

HW1

Name: Naga Satya Silpa Annadevara

Student ID: A20517818

Use the data set in case study 2 and utilize R to answer the following questions.

Note: you need to take snapshots of your R coding & outputs, also use your texts to answer the following questions (if necessary). Just like the example given in our slides. Upload a single PDF document as your submission.

1. Load dataset into R and show the column names, example of values, and the size of the data [10]

ANSWER:

Loading data set:

- Case study 2(Case2_clerical.txt) is loaded in R Studio by using the command:
- `data = read.table ("C:/Users/satya/OneDrive/Desktop/Case2_clerical.txt", header = T, sep = "\t")`
- **Note:** I used separator as "\t" , because I noticed that the separator in the text file you have given is not coma(,).
- The separator between the columns in the text file is space. So, we need to use \t as separator.

The screenshot shows the RStudio interface. In the Environment pane, there is a table named 'data' with 52 observations and 9 variables. In the Console pane, the following R code is visible, with the line `> sep = "\t"` highlighted with a red box:

```
R> data = read.table ("C:/Users/satya/OneDrive/Desktop/Case2_clerical.txt", header = T, sep = "\t")
> View(data)
> |
```

Show the column names:

The screenshot shows the RStudio interface. In the top-left pane, the console window displays R code and its output. A red box highlights the command `> names(data)` and its output, which lists the column names: "day", "hours", "mail", "cert", "acc", "change", and "check". In the top-right pane, the environment browser shows a dataset named "data" with 52 observations and 9 variables. Below the RStudio interface is a file explorer window showing various files and folders on the user's desktop.

```
R 4.3.0 - ~/r
> data = read.table ("C:/users/satya/OneDrive/Desktop/case2_clerical.txt", header = T,
+ sep = "\t")
> View(data)
>
>
>
>
> names(data)
[1] "day"      "hours"    "mail"     "cert"     "acc"      "change"   "check"
[8] "misc"     "tickets"
```

Example of values:

The example of values here are : "M" "T"

128, 114etc.,

The screenshot shows the RStudio interface. In the top-left pane, the console window displays R code and its output. A red box highlights the command `> str(data)` and its output, which provides a detailed structure of the data frame, including variable names and their corresponding types and values. In the top-right pane, the environment browser shows a dataset named "data" with 52 observations and 9 variables. Below the RStudio interface is a file explorer window showing various files and folders on the user's desktop.

```
R 4.3.0 - ~/r
> data = read.table ("C:/users/satya/OneDrive/Desktop/case2_clerical.txt", header = T,
+ sep = "\t")
> View(data)
>
>
>
>
> names(data)
[1] "day"      "hours"    "mail"     "cert"     "acc"      "change"   "check"
[8] "misc"     "tickets"
```

```
'data.frame': 52 obs. of 9 variables:
 $ day : chr "M" "T" "W" "Th" ...
 $ hours : num 128 114 147 124 100 ...
 $ mail : int 7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
 $ cert : int 100 110 61 102 45 144 123 78 172 126 ...
 $ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
 $ change : int 235 388 394 457 577 345 326 161 219 287 ...
 $ check : int 644 589 1081 891 537 563 402 495 823 555 ...
 $ misc : int 56 57 59 57 49 64 60 57 62 86 ...
 $ tickets: int 737 1029 830 1468 335 918 335 962 665 577 ...
```

Size of the data: Size of the data can be known by using `dim` function. There are 52 observations/records & 9 variables/columns.

The screenshot shows the RStudio interface with the following details:

- Console:** Displays R code and its output. A red box highlights the command `> dim(data)` and its result `[1] 52 9`.
- Environment:** Shows a data frame named `data` with 52 observations and 9 variables.
- File Browser:** Shows a list of files in the "Home" directory, including R history files, UML models, Microsoft Word documents, and various application icons.

```

R 4.3.0 - ~/~
> data = read.table ("C:/users/satya/OneDrive/Desktop/case2_clerical.txt", header = T,
> sep = "\t")
> View(data)
>
>
>
>
> names(data)
[1] "day"      "hours"    "mail"     "cert"     "acc"      "change"   "check"
[8] "misc"     "tickets"
>
>
>
>
> str(data)
'data.frame': 52 obs. of 9 variables:
 $ day : chr "M" "T" "W" "Th" ...
 $ hours : num 128 114 147 124 100 ...
 $ mail : int 7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
 $ cert : int 100 110 61 102 45 144 123 78 172 126 ...
 $ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
 $ change : int 235 388 398 457 577 345 326 161 219 287 ...
 $ check : int 644 589 1081 891 537 563 402 495 823 555 ...
 $ misc : int 56 57 59 57 49 64 60 57 62 86 ...
 $ tickets: int 737 1029 830 1468 335 918 335 962 665 577 ...
>
> dim(data)
[1] 52 9
>
```

2. Return a list of records with columns `<day, hours, mail, cert>`, where day is Friday, and the number of mails is larger than 7000 [10]

ANSWER: There are zero observations for the above constraints. Used `Subset` function.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Console Terminal Background Jobs

```
> newdata<-subset(data, day=="F" & mail>7000, select=c("day", "hours", "mail","certificate"))
> str(newdata)
'data.frame': 0 obs. of 4 variables:
$ day      : chr
$ hours    : num
$ mail     : int
$ certificate: int
> |
```

Environment History Connections Tutorial

Project: (None)

Data

- data 52 obs. of 9 variables
- newdata 0 obs. of 4 variables
- \$ day : chr
- \$ hours : num
- \$ mail : int
- \$ certificate: int

Files Plots Packages Help Viewer Presentation

3. Change the column name from “cert” to “certificate” [5]

ANSWER:

The column name changed from "cert" to "certificate" using index:

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Console Terminal Background Jobs

```
> data = read.table ("C:/users/satya/OneDrive/Desktop/Case2_clerical.txt", header
T, sep = "\t")
>
> names(data)
[1] "day"   "hours"  "mail"   "cert"   "acc"   "change" "check"  "misc"
"tickets"
>
>
> str(data)
'data.frame': 52 obs. of 9 variables:
$ day      : chr "M" "T" "W" "Th" ...
$ hours    : num 128 114 147 124 100 ...
$ mail     : int 7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
$ cert     : int 100 110 61 102 45 144 123 78 172 126 ...
$ acc       : int 886 962 1342 1153 803 1127 627 748 876 685 ...
$ change   : int 235 388 398 457 577 345 326 161 219 287 ...
$ check    : int 644 589 1081 891 537 563 402 495 823 555 ...
$ misc     : int 56 57 59 57 49 64 60 57 62 86 ...
$ tickets  : int 737 1029 830 1468 335 918 335 962 665 577 ...
>
>
> dim(data)
[1] 52 9
>
>
> names(data)[4]="certificate"
> names(data)
[1] "day"   "hours"  "mail"   "certificate" "acc"
[6] "change" "check"  "misc"   "tickets"
```

Environment History Connections Tutorial

Project: (None)

Data

data 52 obs. of 9 variables

Files Plots Packages Help Viewer Presentation

4. Use descriptive statistics to understand the column “day”. More specifically, return class frequency, and class relative frequency. Also visualize this column by using bar graph (use class relative frequency as y-axis) and pie chart [20]

ANSWER: Installed and loaded library ‘plyr’ to use and call the functions in it and to calculate the Class frequency and Class relative Frequency.

Class frequency (CF) of the column 'day' using count function & CRF is also calculated using R. Here is the screenshot.

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

Console Terminal Background Jobs x

R 4.3.0 - ~/

```
$ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
$ change : int 235 388 398 457 577 345 326 161 219 287 ...
$ check : int 644 589 1081 891 537 563 402 495 823 555 ...
$ misc : int 56 57 59 57 49 64 60 57 62 86 ...
$ tickets : int 737 1029 830 1468 335 918 335 962 665 577 ...
> options(warn = 0)
> install.packages('plyr')
Warning message:
package ‘plyr’ 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/plyr_1.8.8.zip'
Content type 'application/zip' length 1162905 bytes (1.1 MB)
downloaded 1.1 MB

package 'plyr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:/Users/satya/AppData/Local/Temp/Rtmpwv9AOz/downloaded_packages

> library(plyr)

> day

```
[1] "M"  "W"  "Th" "F"  "S"  "M"  "T"  "W"  "Th" "F"  "S"  "M"  "T"  "W"  "Th"
[17] "M"  "S"  "M"  "T"  "W"  "Th" "F"  "S"  "M"  "T"  "W"  "Th" "F"  "S"  "M"  "T"
[33] "W"  "Th" "F"  "S"  "M"  "T"  "W"  "Th" "F"  "S"  "M"  "T"  "W"  "Th" "E"  "S"
[49] "M"  "T"  "W"  "Th"
```

> count(day)

x	freq	
1	F	8
2	M	9
3	S	8
4	T	9
5	Th	9
6	W	9

> table(day)/nrow(data)

day

	F	M	S	T	Th	W
0.1538462	0.1730769	0.1538462	0.1730769	0.1730769	0.1730769	0.1730769

> |

Environment History Connections Tutorial

R Global Environment

Data

- cf 6 obs. of 2 variables
- data 52 obs. of 9 variables
- h List of 6
- newdata 0 obs. of 4 variables

Values

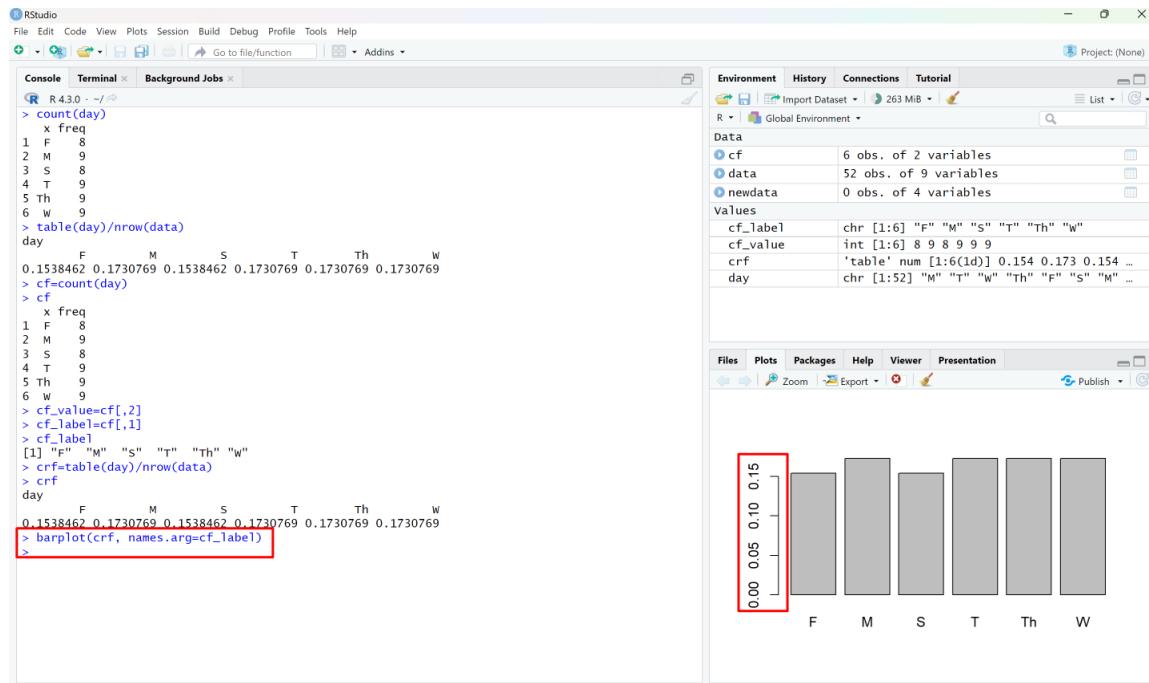
certificate	int [1:52] 100 110 61 102 45 144 123 78 172 ...
cf_label	chr [1:6] "F" "M" "S" "T" "Th" "W"
cf_value	int [1:6] 8 9 8 9 9 9
crf	'table' num [1:6](1d) 0.154 0.173 0.154 0.17 ...
day	chr [1:52] "M" "T" "W" "Th" "F" "S" "M" "T" ...
misc	int [1:52] 56 57 59 57 49 64 60 57 62 86 ...
tickets	int [1:52] 737 1029 830 1468 335 918 335 962 ...
xfit	num [1:40] 126 167 208 249 290 ...
yfit	num [1:40] 1.09 1.3 1.53 1.79 2.06 ...

Files Plots Packages Help Viewer Presentation

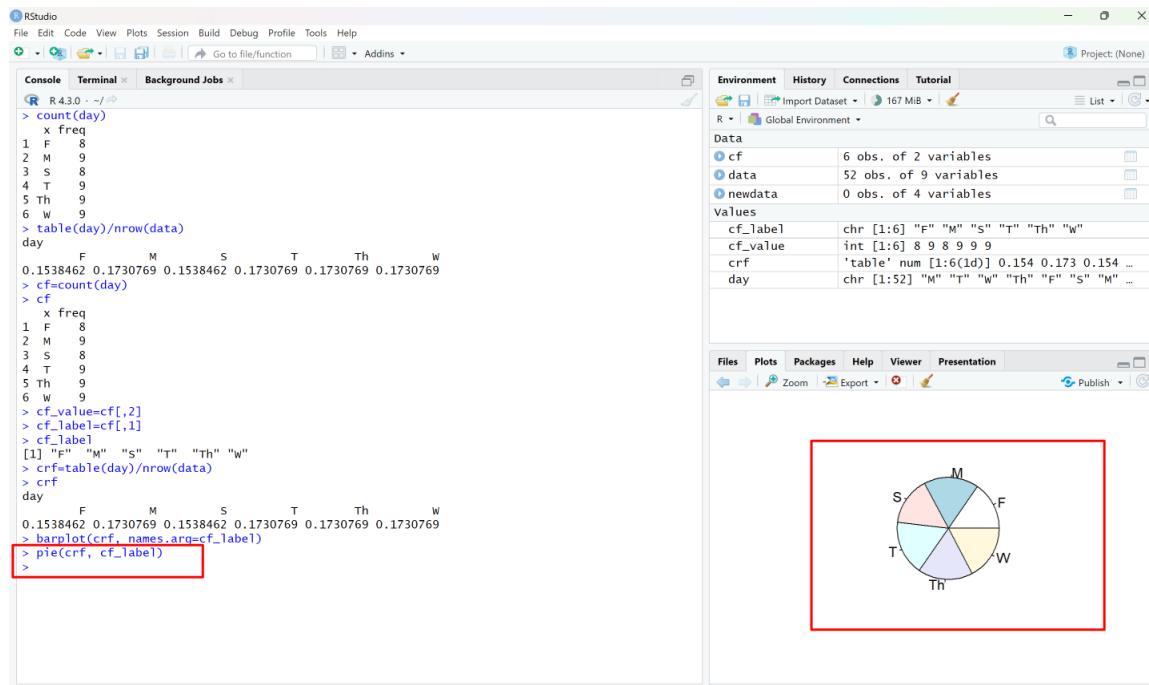
Class relative frequency of the column 'day': (CRF=CF/n) (manual calculation by using formula)

- CRF of F = CF/n = 8/52 = 0.153846
 - CRF of M = CF/n = 9/52 = 0.1730769
 - CRF of S = CF/n = 8/52 = 0.153846
 - CRF of T = CF/n = 9/52 = 0.1730769
 - CRF of Th = CF/n = 9/52 = 0.1730769
 - CRF of W = CF/n = 9/52 = 0.1730769

Visualizing the 'day' column by bar-graph using class relative frequency on y-axis:



Visualizing the 'day' column by pie-chart using class relative frequency:



5. Use descriptive statistics to understand the column “certificate”. More specifically, we want to get q1, q2, q3, average value, variance value. Also, visualize this variable by using histogram, and interpret your histogram [20]

ANSWER:

To calculate the descriptive statistics asked, library ‘psych’ is installed and loaded. Q1, Q2, Q3, average value and variance value can be calculated by using 2 functions.

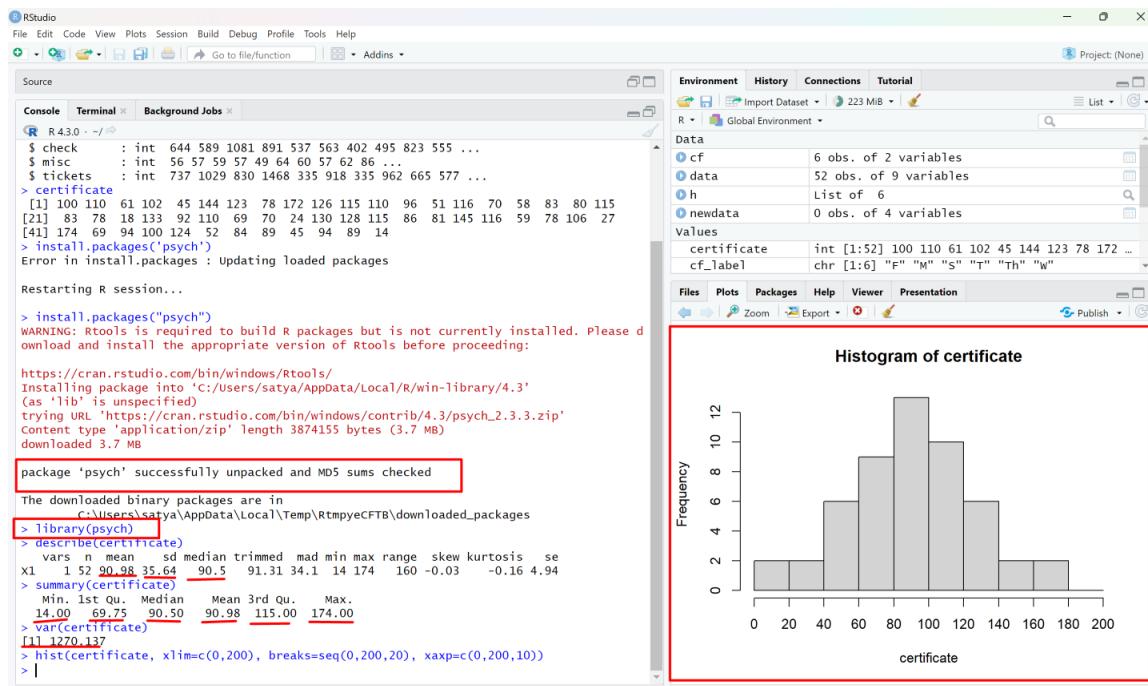
1. `describe()`

2. `summary()`

3. `var()`

The variable ‘certificate’ can be visualized by histogram by using `hist()` function.

Here is the screenshot of the r coding calculating the q1, q2, q3, average value(mean) and variance value and the histogram of the variable ‘certificate’:



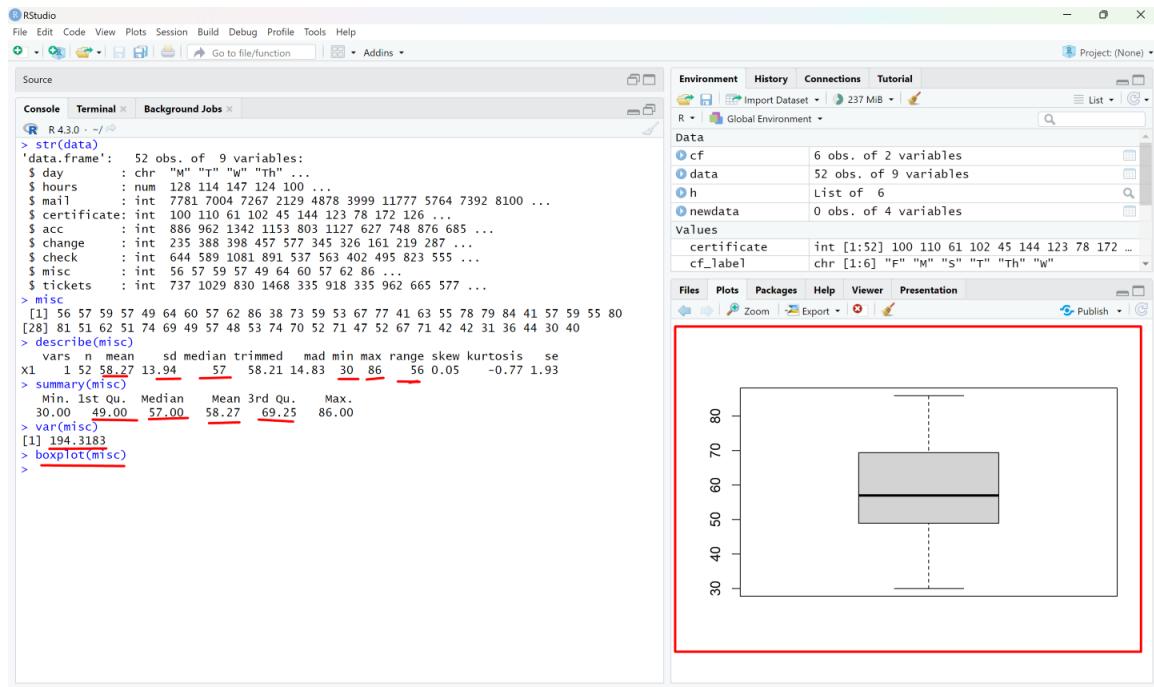
Interpretation of the histogram of certificate: (Analyzing the shape by skewness and outliers)

- **Distribution:** The distribution is “Normal distribution/ Symmetric”. There is no skew in data. Hence the data is evenly distributed around the center.
- **Variance:** Variance is nothing but the variation or the spread of data. The variance is large in this data.
Calculation of variance. = 1270.137

- **Standard deviation:** 35.64. (it's the square root of variance)
- **Potential Outliers:** There are no potential outliers in this distribution of data.
- **Minimum value:** 14
- **Maximum value:** 174
- **1st Quartile(Q1):** 69.75
- **2nd Quartile(Q2):** 90.50 (Note: Q2 is also called Median)
- **3rd Quartile(Q3):** 115
- **Range:** 160 (max-min)
- **Mean/ Average value of the data:** 90.98.
- **Median:** 90.50 (Note: Median is also called Q2)

6. Use descriptive statistics to understand the column “misc”. Visualize it by using boxplot, and interpret your boxplot [20]

ANSWER:



Interpretation of the boxplot of the variable ‘misc’:

- **The distribution:** The box-plot follows “Normal distribution/ Symmetric”. Hence the data is evenly distributed around the center. The median is in the middle.
- **Skewness:** There is no skewness in the boxplot. Because the median is exactly in the middle. So, it follows Normal distribution.
- **Potential Outliers:** There are no potential outliers.
- **The variance.: Variance is nothing but the variation or the spread of data. The variance is small in the boxplot. The data is less**
Calculation of variance: 194.3183

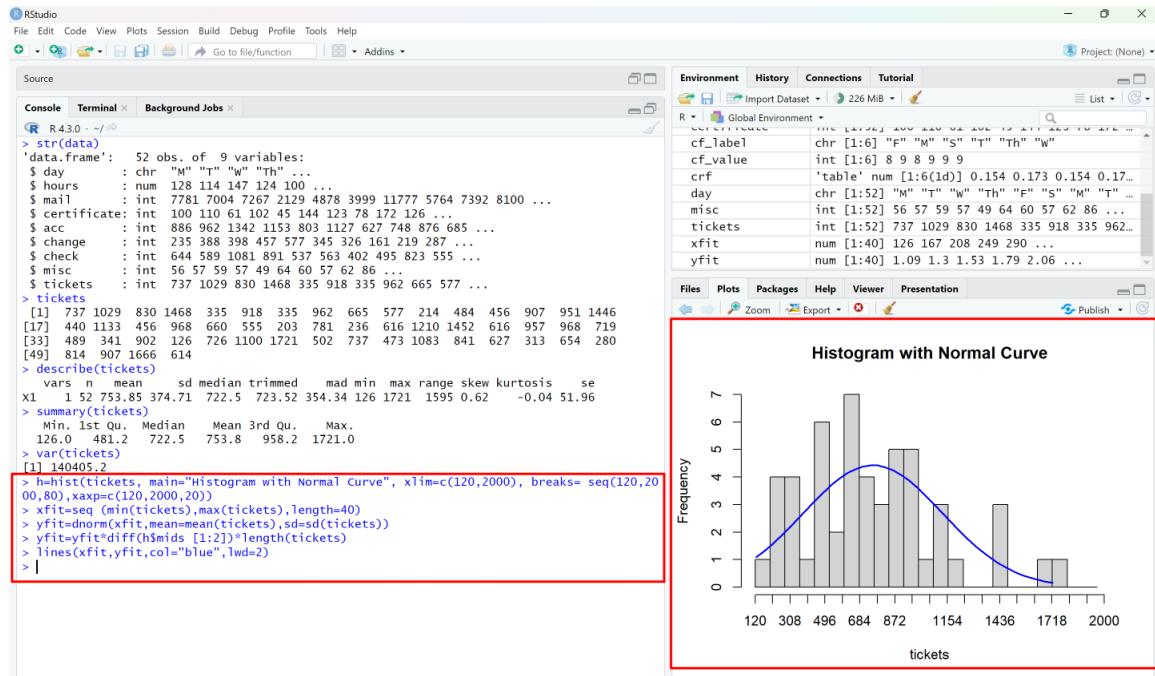
- **Standard deviation:** 13.94 (it's the square root of variance)
- **Minimum value:** 30
- **Maximum value:** 86
- **1st Quartile(Q1):** 49
- **2nd Quartile(Q2):** 57 (Note: Q2 is also called Median)
- **3rd Quartile(Q3):** 69.25
- **Range:** (maximum value – minimum value) = 56
- **Mean value/ Average value of the data:** 58.27
- **Median value:** 57 (Note: Median is also called Q2)

7. Visualize the column “tickets” by using probability curve, interpret it [15]

ANSWER:

To draw the probability curve, first we need to draw the histogram of the variable ‘tickets’ by using ‘hist’ function.

The process is as follows shown in the screenshot:



Interpretation of the distribution curve:

- **The distribution:** Its almost **Normal/Symmetric distribution** where the data is evenly distributed around the center. But the distribution is slightly skewed towards right.
 - ✓ Variable tickets follows normal distribution: $\text{tickets} \sim N(\text{mean}, \text{variance})$
- **Skewness:** There is a slight rightly skewed/positive skewed distribution

- **Potential outliers:** There is 1 outlier towards the maximum value (1721)
 - The maximum value could be the outlier here, according to the formula: $Q3 + 1.5(IQR)$ which calculates the upper boundary. Any data point above upper boundary / lower boundary falls under outliers. (where IQR = Interquartile Range)
 - Here the upper boundary is 1673.7 according to the formula.
 - So, the maximum value is 1721 above the upper boundary 1673.7, which is considered as an outlier in this distribution.
- **Variance:** Variance is nothing but the variation or the spread of data. The variance is large in this data.
 - Calculation of variance: 140405.2
- **Standard deviation:** 374.71 (it's the square root of variance)
- **Minimum value:** 126
- **Maximum value:** 1721
- **Range:** 1595 (max-min)
- **1st Quartile(Q1):** 481.2
- **2nd Quartile(Q2):** 722.5 (Note: Q2 is also called Median)
- **3rd Quartile(Q3):** 958.2
- **Mean value/ Average value of the data:** 753.8
- **Median value:** 722.5 (Note: Median is also called Q2)

Homework 2

Your Name: Naga Satya Silpa Annadevara

Student ID: A20517818

1. (35 points) Manually solve the problem below (do not use R):

Note, if you need either z value or t value, you can find them by using this tool:

http://www.mathcracker.com/z_critical_values.php

http://www.mathcracker.com/t_critical_values.php

<https://www.socscistatistics.com/pvalues/>

A bank branch located in a commercial district of a city has the business objective of improving the process for serving customers during the noon to 1 PM (lunch period). The waiting time (defined as the time the customer enters the line until he or she reaches the teller window) of a random sample of 15 customers is collected, and the results are organized and stored as below:

4.21, 5.55, 3.02, 5.13, 4.77, 2.34, 3.54, 3.20

4.50, 6.10, 0.38, 5.12, 6.46, 6.19, 3.79

- Calculate the mean and standard deviation, and find q1, q3 from the values above. Is the distribution symmetric? Why? [10]

ANSWER:

- The sample size: $n = 15$
- Sample mean = $\sum_{i=0}^n xi/n$ (dividing the sum of all values in a data set by the number of values.)
 - Mean = $(4.21 + 5.55 + 3.02 + 5.13 + 4.77 + 2.34 + 3.54 + 3.20 + 4.50 + 6.10 + 0.38 + 5.12 + 6.46 + 6.19 + 3.79) / 15 = 64.3/15 = 4.28$
 - Mean ($\bar{x} = 4.28$)
- Standard deviation (S) = square root of the variance. ($\sqrt{S^2}$. Where is S^2 variance.)
- So, to calculate the Standard deviation, first we need to calculate the variance of the sample (S^2)
 - Variance = $S^2 = \sum_{i=1}^n \frac{(xi - \bar{x})^2}{n-1}$
 - First, we need to calculate the mean (mean is already calculated above i.e., 4.28)
 1. Subtract the mean from each data point and square the result:
 - $(4.21 - 4.28)^2 = 0.0049$
 - $(5.55 - 4.28)^2 = 1.6129$

- $(3.02 - 4.28)^2 = 1.5876$
- $(5.13 - 4.28)^2 = 0.7225$
- $(4.77 - 4.28)^2 = 0.2401$
- $(2.34 - 4.28)^2 = 3.7636$
- $(3.54 - 4.28)^2 = 0.5476$
- $(3.20 - 4.28)^2 = 1.1664$
- $(4.50 - 4.28)^2 = 0.0484$
- $(6.10 - 4.28)^2 = 3.3124$
- $(0.38 - 4.28)^2 = 15.21$
- $(5.12 - 4.28)^2 = 0.7056$
- $(6.46 - 4.28)^2 = 4.7524$
- $(6.19 - 4.28)^2 = 3.6481$
- $(3.79 - 4.28)^2 = 0.2401$

2. Calculate the average of the squared differences:

$$(0.0049 + 1.6129 + 1.5876 + 0.7225 + 0.2401 + 3.7636 + 0.5476 + 1.1664 + 0.0484 + 3.3124 + 15.21 + 0.7056 + 4.7524 + 3.6481 + 0.2401) \div 15 - 1 \\ = 37.5626 \div 14 = 2.68$$

Therefore, The variance of the given sample data set: $S^2 = 2.68$

- The standard deviation = square root of the variance ($\sqrt{S^2}$)

$$\sqrt{(2.68)} = 1.63$$

- To calculate Q1 , Q2 , Q3 , First we need to re-arrange the sample dataset in Ascending order:
 - 0.38, 2.34, 3.02, 3.20, 3.54, 3.79, 4.21, 4.50, 4.77, 5.12, 5.13, 5.55, 6.10, 6.19, 6.46
 - Q1: (Lower quartile) = 3.20
 - Q2: (Median/Middle value) = 4.50
 - Q3: (Upper quartile) = 5.55
- Is the distribution symmetric?
 - Based on the central limit theorem, If the sample size(n) is less than 30, then it will not follow normal distribution.
 - Here the sample size (n) = 15 which is less than 30, The distribution is not symmetric. It follows t-distribution.
 - To tell whether it is symmetric or not, we can say if the mean and median are equal, then its symmetric. Here the mean is 4.28 & the median is 4.50.
 - Mean and median are not equal, hence we can say the distribution is not symmetric and it is skewed.
 - To say its skewness in detail, we can use a formula:

$$\text{Skewness} = (3 * (\text{Mean} - \text{Median})) / \text{Standard Deviation} \\ = (3 * (4.28 - 4.50)) / 1.63 \\ = (3 * -0.22) / 1.63$$

$$= -0.66 / 1.63 \\ = -0.40$$

Hence, we can say it is slightly left-skewed/ negative-skewed.

- b) As a customer walks into the branch office during the lunch period. She asks the branch manager how long she can expect to wait. If you are the manager, answer this question by using 95% confidence. [10]

ANSWER:

To answer this question, we need to follow the below steps one by one:

1. Calculate sample statistics such as ‘sample mean’ and ‘sample standard deviation’ .

Sample mean: already calculated in the above question.

$$\bar{y} = 4.28$$

Sample standard deviation: already calculated in the above question.

$$S = 1.63$$

2. As sample size (n = 15) is less than 30, it follows t- distribution, hence we need to use t-value in the margin of error.

3. Produce confidence interval [a,b] by using the formula :

$a = \text{sample estimate} - \text{margin of error}$ (sample estimate i.e, sample mean)

$$\bar{y} - t \frac{\alpha}{2} \left(\frac{S}{\sqrt{n}} \right)$$

$b = \text{sample estimate} + \text{margin of error}$

$$\bar{y} + t \frac{\alpha}{2} \left(\frac{S}{\sqrt{n}} \right)$$

- To get the t- value we need to have 2 inputs:

$$1. \alpha = 1 - \text{confidence level} \\ = 1 - 95\%$$

$$\alpha = 0.05$$

$$2. \text{degree of freedom (df)} = n - 1 \\ = 15 - 1 \\ = 14$$

- http://www.mathcracker.com/t_critical_values.php

from this given tool, the t- value is calculated from the t- table and observed as
 $t \frac{\alpha}{2} = 1.761$

Now,

$$\begin{aligned}- \text{ Let's calculate } a &= \bar{y} - t \frac{\alpha}{2} \left(\frac{s}{\sqrt{n}} \right) \\&= 4.28 - 1.761 \left(\frac{1.63}{\sqrt{15}} \right) \\&= 4.28 - 1.761 \left(\frac{1.63}{3.87} \right) \\&= 4.28 - 1.761 (0.421) \\&= 4.28 - 0.741 \quad (0.741 \text{ is the margin of error}) \\&= 3.53\end{aligned}$$

$$\begin{aligned}- \text{ Let's calculate } b &= \bar{y} + t \frac{\alpha}{2} \left(\frac{s}{\sqrt{n}} \right) \\&= 4.28 + 1.761 \left(\frac{1.63}{\sqrt{15}} \right) \\&= 4.28 + 1.761 \left(\frac{1.63}{3.87} \right) \\&= 4.28 + 1.761 (0.421) \\&= 4.28 + 0.741 \\&= 5.02\end{aligned}$$

Conclusion: I have 95% confidence to say that the average waiting time (population mean) will fall in the interval [3.53, 5.02].

- c) We were told the average waiting minute will be 5 minutes. But we think it could be more than 5 minutes. By using 90% as confidence level, validate the hypothesis. Show your steps and calculations [15]

ANSWER:

To answer this question, we need to follow the below steps one by one:

1. Collect sample data & Calculate sample statistics such as ‘sample mean’ and ‘sample standard deviation’:

- ✓ $n = 15$
- ✓ Sample mean $\bar{X} = 4.28$
- ✓ Sample standard deviation = $S = 1.63$

2. State null hypothesis and alternative hypothesis:

$$\begin{aligned}H_0 : \mu &= 5 \\H_a : \mu &> 5\end{aligned}$$

3. Based on H_a , decide it is one-tailed or two-tailed test:

- Based on alternative hypothesis H_a , it is one-tail test.
- As alternative hypothesis is larger than μ , the tail is on the right side. That means the reject region is always on the right.

4. Specify the desired level of significance:

$$\begin{aligned}\alpha &= 1 - \text{confidence level.} \\&= 1 - 90\% \\&= 1 - 0.90 \\&= 0.10\end{aligned}$$

5. Determine the appropriate technique:

- σ is unknown and $n = 15$ which is less than 30.
- As sample size is less, this is a t- test.
- Hence use t- statistics to make the conclusion.

6. Calculate critical value:

- Since it is t- test, to calculate the $t_{\alpha / 2}$ - critical value, we need to have 2 inputs.
 - $\alpha = 0.10$
 - df (degree of freedom) = $n-1 = 15 - 1 = 14$

-
- with the α value and df value available, the t- critical value/ t α value is calculated from the t- table and observed as follows:

$$t_c \text{ or } t_\alpha = 1.345$$

7. Calculate the test statistic value (t-value) by using the formula: (t-score)

$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{4.28 - 5}{\frac{1.63}{\sqrt{15}}}$$

$$= \frac{-0.72}{\frac{1.63}{3.87}}$$

$$= \frac{-0.72}{0.42} \\ = -1.714$$

8. Compare the test- statistic value with the critical value / t α value. 6. Is the test statistic in the rejection region?

$$t_{STAT} < t_c \text{ or } t_\alpha$$

$$-1.714 < 1.345$$

The test statistic is in the non-rejection region.

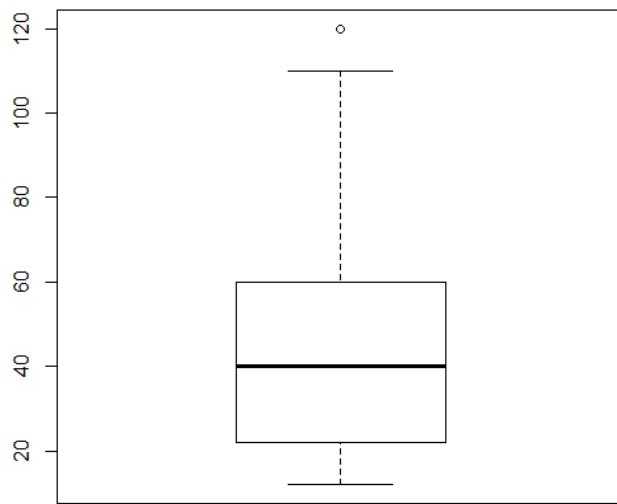
9. Draw the conclusion based on test statistic value:

- Since the test statistic falls in the non- rejection region, we fail to reject the null hypothesis. Therefore, we do not have enough evidence to conclude that the average waiting minute is more than 5 minutes with a 90% confidence level.
- In summary, based on the given data and calculations, we do not have enough evidence to support the hypothesis that the average waiting minute is more than 5 minutes with a 90% confidence level.

2. (55 points) Chicago Ventra Transit Card can be used on both CTA bus, metro and Pace buses. We are going to explore a resident's average monthly cost on CTA transportations. In this case, we performed a survey, and collect monthly cost on CTA transits from 30 people, their monthly cost can be listed as follows:

12, 12, 12, 15, 24, 35, 14, 12, 120, 55, 45, 30, 40, 40, 40, 60, 60, 40, 50, 22, 36, 28, 21, 50, 39, 60, 90, 100, 110, 100

- 1). [5] To further understand the distribution, we draw a boxplot as follows. Interpret the box plot.



ANSWER:

Interpretation of the boxplot:

- **The distribution:** The distribution is not symmetric; the data is not evenly distributed. The median is not exactly in the middle. From the visualization, the box plot clearly shows that it is positively skewed.
- **Skewness:** The box plot clearly follows a positive skew from the visualization.

- To say its skewness in detail, we can use a formula:
$$\text{Skewness} = (3 * (\text{Mean} - \text{Median})) / \text{Standard Deviation}$$
$$= (3 * (45.73 - 40)) / 30.18$$
$$= (3 * 5.73) / 30.18$$
$$= 17.19 / 30.18$$
$$= 0.56$$

Hence, we can say it is positively skewed / right skewed.

- **Potential Outliers:** From the visualization we can clearly see that there is a potential outlier towards the value 120.
 - To confirm the potential outliers, we need to follow the below steps:

- Calculate IQR (Interquartile range):
 - $Q3 - Q1 = 60 - 22 = 38$
- Upper bound value = $Q3 + 1.5 * \text{IQR} = 60 + (1.5 * 38) = 60 + 57 = 117$
- Lower bound value = $Q1 - 1.5 * \text{IQR} = 22 - (1.5 * 38) = 22 - 57 = -35$
- ❖ Any data values not falling between upper bound value and lower bound value are suspected to be potential outliers. Here 120 is the potential outlier that is not falling between the upper and lower bound values.
- **The variance.:** Variance is nothing but the variation or the spread of data. The variance in the boxplot is as follows.
 - $\text{Variance} = S^2 = \sum_{i=1}^n \frac{(xi - \bar{x})^2}{n-1}$
Hence the variance from the above formula: $S^2 = 910.83$
- **Standard deviation:** 30.18 (it's the square root of variance- $(\sqrt{S^2})$)
- **Minimum value:** 12
- **Maximum value:** 110 (Since 120 is the outlier, it cannot be considered as the max value)
- **1st Quartile(Q1):** 22
- **2nd Quartile(Q2):** 40 (Note: Q2 is also called Median)
- **3rd Quartile(Q3):** 60
- **Range:** (maximum value – minimum value) excluding the outliers.
 $= 110 - 12 = 98$
- **Mean value/ Average value of the data:** mean = $\sum_{i=0}^n xi/n$
 (Dividing the sum of all values in a data set by the number of values.)
 $= 45.733$
- **Median value:** Median is the middle value. If the size of the data is odd there is 1 median, If the data size is even, then the sum of middle 2 numbers divided by 2.

Fist we need to arrange the data in Ascending order:

12,12,12,12,14,15,21,22,24,28,30,35,36,39,40,40,40,40,45,50,50,55,60,60,60,90,100,100,110,120

Hence median = $(40+40)/2 = 40$ (Note: Median is also called Q2)

2). [15] Use the sample statistics to estimate the average monthly cost on CTA transits by Chicago residents by using 95% as the confidence level. Assume that we know the population variance is 4. Use R to solve the problem. paste your snapshot of R coding and outputs, also deliver your conclusions.

ANSWER:

Loaded the sample data of CTA and Trains in r-studio:

The screenshot shows the RStudio interface. In the Environment pane, there is a data frame named "data" with 30 observations and 2 variables. The variables are "CTA" and "Trains". The data looks like this:

	CTA	Trains
18	40	60
19	50	60
20	22	80
21	36	40
22	28	25
23	21	25
24	50	40
25	39	25
26	60	25
27	90	120
28	100	120
29	110	120
30	100	100

Below the data frame, the console pane shows the R code used to load the data:

```
R 4.3.0 · ~/r
> data = read.table(file = "C:/Users/satya/OneDrive/Desktop/cta_trains.txt", header = T, f
ill = TRUE, sep = ",")
> view(data)
> data=na.omit(data)
>
```

Collect sample statistics such as sample mean and standard deviation of CTA:

The screenshot shows the RStudio interface. In the Environment pane, there is a data frame named "data" with 30 observations and 2 variables. The variables are "CTA" and "Trains". The data looks like this:

	CTA	Trains
[1]	12	12
[2]	12	12
[3]	15	24
[4]	35	35
[5]	14	12
[6]	120	120
[7]	55	55
[8]	45	40
[9]	30	40
[10]	40	40
[11]	60	60
[12]	60	60
[13]	40	40
[14]	50	50
[15]	22	22

Below the data frame, the console pane shows the R code used to install the "psych" package and then run the "describe" function on the "CTA" variable. The output of the "describe" function is highlighted with a red box.

```
R 4.3.0 · ~/r
> CTA= data$CTA
> CTA
[1] 12 12 12 15 24 35 14 12 120 55 45 30 40 40 60 60 40 50 22
[21] 36 28 21 50 39 60 90 100 110 100
> install.packages('psych')
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/psych_2.3.3.zip'
Content type 'application/zip' length 3873609 bytes (3.7 MB)
downloaded 3.7 MB

package 'psych' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/satya/AppData/Local/Temp/RtmpQDMzGx/downloaded_packages
> library(psych)
> describe(CTA)
vars n  mean   sd median trimmed  mad min max range skew kurtosis    se
X1 1 30 45.73 30.7    40  41.92 27.43 12 120 108 0.95 -0.11 5.6
> summary(CTA)
   Min. 1st Qu. Median  Mean 3rd Qu. Max.
12.00 22.50 40.00 45.73 58.75 120.00
>
```

Produce confidence interval: To Produce confidence interval, “Rmisc” package was installed & called to use the function CI in it.

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ Go to file/function | Addins *
HW_2_RStudio.R
Console Terminal Background Jobs
R 4.3.0 · ~/R
Content type: application/zip length: 507,009 bytes (507 MB)
downloaded 3.7 MB

package 'psych' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\users\satya\AppData\Local\Temp\Rtmpsu8xqi\downloaded_packages
> library(psych)
> describe(CTA)
  vars   n   mean    sd median trimmed   mad min max range skew kurtosis se
X1  1 30 45.73 30.7   40  41.92 27.43 12 120 108 0.95 -0.11 5.6
> summary(CTA)
   Min. 1st Qu. Median 3rd Qu. Max.
12.00 22.50 40.00 45.73 58.75 120.00
> install.packages("Rmisc")
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/Rmisc_1.5.1.zip'
Content type: application/zip length: 52463 bytes (51 KB)
downloaded 51 KB

package 'Rmisc' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\users\satya\AppData\Local\Temp\Rtmpsu8xqi\downloaded_packages
> library(Rmisc)
Loading required package: lattice
Loading required package: plyr
> CI(CTA, ci=0.95)
  upper   mean   lower
57.19599 45.73333 34.27068
>

```

Conclusion: I have 95% confidence to say that the average monthly cost on CTA transits by Chicago residents (population mean) will fall in the interval [34.27, 57.19].

4). In addition, they can use Chicago metro trains in their daily life. We ask the same group of 30 users to use Chicago metro trains only and record their monthly cost on trains. In this case, we get two groups of data as follows. We display it as two tables, since it is not able to put them on a single table. For each table, row 1 is the monthly cost by using CTA for each person, row2 is the monthly cost by using train for each person. Each column contains the costs by a same person but use different transportation (CTA vs Metro Trains) . Assume that we know they have the same population variance of 4.

user	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
row1	12	12	12	15	24	35	14	12	120	55	45	30	40	40	40
row2	10	16	13	14	28	41	16	10	80	40	75	25	41	29	40

user	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
row1	60	60	40	50	22	36	28	21	50	39	60	90	100	110	100
row2	50	50	60	60	80	40	25	25	40	25	25	120	120	120	100

It is told that there are no differences if they are going to use the CTA or Metro trains. However, we believe using CTA is more expensive than using Metro trains. We are going to use hypothesis testing to examine whether the costs by two different means should be the same or not. Assume we use 95% confidence level.

4.1), [15] write down your null and alternative hypothesis, and tell me is it a two-tailed or one-tailed test, and it is two independent or paired samples, why?

ANSWER:

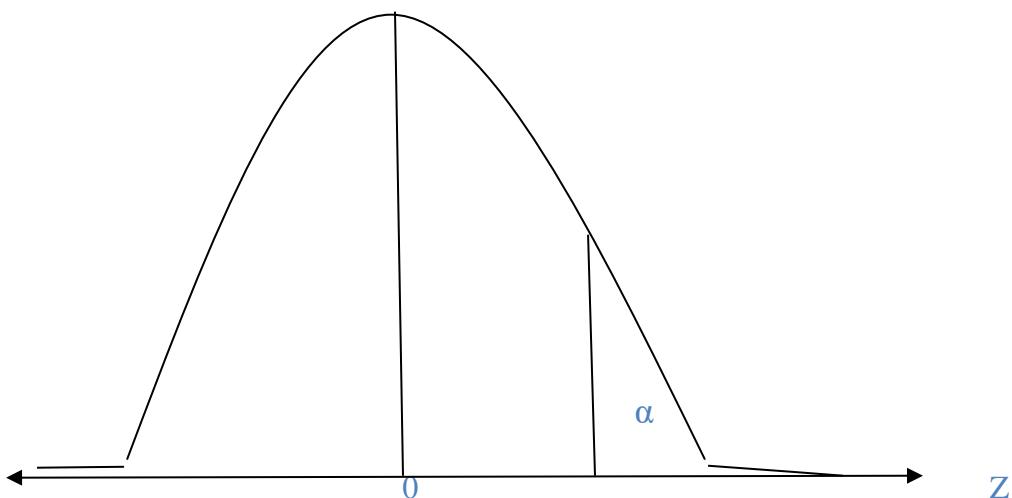
-It is a 2- sample hypothesis testing. Sample 1: CTA (μ_1)
Sample 2: Metro trains. (μ_2)

Null & alternative Hypothesis representation:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

- It is one- tailed hypothesis test, because based on the alternative hypothesis (H_1 or H_a), it says that, CTA is more expensive than Metro trains. So, we can observe there is only 1 reject region (Z_α) towards the right. Z test because the sample size is 30 ($n = 30$) and the population variance is known.



- It is two – paired samples because the user (same group) was measured for the monthly cost twice by CTA & Metro trains. So, the 2 samples are paired & dependent on each other.

4.2), [20] Perform hypothesis testing to tell whether the costs by two different means are the same or not based on 95% confidence level, by using R. Again, give the R coding, snapshot, outputs, deliver your conclusions by referring to/explaining the outputs.

ANSWER: It is 2-sample paired hypothesis testing.

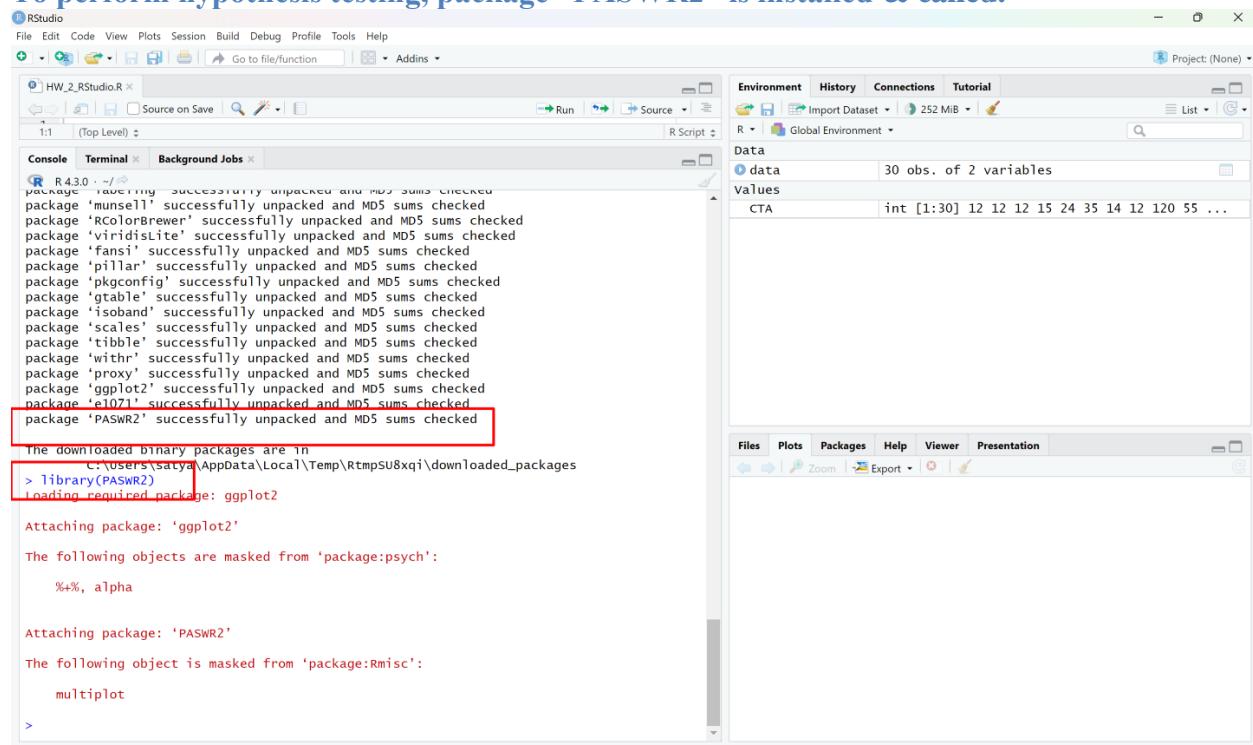
As it is 2- paired sample hypothesis, we can convert it into 1-sample and perform hypothesis testing. For that we need to calculate the difference between CTA & Trains columns & represent null and alternative hypothesis in 1- sample as follows:

$$\mu_{diff} = \mu_1 - \mu_2$$

$$H_0 : \mu_{diff} = 0$$

$$H_1 : \mu_{diff} > 0$$

To perform hypothesis testing, package “PASWR2” is installed & called:



The screenshot shows the RStudio interface with the following details:

- Console:** Displays the command line output of the R session. It shows the download of binary packages from a temporary directory and the successful installation of the 'PASWR2' package.
- Environment:** Shows the global environment with a variable 'data' containing 30 observations of 2 variables. One of the variables is 'CTA'.
- Plots:** No plots are currently displayed.
- Packages:** Shows the loaded packages: 'ggplot2', 'psych', and 'PASWR2'.

```
R 4.3.0 -->
package 'tauri' successfully unpacked and MD5 sums checked
package 'munsell' successfully unpacked and MD5 sums checked
package 'ColorBrewer' successfully unpacked and MD5 sums checked
package 'viridisLite' successfully unpacked and MD5 sums checked
package 'fansi' successfully unpacked and MD5 sums checked
package 'pillar' successfully unpacked and MD5 sums checked
package 'pkgconfig' successfully unpacked and MD5 sums checked
package 'gtable' successfully unpacked and MD5 sums checked
package 'isoband' successfully unpacked and MD5 sums checked
package 'scales' successfully unpacked and MD5 sums checked
package 'tibble' successfully unpacked and MD5 sums checked
package 'withr' successfully unpacked and MD5 sums checked
package 'proxy' successfully unpacked and MD5 sums checked
package 'ggplot2' successfully unpacked and MD5 sums checked
package 'e1071' successfully unpacked and MD5 sums checked
package 'PASWR2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\users\satya\AppData\Local\Temp\Rtmpsu8xqi\downloaded_packages
> library(PASWR2)
Loading required package: ggplot2

Attaching package: 'ggplot2'

The following objects are masked from 'package:psych':
  %++, alpha

Attaching package: 'PASWR2'

The following object is masked from 'package:Rmisc':
  multiplot

>
```

Now, let's perform hypothesis testing: as the sample size is 30 (n=30) which is large, we do z-test. We converted the 2-paired sample to 1-sample by calculating the difference between 2 columns.

Now we have only one column (1-sample) which is the ‘diff’ column. And we can perform hypothesis testing on that one ‘diff’ column,

Note: diff = CTA - Trains

The screenshot shows the RStudio interface. The console pane displays R code and its output. A red box highlights the following code and its output:

```

> CTA
[1] 12 12 12 15 24 35 14 12 120 55 45 30 40 40 40 60 60 40 50 22
[21] 36 28 21 50 39 60 90 100 110 100
> Trains = data$Trains
> Trains
[1] 10 16 13 14 28 41 16 10 80 40 75 25 41 29 40 50 50 60 60 80
[21] 40 25 25 40 25 25 120 120 120 100
> diff = mCTA-mTrains
Error: object 'mCTA' not found
> diff = CTA-Trains
> z.test(diff,NULL,alternative = "two.sided",mu = 0,sigma.x= sd(diff),sigma.y=NULL,conf.level=0.95)

One Sample z-test

data: diff
z = -0.45433, p-value = 0.6496
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-8.148086 5.081419
sample estimates:
mean of x
-1.533333

```

The environment pane shows the global environment with three variables: CTA, diff, and Trains.

Conclusion:

Based on the paired z-test performed at a 95% confidence level, we fail to reject the null hypothesis. The p-value is 0.6496, which is greater than the significance level of 0.05. Therefore, we do not have sufficient evidence to conclude that there is a significant difference in the monthly costs between using CTA and metro trains. The 95% confidence interval for the difference in means is -8.1480 to 5.0814, which includes zero. This further supports the conclusion that the means of the two transportation methods are likely to be equal.

Hence, we do not reject the null hypothesis based on 95% confidence level.

3. (10 points) The z-test and t-test we used for hypothesis testing (including both one-sample and two-sample hypothesis testing) are also known as "Parametric Test". Alternatively, there are other statistical tests which can be used as alternatives, and they are called "Non-Parametric Test". Search online or find other learning materials to answer the questions below.

1). (5 points) what are the differences between Parametric Test and Non-Parametric Test?

ANSWER:

Basis for comparison	Parametric Test	Non-Parametric Test
1. Definition	The parametric test is a hypothesis test that offers generalizations for claiming something about the parent population's mean. A commonly employed t-test in this context is one that is based on Student's t-statistic.	The nonparametric test is characterized as a hypothesis test without underlying assumptions, i.e., it does not call for the population's distribution to be represented by parameters.
2. Meaning	A parametric test is a statistical test in which assumptions about a population parameter are made.	Non-parametric test is a statistical test that is used when there are non-metric independent variables.
3. Population Information	Completely known	Not known
4. Applicability	Only on variables.	Both variables and attributes
5. Central tendency measurement	In general, the measure of central tendency in the parametric test is mean,	<ul style="list-style-type: none"> - While in the case of the nonparametric test is median. - As a result, it is also known as the distribution-free test.
6. Correlation test	The parametric test makes use of Pearson's coefficient of correlation to assess the degree of relationship between two quantitative variables.	while spearman's rank correlation is used in the nonparametric test.
7. Basis of test statistic	Distribution	Arbitrary

2). (5 points) what are the requirements/conditions to use z-test or t-test? If our data does not meet these requirements, which non-parametric test is the alternative for hypothesis testing?

ANSWER:

Requirements to use Z-test / t-test:

- 1) If the sample size is large ($n \geq 30$) & population standard deviation is known => Z-test
- 2) If Sample size is small ($n < 30$) and population STD is unknown => t-test
- 3) If Sample size is large ($n \geq 30$) and population STD is unknown => t-test
- 4) Either a Z-test or a T-test can be used to compare the means of two independent samples, such as two separate groups.

- 5) A paired T-test is often employed when we need to compare the means of paired samples (for instance, before and after measurements).
- 6) Df (degree of freedom must be known to carry t-test
- You can use non-parametric tests for hypothesis testing if your data does not fulfill these conditions or if you would prefer a non-parametric option. Non-parametric tests don't rely on presumptions about how the data are distributed.

Equivalent alternative non-parametric tests for hypothesis testing:

Parametric Test	Non- Parametric Test
1. Independent Sample t Test	Mann-Whitney U test: Utilizing this test, two independent groups' medians are compared. The independent samples t-test can be replaced with this method.
2. Paired samples t test	Wilcoxon signed Rank test: This test is used to determine if the median of one group deviates from a value that is hypothesized or to compare the medians of two groups that are related. It serves as a substitute for the paired samples t-test.
3. One way Analysis of Variance (ANOVA)	Kruskal Wallis Test: To compare the medians of three or more independent groups, apply this test. The one-way analysis of variance (ANOVA) test can be substituted with this method
4. One-way repeated measures Analysis of Variance	Friedman's ANOVA: The medians of three or more related groups are compared using this test. This test serves as an alternative to the repeated measures ANOVA.

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 a dataset, printing its structure, and creating a subset of the data excluding columns for ID, Nationality, Gender, Degree, and GradeLetter. The last few lines of the output show the structure of the remaining 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 10000 rows and 6 columns.

```
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 bottom menu bar includes Files, Plots, Packages, Help, Viewer, and Presentation. The main area has tabs for Untitled1 (Source), Run, Source, and Environment. The Environment tab shows a global environment with variables like data, Age, Grade, Hours.on.Assig..., Hours.on.Games, Hours.on.Inter..., and Hours.on.Readi... The Data tab shows a data frame with 10000 observations and 6 variables: Age, Grade, Hours.on.Assig..., Hours.on.Games, Hours.on.Inter..., and Hours.on.Readi... Below the code editor, there is a red box highlighting the assignment of 'Age' and 'Grade' from the 'data' frame.

```
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 12 13 13 13 ...
 $ Hours.on.Assignments: int 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 12 13 13 13 ...
 $ Hours.on.Assignments: int 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 2nd method and calculate the correlation values.

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

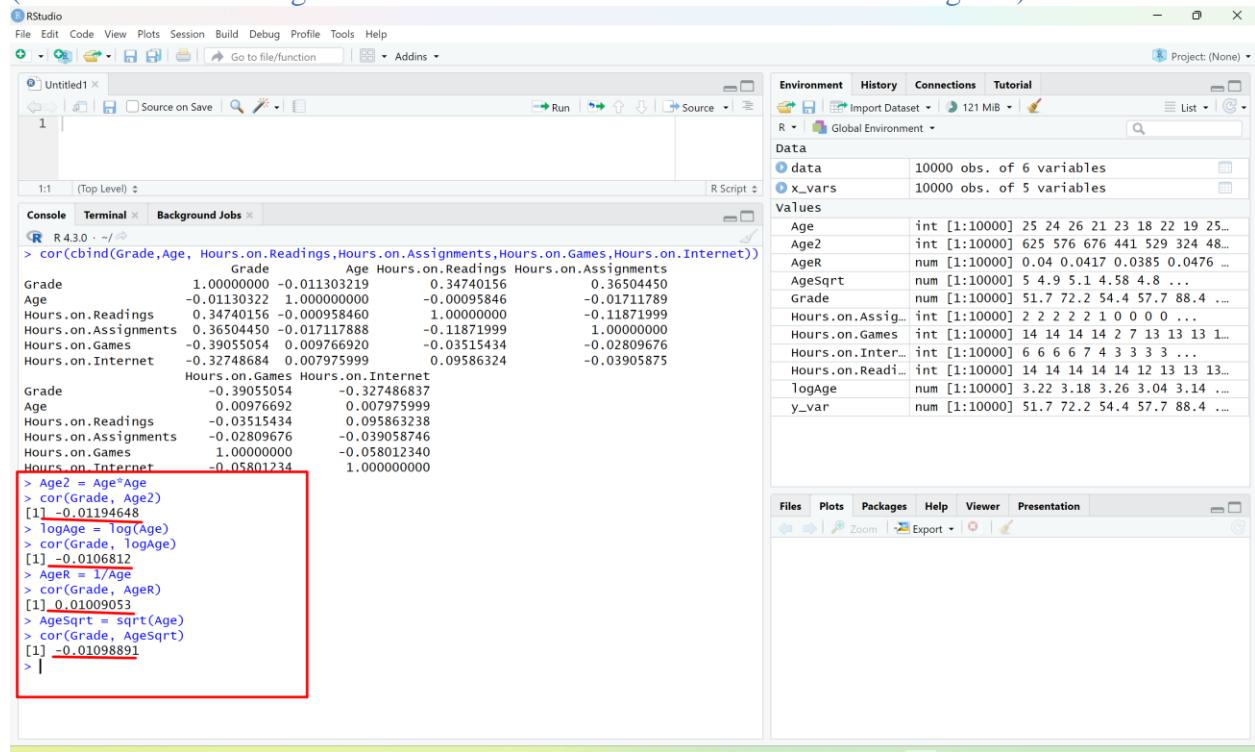
The screenshot shows the RStudio interface with a red box highlighting the command for calculating the correlation matrix. The Data tab shows variables data, test.data, and train.data. The console output shows the correlation matrix for Grade, Age, Hours.on.Readings, Hours.on.Assignments, Hours.on.Games, and Hours.on.Internet.

```
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 1.0000000 -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 -0.11871999 1.000000000
Hours.on.Games 0.355364 -0.00976020 0.00095846 -0.02800797
Hours.on.Internet -0.32748684 0.007975999 0.09586324 -0.03905875
   Hours.on.Games Hours.on.Internet
Grade -0.3905504 -0.327486837
Age 0.009759999 0.009759999
Hours.on.Readings -0.013114 0.009759999
Hours.on.Assignments -0.02809676 -0.039058746
Hours.on.Games 1.000000000 -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 with the following components:

- File Edit Code View Plots Session Build Debug Profile Tools Help**
- Addins** dropdown menu
- Untitled1** script editor pane with code and output.
- Run** button and other toolbar icons.
- Environment**, **History**, **Connections**, **Tutorial** tabs.
- Project**: (None)
- Console** tab showing R session history.
- Terminal** tab.
- Background Jobs** tab.
- Data** pane showing objects in the global environment:

 - data**: 10000 obs. of 5 variables
 - x_vars**: 10000 obs. of 5 variables
 - values** table:

Age	int [1:10000]	25	24	26	21	23	18	22	19	25...
Age2	int [1:10000]	625	576	676	441	529	324	48...		
AgeR	num [1:10000]	0.04	0.0417	0.0385	0.0476...					
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.Assig...	int [1:10000]	2	2	2	2	1	0	0	0	...
Hours.on.Games	int [1:10000]	14	14	14	14	2	7	13	13	1...
Hours.on.Inter...	int [1:10000]	6	6	6	6	7	4	3	3	3...
Hours.on.Readi...	int [1:10000]	14	14	14	14	14	12	13	13	1...
logAge	num [1:10000]	3.22	3.18	3.26	3.04	3.14...				
y_var	num [1:10000]	51.7	72.2	54.4	57.7	88.4...				

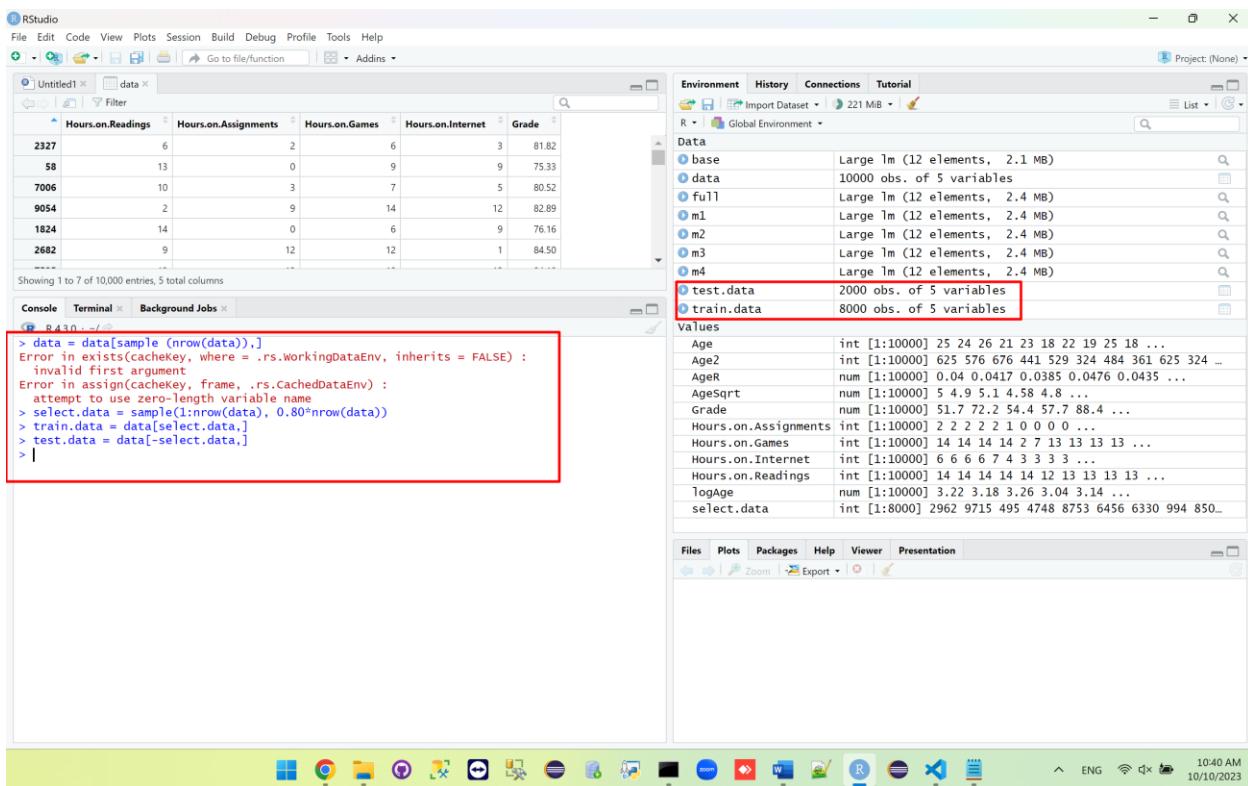
- Files**, **Plots**, **Packages**, **Help**, **Viewer**, **Presentation** tabs.

- 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:

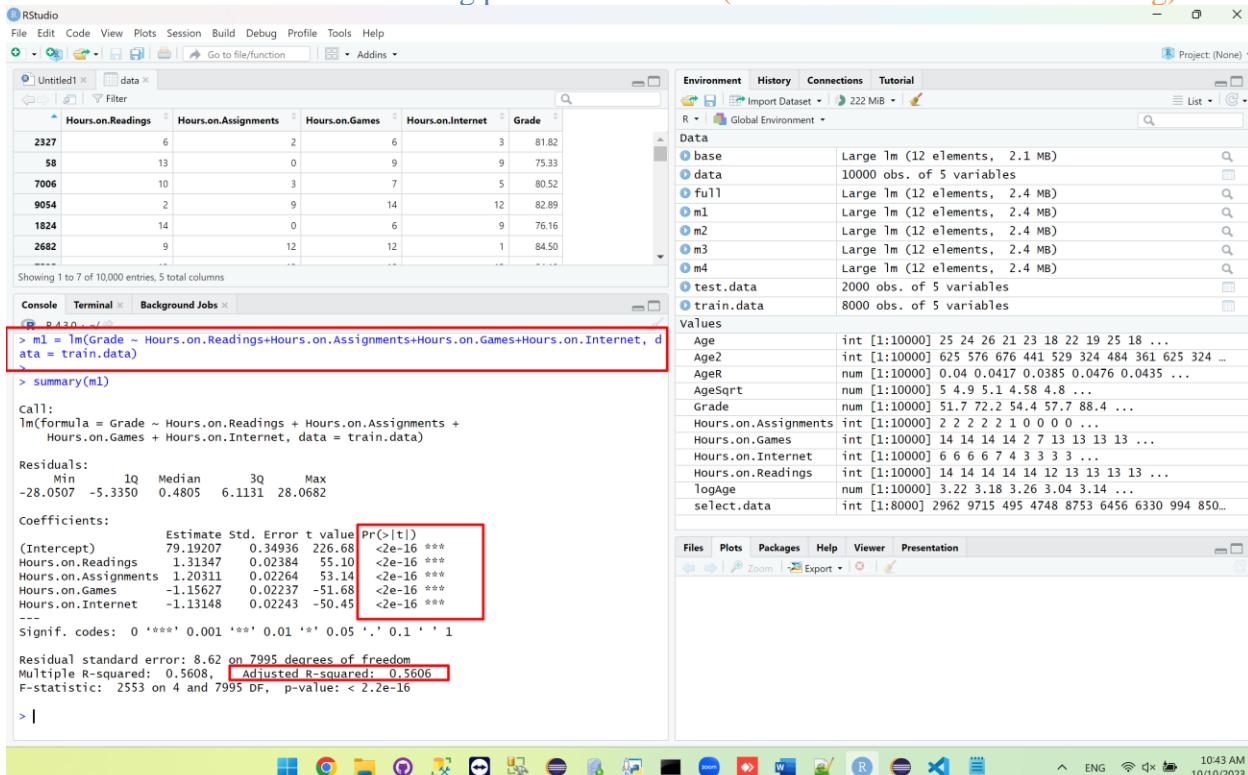
- As mentioned in the question, we need to use **hold out evaluation** for this data split.
 - **Note:** In general, this data split depends on the size of the data.
 - If the data is large (more than 1 million rows) then hold-out evaluation is used.
 - If the data is small (1 million rows /smaller) 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:

- Data View:** Displays a table with columns: Hours.on.Readings, Hours.on.Assignments, Hours.on.Games, Hours.on.Internet, and Grade. The data consists of 7 rows with values ranging from 2327 to 2682.
- Environment View:** Shows the global environment with objects like base, data, full, m1, m2, m3, m4, test.data, and train.data.
- Console View:** Contains the R code and output for the stepwise regression:

```
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
<none>
Df Sum of Sq   RSS   AIC
- Hours.on.Internet  1  189095 783163 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 Environment View showing the results of a regression model fit using the step function with AIC as the metric. The console output shows the steps taken to build the model, including the base model and the addition of variables one by one until the AIC no longer decreases.

```

R > 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   189005 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 Environment View showing the results of a regression model fit using the step function with AIC as the metric, demonstrating the forward selection process. The console output shows the steps taken to build the model, starting from a base model and adding variables one by one until the AIC no longer decreases.

```

R > base = lm(Grade~ Hours.on.Readings, data = train.data)
> step(base, scope=list(lower=NULL, upper=1), 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.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 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

R Project: (None)

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  793566 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 15th model.
- Then we need to look in to the true/false chart and go to the 15th 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:** Shows objects like base, data, m1-m5, test.data, and train.data.
- Console:**

```
$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

> 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

```

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.

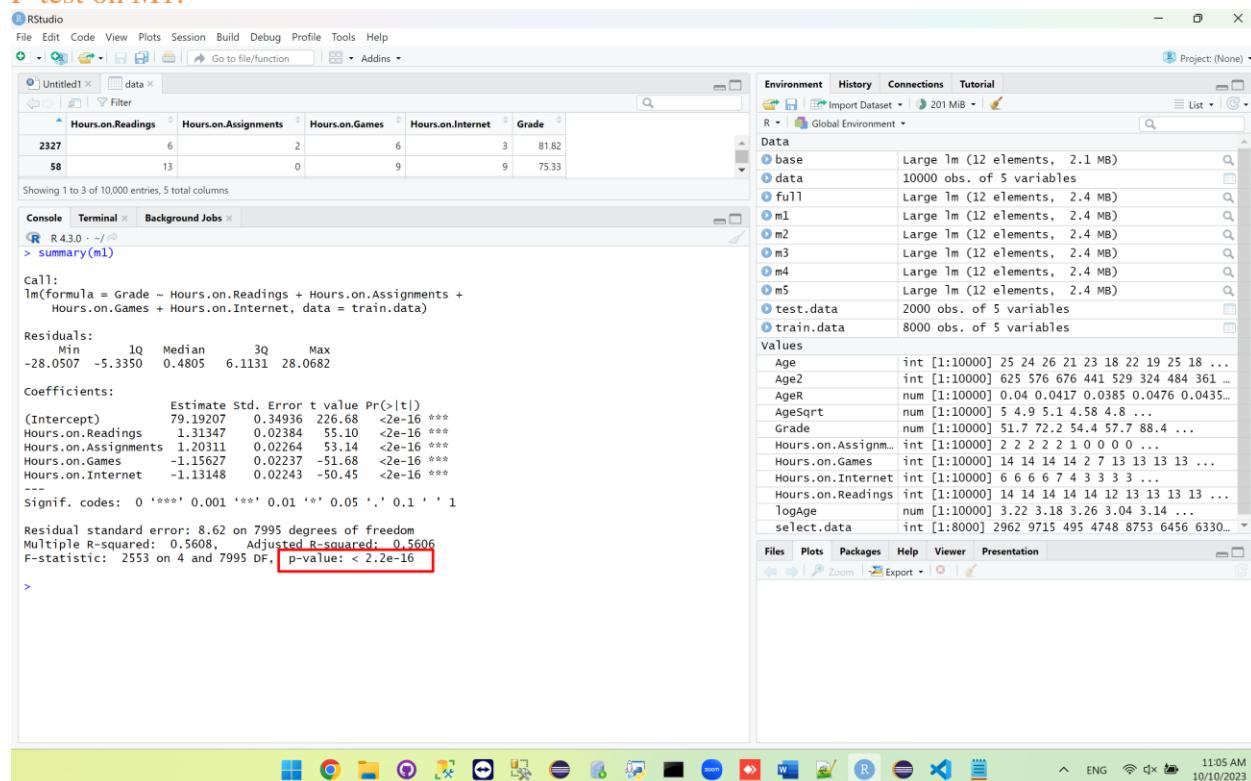
But pls make note that all models are same we can pick any 1 model and perform model diagnosis... Lets pick M1 and go ahead**

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

```

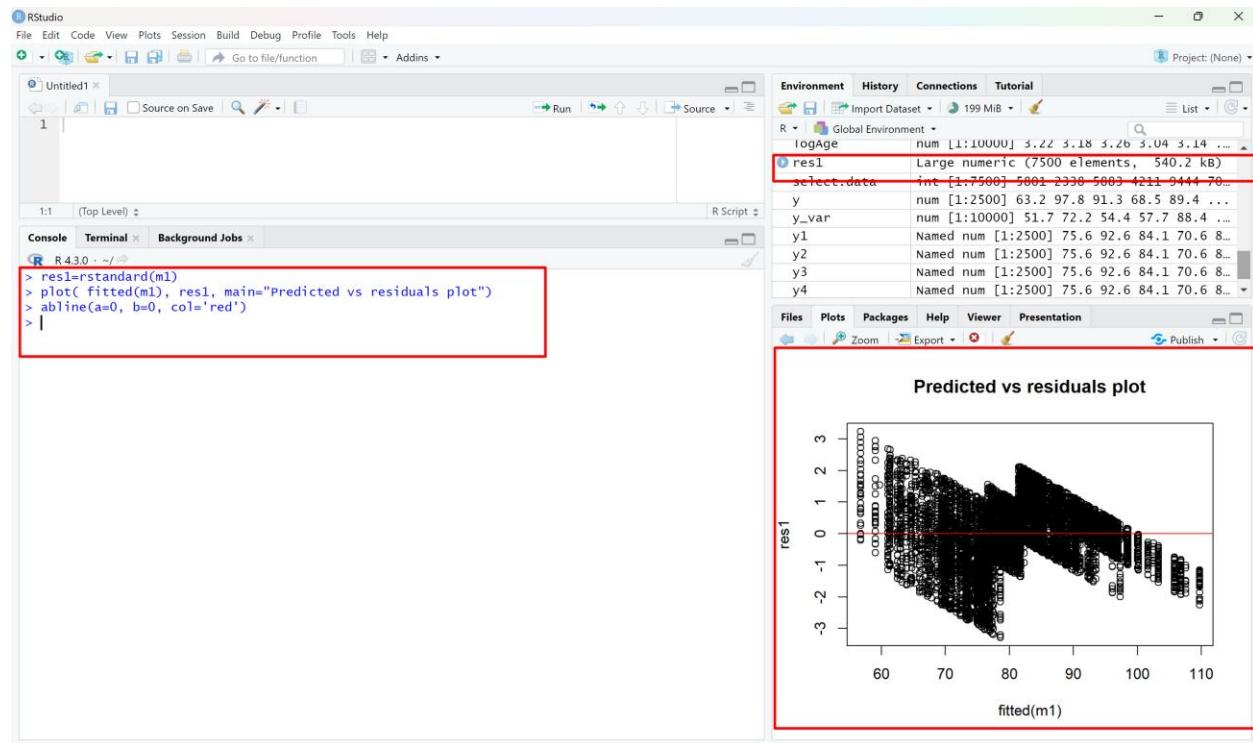
Residual analysis on M1:

- 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 ****

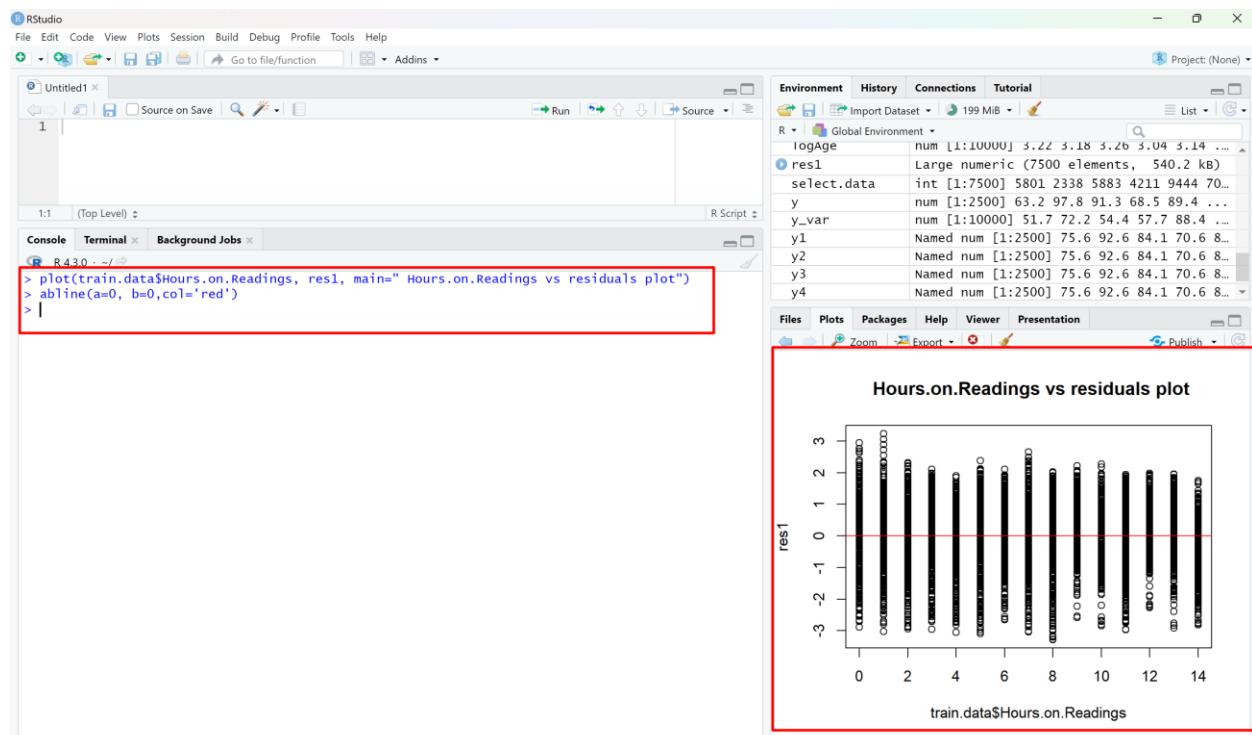
Residual analysis on M1:

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

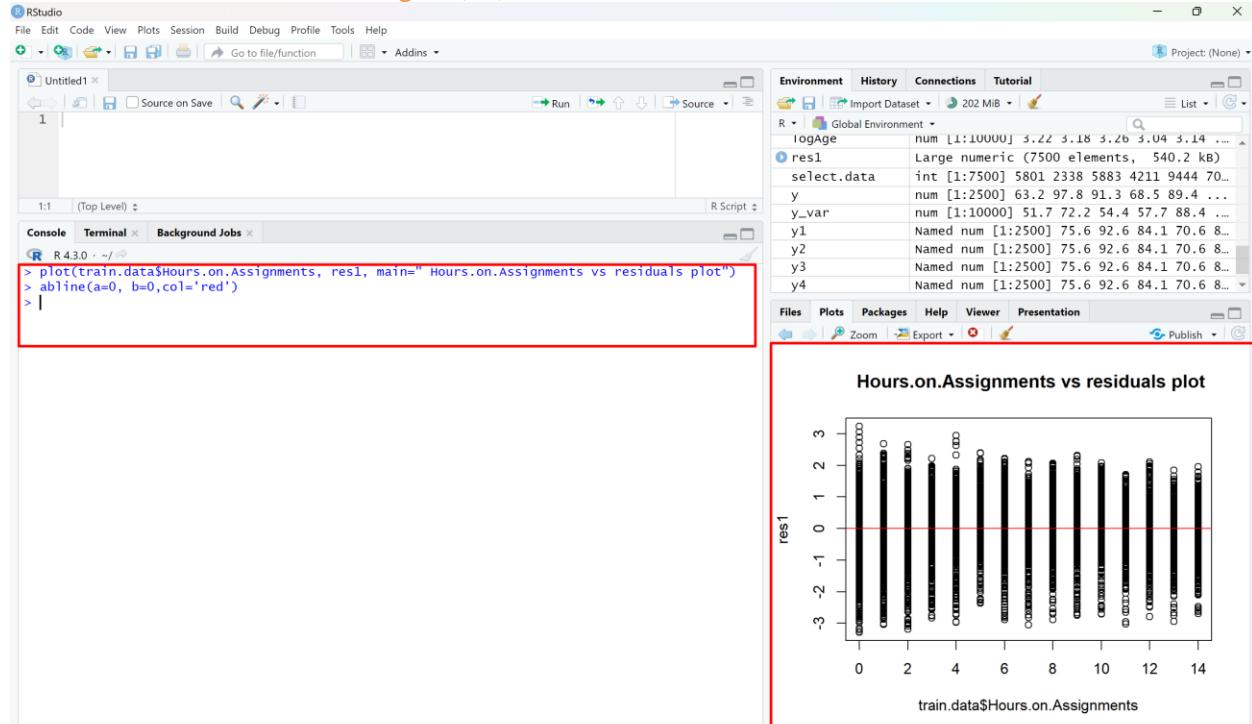


- Validate the linearity relationship: Plot residuals vs each x-variable: To check linearity assumptions for Y and the x-variable.

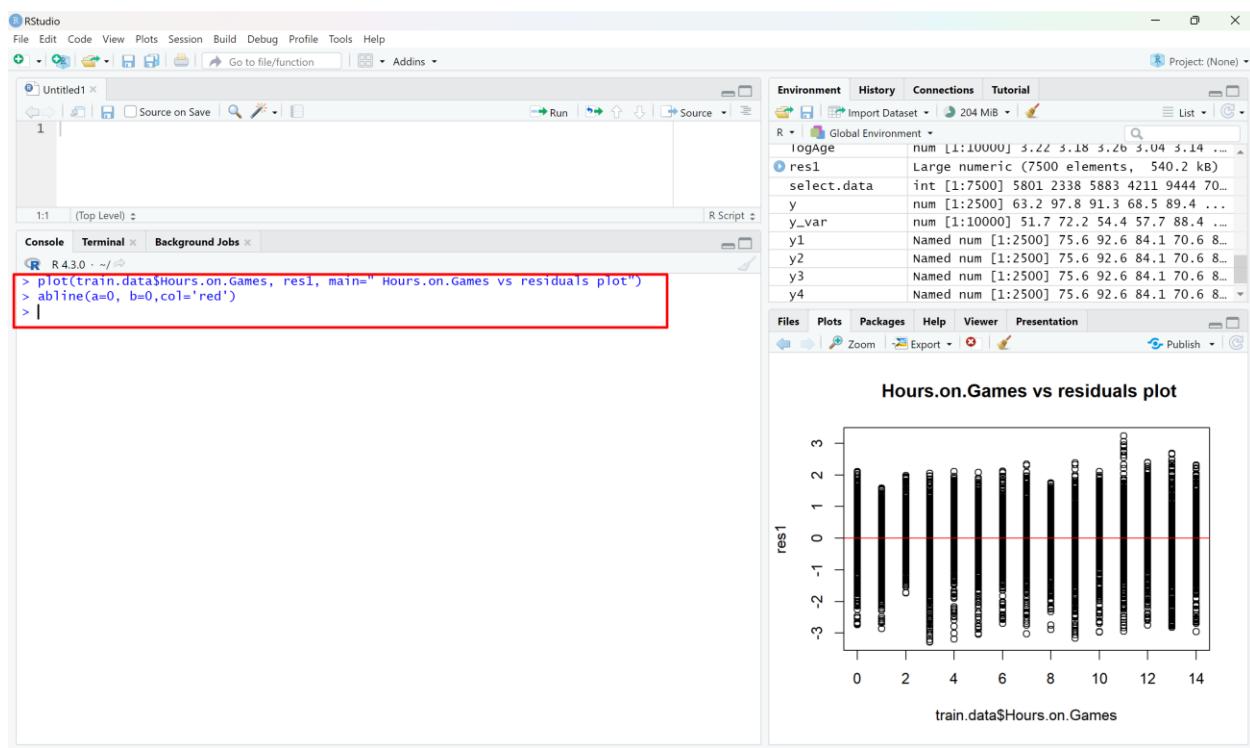
Plot residual vs Hours on Readings(x1)



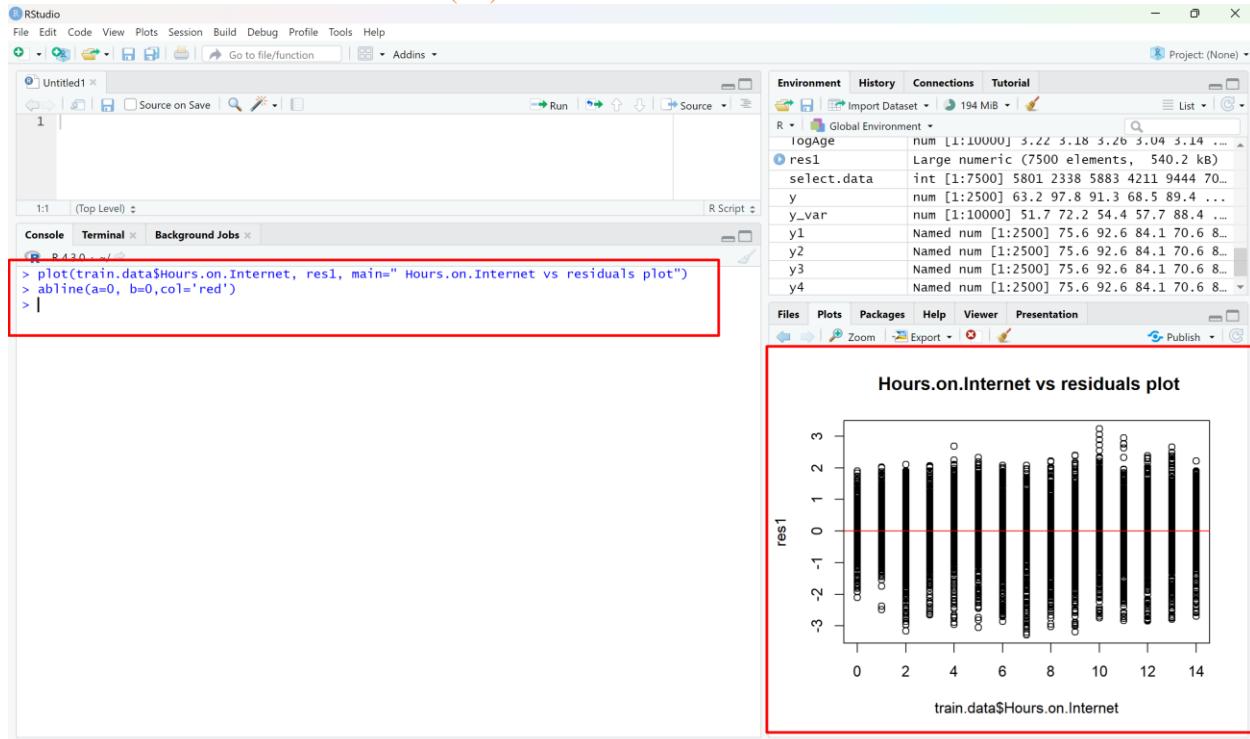
Plot residual vs Hours on Assignm(x2)



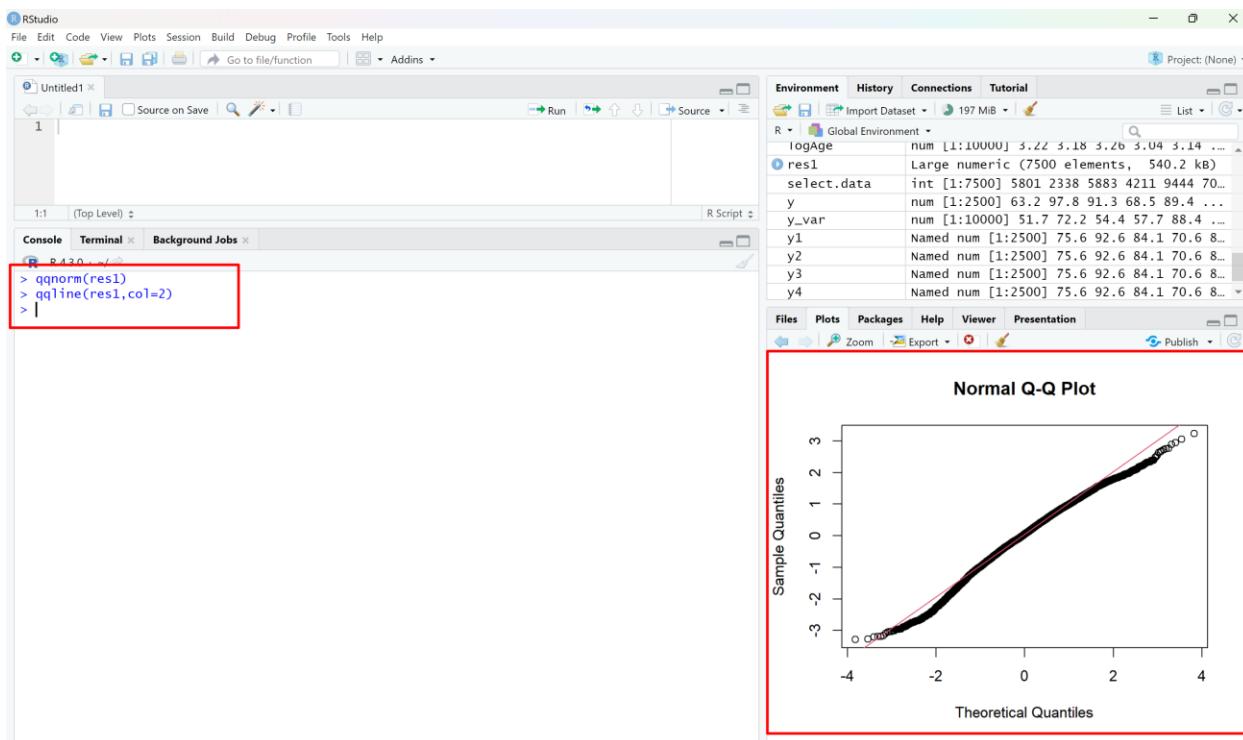
Plot residual vs Hours on Games(x3)



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 2nd method and calculate the correlation values.

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

The screenshot shows the RStudio interface with the following details:

- File menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project:** Project (None).
- Environment pane:** Shows variables: data (10000 obs. of 6 variables), test.data (2000 obs. of 6 variables), train.data (8000 obs. of 6 variables). It also lists values for Age, Grade, Hours.on.Assignments, Hours.on.Games, Hours.on.Internet, Hours.on.Readings, and select.data.
- Console pane:** Displays the command `> cor(cbind(Grade,Age, Hours.on.Readings,Hours.on.Assignments,Hours.on.Games,Hours.on.Internet))` and its output, which is a correlation matrix. The matrix includes columns for Grade, Age, Hours.on.Readings, Hours.on.Assignments, Hours.on.Games, and Hours.on.Internet. Correlation values range from -0.395 to 0.347.
- Plots pane:** Not visible in the screenshot.
- Packages pane:** Not visible in the screenshot.
- Help pane:** Not visible in the screenshot.
- Viewer pane:** Not visible in the screenshot.
- Presentation pane:** Not visible in the screenshot.

- 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)

```

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 : (H₀): 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, β_1 represents the coefficient of "Hours.on.Readings," β_2 represents the coefficient of "Hours.on.Assignments," β_3 represents the coefficient of "Hours.on.Games," and β_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

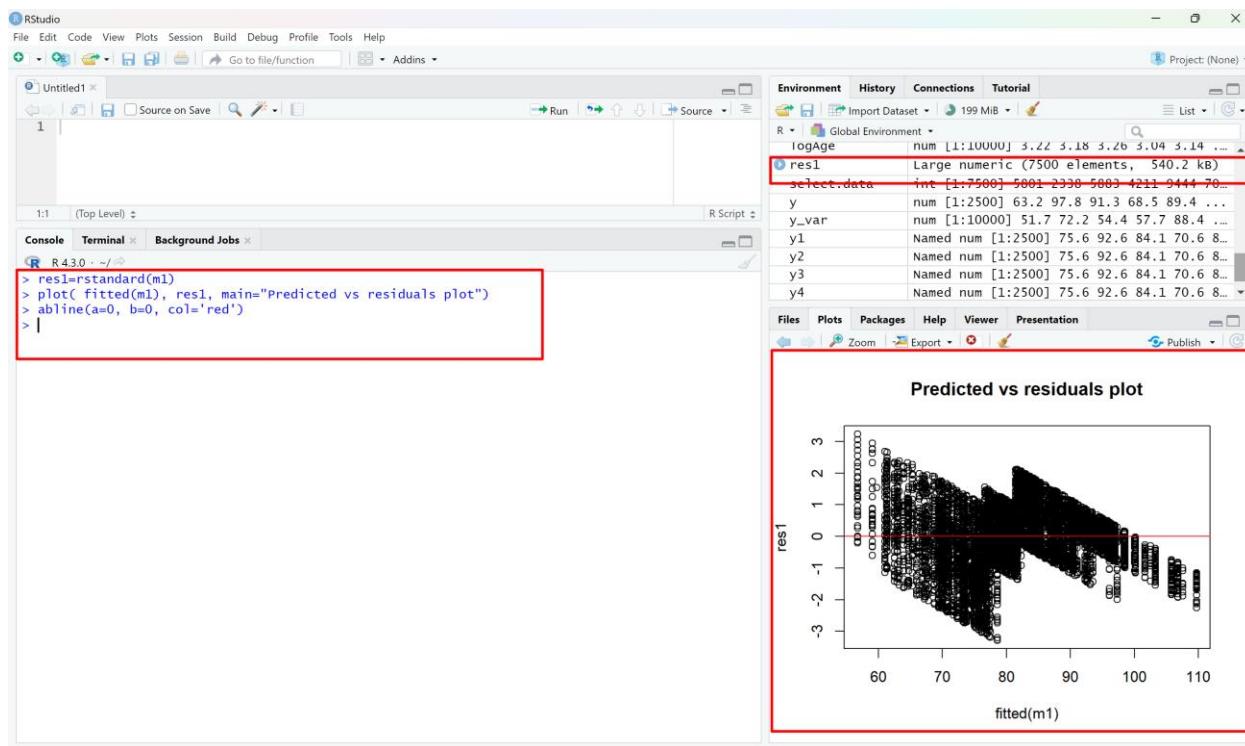
5. Any issues in the residual analysis for the model above?

ANSWER:

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

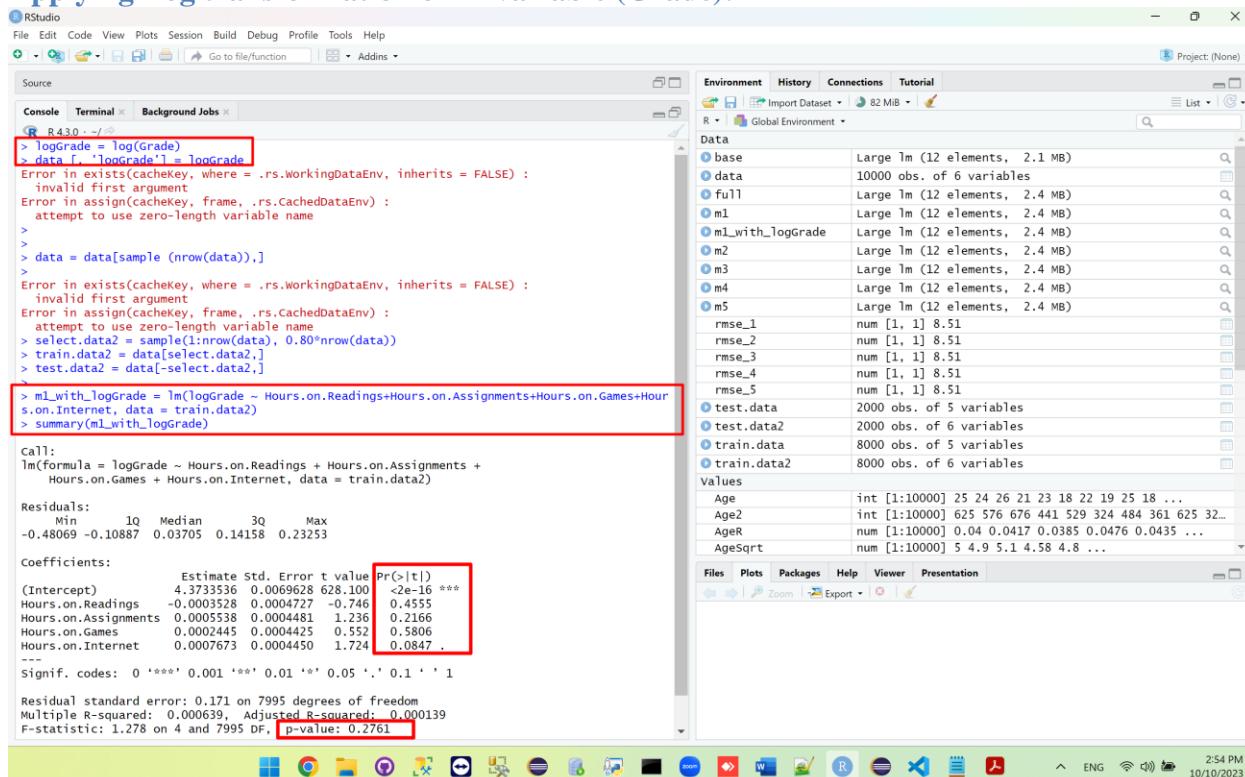
Residual analysis on m1:

- 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
>

```

Environment History Connections Tutorial

R Global Environment

rmse_5 num [1: 1] 8.51
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] 625 576 676 441 529 324 484 361 ...
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.Assignm... int [1:10000] 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 12 13 13 13 13 13 ...
logAge num [1:10000] 3.22 3.18 3.26 3.04 3.14 ...
res1 Large numeric (8000 elements, 576.2 kB)
select.data int [1:8000] 2962 9715 495 4748 8753 6456 6330 ...
y num [1:2000] 95.5 79.8 71.7 65.8 98.3 ...
y1 Named num [1:2000] 87.1 67.3 78.1 72.9 95.1 ...
y2 Named num [1:2000] 87.1 67.3 78.1 72.9 95.1 ...
y3 Named num [1:2000] 87.1 67.3 78.1 72.9 95.1 ...
y4 Named num [1:2000] 87.1 67.3 78.1 72.9 95.1 ...
v5 Named num [1:2000] 87.1 67.3 78.1 72.9 95.1 ...

Files Plots Packages Help Viewer Presentation

11:38 AM 10/10/2023

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 (β_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 (β_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 (β_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 (β_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).

HW4: Multiple Linear Regression Analysis

Note: every step you use R, you should provide the snapshots of your R commands and R outputs.

Note: do not split the data for this assignment, just use all the data to build the models

Problem 1 [40]

A researcher is interested in evaluating the relationship between energy consumption by the homeowner and the difference between the internal and external temperatures. A sample of 30 homes was used in the study. During an extended period of time, the average temperature difference (in °F) (TEMPD) inside and outside the homes was recorded. The average energy consumption (ENERGY) was also recorded for each home. The data are stored in the energytemp.txt data file.

- a) Create a scatterplot of ENERGY (y) versus TEMPD (x) to visualize the association between the two variables. Analyze the association displayed by the scatterplot.
- b) Fit a cubic model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$ (*HINT: create two new variables TEMP2 and TEMP3:*

In SAS DATA STEP use the code: tempd2 = tempd**2;
 tempd3 = tempd**3;

In R use the code: tempd2 = tempd^2;
 tempd3 = tempd^3;

Include the new variables in the regression model)

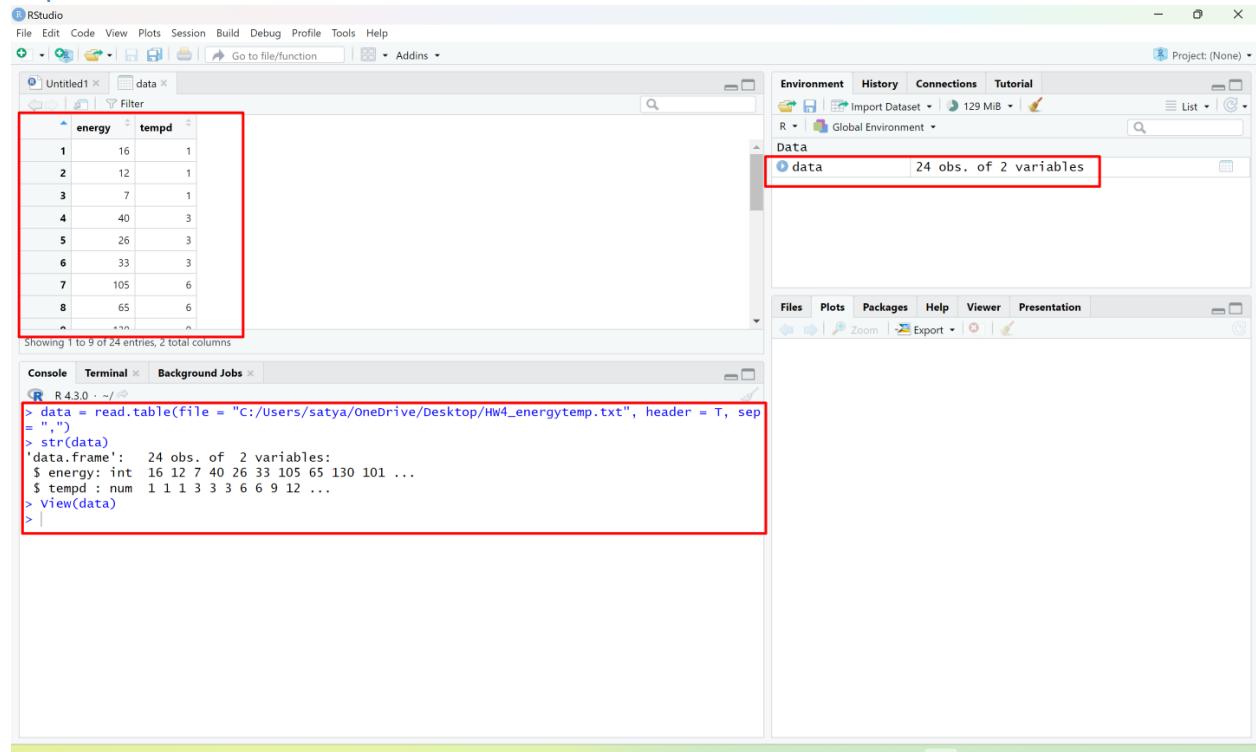
- a) Test the goodness of fit of the model at the 5% significance level.
- b) Are all variables in the model significant?
- c) Create the residual plots (residuals vs predicted; residuals vs x variable; and normal plot of residuals). Analyze residual plots to evaluate the normality and constant variance assumptions. Discuss your findings.
- d) If you are satisfied with the fitted regression model, write down its expression.
- e) Use the fitted regression model to predict the average energy consumption for an average difference in temperature equal to TEMPD=10.

In R, you should use new = data.frame (tempd=c(10), tempd2=c(100), tempd3=c(1000)), and then use the predict() function in R to produce predictions and confidence interval

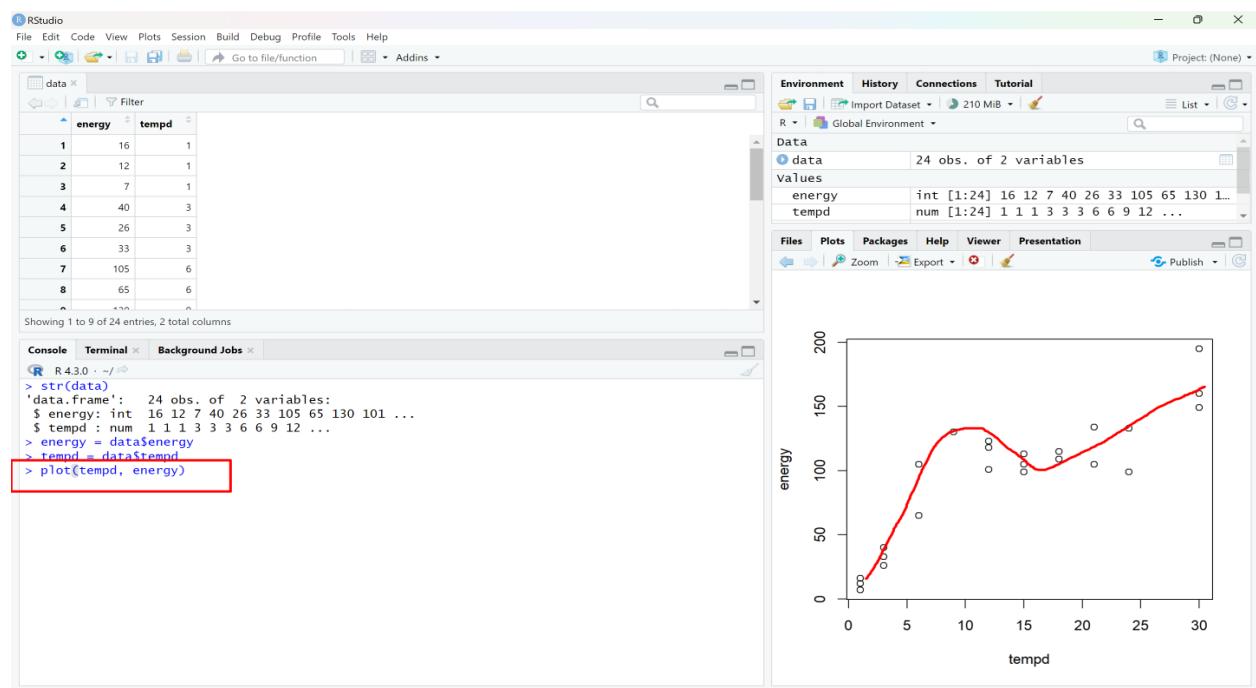
- f) By using influence.measures() function to identify whether there are influential points that can affect your final model. Use cook's distance as the metric to identify the influential points.

ANSWERS:

Imported data into R-studio:



a. Create a scatterplot of ENERGY (y) versus TEMPD (x) to visualize the association between the two variables. Analyze the association displayed by the scatterplot.

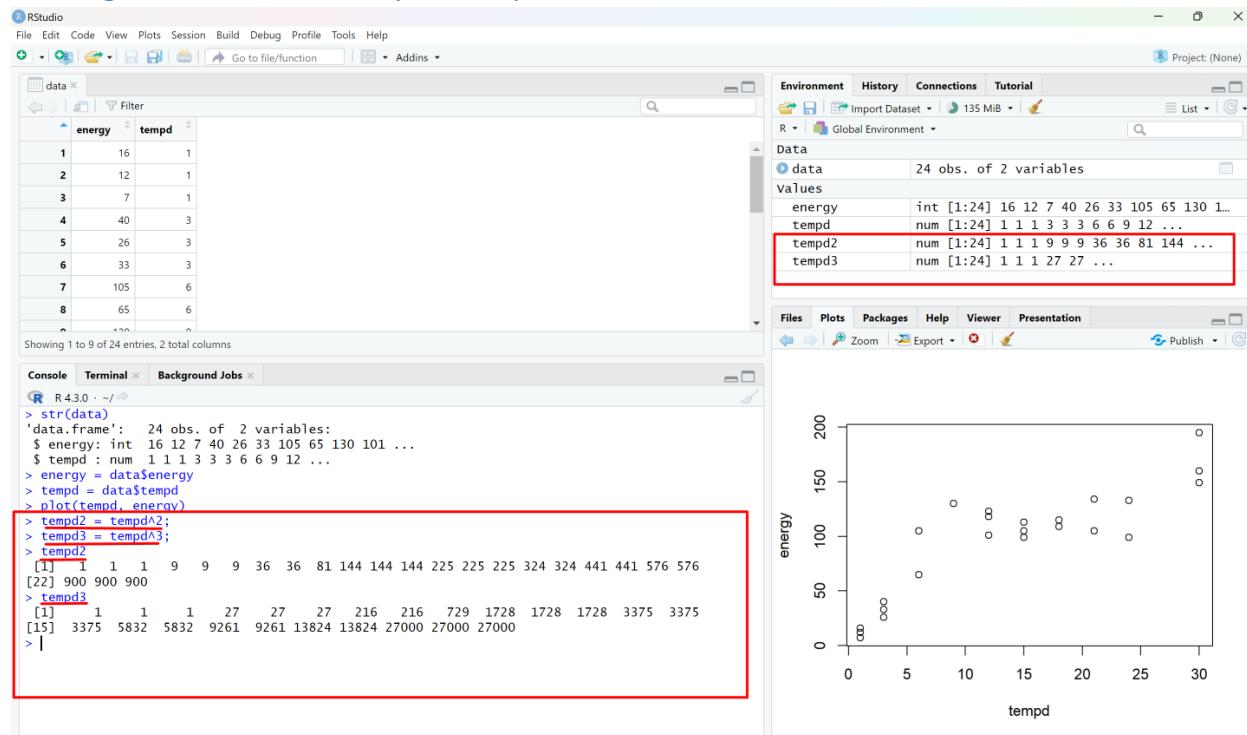


Analyzing the association b/w the 2 variables:

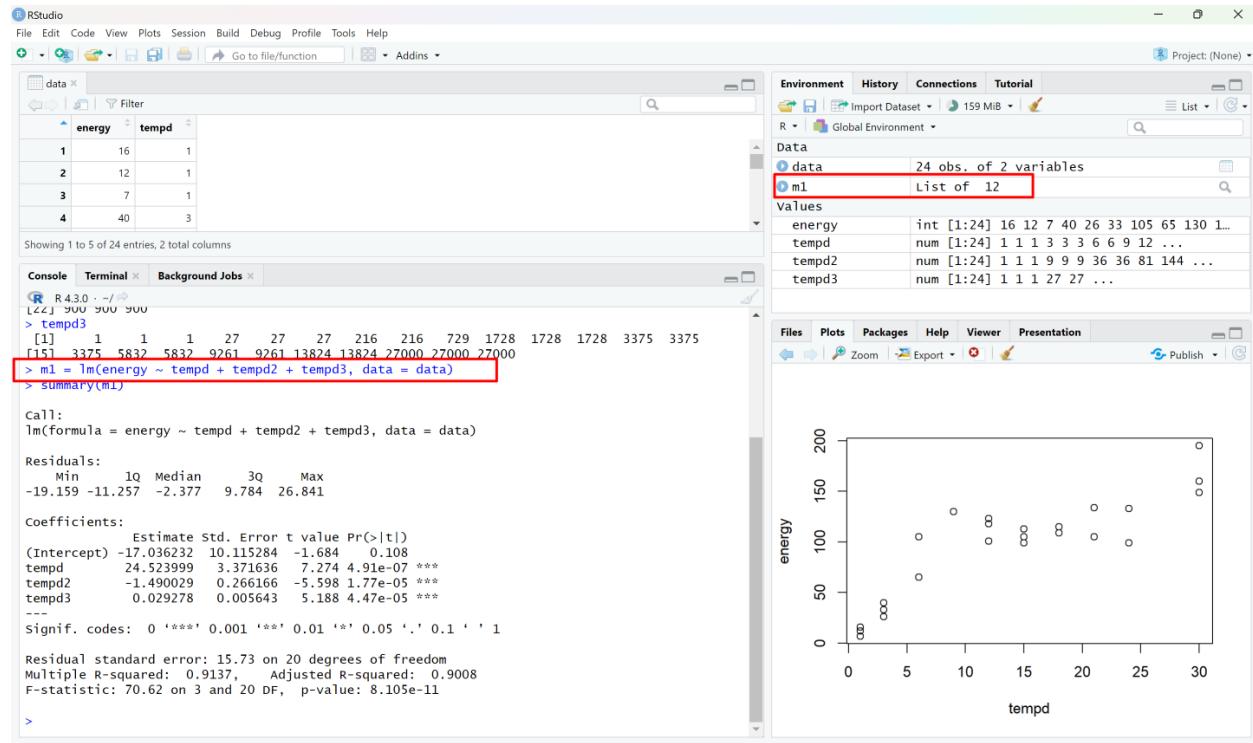
- From the scatterplot, we can observe that there is 1-top & 1-bottom in the curve, and the scatterplot produced S-shaped pattern. Hence, we can conclude that the association between energy (y) and tempd (X) is cubic association / cubic relation.
- As the scatterplot follows cubic relation, we need to add 3rd order term to build polynomial regression model (fit a cubic model)
- But we cannot add 3rd order term alone. We also need to add 2nd and 1st order terms also.
- So, we need to try power transformation on tempd variable & create 2 new variables.
- And add the 2 new variables to the model that we got after power transformation.
- As follows:

b. Fitting a cubic model, and creating 2 new variable tempd2, tempd3 and adding the new 2 variables to the dataset:

creating 2 new variables: tempd2, tempd3:



fitting a cubic regression model: M1



a. Test the goodness of fit of the model at the 5% significance level:

In f-test, we need to write down null and alternative hypothesis for the model:

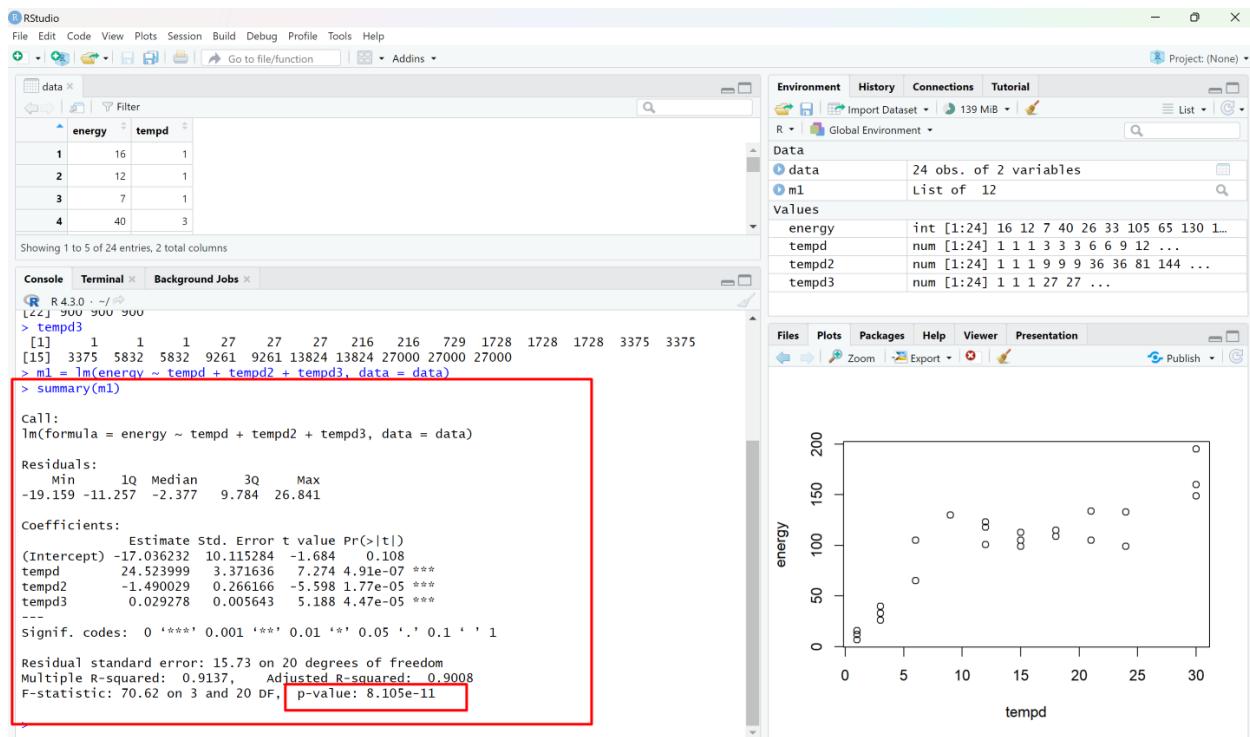
- Null hypothesis: (H_0): The coefficients of all x-variables are zero and there is no linear relationship with energy.
- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- Alternative Hypothesis: At least one of the coefficients of the x-variables (tempd, tempd2, tempd3) is not zero and can affect energy.
- $H_a : \beta_j \neq 0$

Note: In these hypotheses,

β_1 represents the coefficient of "tempd"

β_2 represents the coefficient of "tempd2."

β_3 represents the coefficient of "tempd3."

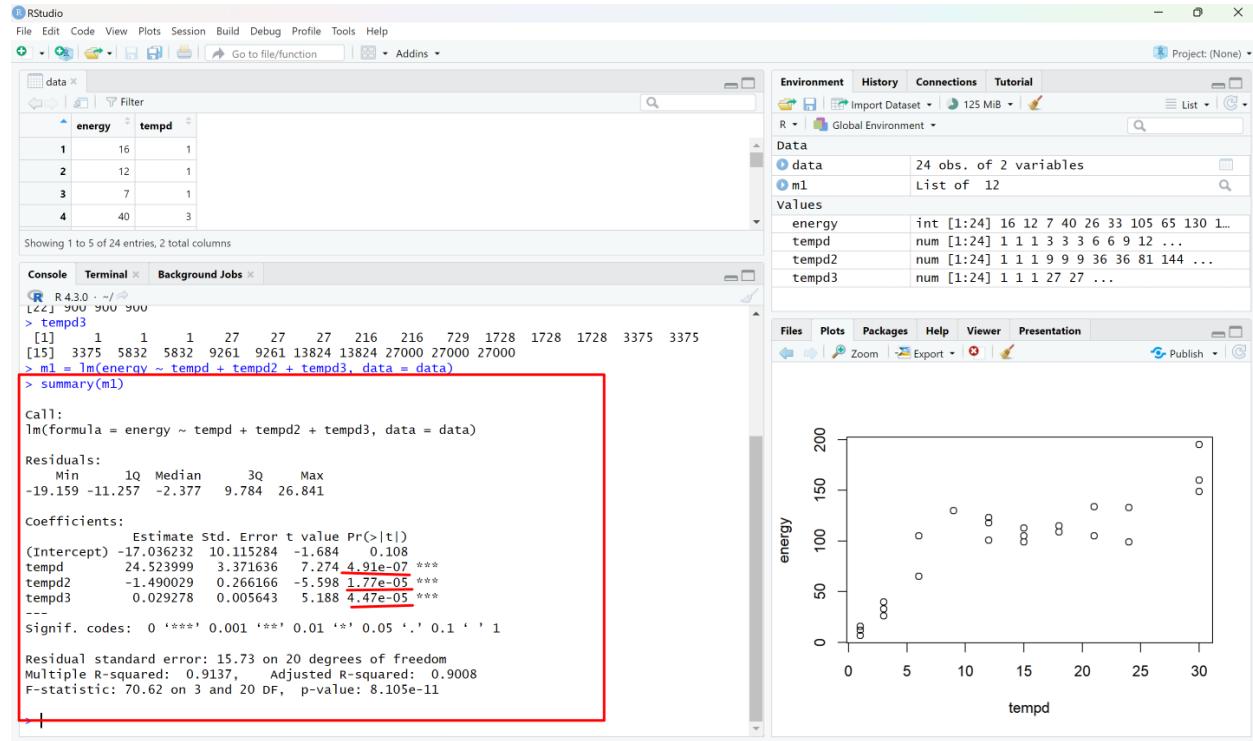


Conclusion on F-test: p-value < 0.05 for the model m1(fitted cubic regression model)

- At 5% significance level(alpha = 5%, i.e., alpha = 0.05) we can say that at least 1 x variables among (tempd, tempd2, tempd3) has a significant linear relationship with energy and can affect the value of y-variable ~ energy.

*****By conducting an F-test, we can observe that there is sufficient evidence to reject the null hypothesis and conclude that independent variables (tempd, tempd2, tempd3) have a significant impact on the dependent variable.(energy)*****

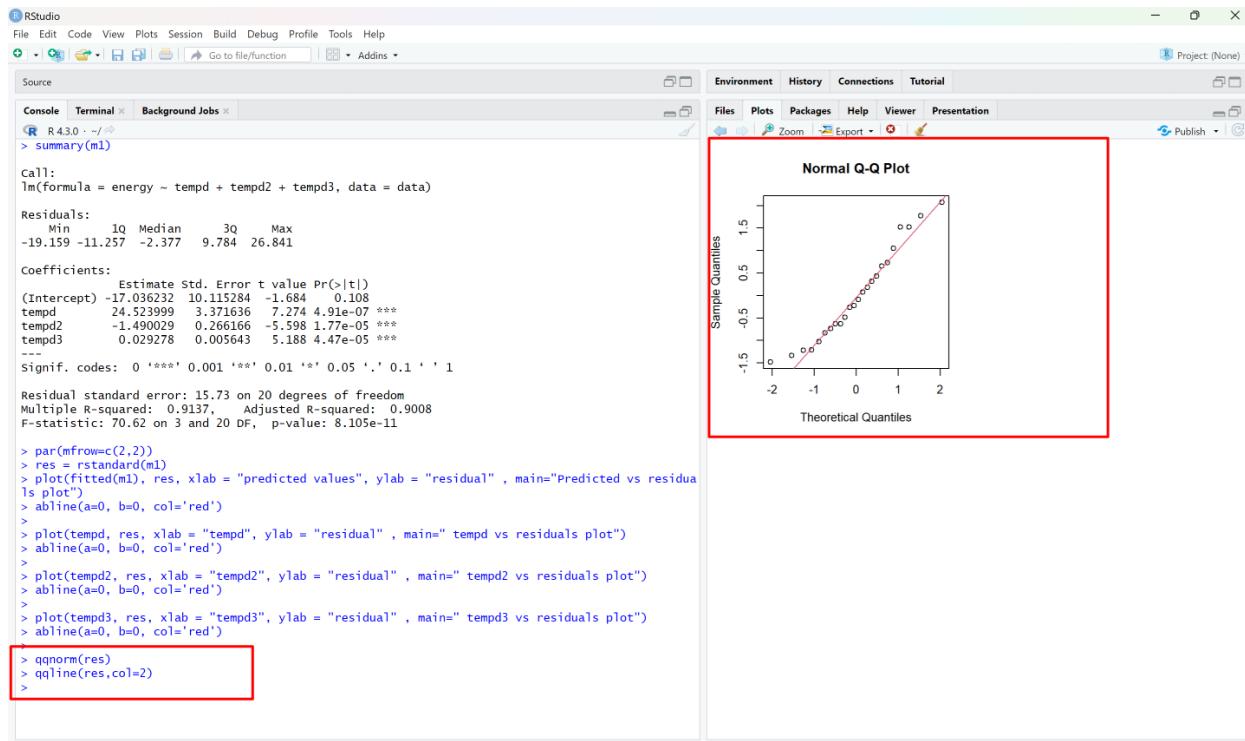
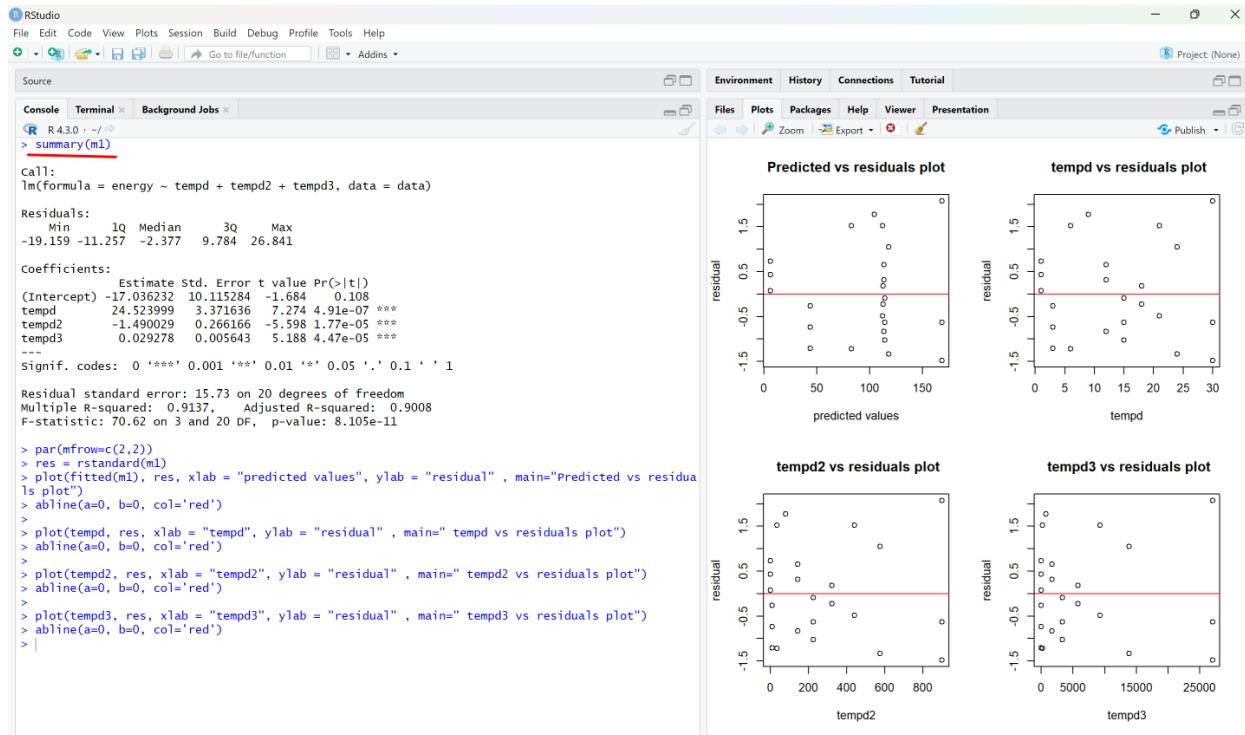
b. Are all variables in the model significant?



- To check whether all the variables are significant or not, we need to look at the individual t-test.
- As all the x-variables (tempd, tempd2,tempd3) have smaller p values than alpha 0.05(Assuming using 95% confidence level), all the x- variables in the model are significant.

c. Perform residual analysis: create residual plots and analyze the plots & discuss findings:

- Plot residuals vs predicted values plot: To check constant variance for the residuals:
- Plot residuals vs each x-variable to validate the linearity relationship:
- qqplot to validate normal distribution of residual:



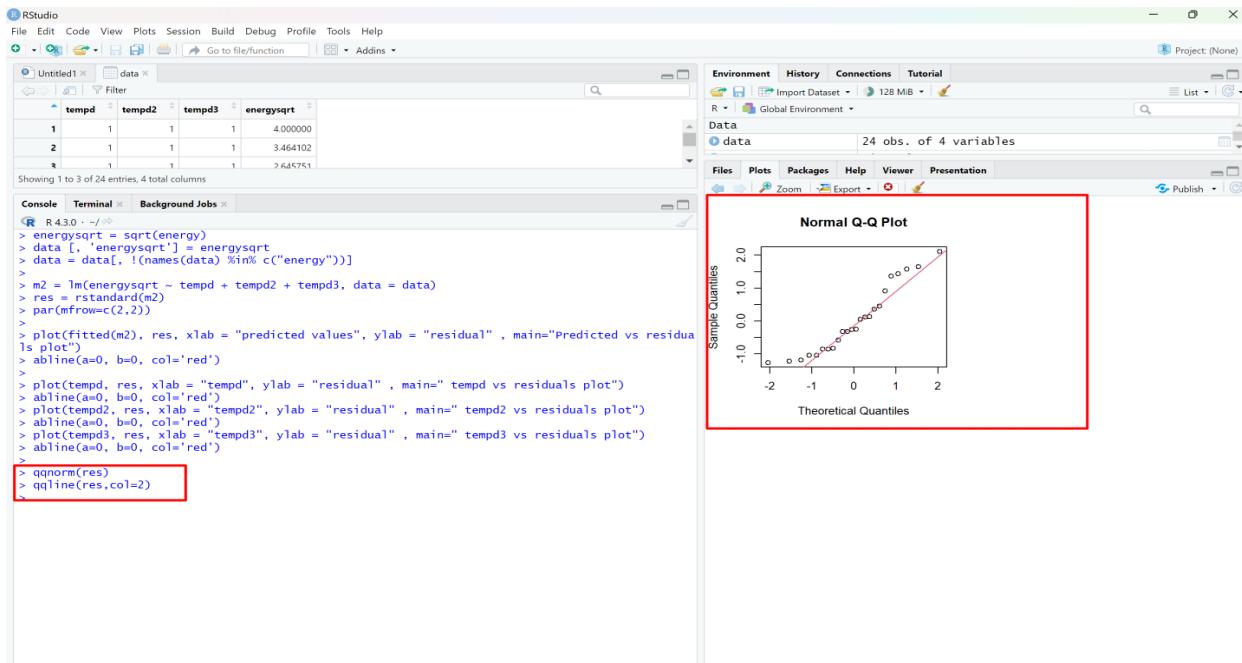
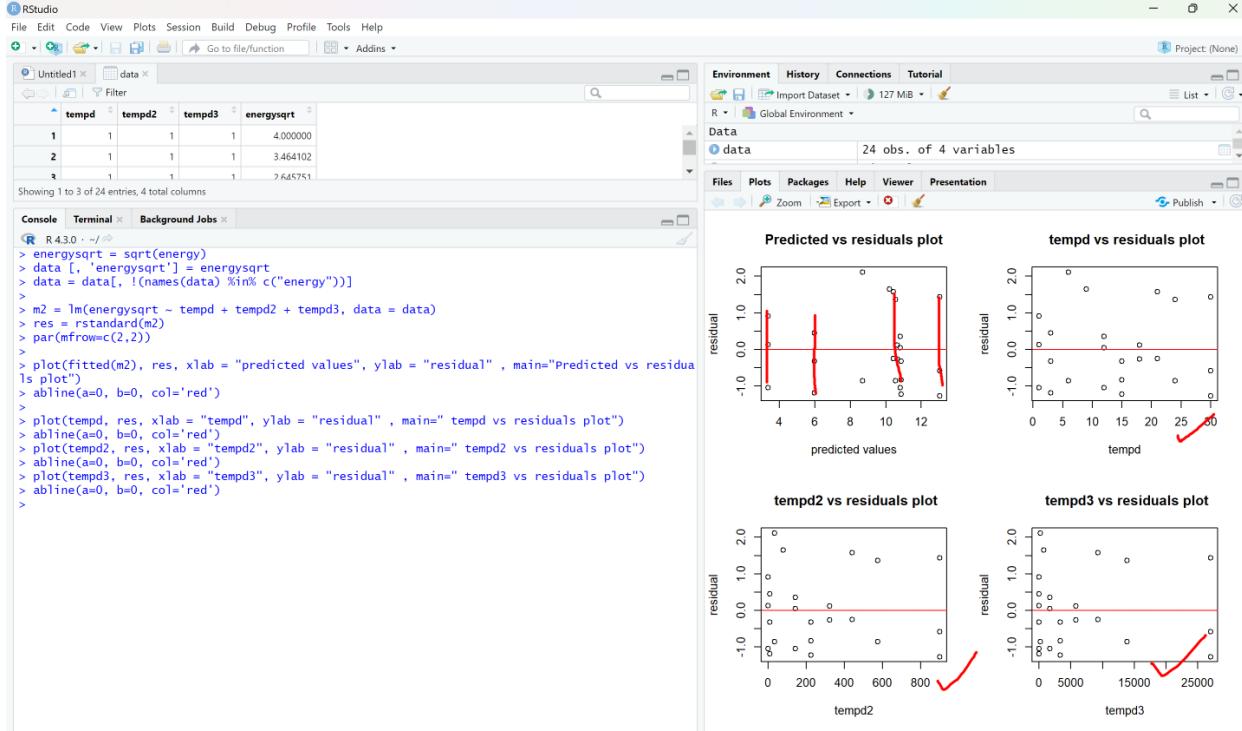
Observation: while performing res analysis for m1, we observed that

- In the res vs fitted values plot, we can observe that there is no constant variance, so we need to do transformation on y-var (energy)

- From the qqplot we can observe that residual follows normal distribution.

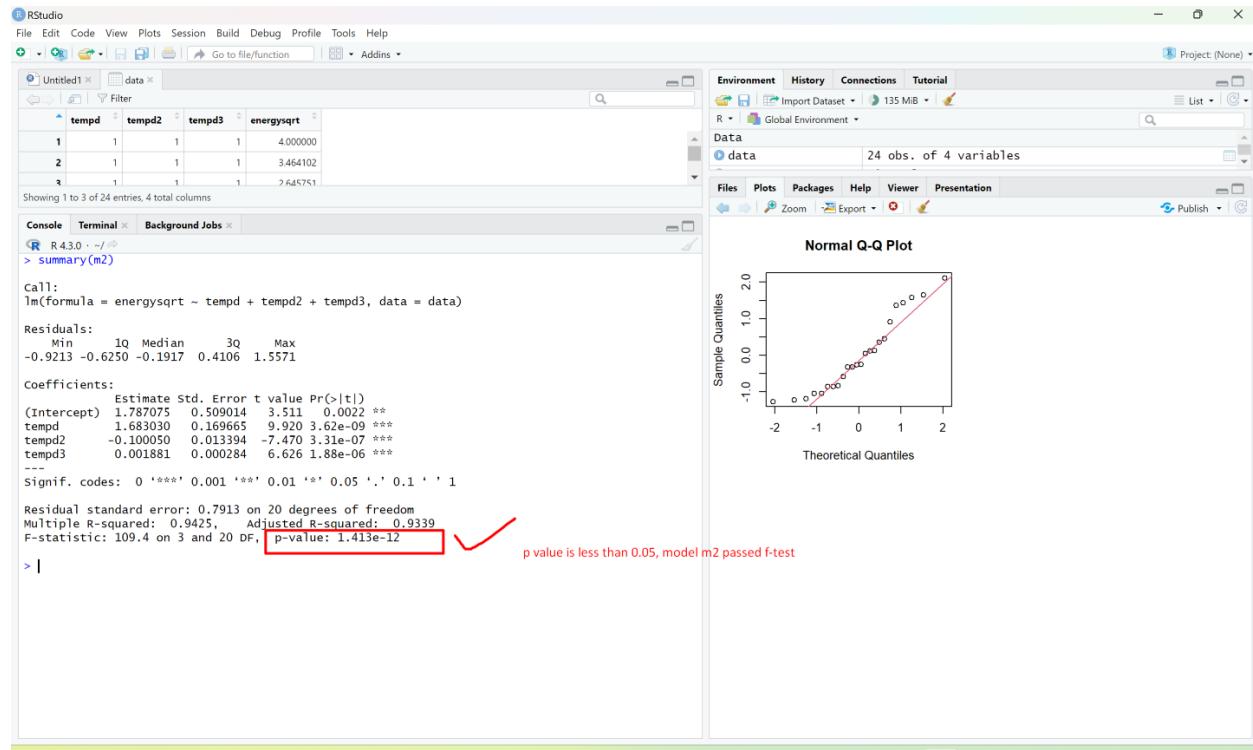
I tried log , reverse, sqrt on energy. I found sqrt transformation is better.

Sqrt tanformation on y-variable and built model m2 and performed residual analysis and f-test for m2:



Passed res analysis.

f-test on m2: passed f-test



Now we need check vif for m2: to remove multi-collinearity problem

- Using vif () function we can calculate the vif of m1, library(car) is installed.
- We can observe there is hight vif value for all the x -variables.
- To know which pair of x-variables has high collinearity, we used cor(data)
- From cor(data) we observed that the pair tempd2 & tempd3 has high collinearity
- build new model m3 by removing tempd2 and tempd3 1 by 1.
- Now we got model m3 with variables energy and tempd

The entire process is carried out in R as follows in the screenshot:

M3

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ - Go to file/function Addins
Untitled1 data
tempd tempd2 tempd3 energysqrt
1 1 1 1 4.000000
2 1 1 1 3.464102
Showing 1 to 2 of 24 entries, 4 total columns
Console Terminal Background Jobs
R 4.3.0 - ~/...
> vif(m2)
tempd tempd2 tempd3
98.12866 590.51231 237.78598
> cor(data)
tempd tempd2 tempd3 energysqrt
tempd 1.000000 0.9595857 0.8962859 0.8389276
tempd2 0.9595857 1.000000 0.9835233 0.7105825
tempd3 0.8962859 0.9835233 1.000000 0.6277228
energysqrt 0.8389276 0.7105825 0.6277228 1.000000
the pair tempd2, tempd3 are having higher correlation
so we need to remove 1 by 1
> m3 = lm(energysqrt ~ tempd, data = data)
> vif(m3)
tempd tempd2
12.62702 12.62702
> m3 = lm(energysqrt ~ tempd, data = data)
> summary(m3)
removed tempd2
Call:
lm(formula = energysqrt ~ tempd, data = data)

Residuals:
    Min      1Q Median      3Q     Max 
-3.2325 -1.0813  0.1466  0.8824  3.3789 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.61021   0.61820  9.075 6.85e-09 ***
tempd       0.26807   0.03708  7.230 3.04e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.713 on 22 degrees of freedom
Multiple R-squared:  0.7038, Adjusted R-squared:  0.6903 
F-statistic: 52.27 on 1 and 22 DF, p-value: 3.035e-07

```

d. If you are satisfied with the fitted regression model, write down its expression:

Though we removed multicollinearity in m2 and built m3, it didn't improve the model. Because we observed that adjr2 is better for m2, we decided to pick m2. To do the prediction.

MODEL	ADJ r2	
M1 = initial model with energy~ tempd,tempd2,tempd3	90.08%	
M2 = after sqrt tranf on energy variable (energysqrt~tempd,tempd2,tempd3)	93.39%	Passed f-test & res analysis.
M3 = after removing mcp for m2 (energysqrt~ tempd)	69.03%	

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ Go to file/function | Addins
Untitled1 data
tempd tempd2 tempd3 energysqrt
1 1 1 1 4.000000
2 1 1 1 3.464102
Showing 1 to 2 of 24 entries, 4 total columns
Console Terminal Background Jobs
R 4.3.0 -/-
> summary(m1)

Call:
lm(formula = energysqrt ~ tempd + tempd2 + tempd3, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9213 -0.6250 -0.1917  0.4106  1.5571 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.787075  0.509014  3.511  0.0022 ***
tempd       1.683030  0.169665  9.920 3.62e-09 ***
tempd2      -0.100050  0.013394 -7.470 3.31e-07 ***
tempd3       0.001881  0.000284  6.626 1.88e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7913 on 20 degrees of freedom
Multiple R-squared:  0.9425, Adjusted R-squared:  0.9339 
F-statistic: 109.4 on 3 and 20 DF,  p-value: 1.413e-12
>

```

From the coefficients of the fitted regression model $m1$, we can write the expression as follows:

- $\text{ENERGYsqrt} = \beta_0 + \beta_1 * \text{TEMPD} + \beta_2 * \text{TEMPD}^2 + \beta_3 * \text{TEMPD}^3 + e$
- i.e.,
- $\text{SqrtEnergy} = 1.78 + 1.683 * \text{tempd} + (-0.10) * \text{tempd}^2 + 0.001 * \text{tempd}^3 + e$

Explaining the affect:

β_0 ~ explaining the intercept:

The intercept, represented by 1.78, is the value of energy when tempd is equal to zero. In other words, it indicates the sqrtenergy level when there is no temperature change ($\text{tempd} = 0$). The intercept represents a constant or baseline sqrtenergy level.

β_1 ~

The slope for the term $1.683 * \text{tempd}$ represents the linear relationship between sqrtenergy and tempd. It indicates how energy changes per unit change in tempd. For every unit increase in tempd, the sqrtenergy increases by 1.683 (Assuming other variables are held constant).

β_2 ~

The slope for the term $(-0.10) * \text{tempd}^2$ represents the quadratic relationship between sqrtenergy and tempd. It indicates how the sqrtenergy changes as the square of tempd

changes. The negative sign suggests a concave-down relationship, meaning the sqrtenergy initially increases at a decreasing rate and eventually decreases as tempd increases. (Assuming other variables are held constant).

$$\beta_3 \sim$$

The slope for the term $0.001 * \text{tempd}^3$ represents the cubic relationship between sqrtenergy and tempd. It indicates how the sqrtenergy changes as the cube of tempd changes. The positive sign suggests a concave-up relationship, where the sqrtenergy increases as tempd increases. (Assuming other variables are held constant).

e. Use the fitted regression model to predict the average energy consumption for an average difference in temperature equal to TEMPD=10:

(In R, you should use `new = data.frame (tempd=c(10), tempd2=c(100), tempd3=c(1000))`, and then use the `predict()` function in R to produce predictions and confidence interval)

The screenshot shows the RStudio interface with the following details:

- Environment Pane:** Shows the global environment with objects `m1`, `m2`, `m3`, `new`, `data`, and `values`. `data` contains 24 observations of 4 variables: `energy` (int [1:24]), `energysqrt` (num [1:24]), `res` (Named num [1:24]), `tempd` (num [1:24]), `tempd2` (num [1:24]), and `tempd3` (num [1:24]).
- Console Pane:** Displays the R command history and output. The commands run were:


```
R 4.3.0 -- / 
> new = data.frame (tempd=c(10), tempd2=c(100), tempd3=c(1000))
> predict(m2, new = new, interval = "confidence")
  fit    lwr     upr
1 10.49382 9.864162 11.12348
```

Observation:

From the output we can observe that

fit value is the predicted value: 10.49 (energy consumption) (energysqrt variable)

Confidence interval is lower ci = 9.86

Upper ci = 11.12

f) By using `influence.measures()` function to identify whether there are influential points that can affect your final model. Use cook's distance as the metric to identify the influential points.

Final model :m2

If we use cook distance, any points with cook distance $> 4/n$ are influential points. We need to remove them. And re-build the model and check the adj r2. Here, n = data size = 24; $4/24 = 0.16$
So, we need to remove any data points with values larger than 0.16:

```
R 4.3.0 - /-
> library(stats)
> influence.measures(m2)
Influence measures of
  lm(formula = energysqrt ~ tempd + tempd2 + tempd3, data = data):
    dfb.II dfb.tmpd dfb.tmpd2 dfb.tmpd3 dffit cov.r cook.d hat inf
1   0.49522 -0.34966  0.28194 -0.24580  0.5077 1.362 6.50e-02 0.239
2   0.07118 -0.05026  0.04053 -0.03533  0.0730 1.603 1.40e-03 0.239 *
3   -0.57680  0.40726 -0.32838  0.28629 -0.5914 1.283 8.69e-02 0.239
4   0.10000 -0.01466 -0.00939  0.01839  0.1591 1.323 6.60e-03 0.114
5   -0.27097  0.03972  0.02543 -0.04984 -0.4313 1.034 4.55e-02 0.114
6   -0.07139  0.01047  0.00670 -0.01313 -0.1136 1.356 3.38e-03 0.114
7   -0.09590  0.58055 -0.64560  0.64257  0.9276 0.513 1.76e-01 0.136
8   0.03469 -0.20999  0.23351 -0.23242 -0.33355 1.223 2.85e-02 0.136
9   -0.27078  0.57029 -0.56882  0.53698  0.7336 0.805 1.22e-01 0.153
10  0.16221 -0.27254  0.24381 -0.21294 -0.3883 1.114 3.75e-02 0.120
11  -0.00718  0.01207 -0.01080  0.00943  0.0172 1.393 7.78e-05 0.120
12  -0.05355  0.08997 -0.08048  0.07029  0.1282 1.361 4.30e-03 0.120
13  0.10119 -0.13435  0.07510 -0.03296 -0.3966 0.993 3.83e-02 0.093
14  0.02553 -0.03389  0.01895 -0.00832 -0.1000 1.321 2.62e-03 0.093
15  0.06711 -0.08910  0.04981 -0.02186 -0.2630 1.171 1.76e-02 0.093
16  -0.00519  0.01708 -0.03281  0.04128 -0.0922 1.364 2.23e-03 0.113
17  0.00225 -0.00741  0.01423 -0.01791  0.0400 1.384 4.21e-04 0.113
18  0.19236 -0.35635  0.45939 -0.49928  0.7200 0.864 1.19e-01 0.161
19  -0.02928  0.05425 -0.06993  0.07600 -0.1000 1.444 3.15e-03 0.161
20  0.22933 -0.38978  0.45326 -0.46248  0.6553 1.361 1.03e-01 0.182
21  -0.13960  0.23575 -0.27518  0.28153 -0.3939 1.291 4.04e-02 0.182
22  -0.10361  0.19320 -0.27047  0.36339  0.0159 1.176 2.44e-01 0.322
23  0.09055 -0.16885  0.23638 -0.31758 -0.8879 1.297 1.91e-01 0.322
24  0.04061 -0.07573  0.10602 -0.14244 -0.3982 1.694 4.10e-02 0.322 *
```

From the output we observed that there are no data points has larger value than 0.16,

Hence,

we can conclude that there are no influential points that can affect the final model m2.

Problem 2 [50]

Variation in gasoline mileage among makes and models of automobiles is influenced substantially by the weight and horsepower of the vehicles. The data in the file *mileage.txt* were provided by the U.S. Environmental Protection Agency and report car models, miles per gallon (MPG), weight and horse power (HP).

- a) Fit a regression model to predict miles per gallon (MPG) from WEIGHT and horsepower (HP). Analyze residual plots to evaluate if the regression model is appropriate. Discuss if there is any evidence that the assumption of non-constant variance is satisfied by the data.
 - b) Apply a transformation to the Y variable (for instance you can try log(y), sqrt(y) or 1/y). Find the regression model that seems more appropriate for the analysis.
 - c) Analyze residual plots to evaluate if the regression model for the transformed Y variables is adequate. Do the plots show a deviation from the assumption of constant variance?
 - d) Write down the expression of the regression model and interpret the estimated values of the regression parameters.
- e). Use the step function to adopt both backward and stepwise forward selection to build a new model, compare the new model with the previous model in terms of the adj-R2 value.

ANSWER:

Imported data into r -studio:

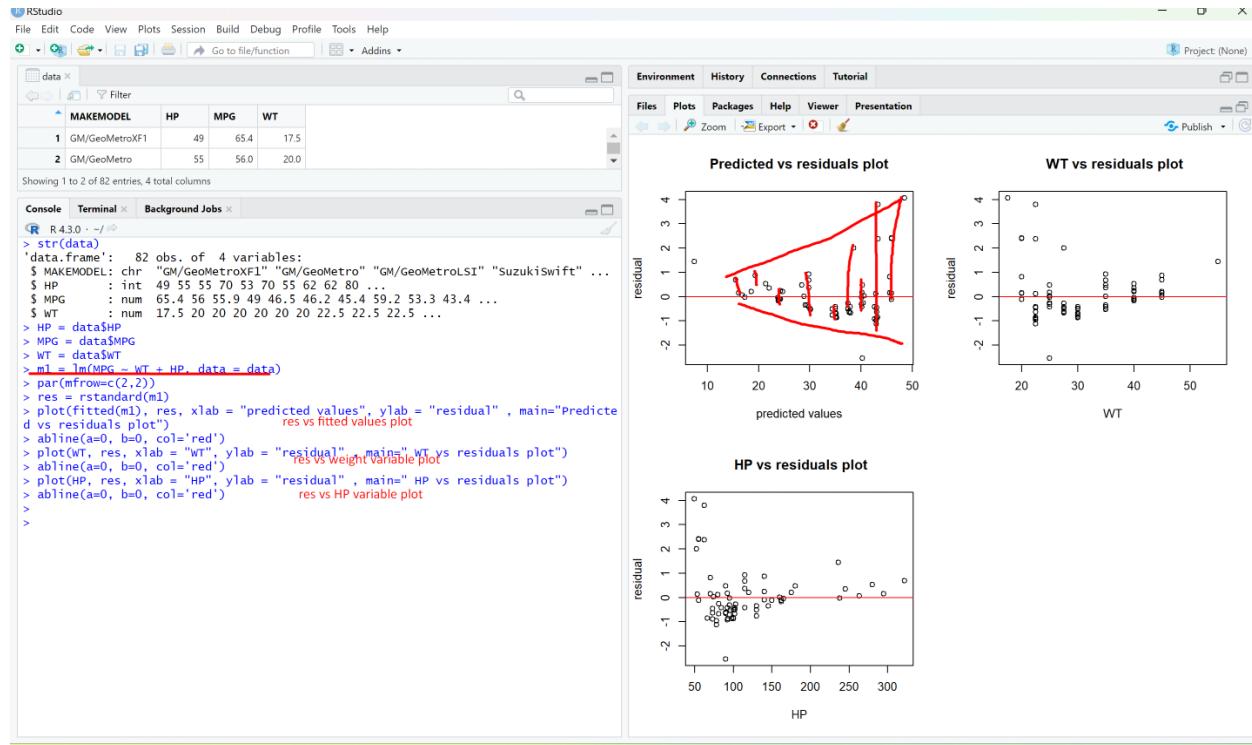
The screenshot shows the RStudio interface with the following components:

- Environment View:** Shows a data frame named "data" with 82 observations and 4 variables.
- Data View:** Displays the first 13 rows of the "data" frame, which includes columns: MAKEMODEL, HP, MPG, and WT.
- Console View:** Shows the R code used to import the data from a file named "HW4_mileage.txt".

MAKEMODEL	HP	MPG	WT
GM/GeoMetroXF1	49	65.4	17.5
GM/GeoMetro	55	56.0	20.0
GM/GeoMetroLSI	55	55.9	20.0
SuzukiSwift	70	49.0	20.0
DaihatsuCharade	53	46.5	20.0
GM/GeoSprintTurbo	70	46.2	20.0
GM/GeoSprint	55	45.4	20.0
HondaCivicCRXHF	62	59.2	22.5
HondaCivicCRXHF	62	53.3	22.5
DaihatsuCharade	80	43.4	22.5
SubaruJusty	73	41.1	22.5
HondaCivicCRX	92	40.9	22.5

```
R 4.3.0 - / 
> data = read.table(file = "C:/Users/satya/OneDrive/Desktop/HW4_mileage.txt", header = T, sep =
"\"")
> View(data)
> str(data)
'data.frame': 82 obs. of 4 variables:
 $ MAKEMODEL: chr "GM/GeoMetroXF1" "GM/GeoMetro" "GM/GeoMetroLSI" "SuzukiSwift" ...
 $ HP        : int 49 55 55 70 53 70 55 62 62 80 ...
 $ MPG       : num 65.4 56 55.9 49 46.5 46.2 45.4 59.2 53.3 43.4 ...
 $ WT        : num 17.5 20 20 20 20 20 20 22.5 22.5 22.5 ...
```

a) Fit a regression model to predict miles per gallon (MPG) from WEIGHT and horsepower (HP). Analyze residual plots to evaluate if the regression model is appropriate. Discuss if there is any evidence that the assumption of non-constant variance is satisfied by the data.



We draw residual plots: (assumptions on residual)

1. res vs fitted value/pred values: To check if res has constant variance or not.
2. res vs wt variable: To check if res has linear r/n with Weight(x1-variable) or not.
3. res vs hp variable: To check if res has linear r/n with HP(x2-variable) or not.

- To say whether the regression model (m1) is appropriate or not, we need to look for patterns or deviations in the residual plots or is there any constant variance in the res vs pred values plot.
- Specifically, we are interested in assessing whether there is evidence of non-constant variance. If the residual plots exhibit a cone or fan-like shape, it suggests non-constant variance.

EVIDENCE:

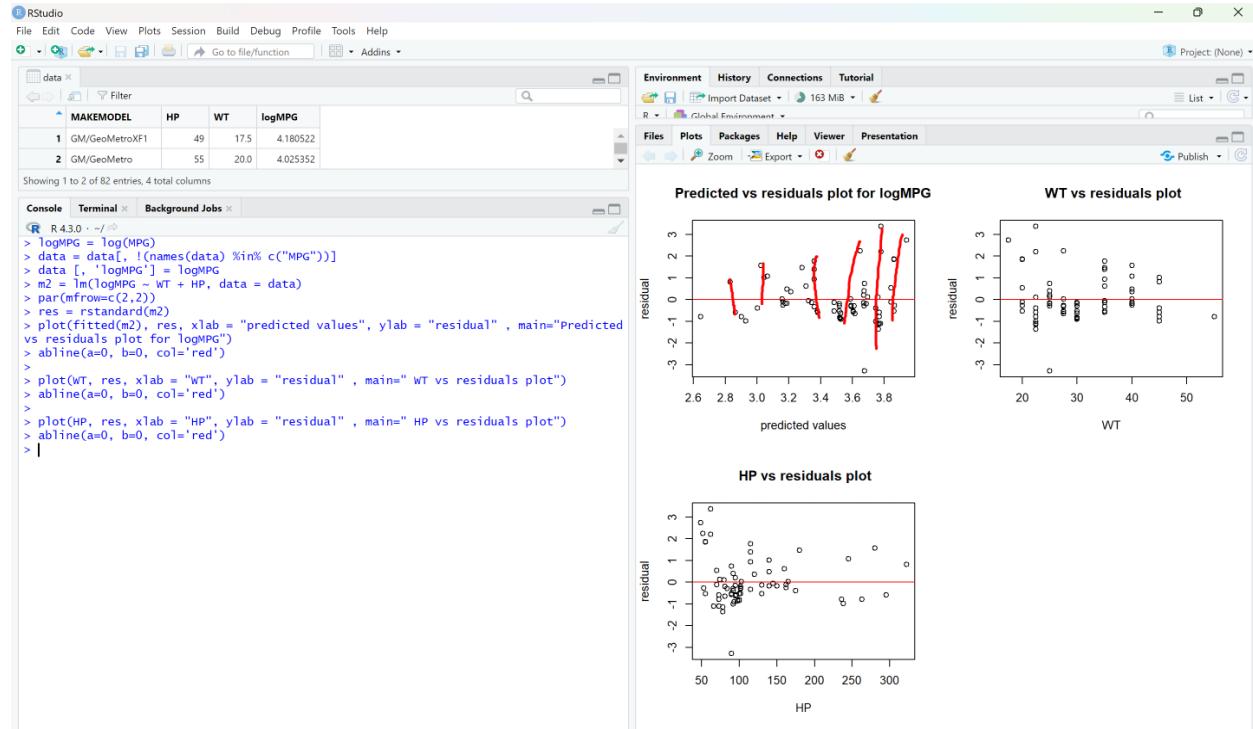
- From the res vs fitted values plot we can clearly observe that the variance is not constant. The variance on the left is small and on the right the variance is large.
- In other words, the variance increased from left to right and exhibits a fan shaped pattern which indicates that there is non-constant variance by the data.
- As the variance is not constant, we need to apply transformation on y-variable (MPG)

ANALYSING THE MODEL IS APPROPRIATE OR NOT:

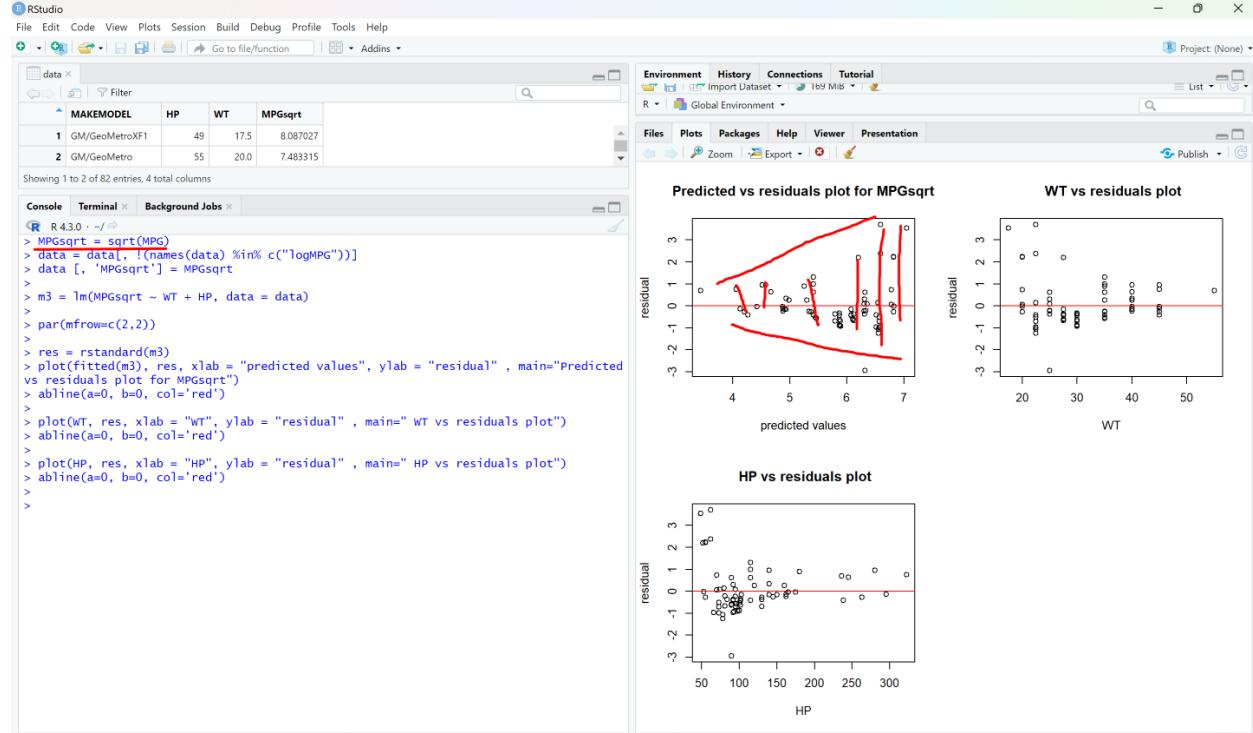
- As, there is no constant variance, **the model (m1) is not appropriate.**
- After transforming y-variable, we need to rebuild the model and check.

b) Apply a transformation to the Y variable (for instance you can try $\log(y)$, \sqrt{y} or $1/y$). Find the regression model that seems more appropriate for the analysis.

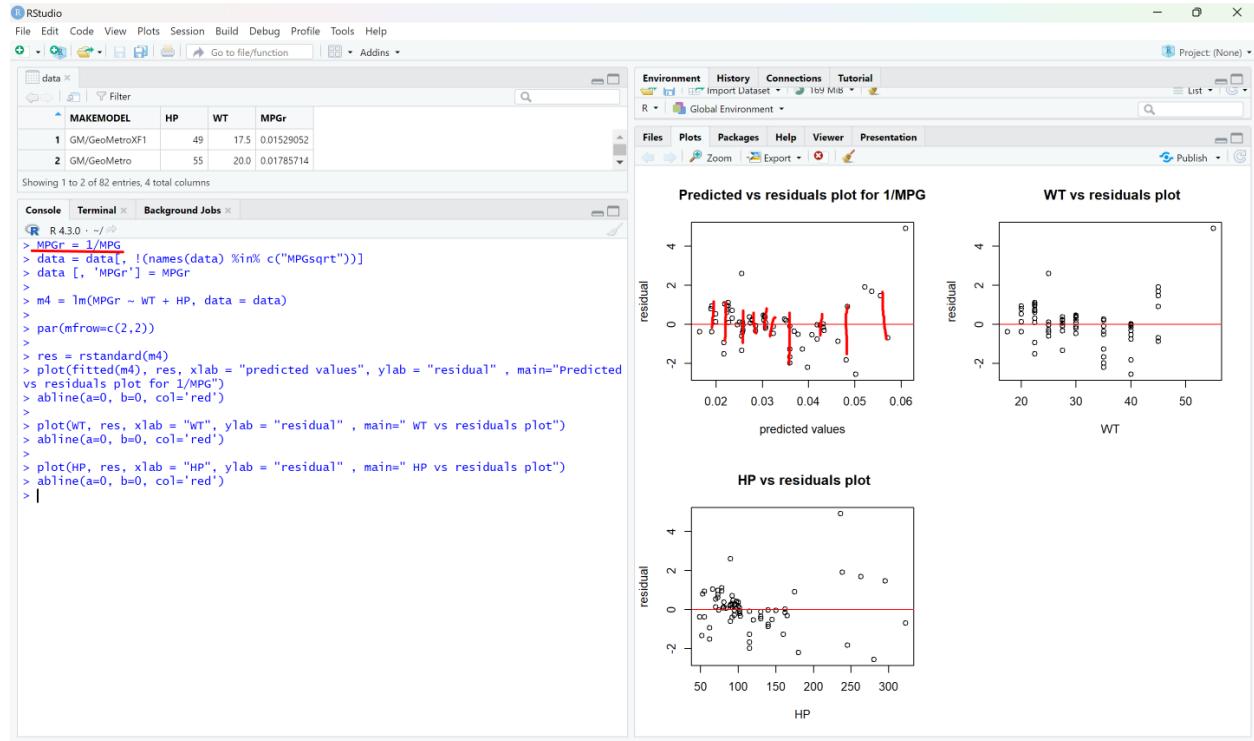
Log transformation: m2 (logmpg)



Sqrt transformation: m3 (sqrtmpg)



Reverse transformation: m4 (1/mpg) mpgr



Observation:

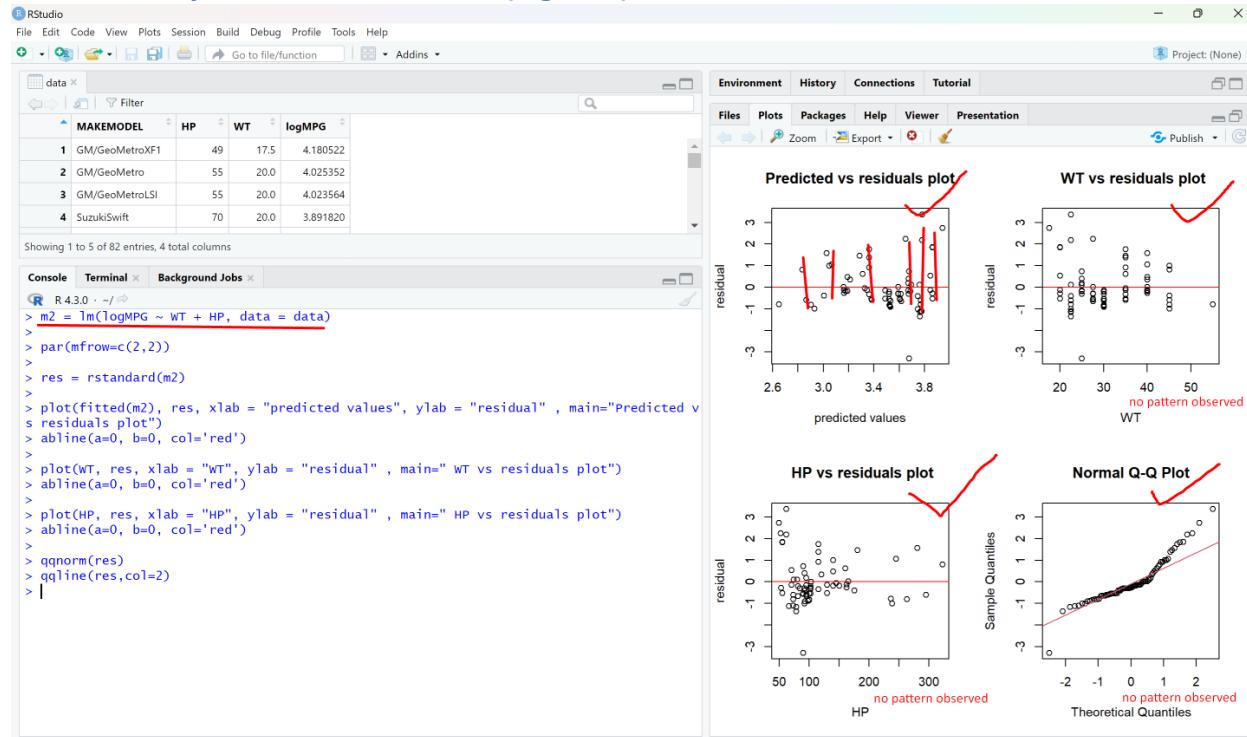
We did transformation on y-variable 'mpg' and built models. And analyzed res plots with all three transformed variables:

No-transformation(M1)	Original mpg	Model m1 (mpg~ wt+hp)	Res plot	
Log transformation (M2)	log mpg	model m2 (logmpg ~ wt + hp)	Res plot	More appropriate
Sqrt transformation (M3)	mpgsqrt	model m3 (mpgsqrt ~ wt + hp)	Res plot	
Rev transformation (M4)	1/mpg(mpgr)	model m4 (mpgr ~ wt + hp)	Res plot	

- We found that model m2 (log mpg model) is more appropriate in terms of res constant variance with x-variables and pred/fitted values & adj r² and co-efficients interpretability.

c) Analyze residual plots to evaluate if the regression model for the transformed Y variables is adequate. Do the plots show a deviation from the assumption of constant variance?

Residual analysis for final model: m2 (logMPG)



No, the plots do not show deviation from constant variance.

** in the qq plot, the residual slightly follows normal distribution, but it depends on the size of the data. As our data size 24 rows, there is no effective solution. *****

d) Write down the expression of the regression model and interpret the estimated values of the regression parameters.

Final model we picked : M2

```

R 4.3.0 · ~/r/
> summary(m2)

Call:
lm(formula = logMPG ~ WT + HP, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.29269 -0.04957 -0.02358  0.03498  0.29937 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.493022  0.0438439 102.621 < 2e-16 ***
WT          -0.0286732  0.0022087 -12.982 < 2e-16 ***
HP          -0.0011710  0.0003164 -3.702 0.000395 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08973 on 79 degrees of freedom
Multiple R-squared:  0.9152, Adjusted R-squared:  0.9131 
F-statistic: 426.5 on 2 and 79 DF,  p-value: < 2.2e-16

> 

```

From the coefficients of the regression model m2, we can write the expression as follows:

- $\text{LogMPG} = \beta_0 + \beta_1 * \text{WT} + \beta_2 * \text{HP} + e$
- i.e.,
- $\text{LogMPG} = 4.49 + (-0.02) * \text{WT} + (-0.0017) * \text{HP} + e$

Explaining the affect:

β_0 ~ explaining the intercept:

The intercept, represented by 4.49, is the value of LogMPG when both Weight and Horsepower are equal to zero. In this context, the intercept represents the baseline or average LogMPG value when there is no weight or horsepower influencing the MPG. It is a constant term in the equation.

β_1, β_2 ~ The slopes represent how the logarithm of MPG changes with respect to the independent variables, Weight and Horsepower:

β_1 ~ The slope for the term (-0.02) * Weight indicates the change in LogMPG for every unit increase in Weight, assuming other variables are held constant. A negative slope suggests that as the weight of the vehicle increases, the LogMPG decreases. This implies that heavier vehicles tend to have lower fuel efficiency.

$$\beta_2 \sim$$

The slope for the term (-0.0017)*Horse Power indicates the change in LogMPG for every unit increase in Horse Power, **assuming other variables are held constant**. A negative slope suggests that as the Horsepower of the vehicle increases, the LogMPG decreases. This implies that more powerful vehicles tend to have lower fuel efficiency.

e) Use the step function to adopt both backward and stepwise forward selection to build a new model, compare the new model with the previous model in terms of the adj-R2 value.

The screenshot shows the RStudio interface with the following details:

- Data View:** Shows a table named "MAKEMODEL" with columns HP, WT, and logMPG. One entry is visible: GM/MetroX1, HP=49, WT=17.5, logMPG=4.180522.
- Console View:**

```
R 4.3.0 -- / 
> base = lm(logMPG ~ WT, data = data)
> full = lm(logMPG ~ WT + HP, data = data)
> step(base, scope=list(upper=full, lower=-1), direction="both", trace=F)

Call:
lm(formula = logMPG ~ WT + HP, data = data)

Coefficients:
(Intercept)          WT          HP
4.499302   -0.028673   -0.001171

> newmodel = lm(logMPG ~ WT + HP, data = data)
> summary(newmodel)

Call:
lm(formula = logMPG ~ WT + HP, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.29269 -0.04957 -0.02358  0.03498  0.29937

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.499302  0.0438439 102.621 < 2e-16 ***
WT         -0.0286732 0.0022087 -12.982 < 2e-16 ***
HP        -0.0011710  0.0003164 -3.702 0.000395 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08973 on 79 degrees of freedom
Multiple R-squared:  0.9152,  Adjusted R-squared:  0.9131
F-statistic: 426.5 on 2 and 79 DF,  p-value: < 2.2e-16
```
- Environment View:** Shows the global environment with objects like base, data, full, m1, m2, m3, m4, and newmodel.
- Plots View:** Not visible in the screenshot.
- Packages View:** Not visible in the screenshot.
- Help View:** Not visible in the screenshot.
- Viewer View:** Not visible in the screenshot.
- Presentation View:** Not visible in the screenshot.

Previous model: (m2) after transformation on y-variable ~ logMPG:

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins Project: (None)

data
MAKEMODEL HP WT logMPG
1 GM/MetroX1 49 17.5 4.180522
Showing 1 to 1 of 82 entries, 4 total columns

Console Terminal Background Jobs
R 4.3.0 · ~/~
> summary(m2)

Call:
lm(formula = logMPG ~ WT + HP, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.29269 -0.04957 -0.02358  0.03498  0.29937 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.4993022  0.0438439 102.621 < 2e-16 ***
WT          -0.0286732  0.0022087 -12.982 < 2e-16 ***
HP          -0.0011710  0.0003164 -3.702 0.000395 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08973 on 79 degrees of freedom
Multiple R-squared:  0.9152,  Adjusted R-squared:  0.9131 
F-statistic: 426.5 on 2 and 79 DF,  p-value: < 2.2e-16

> |

```

Comparison:

MODEL	ADJ r2
M2 = after log transformation on y-variable(logMPG) lm(logMPG ~ WT + HP, data=data)	91.31%
New model = after step(both) lm(logMPG ~ WT + HP, data=data)	91.31%

Both the new-model and previous model(m2) have same adj r2 values. So, both are the same.

Problem 3 [10]

Answer the following concept questions.

- What are the differences between outliers and influential points in linear regression models.

ANSWER:

Basis for comparison	outliers	Influential points
Definition	An outlier is a single data point that drastically deviates from the pattern of the data as a whole. These are observations that are a long way from the majority of the data.	Influential points are the outliers that heavily influence the estimated regression coefficients. They have a high leverage on the model parameters. * You may have many outliers but not all of them are influential points.
Impact	The calculated regression coefficients and the overall model fit can be greatly impacted by outliers. They can change the slope and intercept of the regression line by bringing it closer or farther away from the bulk of the data points.	The slope, intercept, and standard errors of the coefficients may all be dramatically changed by influential points, which also impact the regression line. They have the ability to alter the model fit and the overall findings from the analysis. * If we include influential points and built a model, most of the datapoints will not fall on the regression line. *if we remove them, most of the data points will fall on the line.
Identification	Outliers can be identified by: *Residual plot * Residual vs pred values plot * Res vs each x-variable plot. If any data points goes beyond upper and lower bound values, they can be identified	They can be identified by: *DFFITS () *DFBETAS () *covratio () *hatvalues () *cooks. distance () If cooks.distance is larger than 4/n, they can be identified as

	as outliers.	influential points. (n = size of the data/no.of rows)
Causes	Outliers can occur due to measurement errors, data entry mistakes, natural variation, or genuine extreme observations.	Excessive values of the predictor variables are frequently linked to influential locations. A predictor's extreme value in an observation might have a disproportionately large impact on the calculated regression coefficients.
Treatment	Any data values not falling between upper bound value and lower bound value are suspected to be potential outliers and can be treated and removed.	Influential points can be treated by calculating cooks.distance. if cooks distance for any data point in a data is larger than $4/n$, those data points has to be removed.

2). How to identify a 2nd order and 3rd order terms in linear regression.

ANSWER:

- 2nd & 3rd order terms are called higher order terms.
- First, to identify if the linear regression has higher order terms or not, we need to draw scatter plot with y-variable (dependent variable) and x-variable (predictor variable/independent variables)
- If the plot shows a curve pattern, there are higher order terms in the linear regression.

2nd order term Identification:

If the curve pattern is having 1-top or 1-bottom, we can conclude that the linear regression is having quadratic relation.

In other words, predictor variables might have a quadratic relationship with the response variable.

Hence, the linear regression needs to have 2nd order term in the model.

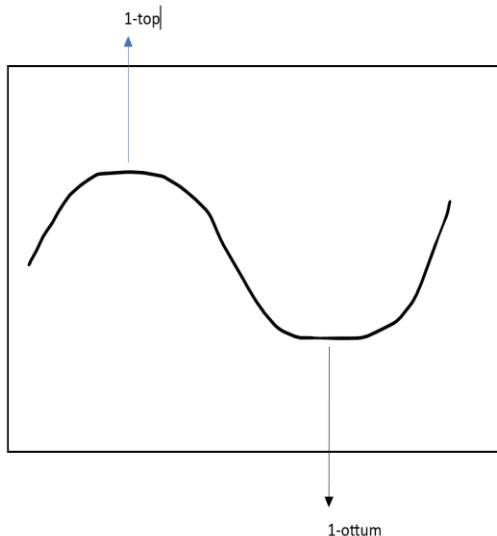


3rd order term identification:

If the curve pattern is having at least 1-top & at least 1-bottom, we can conclude that the linear regression is having cubic relation.

In other words, predictor variables might have a cubic relationship with the response variable.

Hence, the linear regression needs to have 3rd order term in the model.



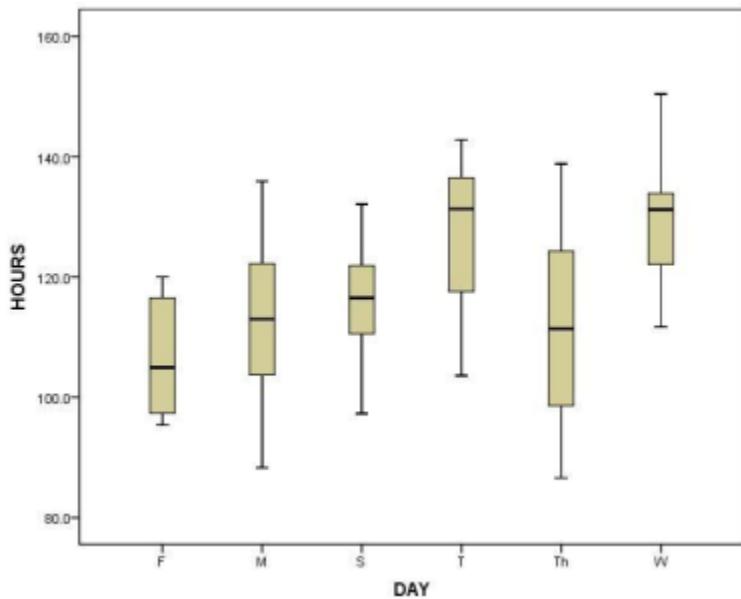
HW5

Name: Naga Satya Silpa Annadevara.

Student ID: A20517818.

let's use the "clerical_Q2.txt" data.

A store manager noticed that the busiest days for clerical staff are Wednesdays and Tuesdays. See enclosed box plot. The manager tries to compare the group means in hours by different days



a). [10] Observe the box plot. Can you confirm that the hours in Tuesday is the highest? Why?

ANSWER:

- From the boxplot, we can observe that there are 6 different groups with days variable on the x-axis and hours variable on the y-axis. For each group, we can observe a 1-box plot.
- From the visualization, we can clearly observe that the variation within the groups (boxplots) are not even/uniform. So, we cannot compare the groups based on their q2 values. Why? Because the within-group variation (IQR) is large. So, we cannot use q2 to compare the groups.
- So, we need to have additional information such as mean values for every group to compare them & give the conclusion.
- As the boxplot cannot provide mean values, we need to use ANOVA technique to say hours on Tuesday is the highest or not.
- So, with the variation difference among the groups & without knowing the mean values for every group, we cannot confirm that hours on Tuesday is the highest.

b). [20] Write down your hypothesis in the ANOVA to compare the group means in hours by different days

ANSWER:

Null hypothesis: All the groups have the same mean. In other words, the means of hours for different days are equal.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$$

- μ_1 = average of hours on Monday group
- μ_2 = average of hours on Tuesday group
- μ_3 = average of hours on Wednesday group
- μ_4 = average of hours on Thursday group
- μ_5 = average of hours on Friday group
- μ_6 = average of hours on Saturday group

Alternative hypothesis: At least 2 groups (1-pair of the mean) have different means. In other words, at least 2 group's means among the means of hours for different days are not equal.

$$H_a = \text{Not all the } \mu_t \text{ are equal} / \mu_i \neq \mu_j$$

b). [30] Using R to build the ANOVA regression model and help the manager to make the decision whether the group means in hours or different days are the same or not.

ANSWER:

Importing `clerical_Q2.txt` data into the r-studio:

The screenshot shows the RStudio interface. In the top-left pane, the 'Console' tab is active, displaying R code and its output. The code reads a dataset from a text file and prints its structure. A red box highlights the output of the `str(data)` command, which shows the data frame has 52 observations and 9 variables, with detailed types for each column. The top-right pane shows the 'Environment' tab with a list of objects: 'data' (52 obs. of 9 variables) and 'anova' (List of 13). The bottom-right pane shows the 'Files' tab.

```

> data = read.table(file = "C:/Users/satya/OneDrive/Desktop/Case2_clerical.txt", header = T, sep = "\t")
> View(data)
> str(data)
'data.frame': 52 obs. of 9 variables:
 $ day : chr "M" "T" "W" "Th" ...
 $ hours : num 128.5 7781 100 886 235 ...
 $ mail : int 7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
 $ cert : int 100 110 61 102 45 144 123 78 172 126 ...
 $ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
 $ change : int 235 388 398 457 577 345 326 161 219 287 ...
 $ check : int 644 589 1081 891 537 563 402 495 823 555 ...
 $ misc : int 56 57 59 57 49 64 60 57 62 86 ...
 $ tickets: int 737 1029 830 1468 335 918 335 962 665 577 ...
>

```

Building anova regression model using aov () function:

The screenshot shows the RStudio interface. The 'Console' tab is active, displaying R code and its output. The code performs an ANOVA regression model using the `aov` function on the 'hours' variable against 'day'. The output includes the ANOVA table, summary statistics, residuals, and coefficients. A red box highlights the F-value and p-value in the ANOVA table. The top-right pane shows the 'Environment' tab with objects 'anova' (List of 13) and 'data' (52 obs. of 9 variables). The bottom-right pane shows the 'Files' tab.

```

> anova = aov(hours ~ day, data = data)
> summary(anova)
Df Sum Sq Mean Sq F value Pr(>F)
day 5 3876 775.2 4.227 0.00303 ***
Residuals 46 8437 183.4
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
>
>
>
> anova = lm(hours ~ day, data = data)
> summary(anova)

Call:
lm(formula = hours ~ day, data = data)

Residuals:
    Min      1Q      Median      3Q      Max 
-25.000 -9.606   1.400   9.402  28.856 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 106.625    4.788  22.269 <2e-16 ***
dayM        6.675    6.581  1.014  0.315722    
dayS        9.288    6.771  1.372  0.176846    
dayT       20.531    6.581  3.120  0.003121 ***
dayTh      3.319    6.581  0.504  0.616367    
dayW       23.342    6.581  3.547  0.000909 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.54 on 46 degrees of freedom
Multiple R-squared:  0.3148, Adjusted R-squared:  0.2403 
F-statistic: 4.227 on 5 and 46 DF,  p-value: 0.003033

```

Decision/ conclusion:

- Using 95% confidence level, as the p-value (0.00303) in F-test is smaller than alpha ($p < 0.05\%$) we do have enough evidence to reject null hypothesis.
- In other words, we have enough evidence to reject that all group means in hours or different days are not same. At least 2 group means are different.
- To know which 2 groups means are different, we need to look at the t-test. We can clearly observe that Tuesday and Wednesday group means are not same.

c). [20] Try to interpret the coefficients you got in the ANOVA regression model from part b).

ANSWER:

We cannot see co-efficients in the model which we built using `aov()` function.

So we need to use other modeling approaches such as linear regression (e.g., `lm()`)

So built model using `lm()` function with same variables to see the co-efficients.

```
R 4.3.0 - /~/
> anova = aov(hours ~ day, data = data)
> summary(anova)
Df Sum Sq Mean Sq F value Pr(>F)
day      5   3876   775.2   4.227 0.00303 ***
Residuals 46   8437   183.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
>
>
> anova = lm(hours ~ day, data = data)
> summary(anova)
Call:
lm(formula = hours ~ day, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.000 -9.606  1.400  9.402 28.856 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 106.625    4.788  22.269 2.2e-16 ***
dayM         6.675    6.581  1.014 0.315722    
dayS         9.288    6.771  1.372 0.176846    
dayT        20.011    6.581  3.120 0.003121 ***
dayTh       3.319    6.581  0.504 0.616367    
dayW        23.342    6.581  3.547 0.000909 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.54 on 46 degrees of freedom
Multiple R-squared:  0.3148, Adjusted R-squared:  0.2403 
F-statistic: 4.227 on 5 and 46 DF,  p-value: 0.003033
```

Interpretation of co-efficients: as there is no dayFriday, we are going to use Friday as baseline.

$$\text{hours} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \beta_5 * X_5 + e$$

- Intercept (+106.625): $\sim \beta_0$

The intercept represents the estimated mean value for the reference group (Friday) when all other dummy variables are zero. In this case, it suggests that the estimated mean value for Friday is 106.625.

- dayM (+6.675): $\beta_1 \dots X_1 = \text{dayM}$

The coefficient for "dayM" (+6.675) represents the difference in mean between Monday and Friday. The positive coefficient indicates that, on average, the mean for Monday is 6.675 units higher than the mean for Friday.

- dayS (+9.288): $\beta_2 \dots X_2 = \text{dayS}$

The coefficient for "dayS" (+9.288) represents the difference in mean between Saturday and Friday. The positive coefficient indicates that, on average, the mean for Saturday is 9.288 units higher than the mean for Friday.

- dayT (+20.531): $\beta_3 \dots X_3 = \text{dayT}$

The coefficient for "dayT" (+20.531) represents the difference in mean between Tuesday and Friday. The positive coefficient suggests that, on average, the mean for Tuesday is 20.531 units higher than the mean for Friday.

- dayTh (+3.319): $\beta_4 \dots X_4 = \text{dayTh}$

The coefficient for "dayTh" (+3.319) represents the difference in mean between Thursday and Friday. The positive coefficient indicates that, on average, the mean for Thursday is 3.319 units higher than the mean for Friday.

- dayW (+23.342): $\beta_5 \dots X_5 = \text{dayW}$

The coefficient for "dayW" (+23.342) represents the difference in mean between Wednesday and Friday. The positive coefficient suggests that, on average, the mean for Wednesday is 23.342 units higher than the mean for Friday.

d). [20] practice for data preprocessing: create N-1 dummy variables for the variable 'DAY'. Convert the variable "mail" to nominal variable by creating 4 groups. Again, paste the codes and snapshots.

ANSWER:

Created N-1 dummy variable for the variable "DAY":

```

R 4.3.0 - ~/ ~
> dummy_data <- cbind(data, model.matrix(~ day - 1, data = data))
> str(dummy_data)
'data.frame': 52 obs. of 15 variables:
 $ day : chr "M" "T" "W" "Th" ...
 $ hours : num 128 114 147 124 100 ...
 $ mail : int 7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
 $ cert : int 100 110 61 102 45 144 123 78 172 126 ...
 $ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
 $ change : int 235 388 398 457 577 345 326 161 219 287 ...
 $ check : int 644 589 1081 891 537 563 402 495 823 555 ...
 $ misc : int 56 57 59 57 49 64 60 57 62 86 ...
 $ tickets: int 737 1029 830 1463 335 918 335 962 665 577 ...
$ dayF : num 0 0 0 0 1 0 0 0 0 0 ...
$ dayM : num 1 0 0 0 0 0 1 0 0 0 ...
$ dayS : num 0 0 0 0 0 1 0 0 0 0 ...
$ dayT : num 0 1 0 0 0 0 0 1 0 0 ...
$ dayTh : num 0 0 0 1 0 0 0 0 0 1 ...
$ dayW : num 0 0 1 0 0 0 0 0 1 0 ...

```

created N-1 dummy variables for 'day'

```

> dummy_data$day <- NULL
> str(dummy_data)
'data.frame': 52 obs. of 14 variables:
 $ hours : num 128 114 147 124 100 ...
 $ mail : int 7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
 $ cert : int 100 110 61 102 45 144 123 78 172 126 ...
 $ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
 $ change : int 235 388 398 457 577 345 326 161 219 287 ...
 $ check : int 644 589 1081 891 537 563 402 495 823 555 ...
 $ misc : int 56 57 59 57 49 64 60 57 62 86 ...
 $ tickets: int 737 1029 830 1463 335 918 335 962 665 577 ...
$ dayF : num 0 0 0 0 1 0 0 0 0 0 ...
$ dayM : num 1 0 0 0 0 0 1 0 0 0 ...
$ dayS : num 0 0 0 0 0 1 0 0 0 0 ...
$ dayT : num 0 1 0 0 0 0 0 1 0 0 ...
$ dayTh : num 0 0 0 1 0 0 0 0 0 1 ...
$ dayW : num 0 0 1 0 0 0 0 0 1 0 ...

```

removed original 'day' variable from the data_set

Note: For creating dummy variable we need to use `model.matrix()` function , As dummies package was removed from the latest version of R.

Convert the variable “mail” to nominal variable by creating 4 groups.:

We need to use `cut()` function to cut the mail into 4 groups:

```

R 4.3.0 - ~/ ~
> data$mail = cut(data$mail, breaks = 4, labels = c("Group 1", "Group 2", "Group 3", "Group 4"))
> data$mail
[1] Group 3 Group 3 Group 3 Group 1 Group 2 Group 1 Group 4 Group 2 Group 2 Group 3 Group 3 Group 2 Group 2 Group 1
[14] Group 3 Group 1 Group 1 Group 2 Group 3 Group 3 Group 2 Group 2 Group 3 Group 1 Group 2 Group 2 Group 2
[27] Group 1 Group 2 Group 2 Group 3 Group 3 Group 2 Group 2 Group 3 Group 2 Group 3 Group 2 Group 1
[40] Group 2 Group 2 Group 2 Group 2 Group 2 Group 1 Group 2 Group 1 Group 2 Group 2 Group 1 Group 1 Group 3
Levels: Group 1 Group 2 Group 3 Group 4

```

Note: as the group values came into scientific notation, I used labels to improve the readability.

