

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 $^{\circ}\text{F}$) (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:

(Include the new variables in the regression model)

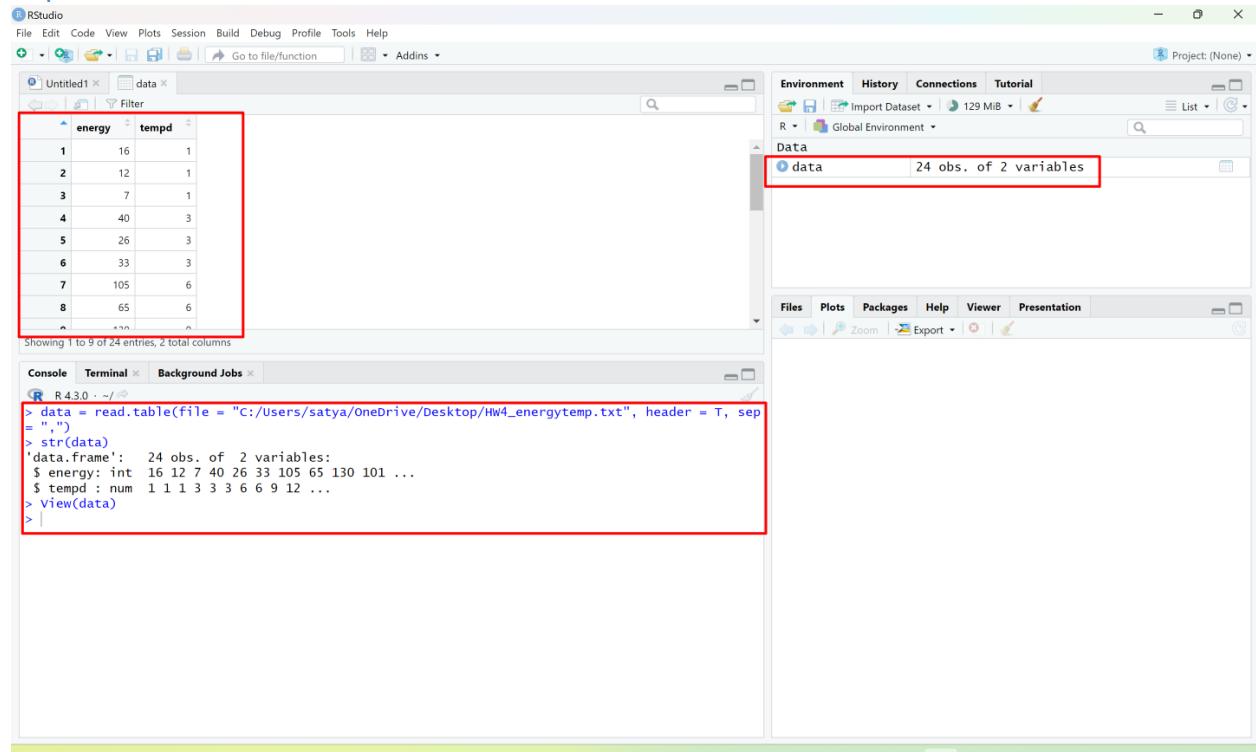
- a) Test the goodness of fit of the model at the 5% significance level.
 - b) Are all variables in the model significant?
 - c) Create the residual plots (residuals vs predicted; residuals vs x variable; and normal plot of residuals). Analyze residual plots to evaluate the normality and constant variance 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 $\text{TEMPD}=10$.

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

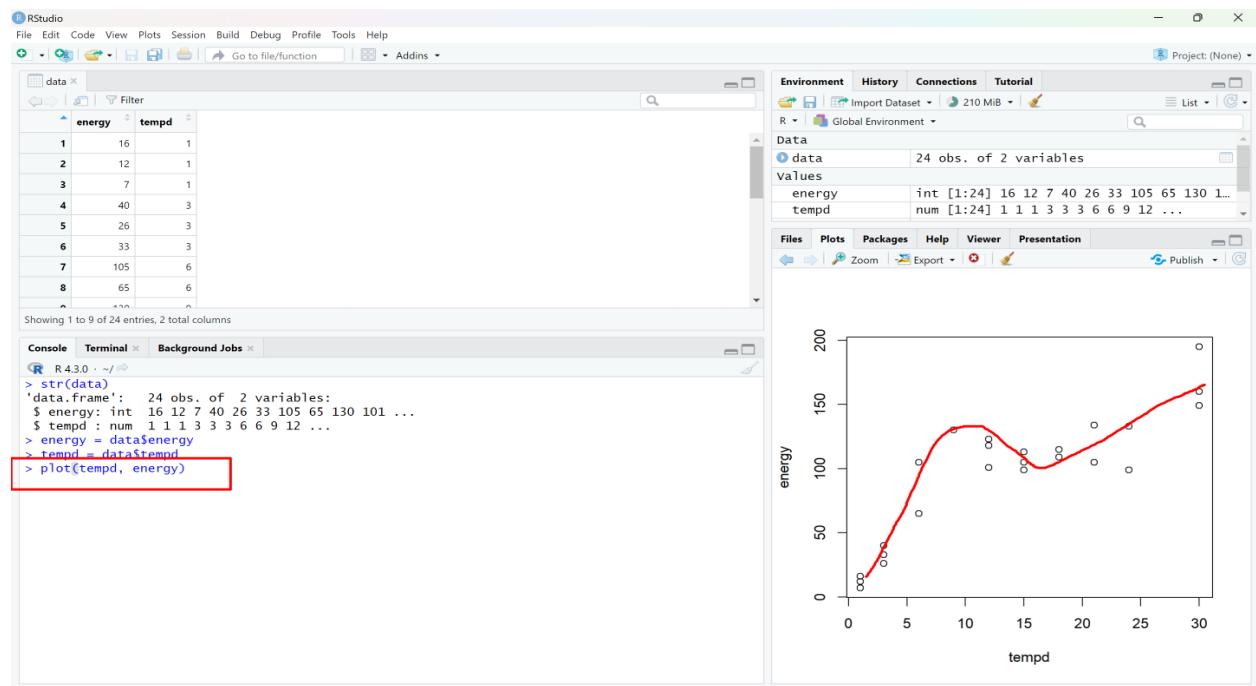
- f) By using influence.measures() function to identify whether there are influential points that can affect your final model. Use cook's distance as the metric to identify the influential points.

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.

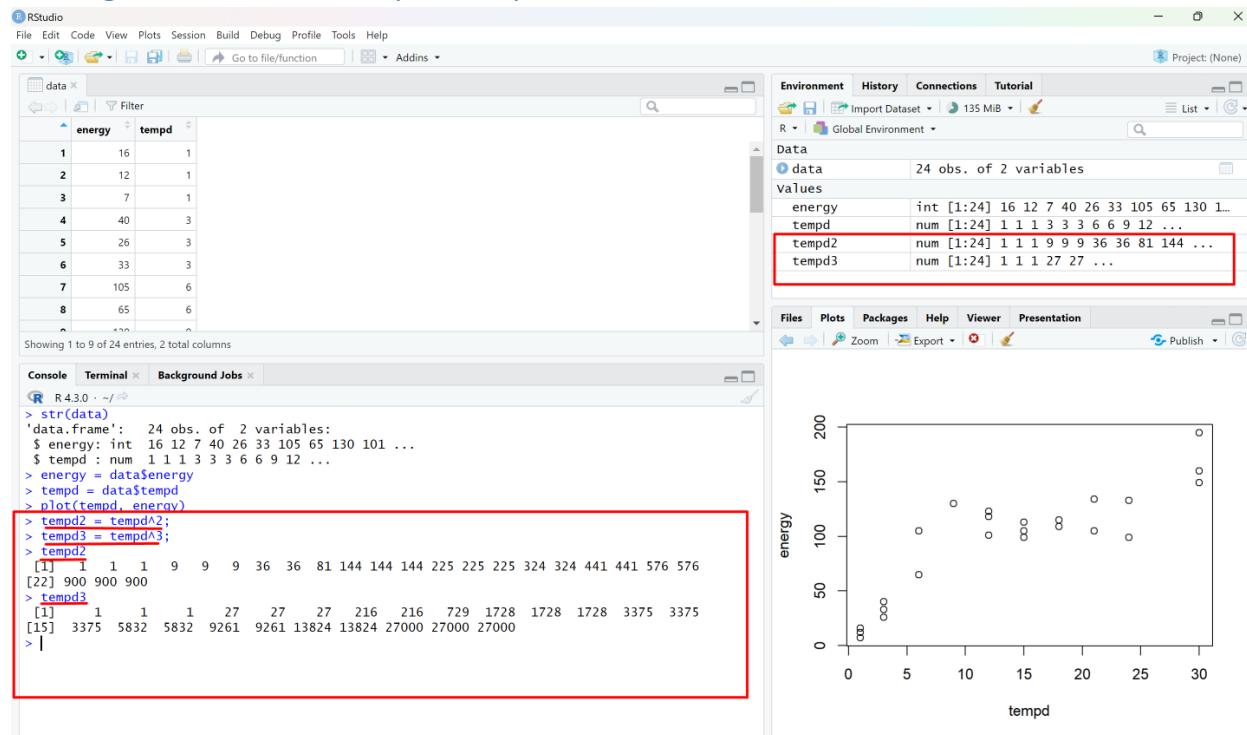


Analyzing the association b/w the 2 variables:

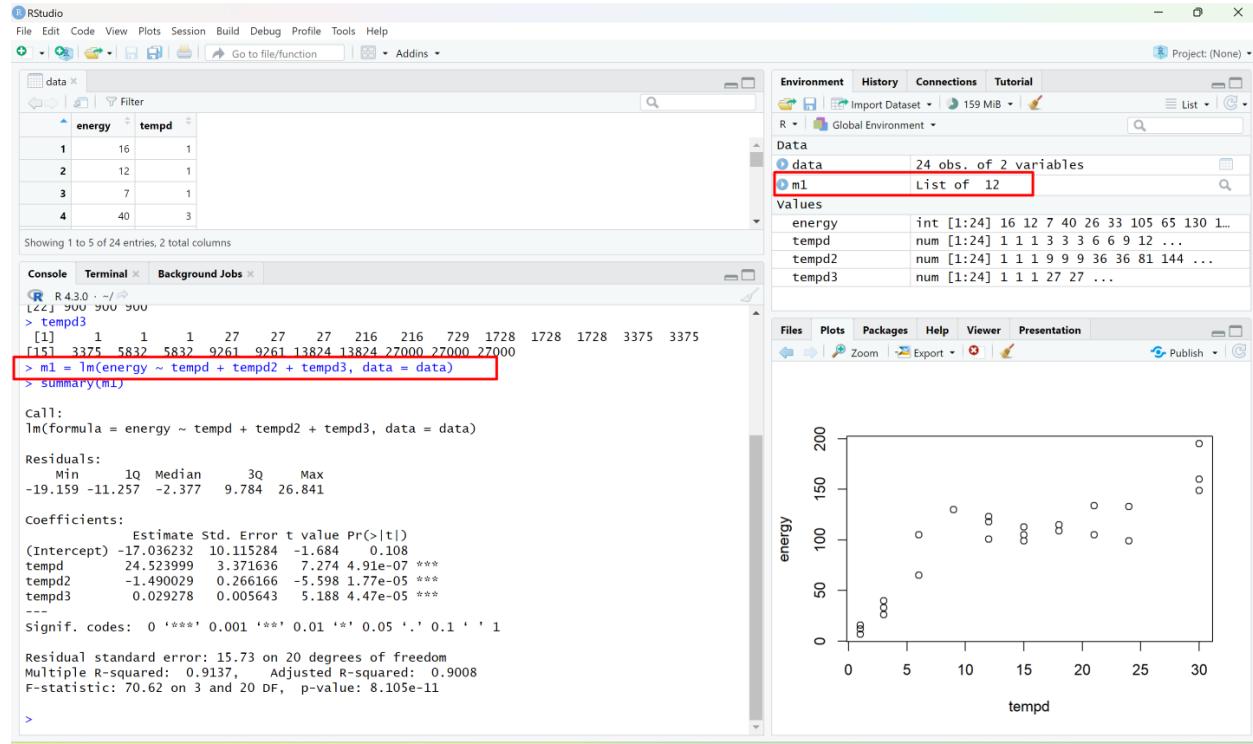
- From the scatterplot, we can observe that there is 1-top & 1-bottom in the curve, and the scatterplot produced S-shaped pattern. Hence, we can conclude that the association between energy (y) and tempd (X) is cubic association / cubic relation.
- As the scatterplot follows cubic relation, we need to 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:

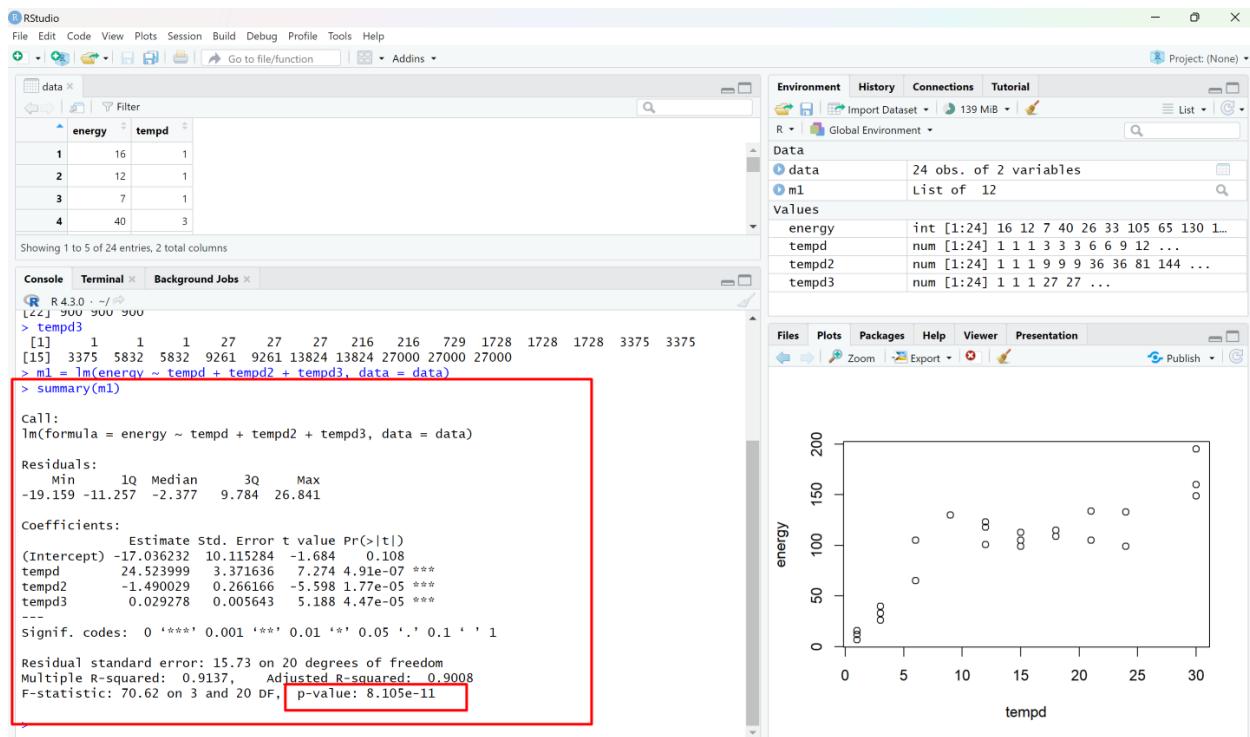
- Null hypothesis: (H_0): The coefficients of all x-variables are zero and there is no linear relationship with energy.
- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- Alternative Hypothesis: At least one of the coefficients of the x-variables (tempd, tempd2, tempd3) is not zero and can affect energy.
-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

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."

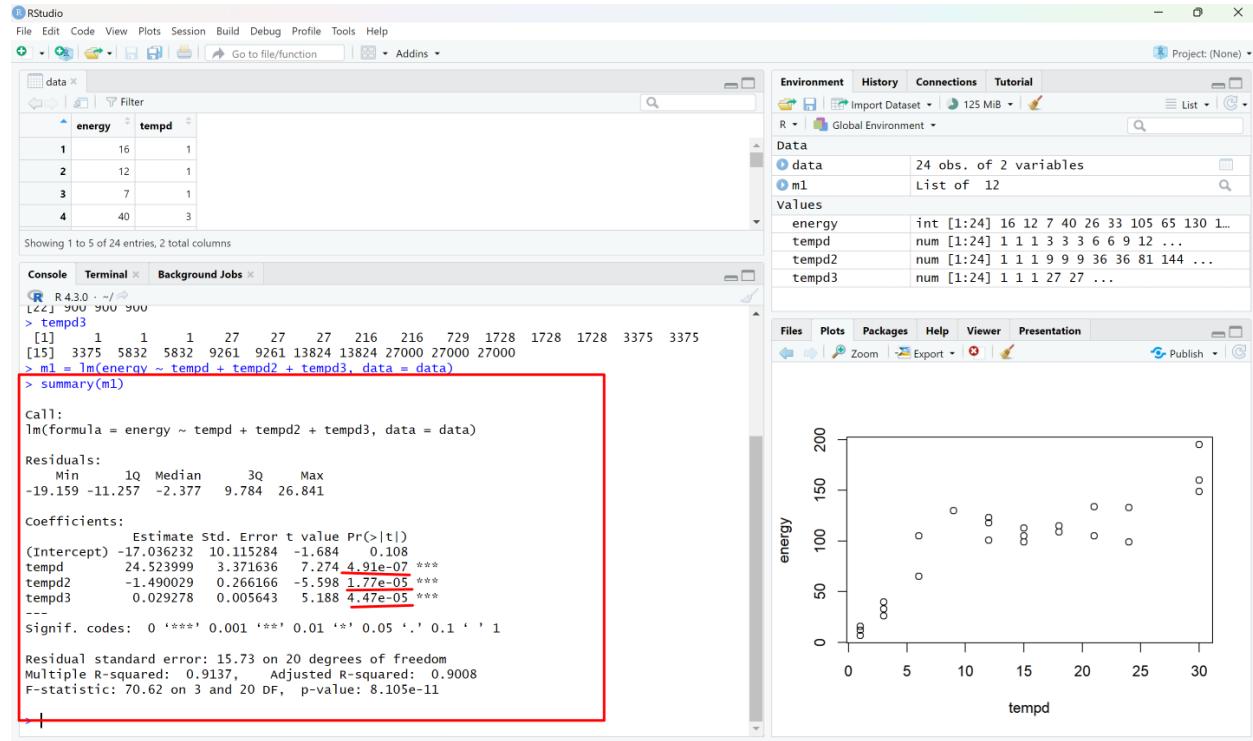


Conclusion on F-test: p-value < 0.05 for the model m1(fitted cubic regression model)

- 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)*****

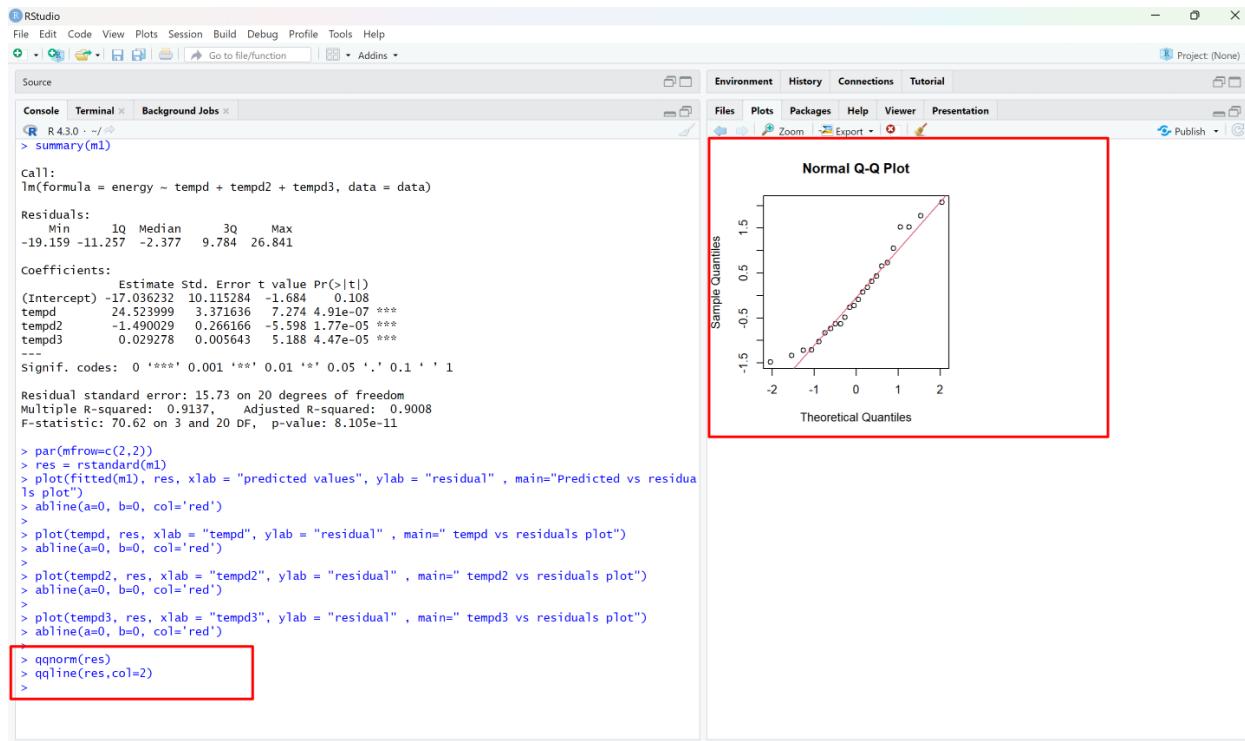
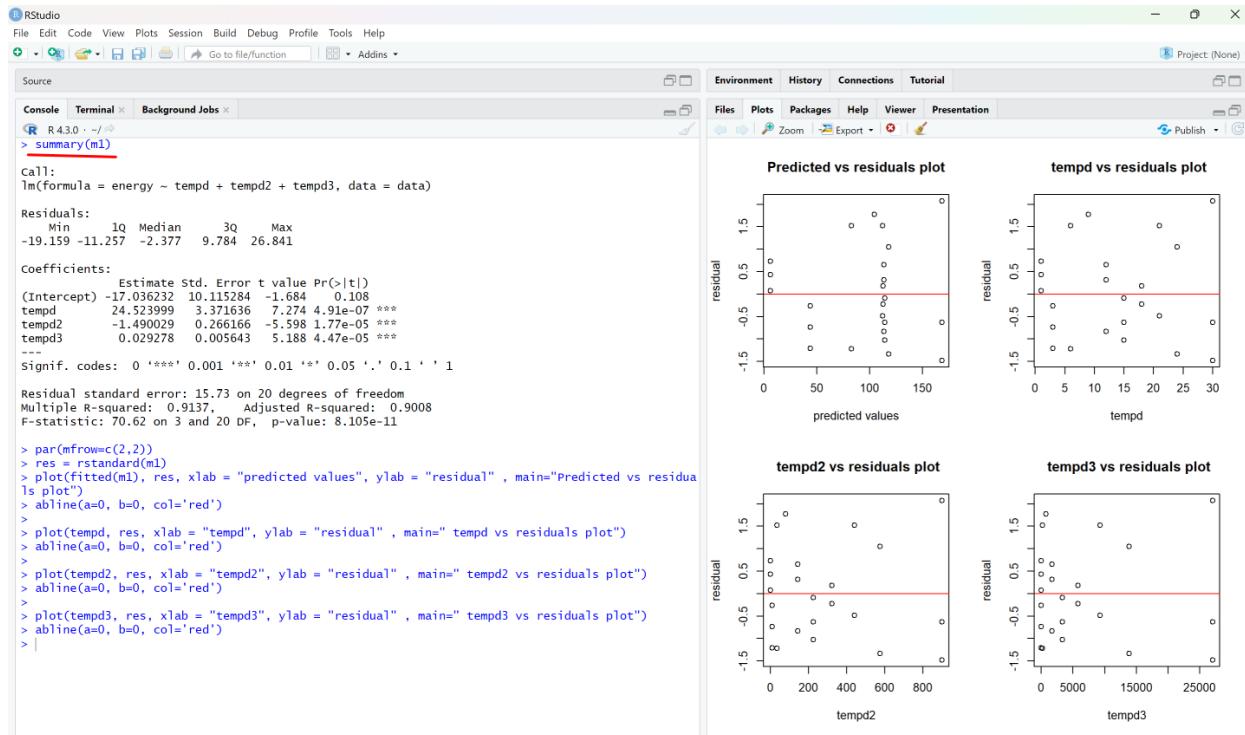
b. Are all variables in the model significant?



- 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:



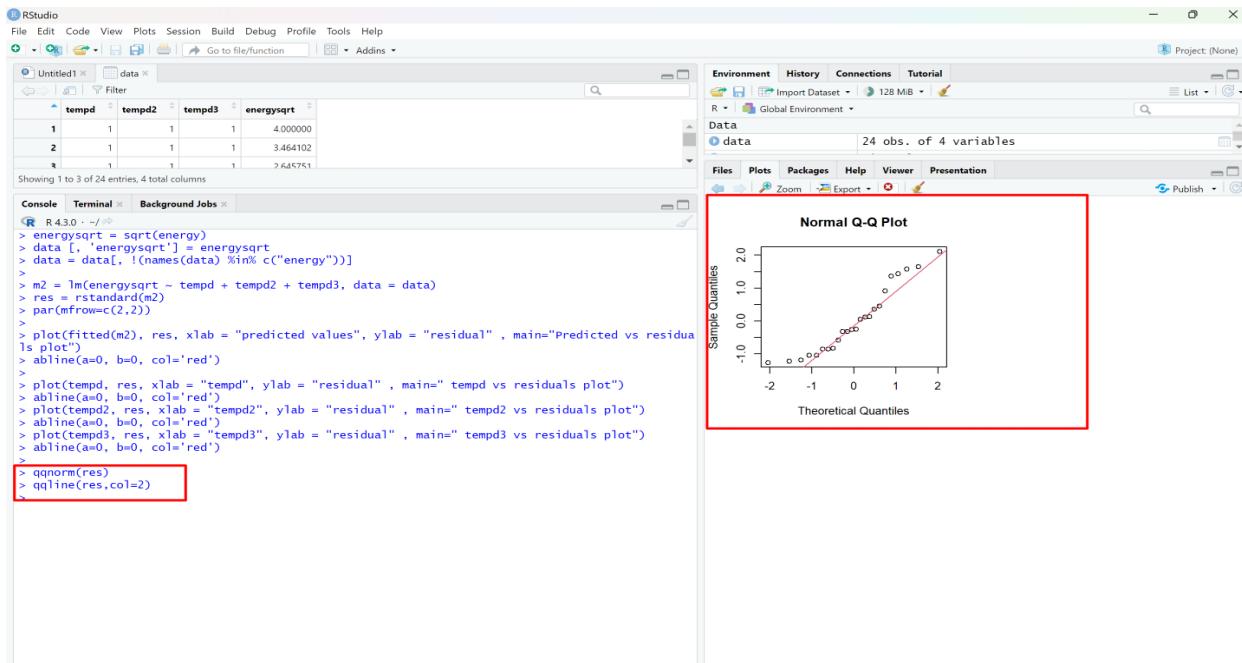
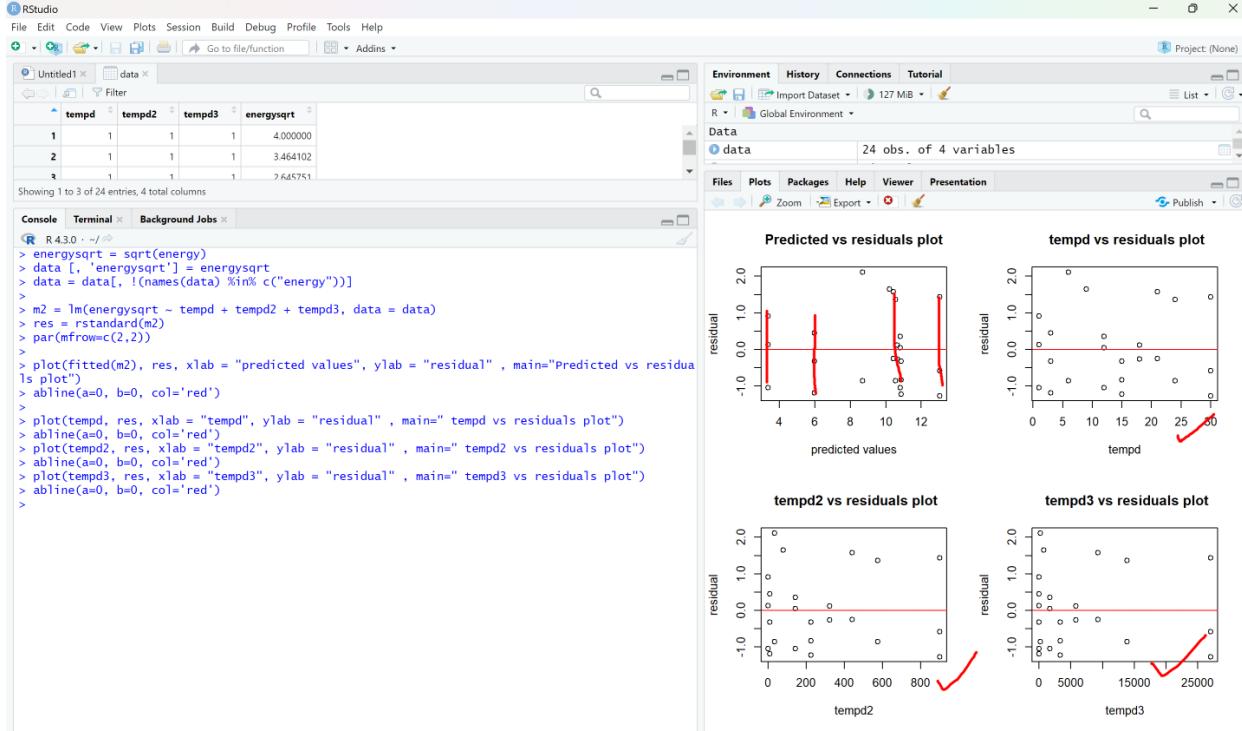
Observation: while performing res analysis for m1, we observed that

- 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.

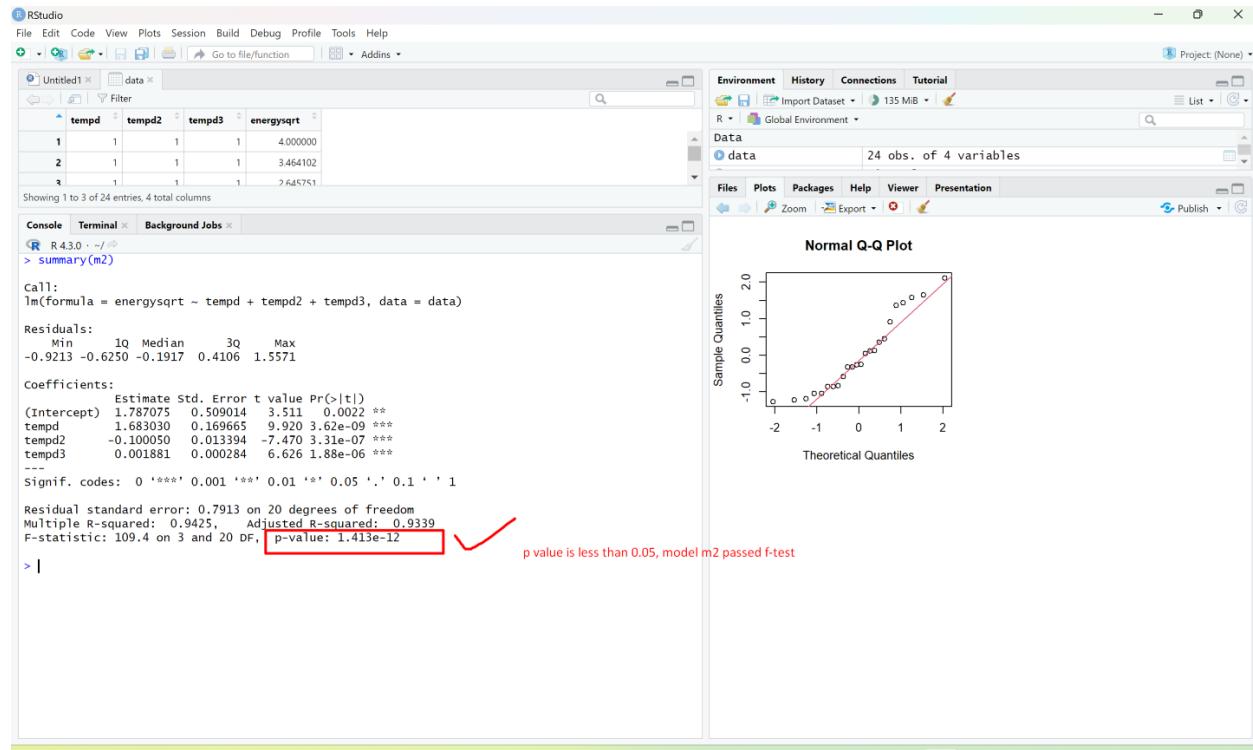
I tried log , reverse, sqrt on energy. I found sqrt transformation is better.

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(m2)

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:

- 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 (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: data (24 obs. of 4 variables), m1, m2, m3, new (1 obs. of 3 variables), and their corresponding structures.
- Data View:** A data frame named "data" is displayed with columns: tempd, tempd2, tempd3, and energysqrt. The data shows values for 10 rows, with the last row being 10, 12, 144, 1728, and 10.049876 respectively.
- Console Pane:**

```
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.I_ 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 -0.3355 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 1.364 1.03e-01 0.182
21 -0.13960  0.23575 -0.27518  0.28153 -0.3939 1.293 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:

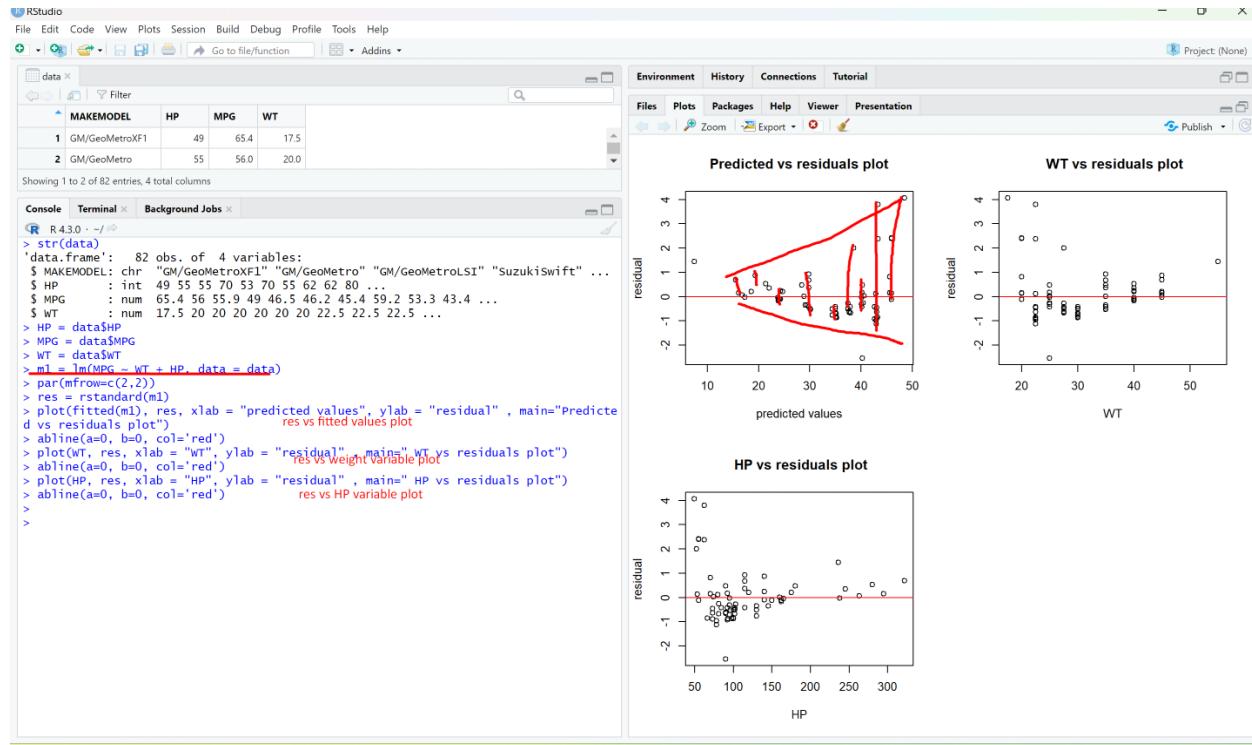
The screenshot shows the RStudio interface with the following components:

- 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 "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:

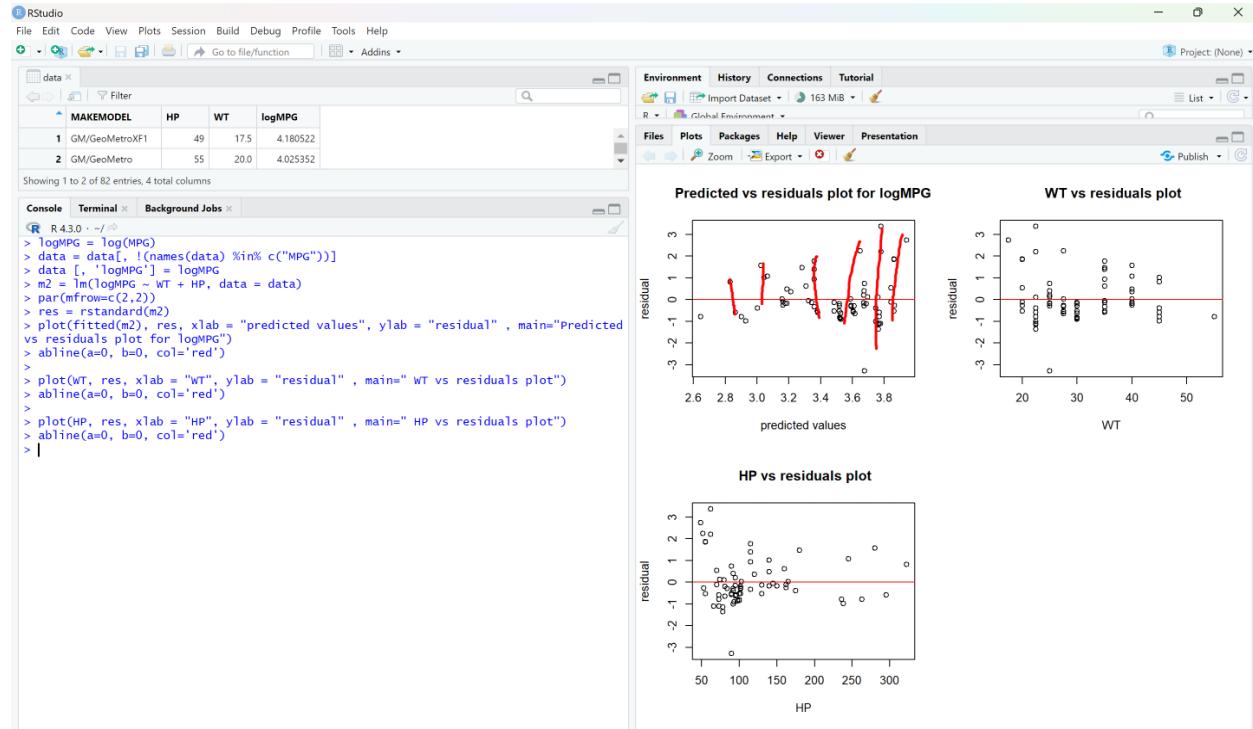
- From the res vs fitted values plot we can clearly observe that the variance is not constant. The variance on the left is small and on the right the variance is large.
- In other words, the variance increased from left to right and exhibits a fan shaped pattern which indicates that there is non-constant 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:

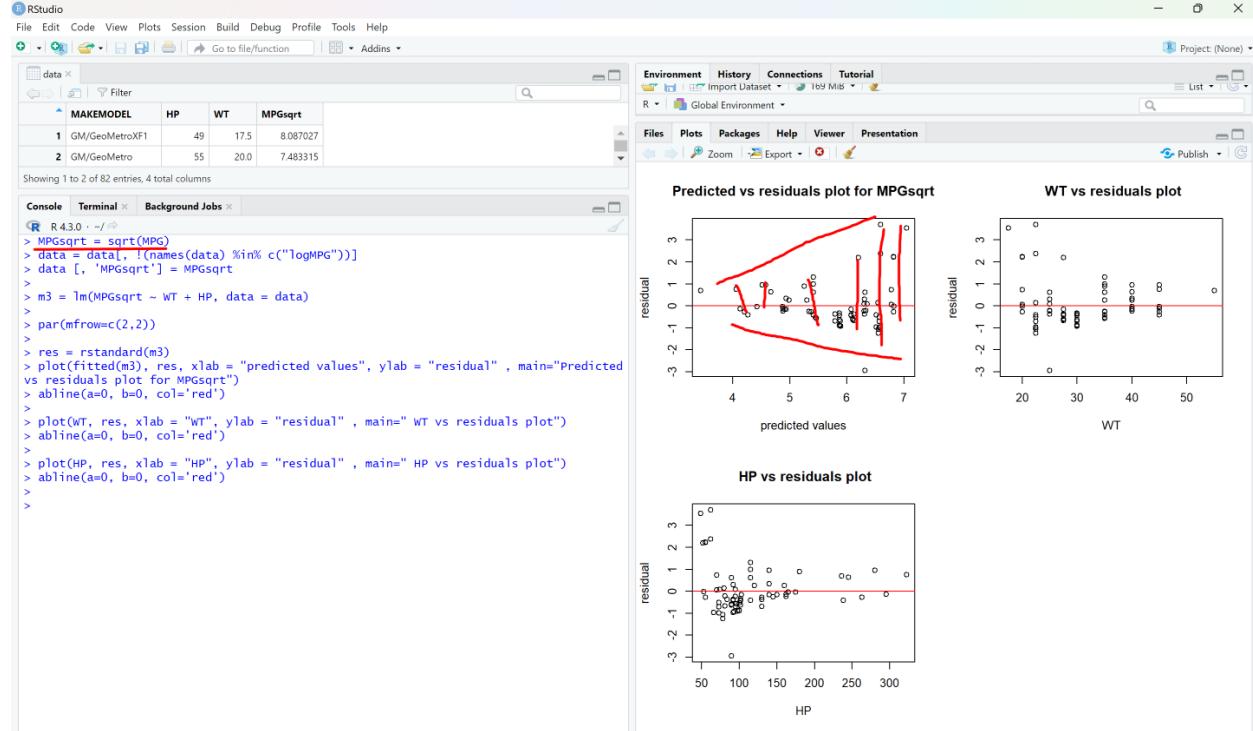
- As, there is no constant variance, **the model (`m1`) is not appropriate.**
- After transforming y-variable, we need to rebuild the model and check.

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.

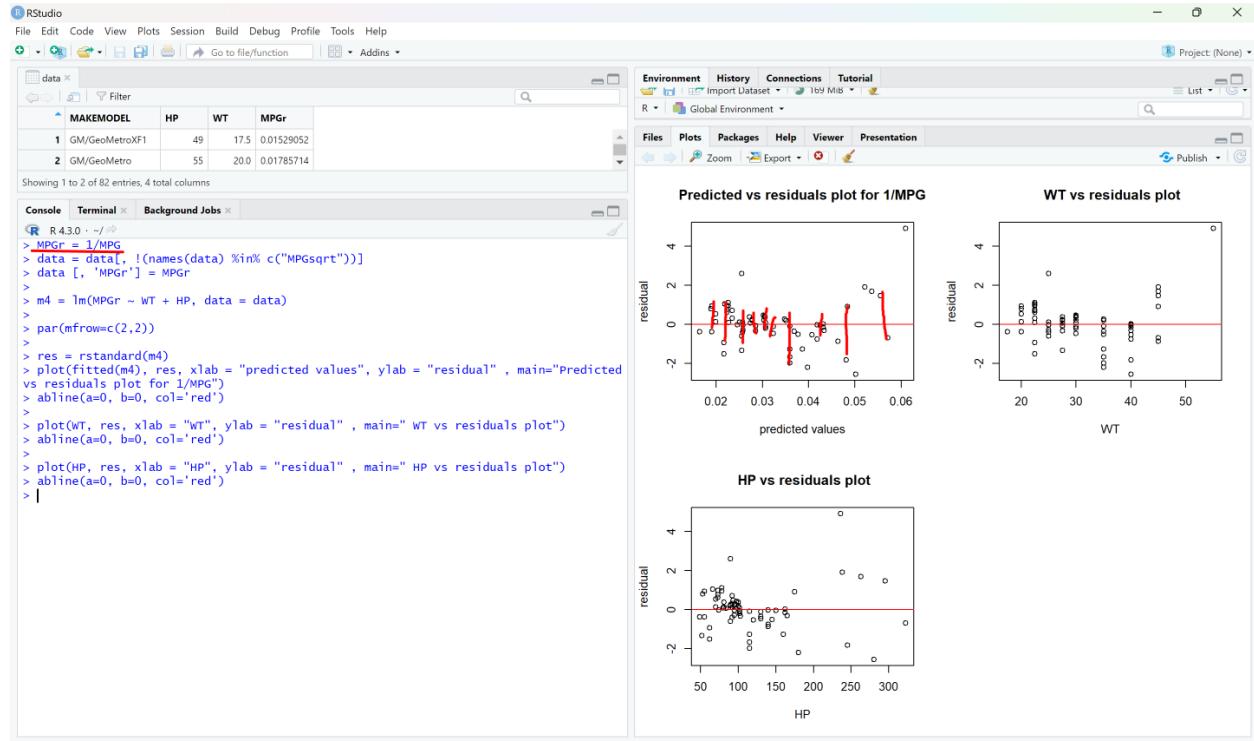
Log transformation: m2 (logmpg)



Sqrt transformation: m3 (sqrtmpg)



Reverse transformation: m4 (1/mpg) mpgr



Observation:

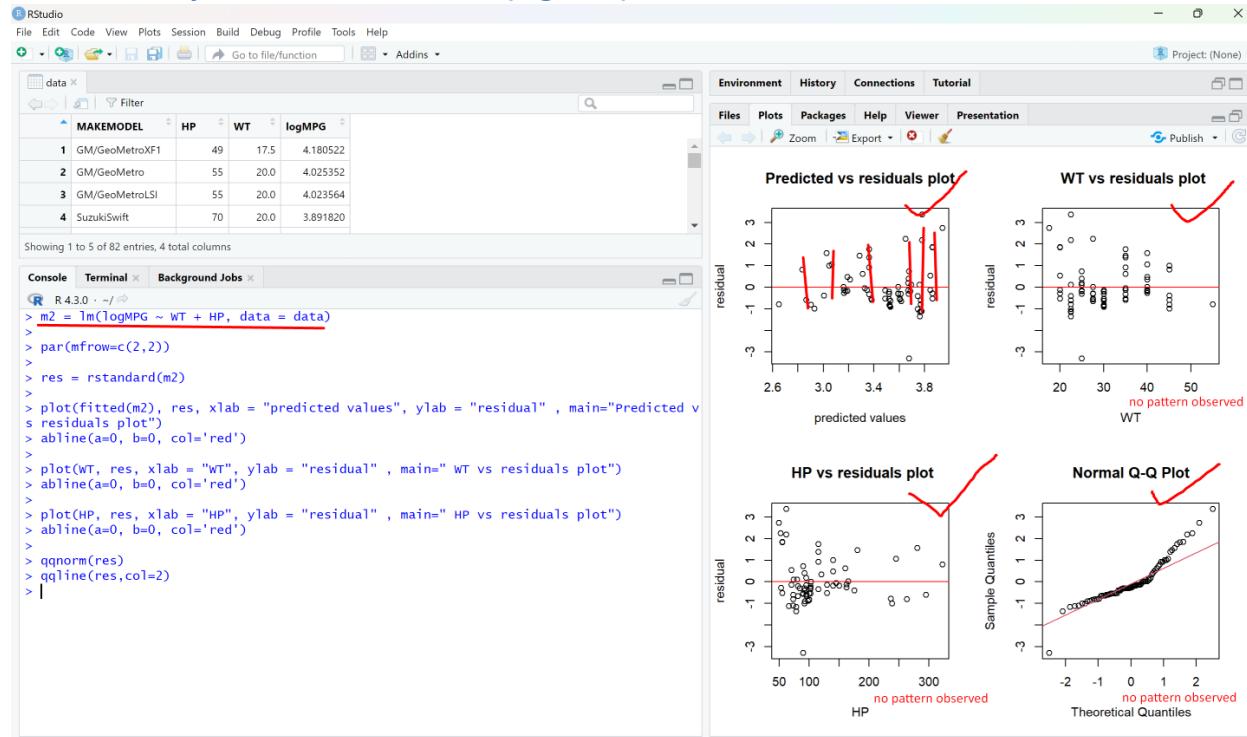
We did transformation on y-variable 'mpg' and built models. And analyzed res plots with all three transformed variables:

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 - ~
> 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:

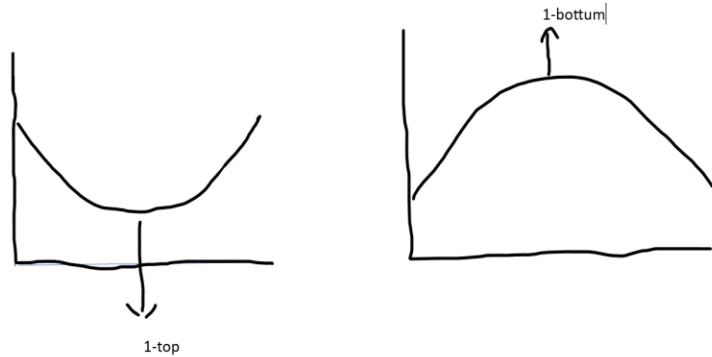
- 2nd & 3rd order terms are called higher order terms.
- First, to identify if the linear regression has higher order terms or not, we need to draw scatter plot with y-variable (dependent variable) and x-variable (predictor variable/independent variables)
- If the plot shows a curve pattern, there are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

