24 26 21 23 18 22 19 25...
Grade	num [1:10000]	51.7 72.2 54.4 57.7 88.4 ...
Hours.on.Assig...	int [1:10000]	2 2 2 2 2 1 0 0 0 0 ...
Hours.on.Games	int [1:10000]	14 14 14 14 2 7 13 13 13 ...
Hours.on.Inter...	int [1:10000]	6 6 6 7 4 3 3 3 3 ...
Hours.on.Readi...	int [1:10000]	14 14 14 14 12 13 13 13 ...

The code in the Source tab is as follows:

```
R 4.3.0 -/
header = T, sep = ",")
> str(data)
'data.frame': 10000 obs. of 12 variables:
 $ ID           : int 1 2 3 4 5 6 7 8 9 10 ...
 $ Nationality : chr "India" "India" "India" "India" ...
 $ Gender       : int 0 0 0 1 1 1 1 1 ...
 $ Age          : int 25 24 26 21 23 18 22 19 25 18 ...
 $ Degree       : chr "BS" "BS" "BS" "BS" ...
 $ Hours.on.Readings: int 14 14 14 14 14 12 13 13 13 ...
 $ Hours.on.Assignments: int 2 2 2 2 2 1 0 0 0 ...
 $ Hours.on.Games: int 14 14 14 14 2 7 13 13 13 ...
 $ Hours.on.Internet: int 6 6 6 7 4 3 3 3 ...
 $ Exam         : num 43.7 62 45 48.9 80.4 ...
 $ Grade        : num 51.7 72.2 54.4 57.7 88.4 ...
 $ GradeLetter  : chr "F" "C" "F" "F" ...
> data = data[, !(names(data) %in% c("ID", "Nationality", "Gender", "Degree", "GradeLetter", "Exam"))]
> str(data)
'data.frame': 10000 obs. of 6 variables:
 $ Age          : int 25 24 26 21 23 18 22 19 25 18 ...
 $ Hours.on.Readings: int 14 14 14 14 14 12 13 13 13 ...
 $ Hours.on.Assignments: int 2 2 2 2 2 1 0 0 0 ...
 $ Hours.on.Games: int 14 14 14 14 2 7 13 13 13 ...
 $ Hours.on.Internet: int 6 6 6 7 4 3 3 3 ...
 $ Grade        : num 51.7 72.2 54.4 57.7 88.4 ...
> Age = data$Age
> Hours.on.Readings = data$Hours.on.Readings
> Hours.on.Assignments = data$Hours.on.Assignments
> Hours.on.Games = data$Hours.on.Games
> Hours.on.Internet = data$Hours.on.Internet
> Grade = data$Grade
> |
```

## Step 2: Examine the linear relationship between x and y variables.

- We can examine linearity by 2 methods.
  - Produce a scatter plot for each x & y variable / produce a single plot with every pair of variables.
  - Calculate correlation values for all the x variables with y variable.
- As the plot method is not clear & reliable all the time, we can perform 2<sup>nd</sup> method and calculate the correlation values.

Here is the screenshot of R coding: (calculating correlation values)

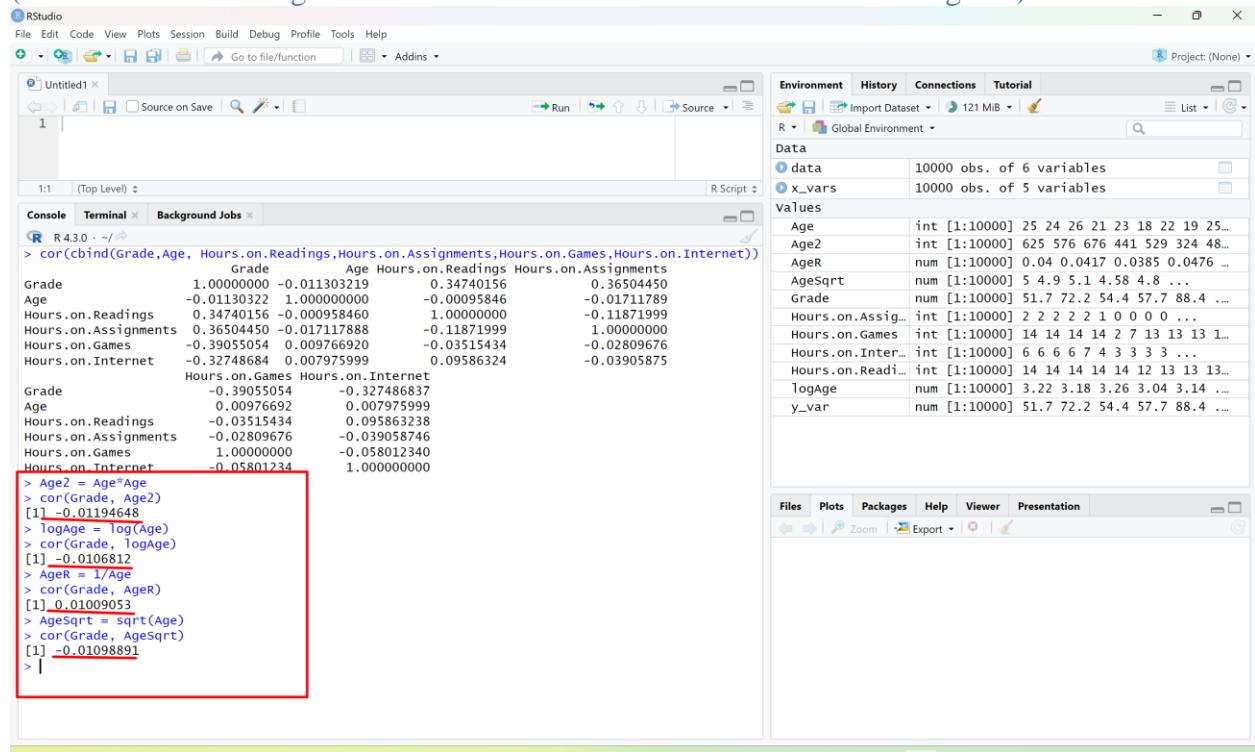
The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Addins. The main area has tabs for Untitled1 (data), Run, Source, and Environment. The Environment tab is active, showing the Global Environment and Data pane. The Data pane displays tables for data, test.data, and train.data. The code in the Source tab is as follows:

```
R 4.3.0 -/
> cor(cbind(Grade, Age, Hours.on.Readings, Hours.on.Assignments, Hours.on.Games, Hours.on.Internet))
   Grade    Age Hours.on.Readings Hours.on.Assignments Hours.on.Games Hours.on.Internet
Grade 1.00000000 -0.01130322 1.000000000 -0.00095846 -0.01711789
Age -0.01130322 1.000000000 -0.00095846 -0.01711789
Hours.on.Readings 0.34740156 -0.00095846 1.000000000 -0.11871999
Hours.on.Assignments 0.36504450 -0.00095846 0.365044500 -0.11871999
Hours.on.Games 0.355364 -0.00976020 0.355364 -0.02809676
Hours.on.Internet -0.32748684 0.007975999 0.05986324 -0.03905875
   Hours.on.Games Hours.on.Internet
Grade -0.39055054 -0.327486837
Age 0.00976692 0.009759999
Hours.on.Readings -0.031314 0.039058738
Hours.on.Assignments -0.02809676 -0.039058746
Hours.on.Games 1.00000000 -0.058012340
Hours.on.Internet -0.05801234 1.000000000
```

- We noticed that the correlation values are as follows:
  - Grade & Hours on readings = **0.34740156** - weak Correlation - +ve correlation
  - Grade & Hours on Assignments = **0.36504450** - weak Correlation - +ve correlation
  - Grade & Hours on games = **-0.39055054** - weak Correlation - -ve correlation
  - Grade & Hours on internet = **-0.32748684** - weak Correlation - -ve correlation
  - Grade & Age = **-0.01130322** – no Correlation – Try transformation to improve correlation.
- As Age has no correlation with Grade(Y-variable), we can try transformation on this Age (x-variable) and re-calculate the correlation with Grade (y-variable) again.
  - Square transformation:  $X' = X * X$
  - Log transformation:  $X' = \log X$
  - Inversion transformation:  $X' = 1/X$
  - Square root transformation:  $X' = \sqrt{X}$

Here is the screenshots of R coding:

(Transformation on Age variable & re-correlation value calculation with grade)



The screenshot shows the RStudio interface with the following details:

- Console:**

```
> cor(cbind(Grade,Age, Hours.on.Readings,Hours.on.Assignments,Hours.on.Games,Hours.on.Internet))
   Grade          Age Hours.on.Readings Hours.on.Assignments
Grade 1.00000000 -0.011303219  0.34740156  0.36504450
Age -0.01130322  1.000000000 -0.00095846 -0.01711789
Hours.on.Readings 0.34740156 -0.000958460  1.00000000 -0.11871999
Hours.on.Assignments 0.36504450 -0.017117888 -0.11871999  1.00000000
Hours.on.Games -0.39055054  0.009766920 -0.03515434 -0.02809676
Hours.on.Internet -0.32748684  0.007975999  0.09586324 -0.03905875
   Hours.on.Games Hours.on.Internet
Grade -0.39055054 -0.327486837
Age 0.00976692  0.007975999
Hours.on.Readings -0.03515434  0.095863238
Hours.on.Assignments -0.02809676 -0.039058746
Hours.on.Games 1.00000000 -0.058012340
Hours.on.Internet -0.05801234  1.000000000
```

```
> Age2 = Age*Age
> cor(Grade, Age2)
[1] -0.01194648
> logAge = log(Age)
> cor(Grade, logAge)
[1] -0.0106812
> AgeR = 1/Age
> cor(Grade, AgeR)
[1] -0.01009053
> Agesqrt = sqrt(Age)
> cor(Grade, Agesqrt)
[1] -0.01098891
> |
```
- Environment:** Shows the global environment with variables like `data`, `x_vars`, `Age`, `Age2`, `AgeR`, `Agesqrt`, `Grade`, `Hours.on.Assignments`, `Hours.on.Games`, `Hours.on.Internet`, `Hours.on.Readings`, `logAge`, and `y_var`.

- As the correlation didn't improve after transformation on 'Age' variable, we decided to drop the age variable.

Here Is the screenshot of r coding: (dropping the age variable)

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and a Go to file/function search bar. The bottom menu bar includes Console, Terminal, and Background Jobs. The main area has tabs for Source on Save, Run, Source, and R Script. The Environment tab is active, showing a global environment with variables like data, x\_vars, Age, Age2, AgeR, Agesqrt, Grade, Hours.on.Assignments, Hours.on.Games, Hours.on.Internet, Hours.on.Readings, LogAge, and y\_var. The Data tab shows a data frame with 10000 observations and 5 variables. The Values tab displays the first few rows of the data frame. The bottom navigation bar includes Files, Plots, Packages, Help, Viewer, and Presentation.

```

R 4.3.0: ~/r/
Hours.on.Readings  0.34741156 -0.000958460  1.000000000 -0.11871999
Hours.on.Assignments  0.36504450 -0.017117888 -0.11871999  1.000000000
Hours.on.Games     -0.39055054  0.009766920 -0.03515434 -0.02809676
Hours.on.Internet   -0.32748684  0.007975999  0.09586324 -0.03905875
Hours.on.Games.Hours.on.Internet
Grade              -0.39055054 -0.327486837
Age                0.00976692  0.007975999
Hours.on.Readings  -0.03515434  0.095863238
Hours.on.Assignments -0.02809676 -0.039058746
Hours.on.Games     1.000000000 -0.058012340
Hours.on.Internet   -0.05801234  1.000000000
> Age2 = Age*Age
> cor(Grade, Age2)
[1] -0.01194648
> logAge = log(Age)
> cor(Grade, logAge)
[1] -0.0106812
> AgeR = 1/Age
> cor(Grade, AgeR)
[1] 0.01009053
> Agesqrt = sqrt(Age)
> cor(Grade, Agesqrt)
[1] -0.01098891
> data = data[, !(names(data) %in% c("Age"))]
> str(data)
'data.frame': 10000 obs. of 5 variables:
$ Hours.on.Readings : int 14 14 14 14 14 12 13 13 13 ...
$ Hours.on.Assignments: int 2 2 2 2 2 1 0 0 0 ...
$ Hours.on.Games    : int 14 14 14 14 2 7 13 13 13 ...
$ Hours.on.Internet : int 6 6 6 6 7 4 3 3 3 ...
$ Grade            : num 51.7 72.2 54.4 57.7 88.4 ...

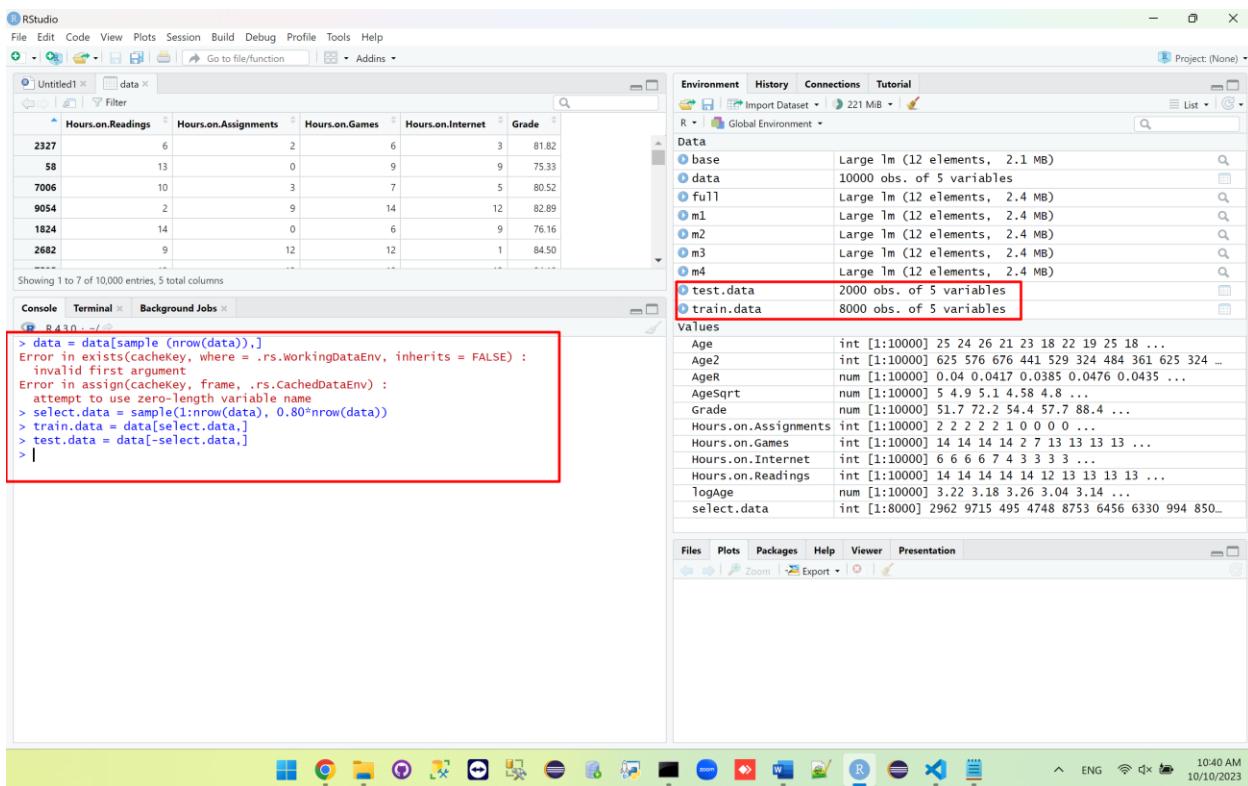
```

- Now the list of X-variables (Age is removed) (5 variables)
  - Hours.on.Readings
  - Hours.on.Assignments
  - Hours.on.Games
  - Hours.on.Internet
- Y-variable:
  - Grade

### Step 3: Decision on evaluation strategy and data splits:

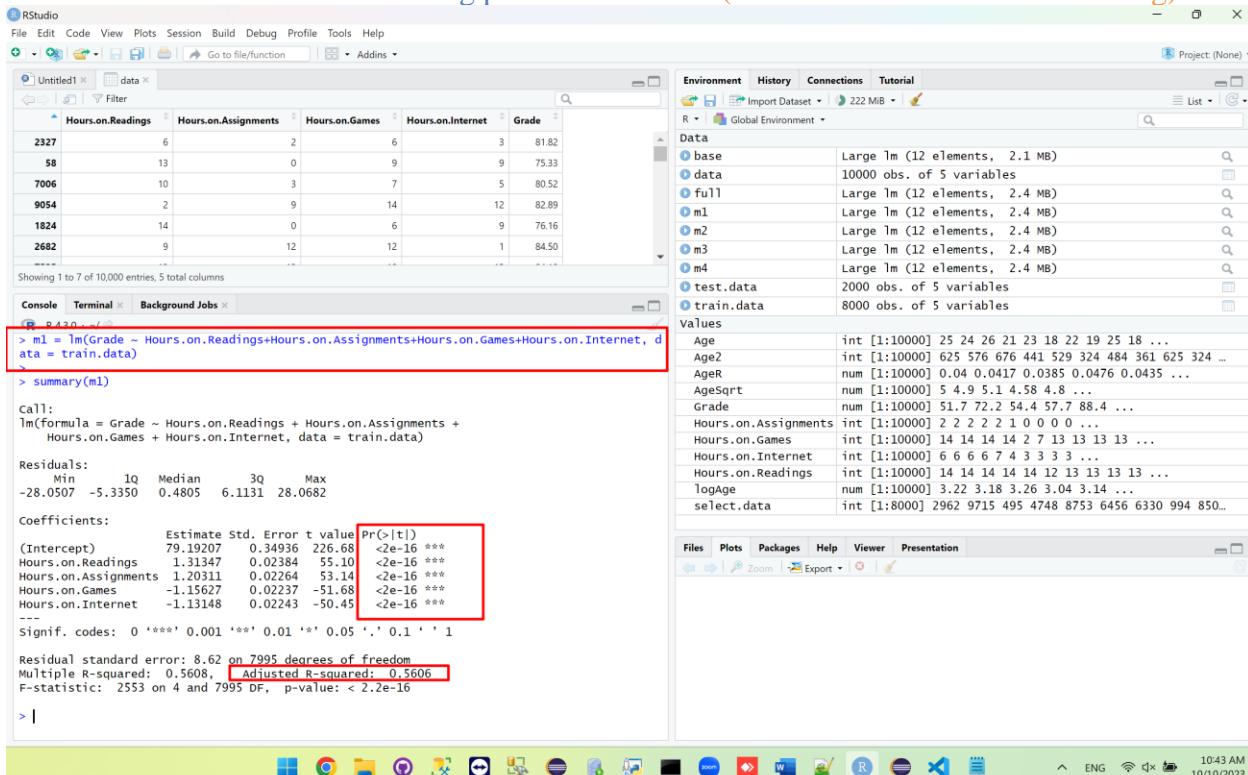
- As mentioned in the question, we need to use **hold out evaluation** for this data split.
- Note:** In general, this data split depends on the size of the data.
  - If the data is large (more than 1 million rows) then hold-out evaluation is used.
  - If the data is small (1 million rows /smaller) then N- fold cross evaluation is used.
  - If your computer is powerful enough, choose N- fold cross evaluation even for the larger data.
  - If the computer is not powerful, then the process may take 5-6 hours, so use hold-out evaluation for the larger data.
  - If your data is small, then definitely choose N-fold cross evaluation.
- Using hold-out evaluation only, **80% as training set and 20% as testing set.**
- First, we need to shuffle the rows. \*\*\*\*\*important point

Here is the screenshot of R coding:



## Step 4: Building Multiple linear regression models using ‘feature selection’ process:

### 1. Based on Backward method using p-value as metric: (here is the screenshot of R-coding)- M1



- As all the x'variables have smaller p values than alpha (Assuming using 95% confidence level), all the x-var's are useful to make prediction on Y-var (Grade).
- As no variable is having larger p-value than alpha, there is no need to eliminate any variable 1 by 1 here manually.

## 2. Based on the Backward method using AIC as metric:

The screenshot shows the RStudio interface with the following details:

- Environment Tab:** Shows objects like base, data, full, m1, m2, m3, m4, test.data, and train.data.
- Data View:** Displays a sample of the 'data' dataset with columns: Hours.on.Readings, Hours.on.Assignments, Hours.on.Games, Hours.on.Internet, and Grade.
- Console Tab:** Contains the R code for the stepwise regression:
 

```
R> full = lm(Grade ~ Hours.on.Readings+Hours.on.Assignments+Hours.on.Games+Hours.on.Internet,
R> data = train.data)
R> m2 = step(full, direction="backward", trace=T)
```
- Output View:** Shows the stepwise regression output with columns: Df, Sum of Sq, RSS, and AIC. The final model is:
 

	Df	Sum of Sq	RSS	AIC
<none>		594067	34470	
- Hours.on.Internet	1	189095	783161	36679
- Hours.on.Games	1	198469	792536	36774
- Hours.on.Assignments	1	209826	803892	36888
- Hours.on.Readings	1	225609	819676	37044

- All the x-variables became useful, so no x-variables got eliminated automatically.
- This is an automatic process.
- So, we got just 1 final step in the output.
- In this output, we can't see p-value, adj r2.
- So, build another model with all the var's that became useful to see adj r2, p-value. **M2**

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R 4.3.0 - /-
> full = lm(Grade ~ Hours.on.Readings+Hours.on.Assignments+Hours.on.Games+Hours.on.Internet, data = train.data)
> m2 = step(full, direction="backward", trace=T)
Start: AIC=34470.4
Grade ~ Hours.on.Readings + Hours.on.Assignments + Hours.on.Games +
Hours.on.Internet

Df Sum of Sq RSS AIC
<none> 594067 34470
- Hours.on.Internet 1 189095 783161 36679
- Hours.on.Games 1 198469 792536 36774
- Hours.on.Assignments 1 209826 803892 36888
- Hours.on.Readings 1 225609 819676 37044
> m2 = lm(Grade ~ Hours.on.Readings+Hours.on.Assignments+Hours.on.Games+Hours.on.Internet, data = train.data)
> summary(m2)

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
    Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q Median      3Q     Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207   0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347   0.02384 55.10 <2e-16 ***
Hours.on.Assignments 1.20311   0.02264 53.14 <2e-16 ***
Hours.on.Games -1.15627   0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148   0.02243 -50.45 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608, Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF, p-value: < 2.2e-16

```

### 3. Based on Forward method using AIC as metric.

- First build a base model with just 1 X-var and y-var, then use step () function to automatically add remaining variables which are useful one by one.
- It is also an automatic process.

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R 4.3.0 - /-
> base = lm(Grade~ Hours.on.Readings, data = train.data)
> step(base, scope=list(lower=NULL, upper=full), direction ="forward", trace=T)
Start: AIC=40043.32
Grade ~ Hours.on.Readings

Df Sum of Sq RSS AIC
+ Hours.on.Assignments 1 233800 959356 38301
+ Hours.on.Games 1 188774 1004381 38667
+ Hours.on.Internet 1 177086 1016070 38760
<none> 1193156 40043

Step: AIC=38300.56
Grade ~ Hours.on.Readings + Hours.on.Assignments

Df Sum of Sq RSS AIC
+ Hours.on.Games 1 176195 783161 36679
+ Hours.on.Internet 1 166820 792536 36774
<none> 959356 38301

Step: AIC=36679.18
Grade ~ Hours.on.Readings + Hours.on.Assignments + Hours.on.Games

Df Sum of Sq RSS AIC
+ Hours.on.Internet 1 180905 594067 34470
<none> 783161 36679

Step: AIC=34470.4
Grade ~ Hours.on.Readings + Hours.on.Assignments + Hours.on.Games +
Hours.on.Internet

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
    Hours.on.Games + Hours.on.Internet, data = train.data)

Coefficients:
            (Intercept) Hours.on.Readings Hours.on.Assignments
                79.192                  1.313                  1.203
                Hours.on.Internet
                    -1.156

```

- In this output, we can't see p-value, adj r2.
- So, build another model with all the var's that became useful to see adj r2, p-value. M3

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Console Terminal Background Jobs

R 4.3.0 - ~

```

Step: AIC=34470.4
Grade ~ Hours.on.Readings + Hours.on.Assignments + Hours.on.Games +
  Hours.on.Internet

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
  Hours.on.Games + Hours.on.Internet, data = train.data)

Coefficients:
(Intercept) Hours.on.Readings Hours.on.Assignments
    79.192          1.313           1.203
Hours.on.Games Hours.on.Internet
   -1.156         -1.131

> m3 = lm(Grade ~ Hours.on.Readings+Hours.on.Assignments+Hours.on.Games+Hours.on.Internet,
  data = train.data)
> summary(m3)

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
  Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207  0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347  0.02384 55.10 <2e-16 ***
Hours.on.Assignments 1.20311  0.02264 53.14 <2e-16 ***
Hours.on.Games -1.15627  0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148  0.02243 -50.45 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608, Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF, p-value: < 2.2e-16
  
```

Environment History Connections Tutorial

Data

base	Large lm (12 elements, 2.1 MB)
data	10000 obs. of 5 variables
full	Large lm (12 elements, 2.4 MB)
m1	Large lm (12 elements, 2.4 MB)
m2	Large lm (12 elements, 2.4 MB)
m3	Large lm (12 elements, 2.4 MB)
m4	Large lm (12 elements, 2.4 MB)
test.data	2000 obs. of 5 variables
train.data	8000 obs. of 5 variables

Values

Age	int [1:10000] 25 24 26 21 23 18 22 19 25 18 ...
Age2	int [1:10000] 625 576 441 529 324 484 361 625 324 ...
AgeR	num [1:10000] 0.04 0.0417 0.0385 0.0476 0.0435 ...
AgeSqrt	num [1:10000] 5 4.9 5.1 4.58 4.8 ...
Grade	num [1:10000] 51.7 72.2 54.4 57.7 88.4 ...
Hours.on.Assignments	int [1:10000] 2 2 2 2 2 1 0 0 0 0 ...
Hours.on.Games	int [1:10000] 14 14 14 14 2 7 13 13 13 13 ...
Hours.on.Internet	int [1:10000] 6 6 6 7 4 3 3 3 3 ...
Hours.on.Readings	int [1:10000] 14 14 14 14 14 12 13 13 13 13 ...
LogAge	num [1:10000] 3.22 3.18 3.26 3.04 3.14 ...
select.data	int [1:8000] 2962 9715 495 4748 8753 6456 6330 994 850 ...

Files Plots Packages Help Viewer Presentation

10:49 AM 10/10/2023

#### 4. Based on Stepwise method using ACI as metric.

- First build a base model with just 1 X-var and y-var, then use step() function to automatically add remaining variables which are useful one by one.
- Set the direction as “both”.
- It is also an automatic process.

```

> base = lm(Grade ~ Hours.on.Readings, data = train.data)
> step(base, scope=list(upper=full, lower=-1), direction="both", trace=T)
Start: AIC=40043.32
Grade ~ Hours.on.Readings
          Df Sum of Sq  RSS   AIC
+ Hours.on.Assignments  1  233800  959356 38301
+ Hours.on.Games        1  188774  1004381 38667
+ Hours.on.Internet     1  177086  1016070 38600
<none>
- Hours.on.Readings     1  159592  1352747 31046
Step: AIC=38300.56
Grade ~ Hours.on.Readings + Hours.on.Assignments
          Df Sum of Sq  RSS   AIC
+ Hours.on.Games        1  176195  781617 36679
+ Hours.on.Internet     1  166820  793536 36774
<none>
- Hours.on.Readings     1  204995  1164351 39848
- Hours.on.Assignments  1  233800  1193156 30043
Step: AIC=36679.18
Grade ~ Hours.on.Readings + Hours.on.Assignments + Hours.on.Games
          Df Sum of Sq  RSS   AIC
+ Hours.on.Internet     1  189095  594067 34470
<none>
- Hours.on.Games        1  176195  959356 38301
- Hours.on.Readings     1  190523  973685 38419
- Hours.on.Assignments  1  221220  1004381 38667
Step: AIC=34470.4
Grade ~ Hours.on.Readings + Hours.on.Assignments + Hours.on.Games +
Hours.on.Internet
          Df Sum of Sq  RSS   AIC
<none>
- Hours.on.Internet     1  189095  783161 36679
- Hours.on.Games        1  198469  792536 36774
- Hours.on.Assignments  1  209826  803892 36888
- Hours.on.Readings     1  225609  819676 37044
call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
  Hours.on.Games + Hours.on.Internet, data = train.data)
Coefficients:
(Intercept) Hours.on.Readings Hours.on.Assignments
    79.192           1.313           1.203
  Hours.on.Games Hours.on.Internet
    -1.156          -1.131
> |

```

- In this output, we can't see p-value, adj r2.
- So, build another model with all the vari's that became useful to see adj r2, p-value. M4

```

File Edit Code View Plots Session Build Debug Profile Tools Help
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R 4.3.0 ~/~/
          Df Sum of Sq  RSS   AIC
+ Hours.on.Internet  1  189095  594067 34470
<none>
- Hours.on.Games     1  176195  959356 38301
- Hours.on.Readings   1  190523  973685 38419
- Hours.on.Assignments 1  221220  1004381 38667
Step: AIC=34470.4
Grade ~ Hours.on.Readings + Hours.on.Assignments + Hours.on.Games +
Hours.on.Internet
          Df Sum of Sq  RSS   AIC
<none>
- Hours.on.Internet  1  189095  783161 36679
- Hours.on.Games     1  198469  792536 36774
- Hours.on.Assignments 1  209826  803892 36888
- Hours.on.Readings   1  225609  819676 37044
call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
  Hours.on.Games + Hours.on.Internet, data = train.data)
Coefficients:
(Intercept) Hours.on.Readings Hours.on.Assignments
    79.192           1.313           1.203
  Hours.on.Games Hours.on.Internet
    -1.156          -1.131
> m4 = lm(Grade ~ Hours.on.Readings+Hours.on.Assignments+Hours.on.Games+Hours.on.Internet, data = train.data)
> summary(m4)

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
  Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207  0.34938 226.68 <2e-16 ***
Hours.on.Readings 1.31347  0.02384 55.10 <2e-16 ***
Hours.on.Assignments 1.20311  0.02264 53.14 <2e-16 ***
Hours.on.Games -1.15627  0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148  0.02243 -50.45 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608, Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF,  p-value: < 2.2e-16
>

```

## 5. Based on Best subset method using Adj-R2 as metric:

- It gives you all the combination of x-variables, then we are going to build a model by calculating a specific metric like adjr2.
- To build this model, we need to install package ‘leaps’ and use leaps () function.

The screenshot shows two RStudio sessions. The top session is a new R session where the 'leaps' package is being installed. The bottom session is an existing session where a regression model is being built using the 'leaps' package.

```

R 4.3.0 - ~/R> >install.packages("leaps")
Error in install.packages : updating loaded packages
Restarting R session...
> install.packages("leaps")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:
https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/satya/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/leaps_3.1.zip'
Content type 'application/zip' length 86991 bytes (84 KB)
downloaded 84 KB

package 'leaps' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
  C:\Users\satya\AppData\Local\Temp\Rtmp04cUNh\downloaded_packages
> library('leaps')
warning message:
package 'leaps' was built under R version 4.3.1
> library('leaps')
> |

```

```

R 4.3.0 - ~/R> >leaps(y=train.data[,5],x=train.data[,cbind(1,2,3,4)],method="adjr2")
SWITCH
  Hours.on.Readings Hours.on.Assignments Hours.on.Games Hours.on.Internet
1 FALSE           FALSE           TRUE           FALSE
1 FALSE           TRUE           FALSE           FALSE
1 TRUE            FALSE           FALSE           TRUE
2 FALSE           TRUE           FALSE           FALSE
2 FALSE           TRUE           TRUE            TRUE
2 FALSE           FALSE          TRUE            TRUE
2 TRUE            FALSE          TRUE            FALSE
2 FALSE           TRUE           FALSE           TRUE
3 TRUE            TRUE           TRUE           FALSE
3 TRUE            TRUE           FALSE           TRUE
3 FALSE           TRUE           TRUE            TRUE
4 TRUE            TRUE           TRUE           TRUE

$label
[1] "(Intercept)"      "Hours.on.Readings"   "Hours.on.Assignments" "Hours.on.Games"
[5] "Hours.on.Internet"

$size
[1] 2 2 2 2 3 3 3 3 3 4 4 4 4 5

$adjr2
[1] 0.1483008 0.1391616 0.1178656 0.1062783 0.2906319 0.2800369 0.2715938 0.2573392 0.2486963
[10] 0.2367854 0.4208416 0.4139091 0.4055106 0.3938385 0.5606247
>

```

Interpretation of the output:

Subset of x-variables selected are: (true / false output)

- 1- true -Hours.on.games
- 1 -true- hours.on.assignments
- ..
- ..... selected the variable which has true value in each row.
- ..
- 2 – true – hours on readings & hours on assignments
- ..
- ..
- .....selected the subset of 2 variables which has true values in each row.
- ..
- ..

### Adj r2 output explanation:

- We need to find the best and largest value in adj r2 output.
- The best value is = 0.5608489
- The index of this model is = 15. That means this adj r2 value belongs to 15<sup>th</sup> model.
- Then we need to look in to the true/false chart and go to the 15<sup>th</sup> row and check which variables are selected true.
- As all the variables are selected as true, all the variables are useful.
- So, build a model with all these x-var's and y-var Grade and look for the adj r2 in the output.

M5

The screenshot shows the RStudio interface with the following details:

- Environment Pane:** Shows objects like base, data, m1-m5, test.data, and train.data.
- Console Pane:**

```

> m5 = lm(Grade ~ Hours.on.Readings+Hours.on.Assignments+Hours.on.Games+Hours.on.Internet, data = train.data)
> summary(m5)

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
    Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207   0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347   0.02384  55.10 <2e-16 ***
Hours.on.Assignments 1.20311   0.02264  53.14 <2e-16 ***
Hours.on.Games -1.15627   0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148   0.02243 -50.45 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608,  Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF,  p-value: < 2.2e-16
  
```
- Project Pane:** Shows a project named 'M5' with files like 'data.R', 'data.RData', and 'train.data'.
- Status Bar:** Shows the time as 11:01 AM and the date as 10/10/2023.

Now we got 5 models : M1, M2, M3, M4,M5 (same results)

MODEL
M1 : Based on Backward method using p-value as metric:
M2: Based on Backward method using AIC as metric:
M3: Based on Forward method using AIC as metric.
M4: Based on Stepwise method using ACI as metric:
M5: Based on Best subset method using Adj-R2 as metric

**Step5 : Model diagnosis:** We need to perform model diagnosis for all the models above and compare and see which model is qualified.

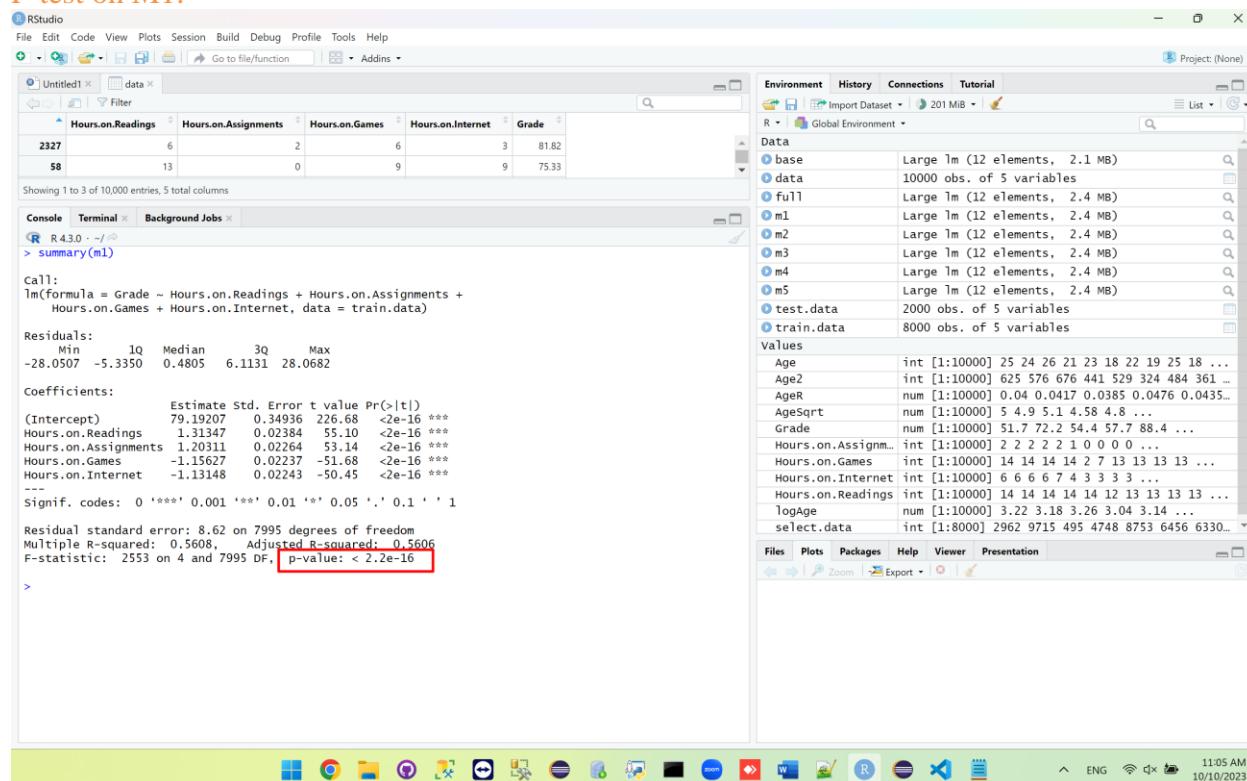
\*\*\*But pls make note that all models are same we can pick any 1 model and perform model diagnosis... Lets pick M1 and go ahead\*\*\*\*\*

There are 2 components in model diagnosis:

1. F-test (goodness of fit test)
2. Residual analysis

**F-test:** is a statistical test for hypothesis testing. We need to write down null hypothesis and alternative hypothesis:

F-test on M1:



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled1 data
Hours.on.Readings Hours.on.Assignments Hours.on.Games Hours.on.Internet Grade
2327 6 2 6 3 81.82
58 13 0 9 9 75.33
Showing 1 to 3 of 10000 entries, 5 total columns

Console Terminal Background Jobs
R 4.3.0 - ~/ ~/
> summary(m1)

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
    Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207  0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347  0.02384 55.10 <2e-16 ***
Hours.on.Assignments 1.20311  0.02264 53.14 <2e-16 ***
Hours.on.Games -1.15627  0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148  0.02243 -50.45 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608, Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF,  p-value: < 2.2e-16

```

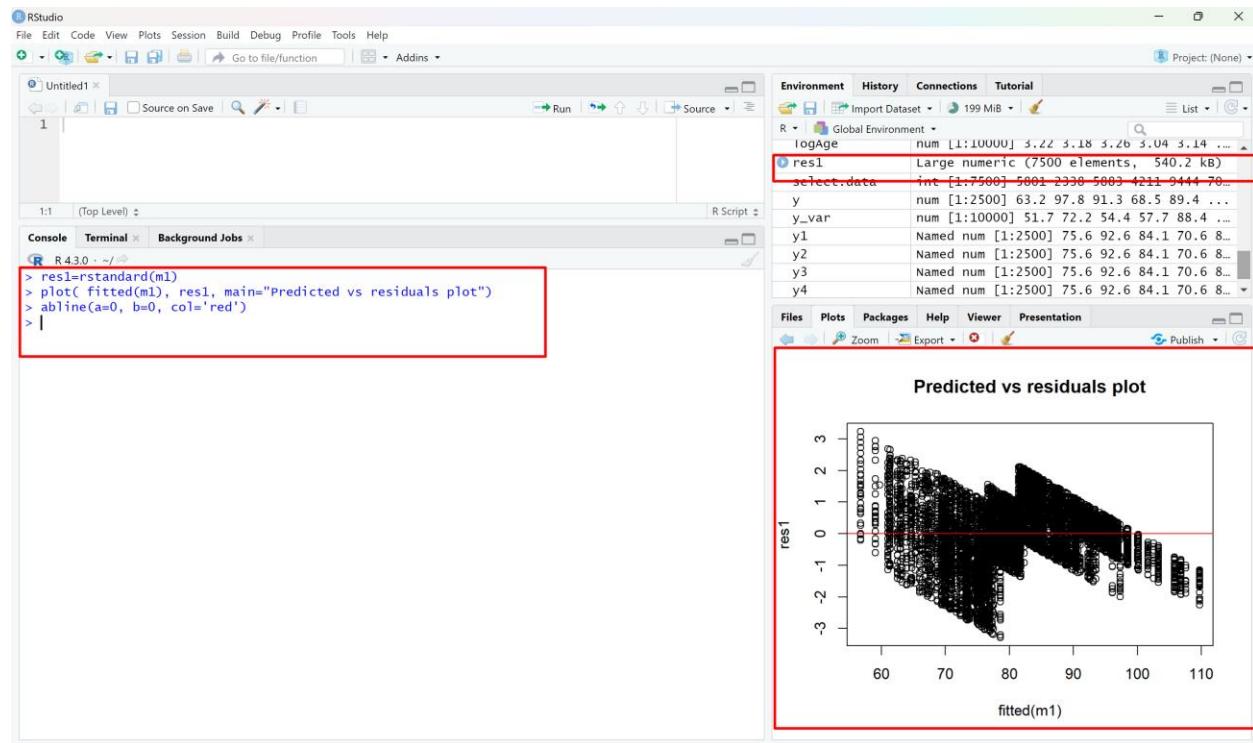
## Residual analysis on M1:

- As we built 5 models above through the feature selection process(M1, M2, M3, M4, M5), we need to perform residual analysis for all the models.

\*\*\*\* but. Please note that all the above models are same, So, we can perform residual analysis on any 1 model. As we already picked M1 for model diagnosis and performed f-test on it, now lets perform residual analysis for the same model M1 \*\*\*\*

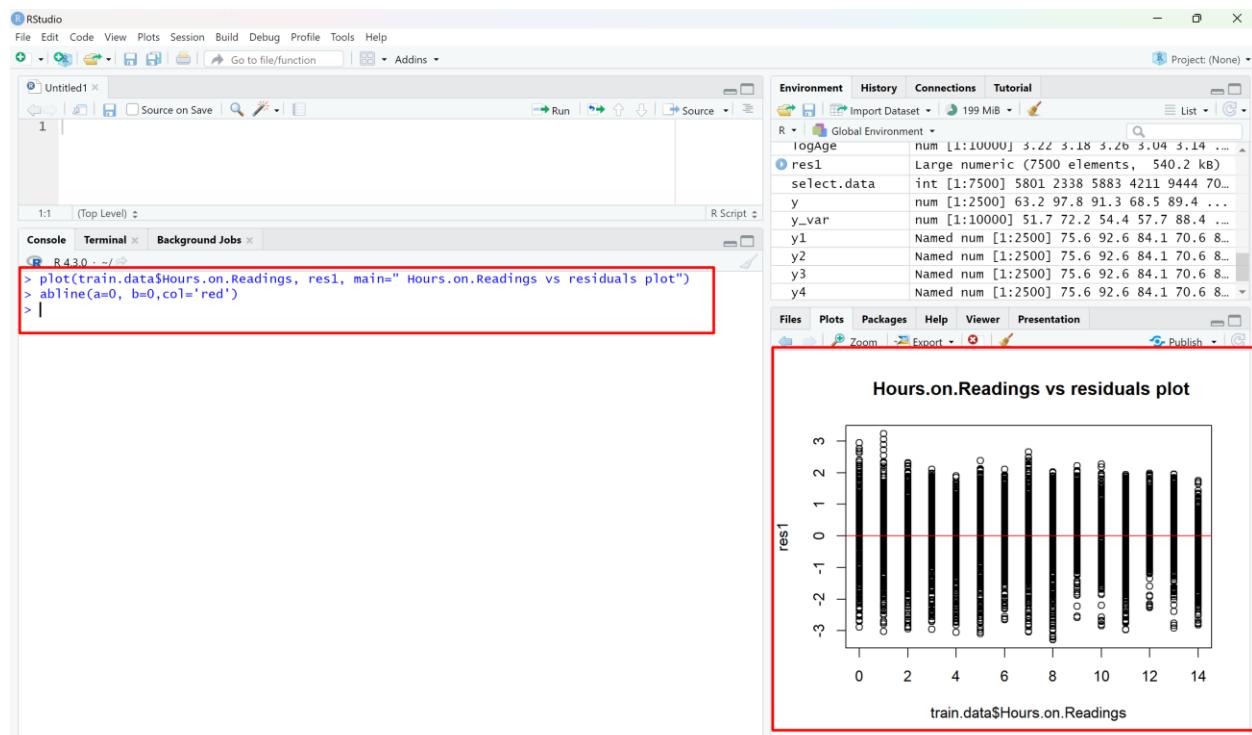
### Residual analysis on M1:

- Validate the constant variance: Plot residuals vs predicted values: To check constant variance for the residuals

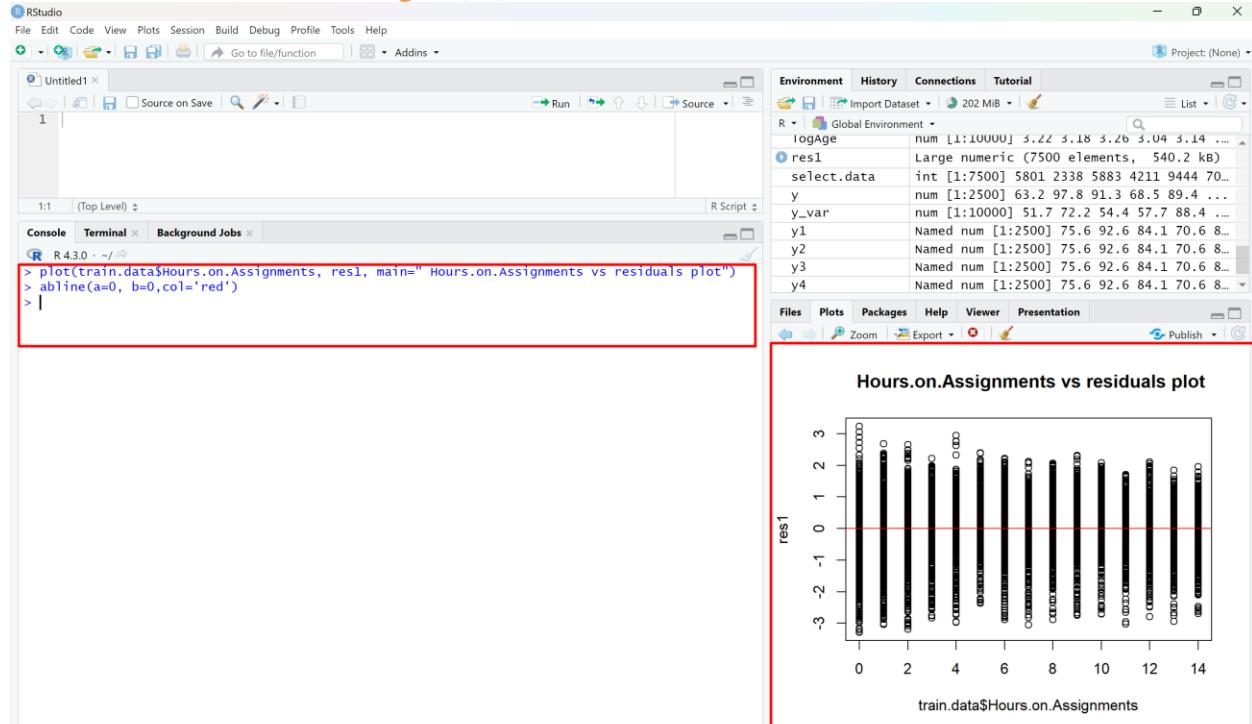


- Validate the linearity relationship: Plot residuals vs each x-variable: To check linearity assumptions for Y and the x-variable.

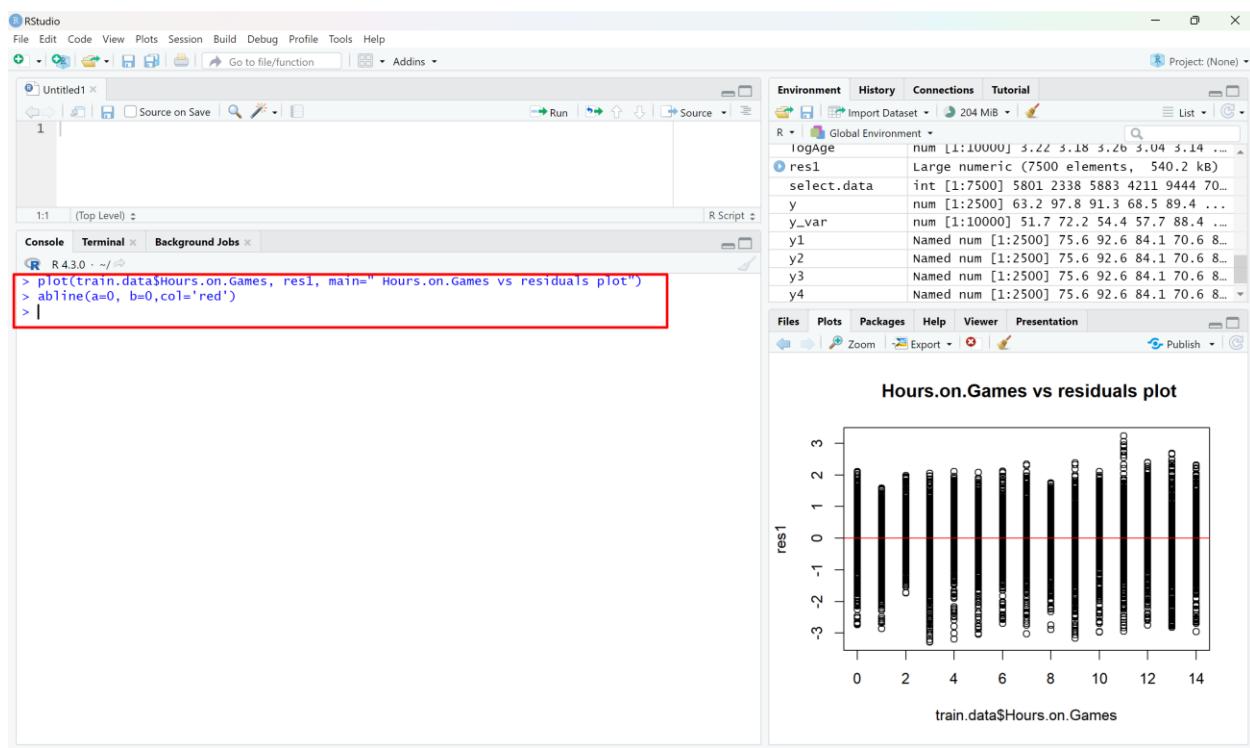
Plot residual vs Hours on Readings(x1)



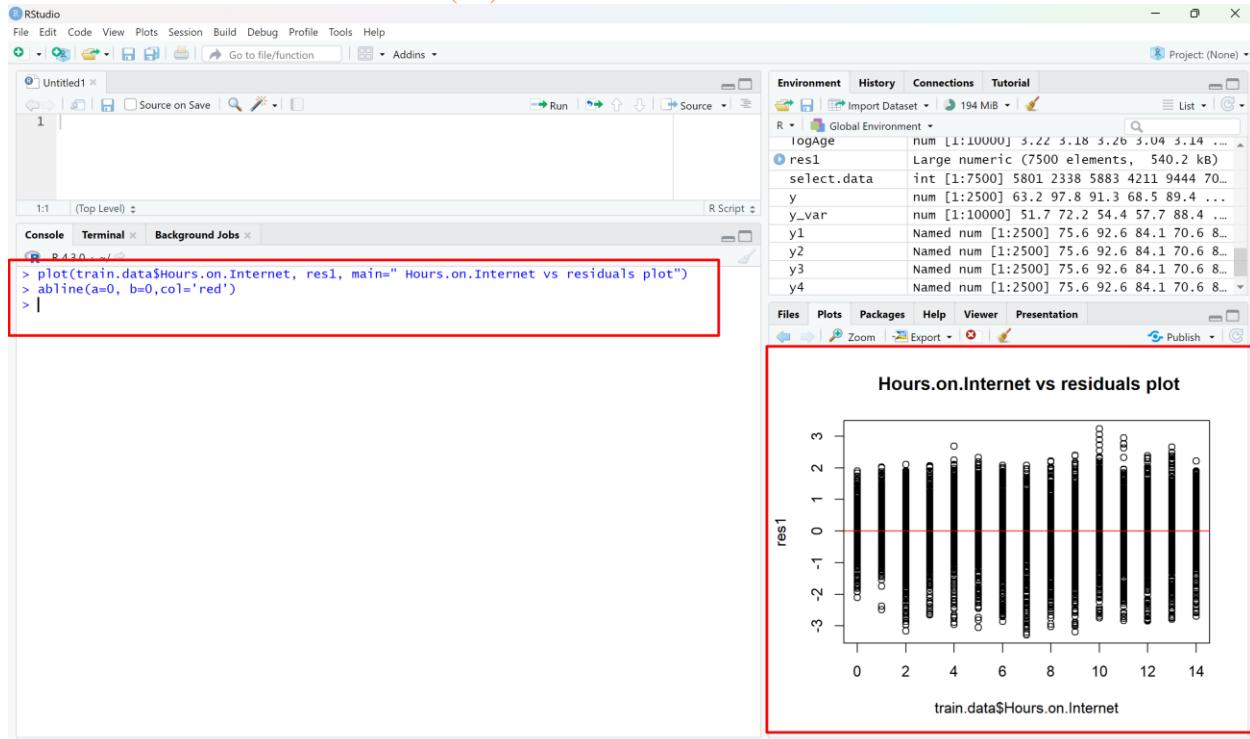
### Plot residual vs Hours on Assignm(x2)



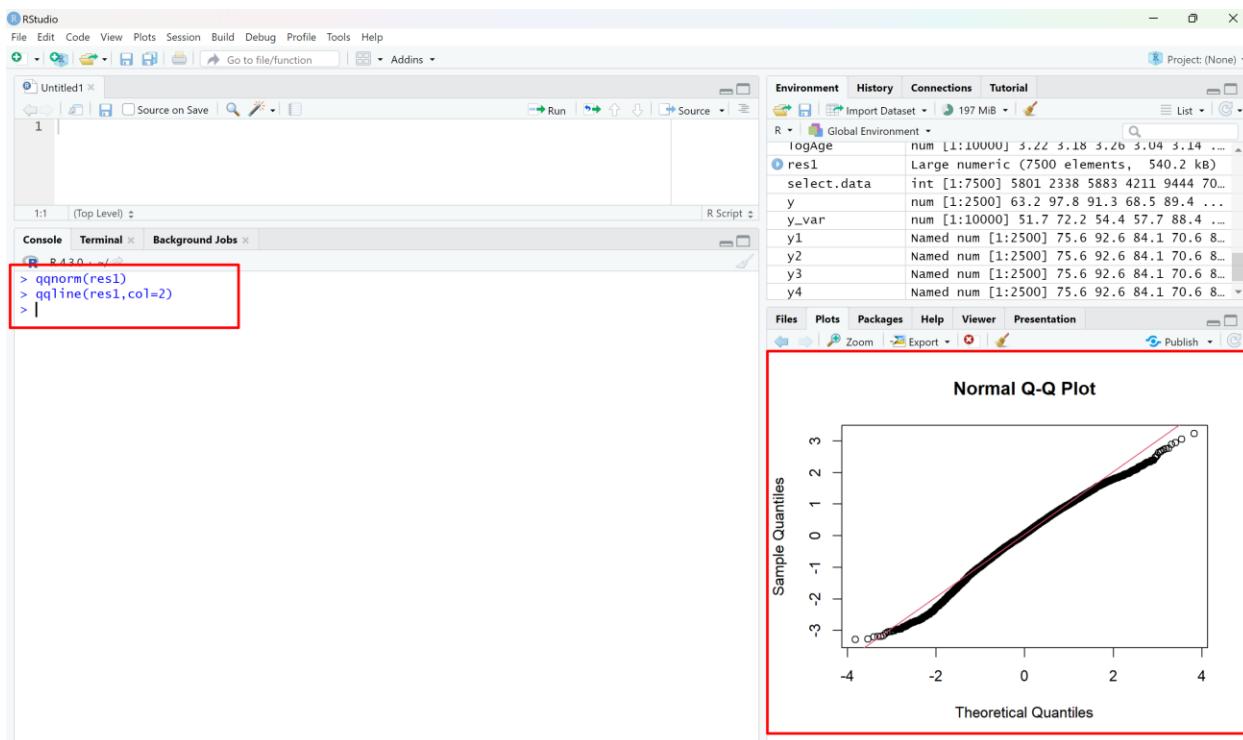
### Plot residual vs Hours on Games(x3)



### Plot residual vs Hours on Internet (x4)



3. Validate normal distribution of residuals: Draw normal probability plot of residuals: To check normality assumption for the error terms; if points lie close to a line, the errors can be assumed to be approximately normal. Otherwise the assumption of normality is not satisfied.



**Conclusion on residual analysis:** We can see there is **an issue in the constant variance (step1)**, It is Constance in the beginning, but variance became smaller at the end. We can try to apply log transformation on y-variable to see whether we can get a better model.

\*\*\*\*\*We will solve this issue in Quest 2.\*\*\*\*\*

Q2 Particularly, you should answer the following questions (at the end of your homework submisson) by using your R coding and outputs.

\*\*\*\*Note: As I performed all the steps in linear regression and pasted r-coding screenshots above already, I am using same screenshots here again to answer the below questions\*\*\*\*

1. Do all x variables have linear relationship with y?

No, 'Age' has no linear relationship with Y. Here is the process of examining & solving the issue:

#### Examine the linear relationship between x and y variables.

- We can examine linearity by 2 methods.
  3. Produce a scatter plot for each x & y variable / produce a single plot with every pair of variables.
  4. Calculate correlation values for all the x variables with y variable.
- As the plot method is not clear & reliable all the time, we can perform 2<sup>nd</sup> method and calculate the correlation values.

Here is the screenshot of R coding: (calculating correlation values)

The screenshot shows the RStudio interface. In the top-left, there's a data grid with columns labeled 'Age', 'Hours.on.Readings', 'Hours.on.Assignments', 'Hours.on.Games', 'Hours.on.Internet', and 'Grade'. Below the grid, the console window displays the command: `> cor(cbind(Grade, Age, Hours.on.Readings, Hours.on.Assignments, Hours.on.Games, Hours.on.Internet))`. The output is a correlation matrix:

	Grade	Age	Hours.on.Readings	Hours.on.Assignments	Hours.on.Games	Hours.on.Internet
Grade	1.00000000	-0.01130322	0.34740156	0.36504450	-0.39055054	-0.32748684
Age	-0.01130322	1.00000000	-0.00095846	-0.01711789	-0.34740156	0.34740156
Hours.on.Readings	0.34740156	-0.00095846	1.00000000	-0.11871999	0.09797599	-0.03905505
Hours.on.Assignments	0.36504450	-0.01711788	-0.11871999	1.00000000	0.09863238	-0.03905505
Hours.on.Games	-0.39055054	0.09797599	-0.03515434	-0.02809676	1.00000000	-0.05801234
Hours.on.Internet	-0.32748684	-0.03905505	-0.09586324	-0.03905873	-0.05801234	1.00000000

- We noticed that the correlation values are as follows:
  - Grade & Hours on readings = **0.34740156** - weak Correlation - +ve correlation
  - Grade & Hours on Assignments = **0.36504450** - weak Correlation - +ve correlation
  - Grade & Hours on games = **-0.39055054** - weak Correlation - -ve correlation
  - Grade & Hours on internet = **-0.32748684** - weak Correlation - -ve correlation
- As Age has no correlation with Grade(Y-variable), we can try transformation on this Age (x-variable) and re-calculate the correlation with Grade (y-variable) again.
  - Square transformation:  $X' = X * X$
  - Log transformation:  $X' = \log X$
  - Inversion transformation:  $X' = 1/X$
  - Square root transformation:  $X' = \sqrt{X}$

Here is the screenshots of R coding:

(Transformation on Age variable & re-correlation value calculation with grade)

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Untitled1
Source on Save Run Source
1 | R Script

Console Terminal Background Jobs
R 4.3.0 - ~/ 
> cor(cbind(Grade, Age, Hours.on.Readings, Hours.on.Assignments, Hours.on.Games, Hours.on.Internet))
   Grade      Age Hours.on.Readings Hours.on.Assignments
Grade 1.00000000 -0.011303219 0.34740156 0.36504450
Age -0.01130322 1.00000000 -0.00095846 -0.01711789
Hours.on.Readings 0.34740156 -0.00095846 1.00000000 -0.11871999
Hours.on.Assignments 0.36504450 -0.017117888 -0.11871999 1.00000000
Hours.on.Games -0.39055054 0.009766920 -0.03515434 -0.02809676
Hours.on.Internet -0.32748684 0.007975999 0.09586324 -0.03905875
   Hours.on.Games Hours.on.Internet
Grade -0.39055054 -0.327486837
Age 0.00976692 0.007975999
Hours.on.Readings -0.03515434 0.095863238
Hours.on.Assignments -0.02809676 -0.039058746
Hours.on.Games 1.00000000 -0.058012340
Hours.on.Internet -0.05801234 1.00000000

> Age2 = Age*Age
> cor(Grade, Age2)
[1] -0.01194648
> logAge = log(Age)
> cor(Grade, logAge)
[1] -0.0106812
> AgeR = 1/Age
> cor(Grade, AgeR)
[1] 0.01009053
> Agesqrt = sqrt(Age)
> cor(Grade, Agesqrt)
[1] -0.01098891
>

```

- As the correlation didn't improve after transformation on 'Age' variable, we decided to drop the age variable.

Here Is the screenshot of r coding: (dropping the age variable)

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Untitled1
Source on Save Run Source
1 | R Script

Console Terminal Background Jobs
R 4.3.0 - ~/ 
> Hours.on.Readings 0.34/4015b -0.00095846b 1.00000000 -0.118/1999
> Hours.on.Assignments 0.36504450 -0.017117888 -0.11871999 1.00000000
> Hours.on.Games -0.39055054 0.009766920 -0.03515434 -0.02809676
> Hours.on.Internet -0.32748684 0.007975999 0.09586324 -0.03905875
   Hours.on.Games Hours.on.Internet
Grade -0.39055054 -0.327486837
Age 0.00976692 0.007975999
Hours.on.Readings -0.03515434 0.095863238
Hours.on.Assignments -0.02809676 -0.039058746
Hours.on.Games 1.00000000 -0.058012340
Hours.on.Internet -0.05801234 1.00000000

> Age2 = Age*Age
> cor(Grade, Age2)
[1] -0.01194648
> logAge = log(Age)
> cor(Grade, logAge)
[1] -0.0106812
> AgeR = 1/Age
> cor(Grade, AgeR)
[1] 0.01009053
> Agesqrt = sqrt(Age)
> cor(Grade, Agesqrt)
[1] -0.01098891
> data = data[, !names(data) %in% c("Age")]
> str(data)
'data.frame': 10000 obs. of 5 variables:
 $ Hours.on.Readings : int 14 14 14 14 12 13 13 13 ...
 $ Hours.on.Assignments: int 2 2 2 2 2 1 0 0 0 ...
 $ Hours.on.Games : int 14 14 14 14 2 7 13 13 13 ...
 $ Hours.on.Internet : int 6 6 6 7 4 3 3 3 ...
 $ Grade : num 51.7 72.2 54.4 57.7 88.4 ...

```

- Write down the null and alternative hypothesis of F-test, use your outputs to draw conclusions of F-test.

**ANSWER:**

**F-test:** is a statistical test for hypothesis testing. We need to write down null hypothesis and alternative hypothesis:

- Null hypothesis : (H<sub>0</sub>): The coefficients of all x-variables are zero and there is no linear relationship with Grade.
- $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
- Alternative Hypothesis: At least one of the coefficients of the x-variables (Hours.on.Readings, Hours.on.Assignments, Hours.on.Games, Hours.on.Internet) is not zero and can affect Grade.
- $H_a : \beta_j \neq 0$

Note : In these hypotheses,  $\beta_1$  represents the coefficient of "Hours.on.Readings,"  $\beta_2$  represents the coefficient of "Hours.on.Assignments,"  $\beta_3$  represents the coefficient of "Hours.on.Games," and  $\beta_4$  represents the coefficient of "Hours.on.Internet."

- As we built multiple models above through the feature selection process (M1, M2, M3, M4, M5), we need to do the f- test for all the models. But as all the models are same we can pick any 1 model and perform F-test. Lets pick M1:

#### F-test on M1:

```
R 4.3.0 - ~/RStudio
> summary(m1)

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
    Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q      Median      3Q      Max 
-28.0507 -5.3350  0.4805  6.1131  28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207  0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347  0.02384  55.10 <2e-16 ***
Hours.on.Assignments 1.20311  0.02264  53.14 <2e-16 ***
Hours.on.Games   -1.15627  0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148  0.02243 -50.45 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608, Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF,  p-value: < 2.2e-16
```

#### F-test on M2:

```

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
    Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207   0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347   0.02384 55.10 <2e-16 ***
Hours.on.Assignments 1.20311   0.02264 53.14 <2e-16 ***
Hours.on.Games    -1.15627   0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148   0.02243 -50.45 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608, Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF,  p-value: < 2.2e-16

```

## F-test on M3:

```

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
    Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207   0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347   0.02384 55.10 <2e-16 ***
Hours.on.Assignments 1.20311   0.02264 53.14 <2e-16 ***
Hours.on.Games    -1.15627   0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148   0.02243 -50.45 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608, Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF,  p-value: < 2.2e-16

```

## F-test on M4:

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Untitled1 data
Hours.on.Readings Hours.on.Assignments Hours.on.Games Hours.on.Internet Grade
2327 6 2 6 3 81.82
58 13 0 9 9 75.33

Showing 1 to 3 of 10,000 entries, 5 total columns

Console Terminal Background Jobs
R 4.3.0 - ~/ 
> summary(m4)

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
  Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q   Median      3Q     Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207  0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347  0.02384 55.10 <2e-16 ***
Hours.on.Assignments 1.20311  0.02264 53.14 <2e-16 ***
Hours.on.Games -1.15627  0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148  0.02243 -50.45 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608, Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF,  p-value: < 2.2e-16

```

## F-test on M5:

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Untitled1 data
Hours.on.Readings Hours.on.Assignments Hours.on.Games Hours.on.Internet Grade
2327 6 2 6 3 81.82
58 13 0 9 9 75.33

Showing 1 to 3 of 10,000 entries, 5 total columns

Console Terminal Background Jobs
R 4.3.0 - ~/ 
> summary(m5)

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
  Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q   Median      3Q     Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207  0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347  0.02384 55.10 <2e-16 ***
Hours.on.Assignments 1.20311  0.02264 53.14 <2e-16 ***
Hours.on.Games -1.15627  0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148  0.02243 -50.45 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608, Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF,  p-value: < 2.2e-16

```

MODEL	F-test -P-VALUE
M1 : Based on Backward method using p-value as metric:	2.2e-16
M2: Based on Backward method using AIC as metric:	2.2e-16
M3: Based on Forward method using AIC as metric.	2.2e-16
M4: Based on Stepwise method using ACI as metric:	2.2e-16
M5: Based on Best subset method using Adj-R2 as metric	2.2e-16

Conclusion on F-test: p-value < 0.05 for all the models

- At 95% confidence level, we can say that at least 1 x variables among ((Hours.on.Readings, Hours.on.Assignments, Hours.on.Games, Hours.on.Internet) has significant linear relationship with Grade and can affect the value of y-variable ~ Grade.

\*\*\*\*\*By conducting an F-test, we can observe that there is sufficient evidence to reject the null hypothesis and conclude that the independent variables have a significant impact on the dependent variable.\*\*\*\*\*

### 3. Which model is the best in terms of Adj-R2?

From the screenshots above:

MODEL	Adj r2
M1 : Based on Backward method using p-value as metric:	56.06%
M2: Based on Backward method using AIC as metric:	56.06%
M3: Based on Forward method using AIC as metric.	56.06%
M4: Based on Stepwise method using ACI as metric:	56.06%
M5: Based on Best subset method using Adj-R2 as metric	56.06%

Conclusion:

- All models have the same adj r2 values and used same set of x-variables, so all models are best.

\*\*\*\*\*Note: But Adj r2 is not the accurate, we need to evaluate based on testing data set by RMSE\*\*\*\*

#### 4. Interpret the Adj-R2 in the best model above.

**ANSWER:**

- Since all models have the same adj r2 values and use same set of x-variables, so all models are best. So, we can interpret any model's adj r2. Lets take M1 and interpret adj r2:

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled1 data
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
> R 4.3.0 ...
> summary(m1)

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
    Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      Q1      Median      Q3      Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207  0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347  0.02384 55.10 <2e-16 ***
Hours.on.Assignments 1.20311  0.02644 53.14 <2e-16 ***
Hours.on.Games -1.15627  0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148  0.02243 -50.45 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608,  Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF,  p-value: < 2.2e-16

```

Interpretation of adj r2 of m1:

- 56.06% variation in Grade can be explained by the variations in X variables (Hours.on.Readings+Hours.on.Assignments+Hours.on.Games+Hours.on.Internet) based on our fitted regression model

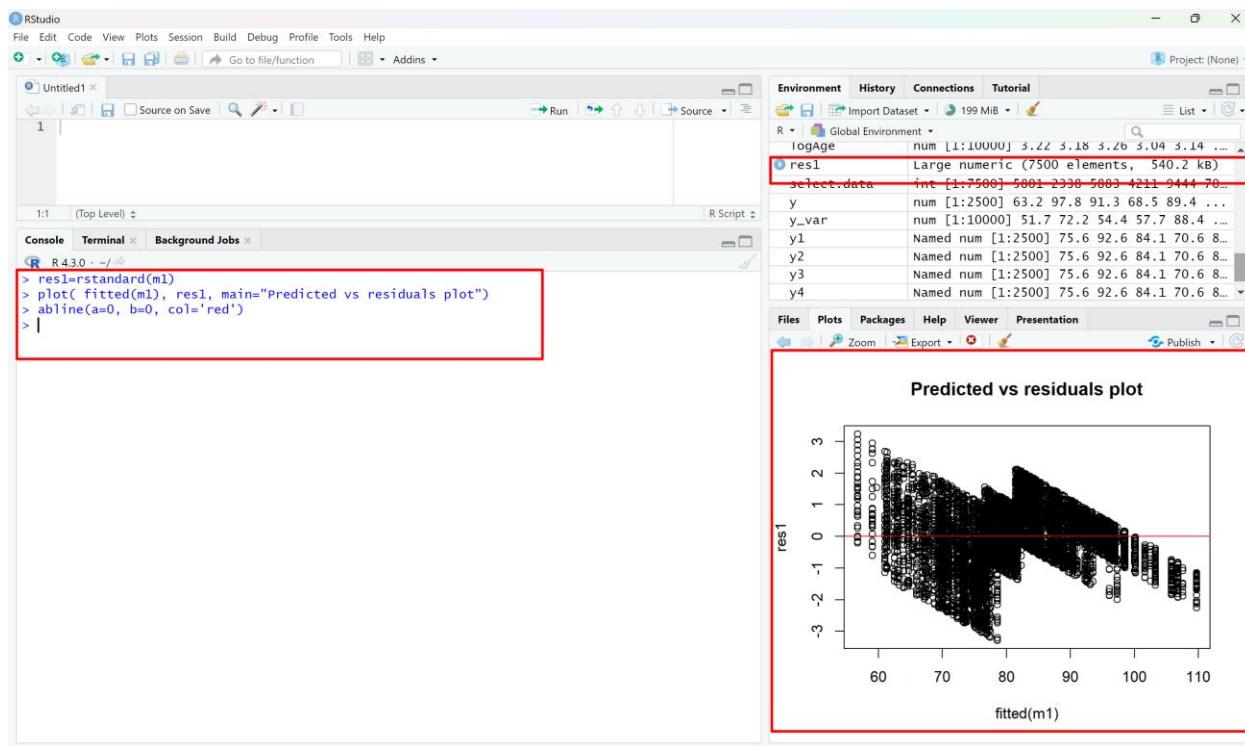
#### 5. Any issues in the residual analysis for the model above?

**ANSWER:**

Yes, there is an issue about constant variance. As all models are same lets take m1 to check:

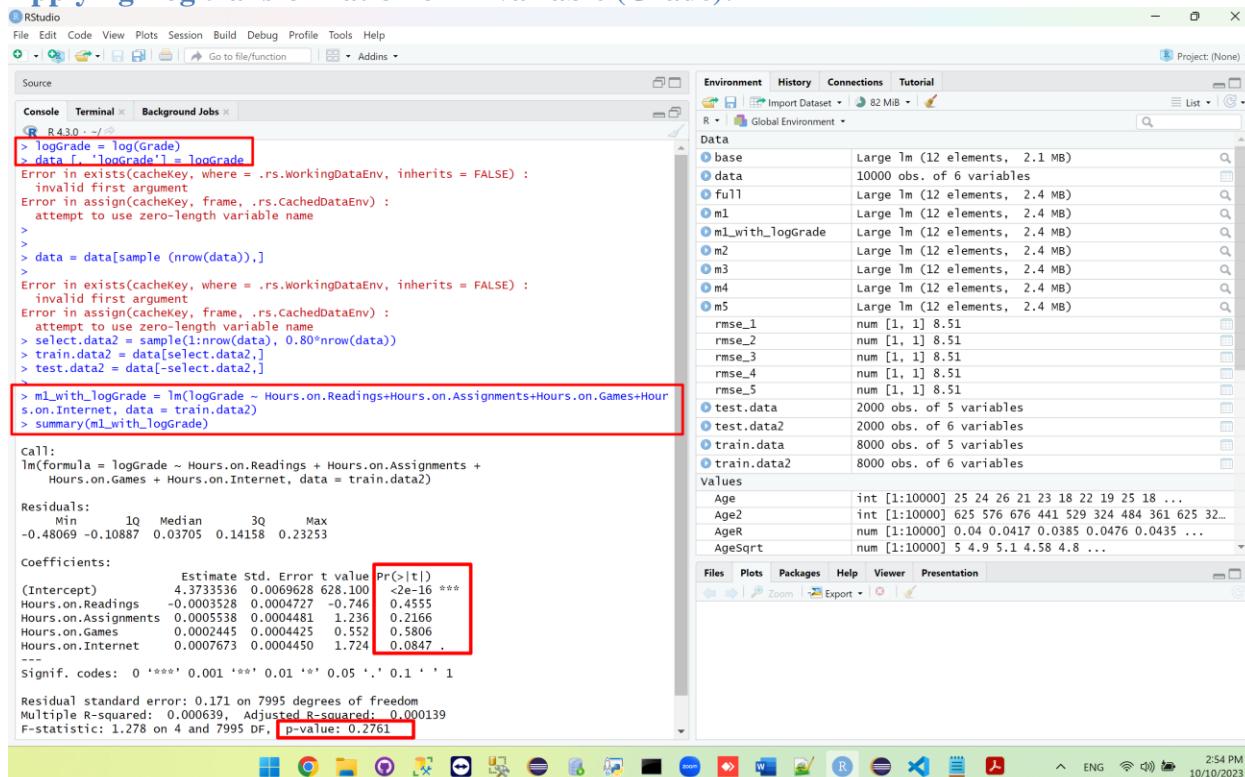
**Residual analysis on m1:**

- Validate the constant variance: Plot residuals vs predicted values: To check constant variance for the residuals



It is Constance in the beginning, but variance became smaller at the end. We can try to apply log transformation on Y-variable(Grade) to see whether we can get a better model.

## Applying Log transformation on Y-variable (Grade):



After applying the log transformation on the variable Grade, I rebuilt the model M1 with logGrade, but it didn't pass the Goodness of fit test as the p - value (0.2761) is greater than 95%(0.05). From this observation I would like to say that we can stick to the old model M1.

## 6. Which model is the best in terms of RMSE?

**ANSWER:**

**Evaluate based on testing data set based on RMSE:**

- To evaluate based on testing data set , we need to calculate the RMSE for all the models Based on RMSE we need to compare all the models and make a conclusion about which model is best.

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Untitled 1 data
Filter
Hours.on.Readings Hours.on.Assignments Hours.on.Games Hours.on.Internet Grade
2327 6 2 6 3 81.82
58 13 0 9 9 75.33
Showing 1 to 3 of 10,000 entries, 5 total columns

Console Terminal Background Jobs
R 4.3.0 - ~/Documents
> names(test.data)
[1] "Hours.on.Readings" "Hours.on.Assignments" "Hours.on.Games" "Hours.on.Internet" "Grade"
[5] "Grade"
> y1<-predict.glm(m1,test.data)
> y2<-predict.glm(m2,test.data)
> y3<-predict.glm(m3,test.data)
> y4<-predict.glm(m4,test.data)
> y5<-predict.glm(m5,test.data)
> y<-test.data[,5]
>
> rmse_1 = sqrt((y-y1)^2/(nrow(test.data)))
> rmse_2 = sqrt((y-y2)^2/(nrow(test.data)))
> rmse_3 = sqrt((y-y3)^2/(nrow(test.data)))
> rmse_4 = sqrt((y-y4)^2/(nrow(test.data)))
> rmse_5 = sqrt((y-y5)^2/(nrow(test.data)))
>
> rmse_1
[1,] 8.512977
> rmse_2
[1,] 8.512977
> rmse_3
[1,] 8.512977
> rmse_4
[1,] 8.512977
> rmse_5
[1,] 8.512977
>

```

MODEL	RMSE
M1 : Based on Backward method using p-value as metric:	8.512
M2: Based on Backward method using AIC as metric:	8.512
M3: Based on Forward method using AIC as metric.	8.512
M4: Based on Stepwise method using ACI as metric:	8.512

M5: Based on Best subset method using Adj-R2 as metric	8.512
--	-------

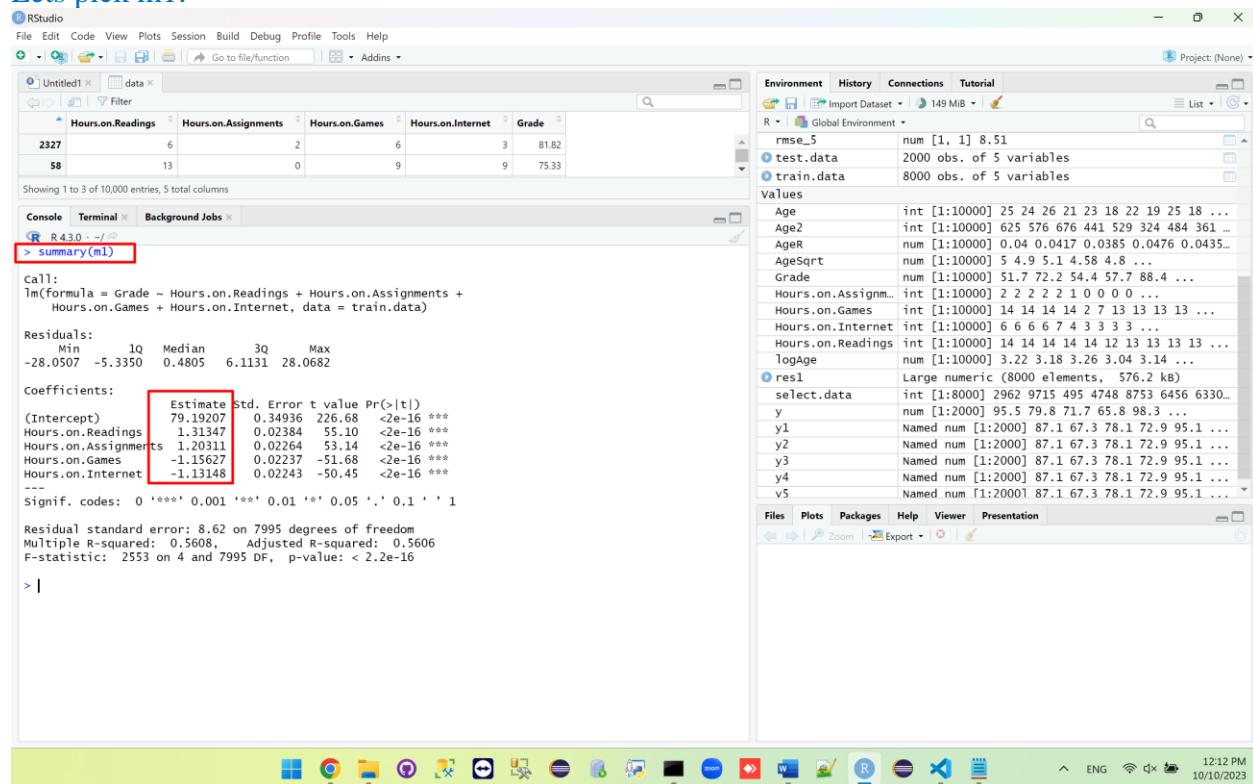
Conclusion: As the RMSE is same for all the models, any model Is best. We can pick any model and write the final model.

7. Write down the best model (identified by RMSE), and explain the intercepts and coefficients in the model.

### ANSWER:

As the RMSE is same for all the models, Any model Is best  
So, we can pick any 1 model and write down the best model by RMSE

Lets pick m1:



```

Call:
lm(formula = Grade ~ Hours.on.Readings + Hours.on.Assignments +
    Hours.on.Games + Hours.on.Internet, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.0507 -5.3350  0.4805  6.1131 28.0682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79.19207   0.34936 226.68 <2e-16 ***
Hours.on.Readings 1.31347   0.02384 55.10 <2e-16 ***
Hours.on.Assignments 1.20311   0.02264 53.14 <2e-16 ***
Hours.on.Games -1.15627   0.02237 -51.68 <2e-16 ***
Hours.on.Internet -1.13148   0.02243 -50.45 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.62 on 7995 degrees of freedom
Multiple R-squared:  0.5608, Adjusted R-squared:  0.5606 
F-statistic: 2553 on 4 and 7995 DF, p-value: < 2.2e-16
  
```

$$Y = 79.1920 + 1.31347 X_1 + 1.20311 X_2 + (-1.15627) X_3 + (-1.13148) X_4$$

Whereas :

Y = Grade

X1 = Hours.on.Readings

X2 = Hours.on.Assignments

X3 = Hours.on.Games

X4 = Hours.on.Internet

Explanation of co-efficient and Intercept (explaining effect):

---

$\beta_0 \sim$

Intercept: 79.1920:

The intercept, represented by 79.1920, is the value of Grade when X1, X2, X3, X4 (all the x-var's) is equal to zero. The intercept represents a constant or baseline for grade when all the x var's are at 0.

$\beta_1 \sim$  It is the co-efficient of X1 variable: (Hours.on.Readings)

1.31347 ( $\beta_1$ ) measures the changes in Grade variable for a unit increase of the variable Hours.on.Readings , (Assuming other variables are held constant).

$\beta_2 \sim$  It is the co-efficient of X2 variable: (Hours.on.Assignments)

1.20311 ( $\beta_2$ ) measures the changes in Grade variable for a unit increase of the variable Hours.on.Assignments, (Assuming other variables are held constant).

$\beta_3 \sim$  It is the co-efficient of X3 variable: (Hours.on.Games)

-1.15627 ( $\beta_3$ ) measures the changes in Grade variable for a unit increase of the variable Hours.on.Games. The negative sign suggests a concave-down relationship, meaning the Grade initially increases at a decreasing rate and eventually decreases as X3 increases. (Assuming other variables are held constant).

$\beta_4 \sim$  It is the co-efficient of X4 variable: (Hours.on.Internet)

(-1.13148 ( $\beta_4$ ) measures the changes in Grade variable for a unit increase of the variable Hours.on.Internet. The negative sign suggests a concave-down relationship, meaning the Grade initially increases at a decreasing rate and eventually decreases as X3 increases. (Assuming other variables are held constant).