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.

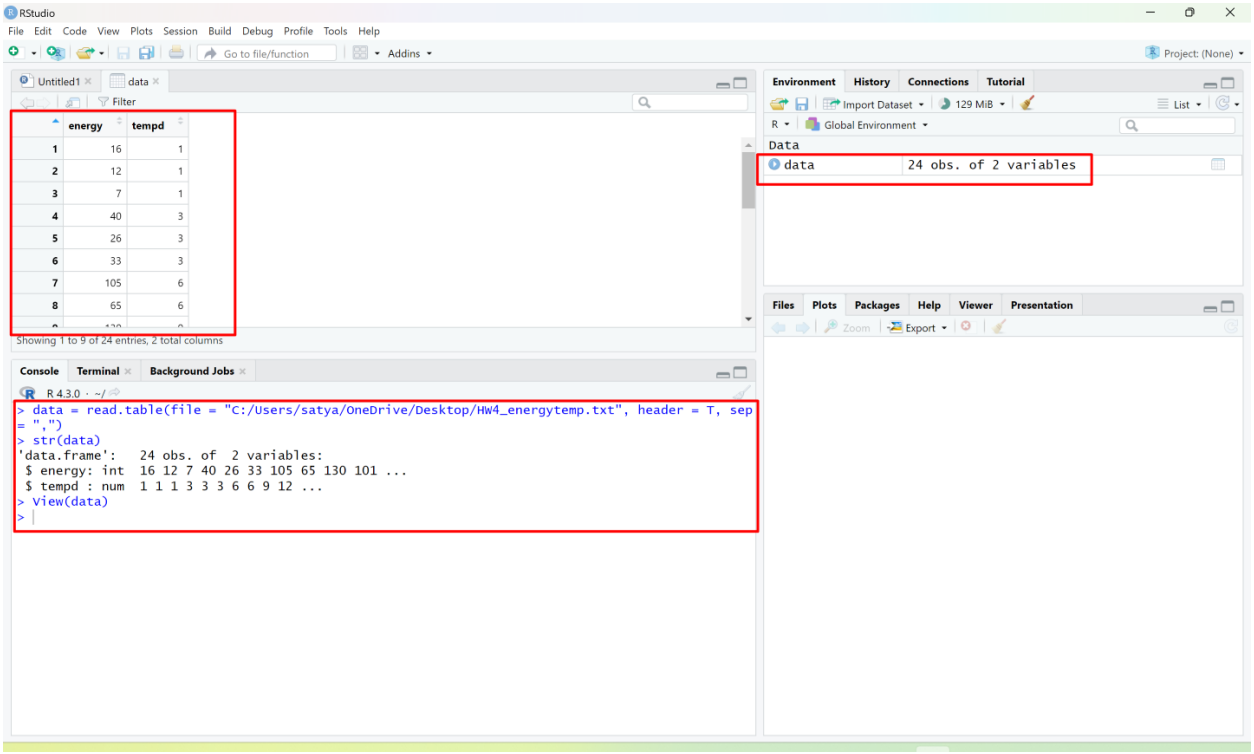
e) Use the fitted regression model to predict the average energy consumption for an average difference in temperature equal to TEMPD=10.

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

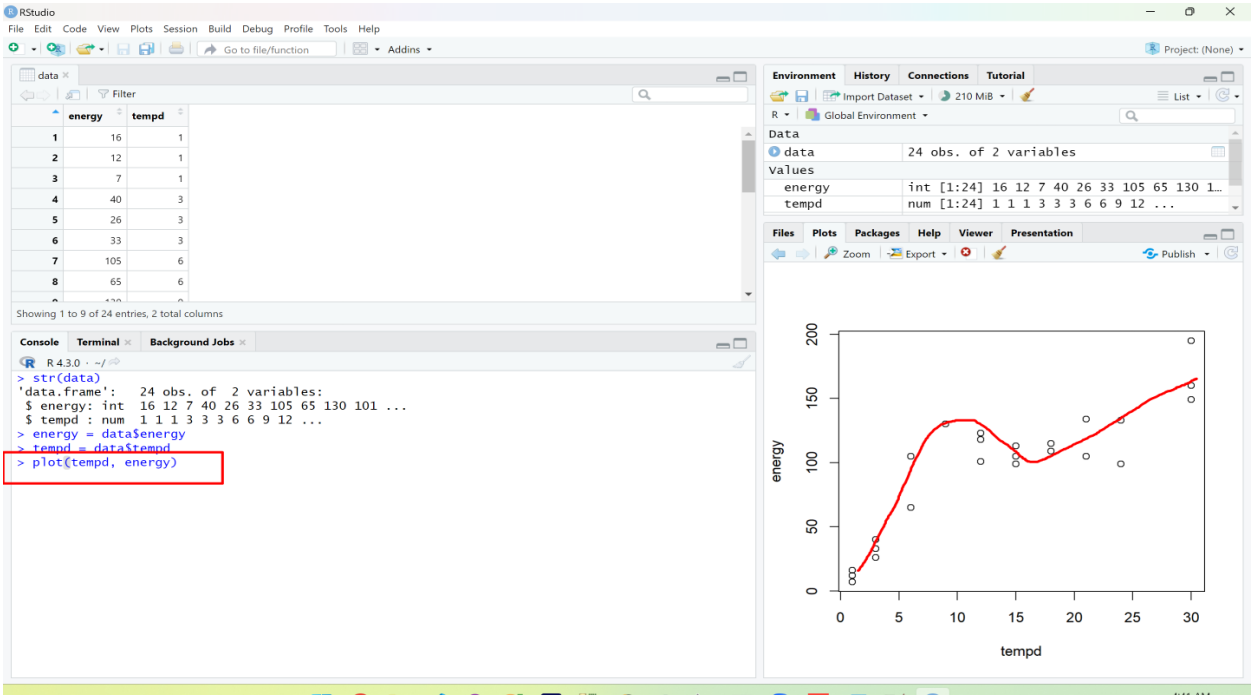
f) By using influence.measures() function to identify whether there are influential points that can affect your final model. Use cook's distance as the metric to identify the influential points.

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.

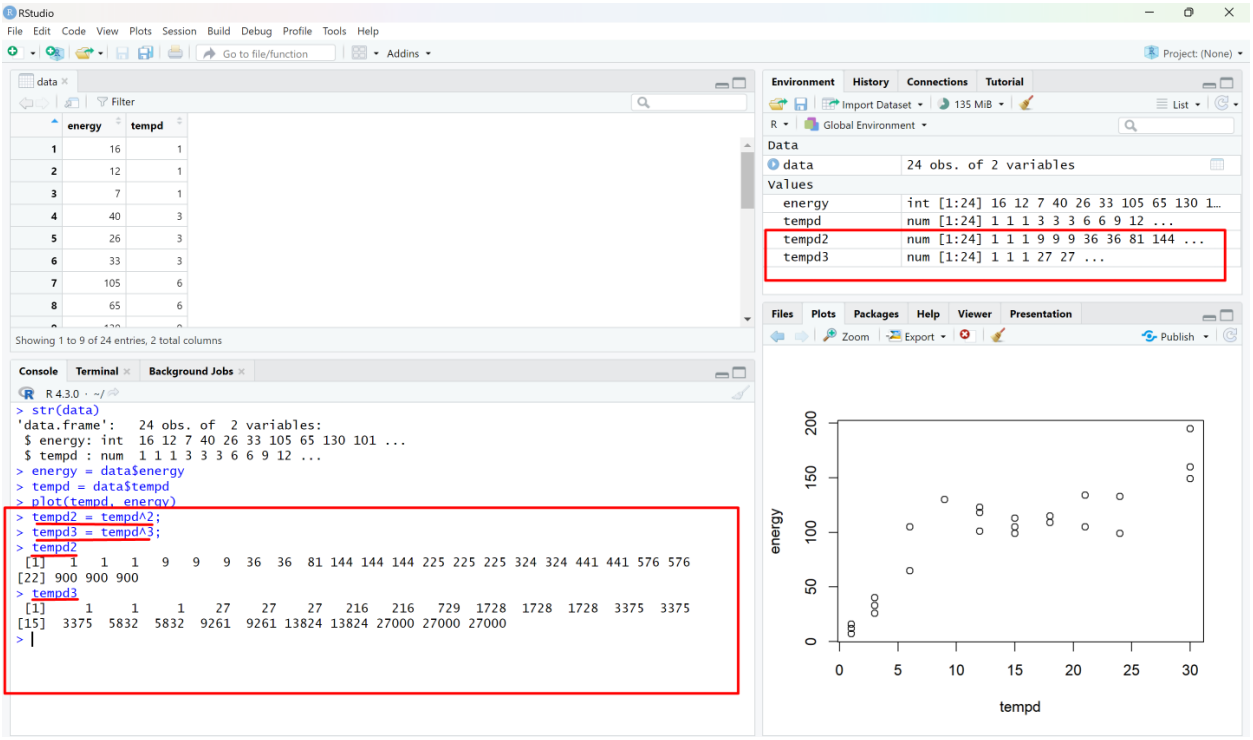


Analyzing the association b/w the 2 variables:

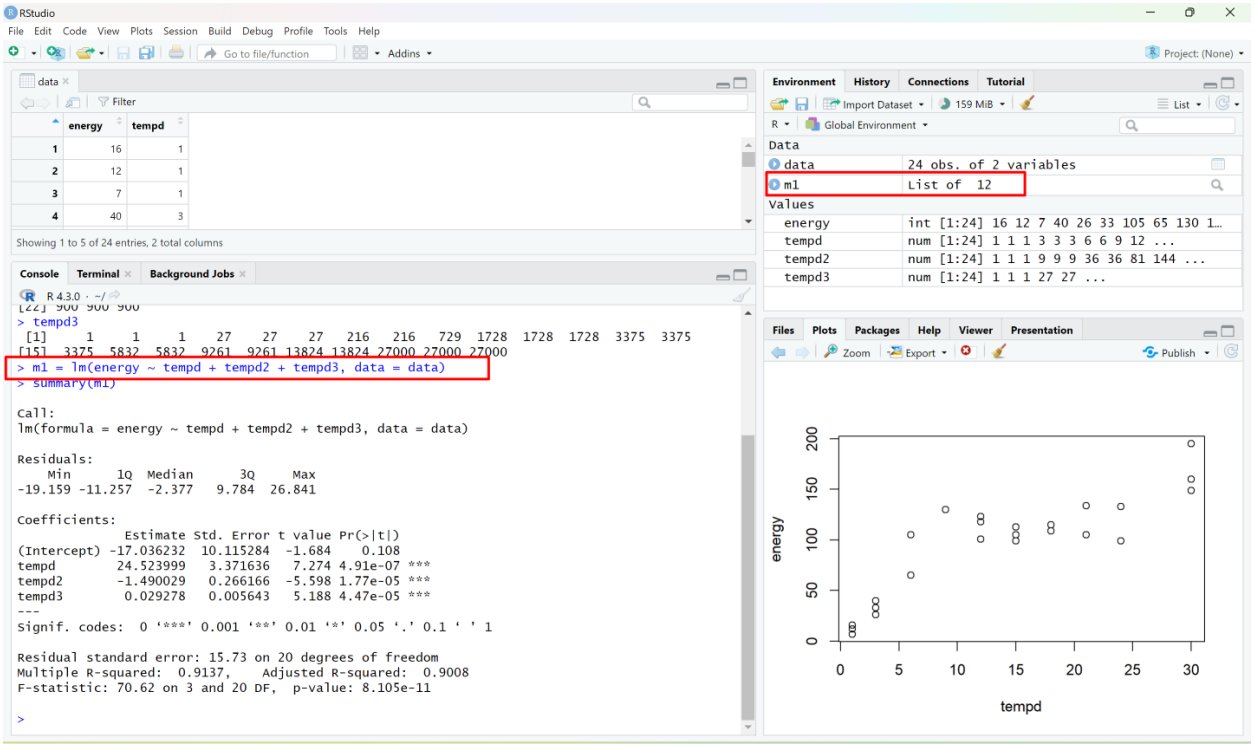
- From the scatterplot, we can observe that there is 1-top & 1-bottum in the curve, and the scatterplot produced S-shaped pattern. Hence, we can conclude that the association between energy (y) and tempd (X) is cubic association / cubic relation.
- As the scatterplot follows cubic relation, we need to 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: (H0): The coefficients of all x-variables are zero and there is no linear relationship with energy.
- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

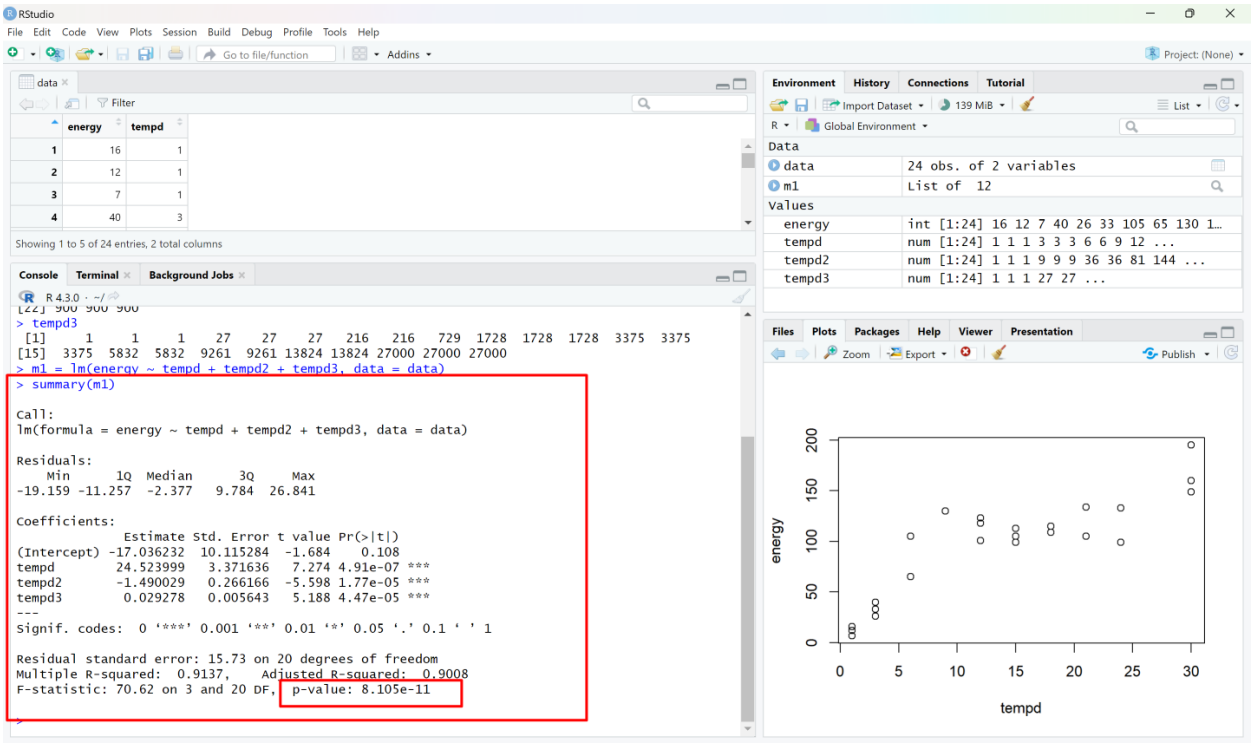
- Alternative Hypothesis: At least one of the coefficients of the x-variables (tempd, tempd2, tempd3) is not zero and can affect energy.
- $H_a : \beta_j \neq 0$

Note: In these hypotheses,

β1 represents the coefficient of "tempd"

β2 represents the coefficient of "tempd2."

β3 represents the coefficient of "tempd3."

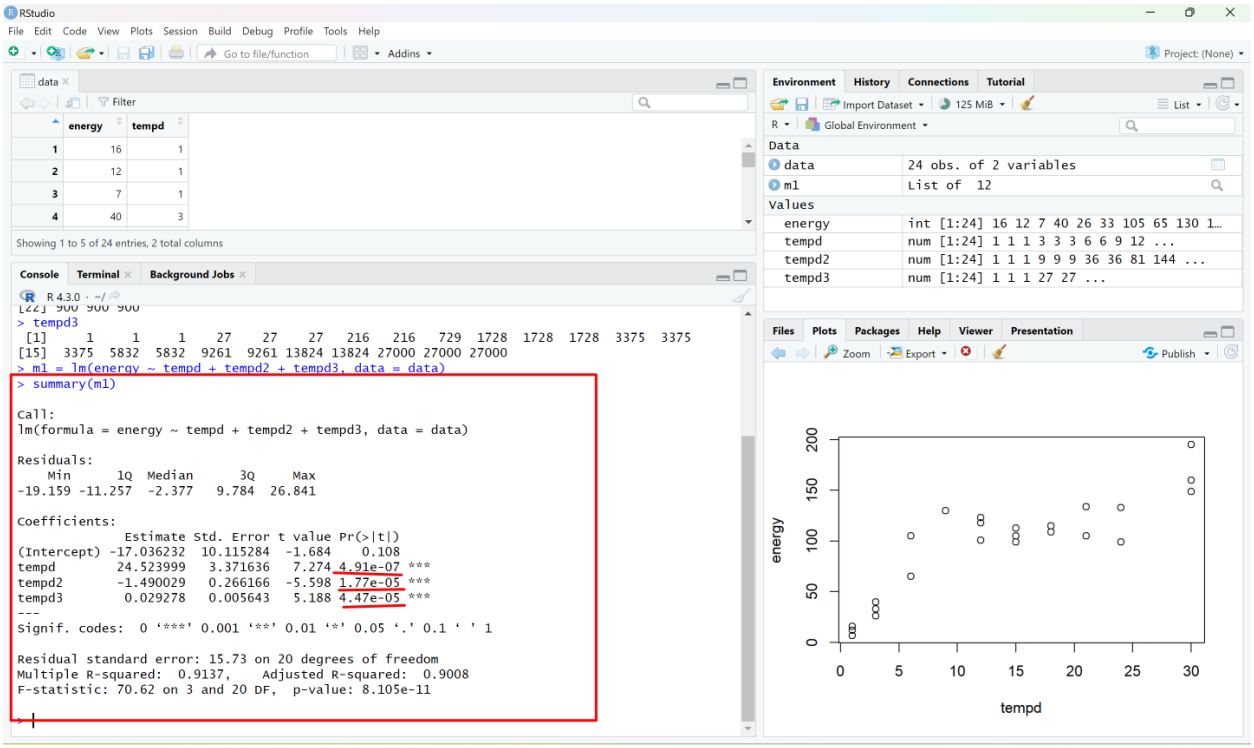


Conclusion on F-test: p-value < 0.05 for the model m1(fitted cubic regression model)

- At 5% signicance 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)*********

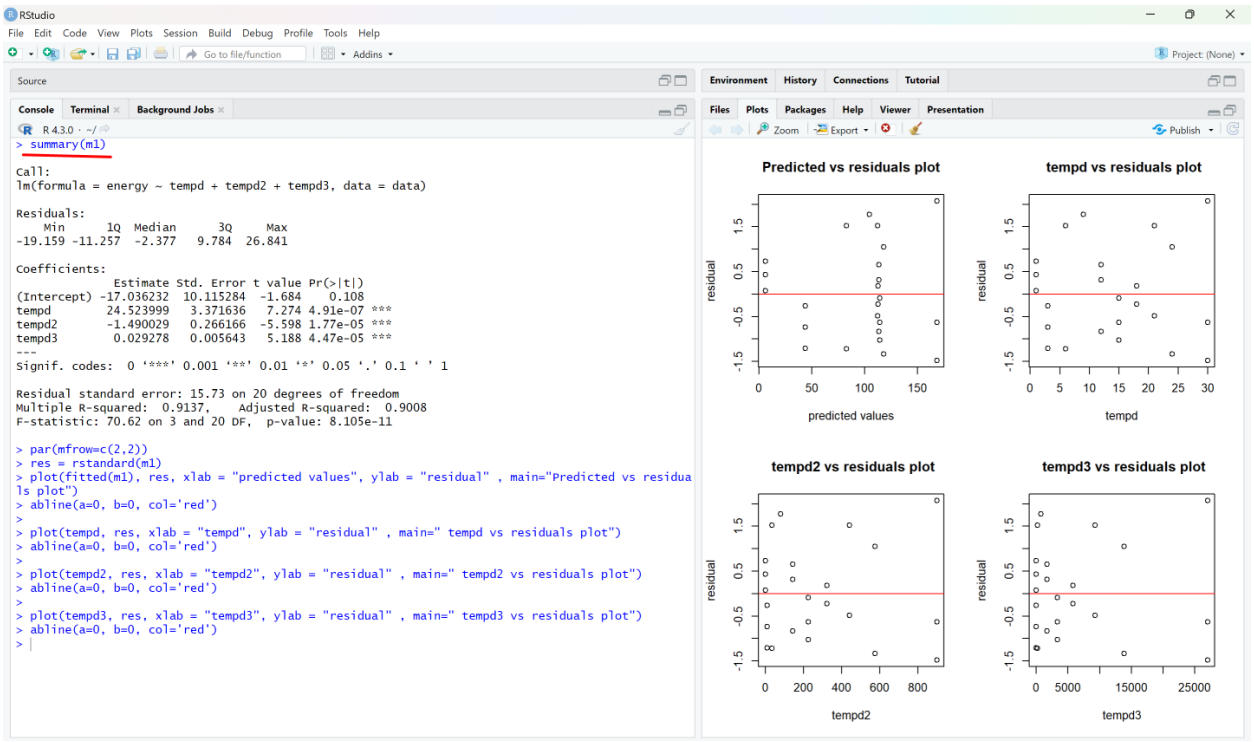
b. Are all variables in the model significant?

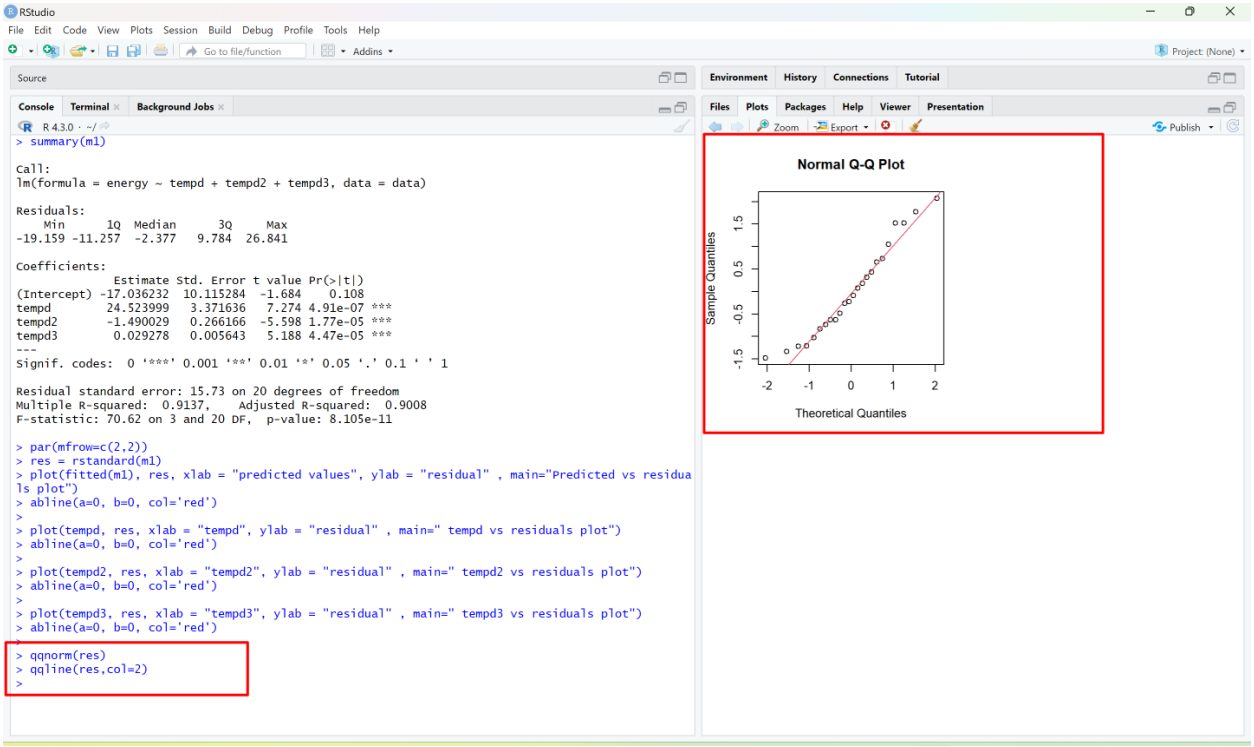


- To check whether all the variables are significant or not, we need to look at the individual t-test.
- As all the x-variables (tempd, tempd2,tempd3) have smaller p values than alpha 0.05(Assuming using 95% confidence level), all the x- variables in the model are significant.

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

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



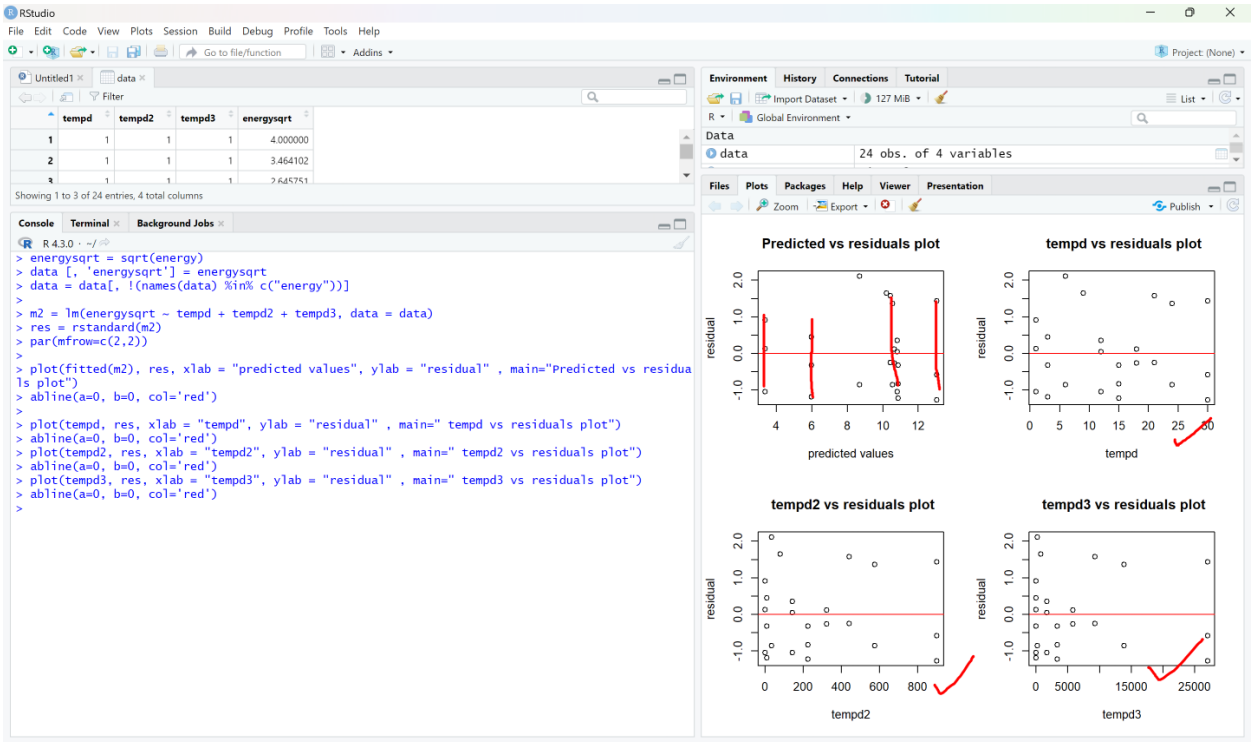


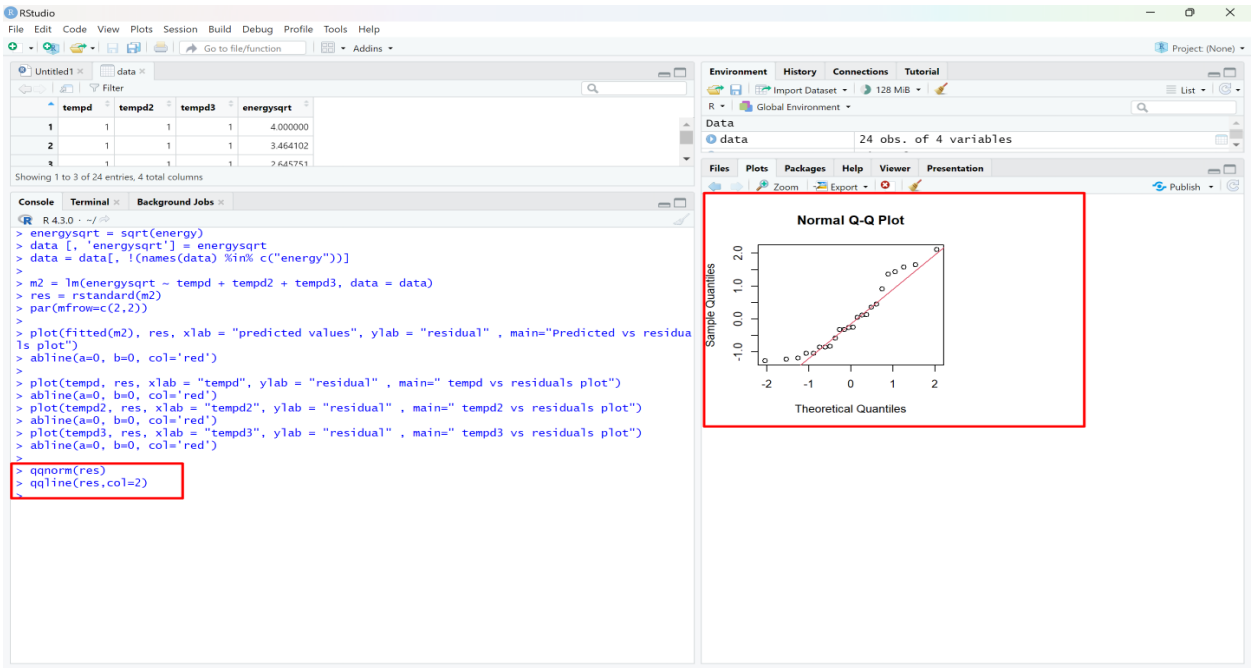
Observation: while performing res analysis for m1, we observed that

- 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.

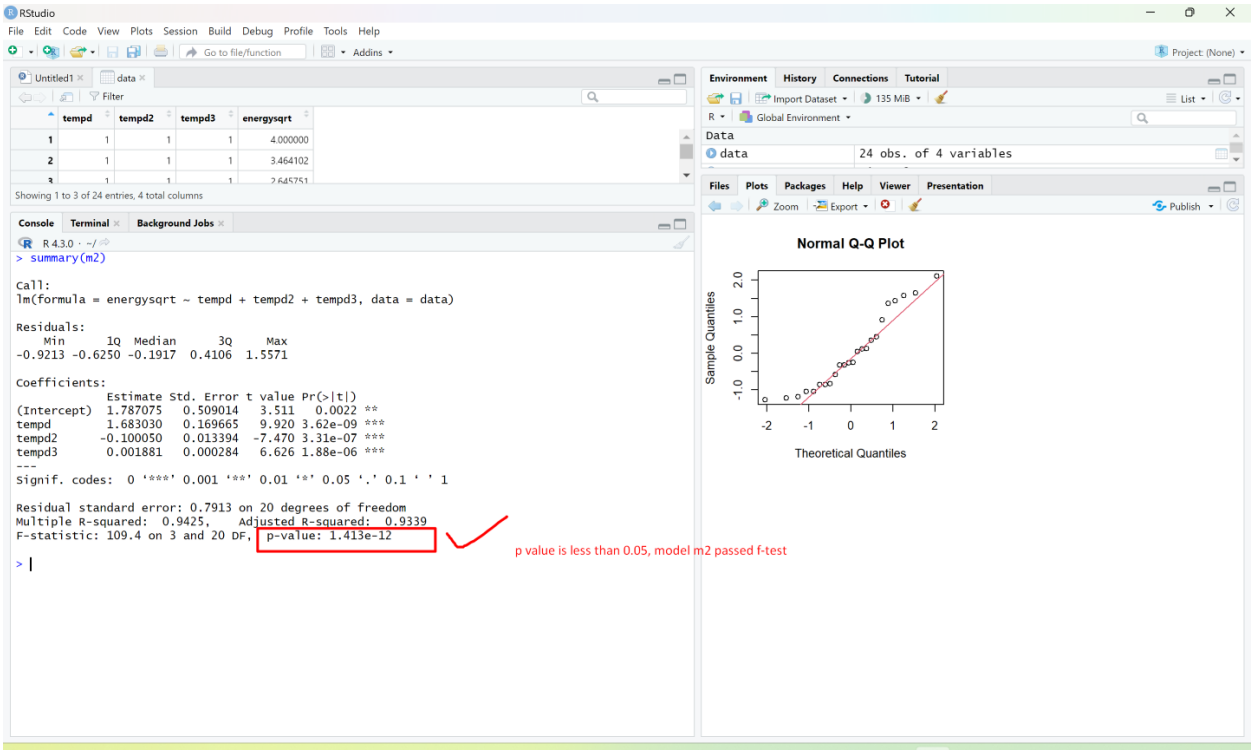
I tried log , reverse, sqrt on energy. I found sqrt transformation is better.
Sqrt tanformation on y-variable and built model m2 and performed residual analysis and f-test for **m2:**





Passed res analysis.

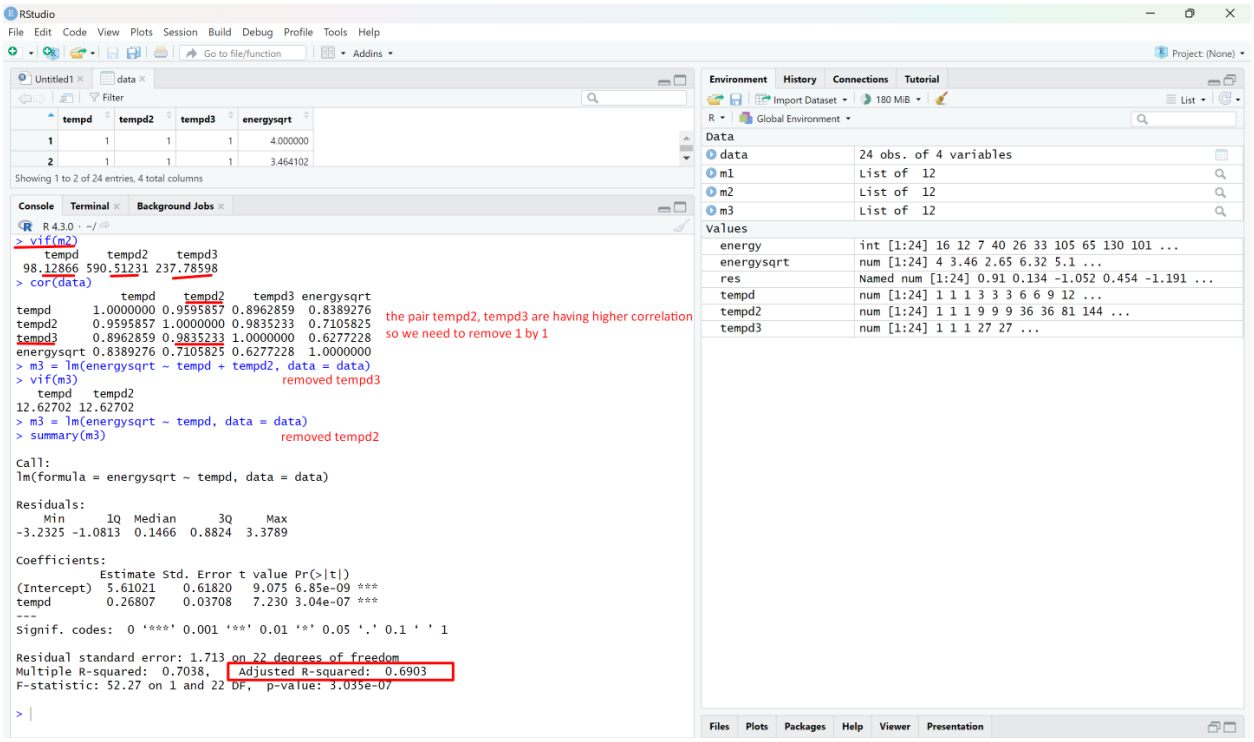
f-test on m2: passed f-test



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

- Using vif () function we can calculate the mcp 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

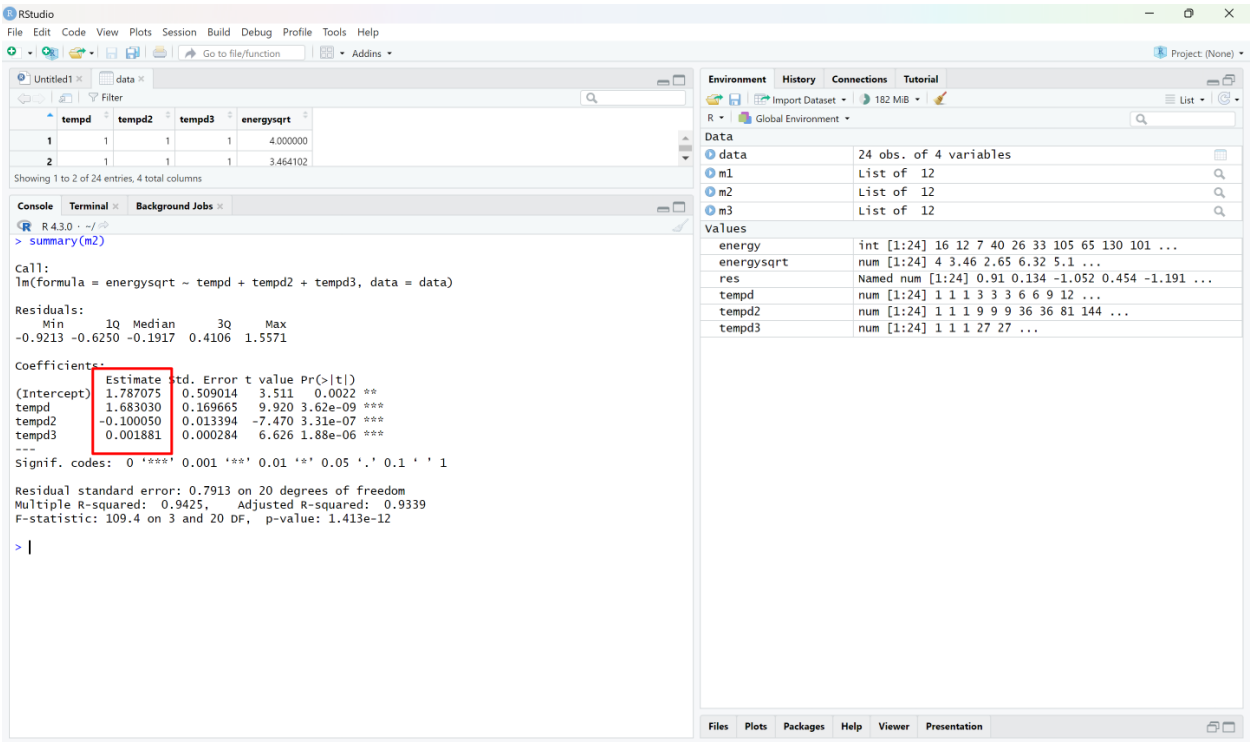
The entire process is carried out in R as follows in the screenshot: **M3**



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%	



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

- ENERGYsqrt = β0 + β1 * TEMPD + β2 * TEMPD^2 + β3 * TEMPD^3 + e
- i.e.,

- SqrtEnergy = 1.78 + 1.683*tempd + (-0.10)*tempd2 + 0.001*tempd3 + 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*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)*tempd^2 represents the quadratic relationship between sqrtenergy and tempd. It indicates how the sqrtenergy changes as the square of tempd

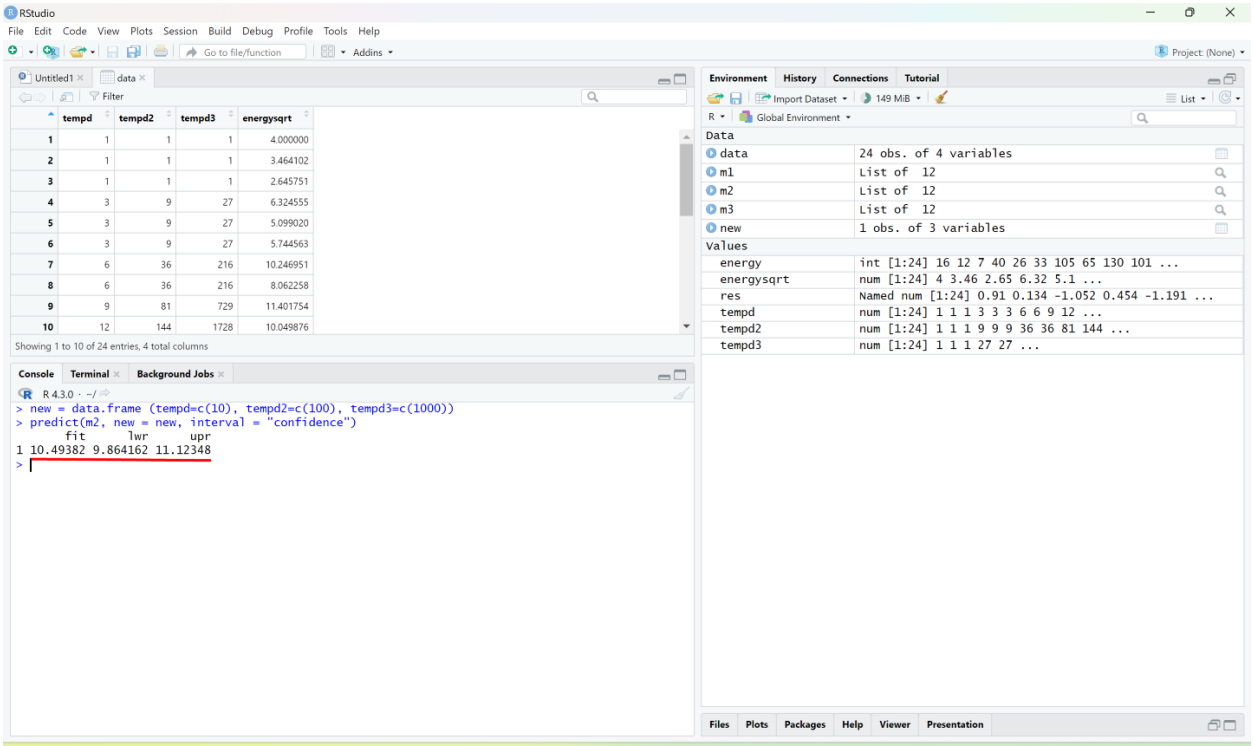
changes. The negative sign suggests a concave-down relationship, meaning the sqrtenergy initially increases at a decreasing rate and eventually decreases as tempd increases. (Assuming other variables are held constant).

β_3 ~

The slope for the term 0.001*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=IO:

(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)



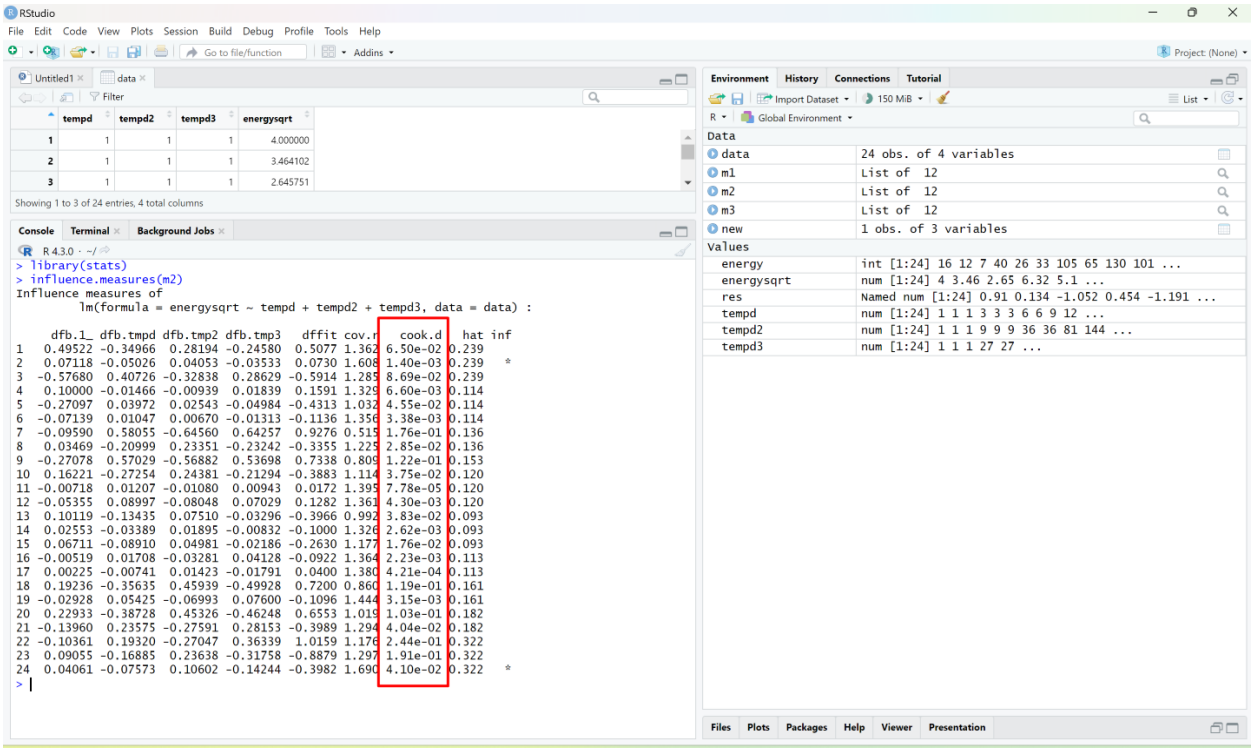
Observation:
From the output we can observe that
fit value is the predicted value: 10.49 (energy concumption) (energysqrt variable)
Confidence interval is lower ci = 9.86
Upper ci = 11.12

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:



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.

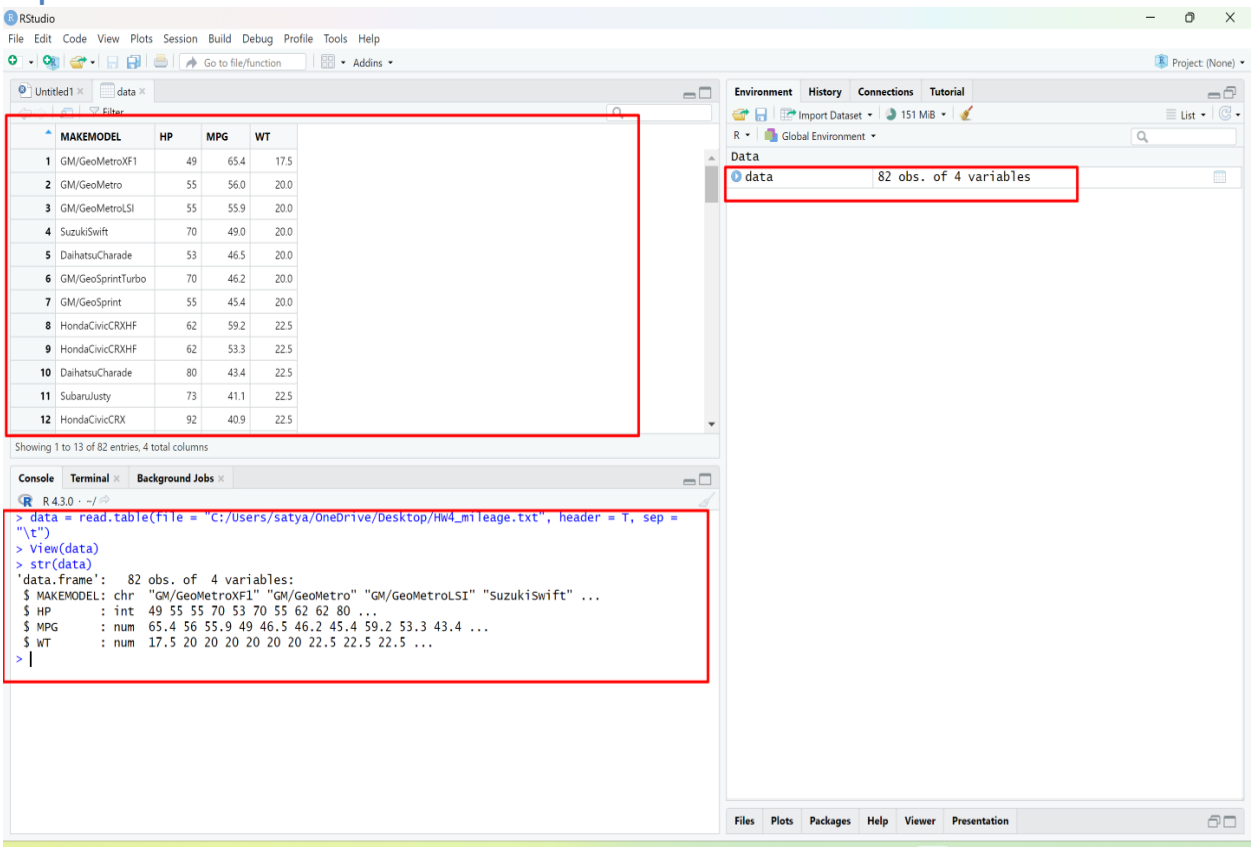
c) Analyze residual plots to evaluate if the regression model for the transformed Y variables is adequate. Do the plots show a deviation from the assumption of constant variance?

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

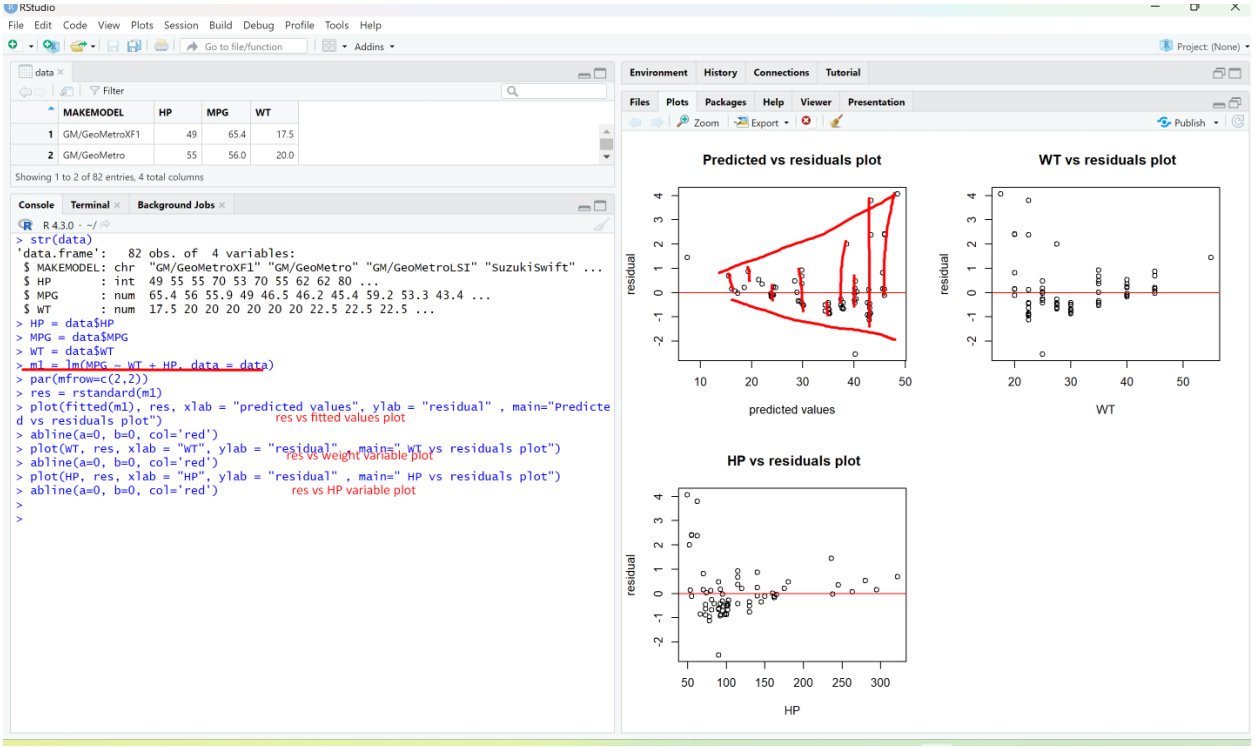
e). Use the step function to adopt both backward and stepwise forward selection to build a new model, compare the new model with the previous model in terms of the adj-R2 value.

ANSWER:

Imported data into r -studio:



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:

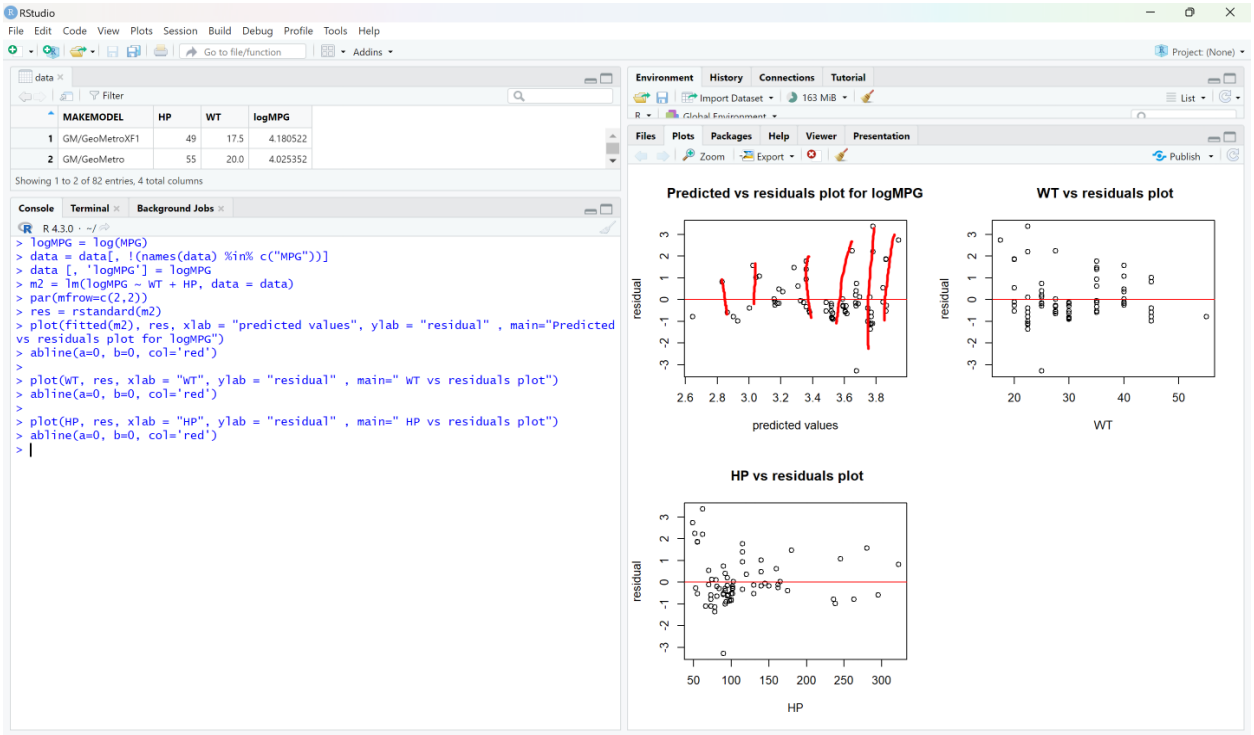
- From the res vs fitted values plot we can clearly observe that the variance is not constant. The variance on the left is small and on the right the variance is large.
- In other words, the variance increased from left to right and exhibits a fan shaped pattern which indicates that there is non- constant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

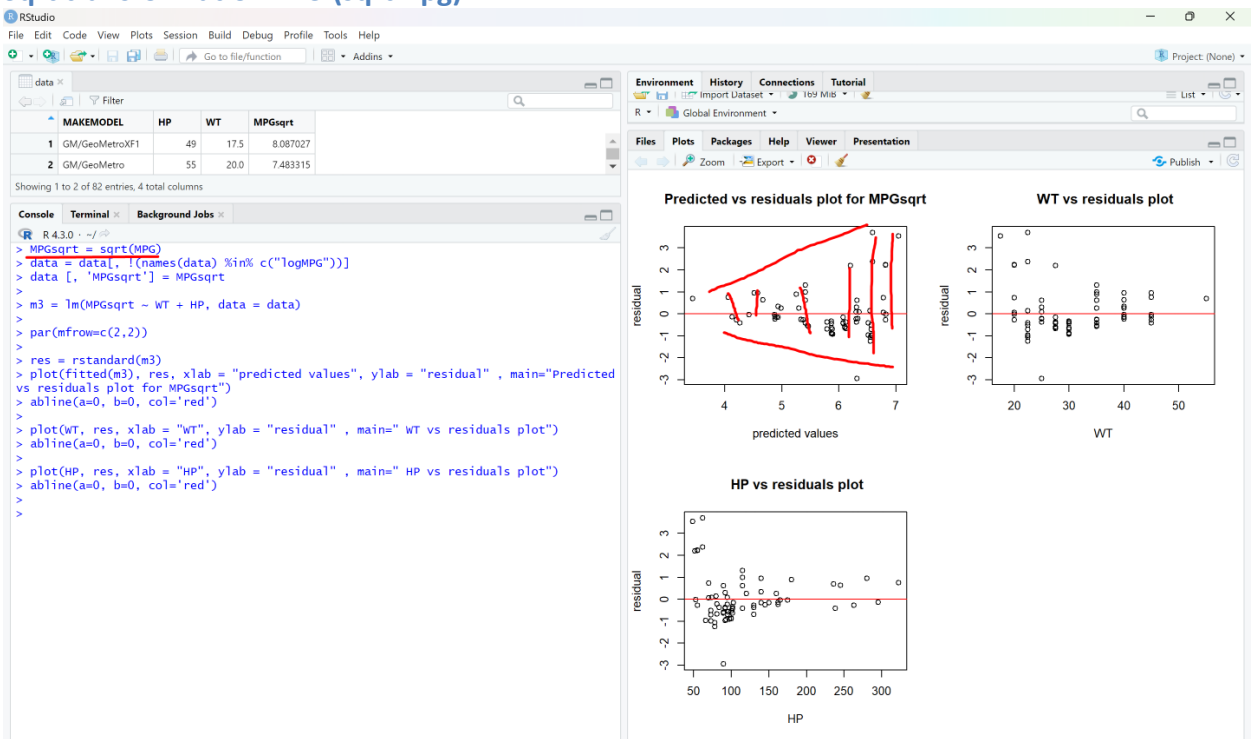
- As, there is no constant variance, **the model (m1) is not appropriate.**
- After transforming y-variable, we need to rebuild the model and check.

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

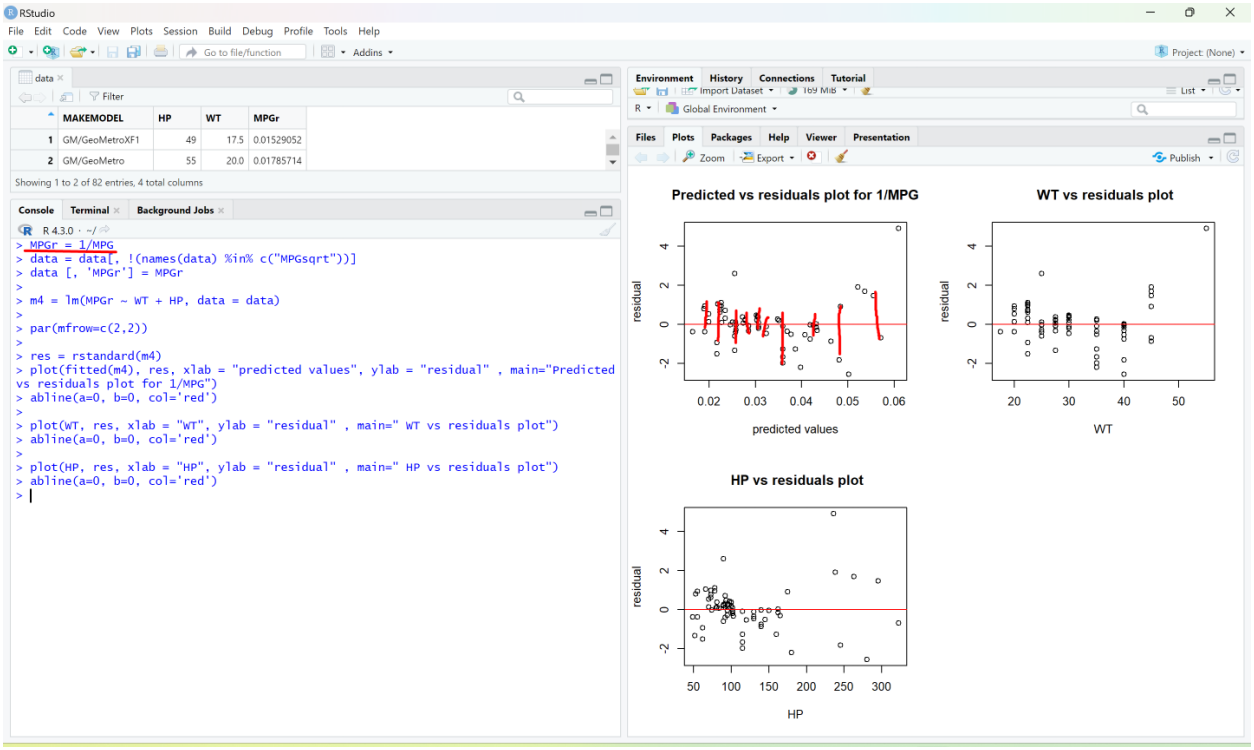
Log transformation: m2 (logmpg)



Sqrt transformation: m3 (sqrtmpg)



Reverse transformation: m4 (1/mpg) mpgr



Observation:

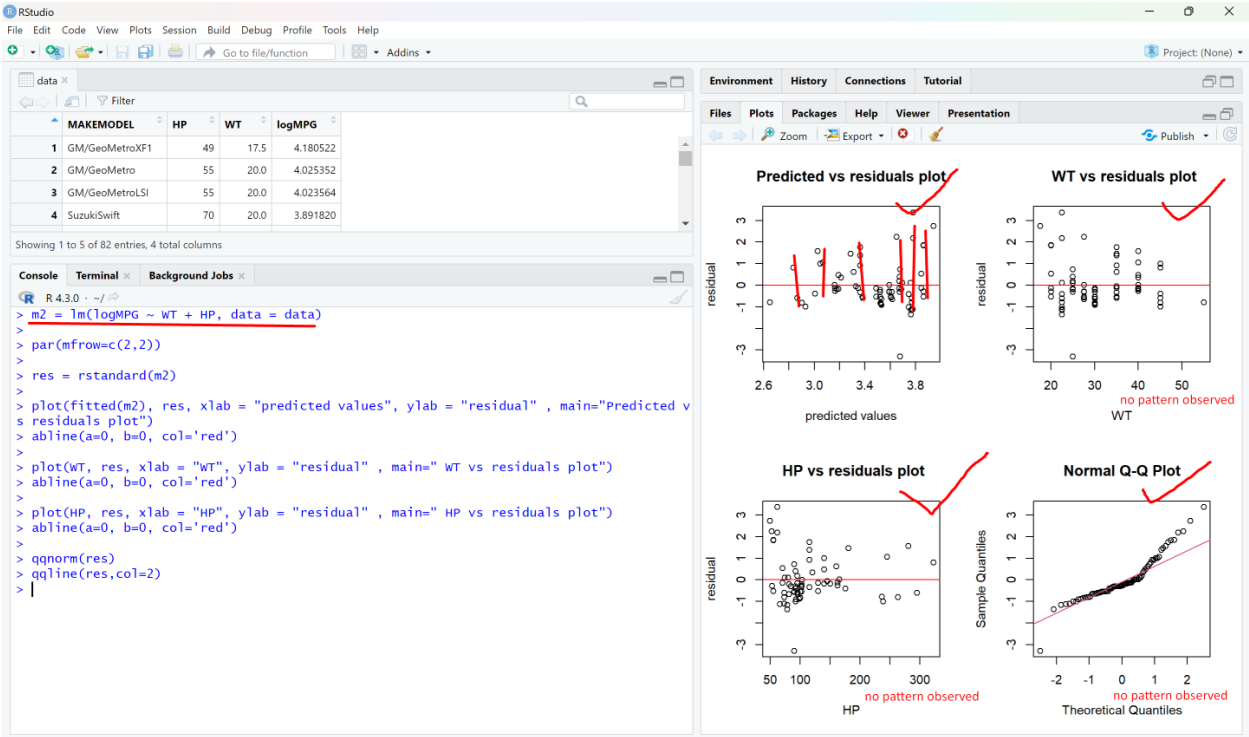
We did transformation on y-variable 'mpg' and built models. And analyzed res plots with all three transformed variables:

No-transormation(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?

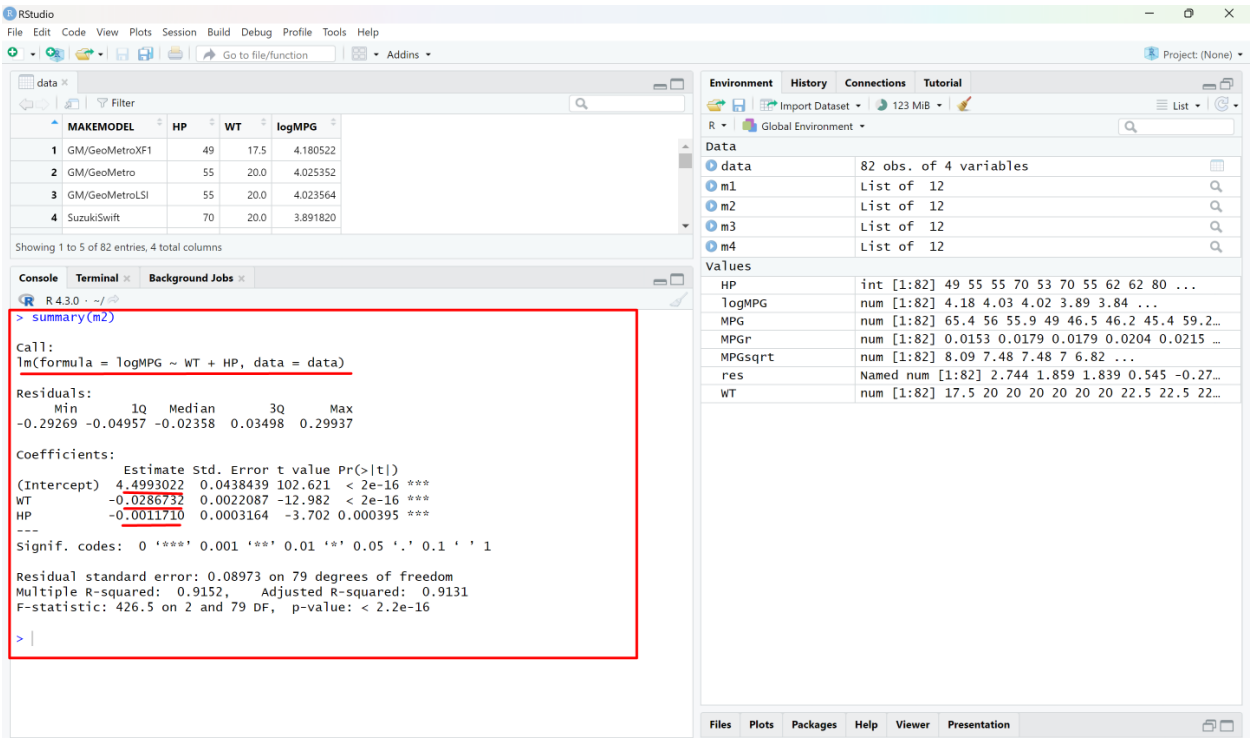
Residual analysis for final model: m2 (logMPG)



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



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

- LogMPG = β0 + β1*WT + β2 * HP + e
- i.e.,

- LogMPG = 4.49 + (-0.02)*WT + (-0.0017)*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.

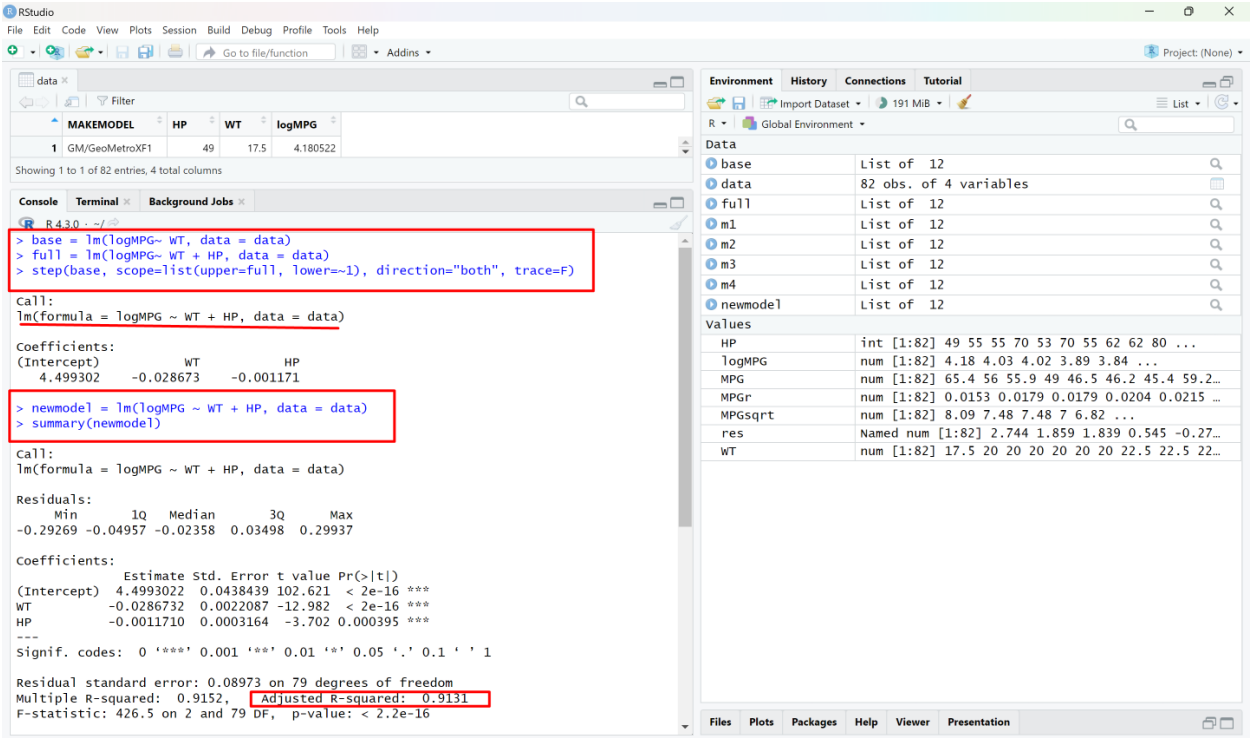
β_1, β_2 ~ The slopes represent how the logarithm of MPG changes with respect to the independent variables, Weight and Horsepower:

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.

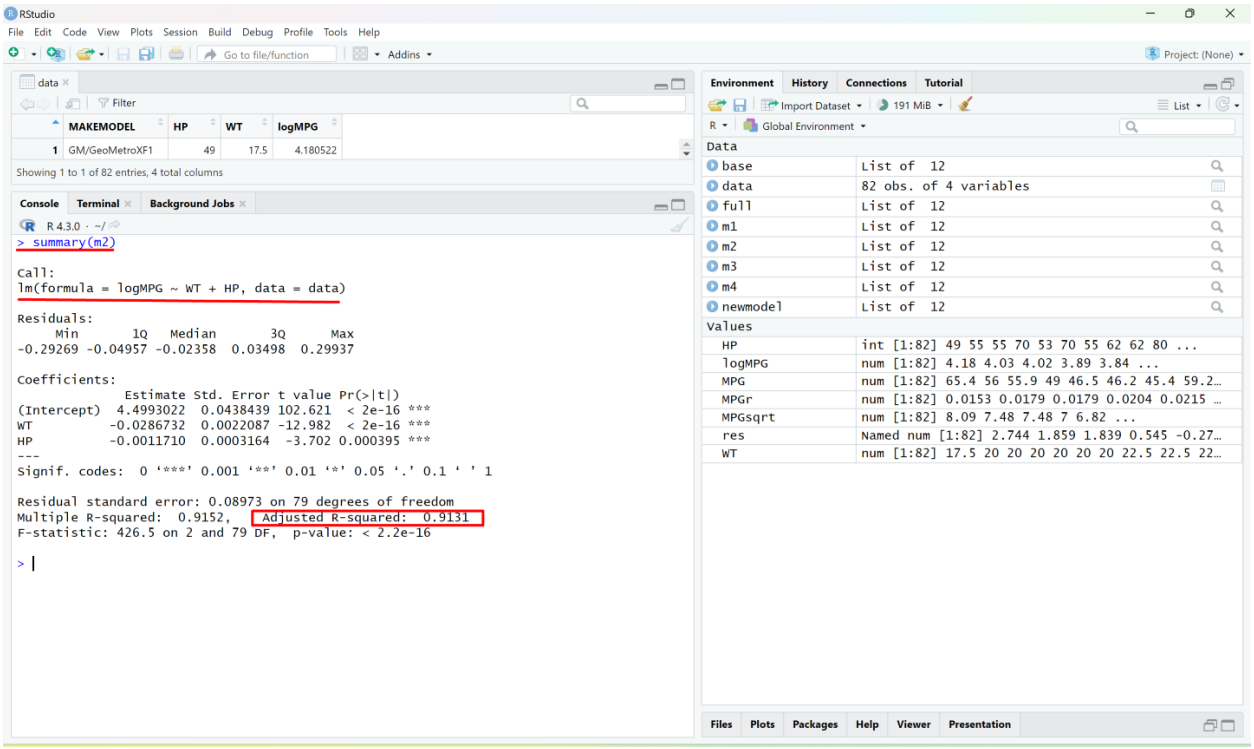
β_2 ~

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.



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



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.

1). 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, they those data points has to be removed.

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

ANSWER:

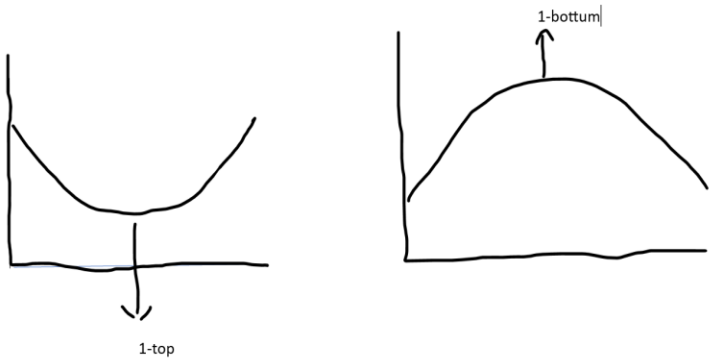
- 2nd & 3rd order terms are called higher order terms.
- First, to identify if the linear regression has higher order terms or not, we need to draw scatter plot with y-variable (dependent variable) and x-variable (predictor variable/independent variables)
- If the plot shows a curve pattern, there are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