4/27/2020

Anupama Kamepalli

University of North Texas

**Prediction Of Cars Miles Per Gallon(MPG)**

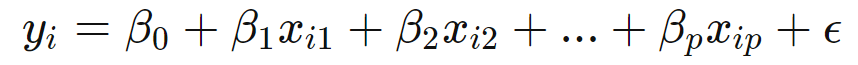
-Prof: Fathi, Madhi

**INTRODUCTION**

**Multiple Liner Regression:**

Multiple linear regression (MLR), also known as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between the explanatory (independent) variables and response (dependent) variable.

In other words, multiple regression is the extension of ordinary least-squares (OLS) [regression](https://www.investopedia.com/terms/r/regression.asp) that involves more than one explanatory variable.

The Formula for Multiple Linear Regression Is

**where, for***i*=*n***observations:**

*yi*​=dependent variable

*xi*​=explanatory variables

*β*0​=y-intercept (constant term)

*βp*=slope coefficients for each explanatory variable

*ϵ*=the model’s error term (also known as the residuals) ​

**Types of Regression:**

A simple linear regression is a function that allows to make predictions about one variable based on the information that is available about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

**Assumptions of Multiple Linear Regression**

The multiple regression model is based on the following assumptions:

There is a [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between the dependent variables and the independent variables.

* The independent variables are not too highly [correlated](https://www.investopedia.com/terms/c/correlation.asp) with each other.
* yi observations are selected independently and randomly from the population.
* Residuals should be [normally distributed](https://www.investopedia.com/terms/n/normaldistribution.asp) with a mean of 0 and [variance](https://www.investopedia.com/terms/v/variance.asp) σ.

The [coefficient of determination](https://www.investopedia.com/terms/c/coefficient-of-determination.asp) (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R2 always increases as more predictors are added to the MLR model even though the predictors may not be related to the outcome variable.

R2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

When interpreting the results of a multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

The purpose of this research is to identify the dependency of the output variable on the input variables. The input variables used in the analysis are Speed, Weight, Horsepower and Volume. The output variable is the Mileage per Gallon (MPG).

The reason for choosing this model is input variables are single and continuous and the output variable is continuous. We in this research are going to derive the impact of input variables Volume, horsepower, Speed and Weight on the mileage per gallon (MPG) of a car. We are going to use multiple Graphical Representations to analyze the impact of input variables on the output variable.

As the input and output variables are continuous in nature, the multi-linear regression analysis best suits our dataset analysis.

The below mentioned techniques were used for the purpose of analysis:

* **Box Plot:** To check the outliers.
* **Scatter Plot:** Correlation between Output Variable (MPG) and Input variables (HP, VOL, SP & WT).
* **Correlation Coefficient Matrix:** To find the strength and direction of the correlation.
* **Partial Correlation Matrix:** It is a pure correlation between variables.
* **Multi Collinearity Check:** On the impacted variables.
* **Variation Inflation Factor (VIF):** Model built on all the variables (collinearity check).
* **AV Plots Evaluation**
* **Influential Index Plot:** Plotting influential measures.
* **Check for line assumptions:** Residual Plots, QQ Plots, Standardization Plots.
* **Adjusted R-Square Validation**

**LITERATURE REVIEW**

In some situations, it appears that consumers choosing among some products are less attentive that to ancillary costs than to purchase prices. It is often asserted that gasoline costs are not fully salient to automobile consumers when they choose among automobiles with different fuel economy ratings. If this is true, consumers buy vehicles with lower fuel economy and higher resulting fuel costs than they would in their private optima. In 2007, the median income American household spent $2,400 on gasoline and consumers spent $286 billion in total (U.S. BLS, 2007). Misoptimization over such a large expenditure class could cause substantial welfare losses. For the cross-sectional estimator to be un-biased, the functional form for how other observed product characteristics enter utility must be correctly specified and any unobserved characteristics must be uncorrelated with energy efficiency. Especially for automobiles, these assumptions appear problematic. Fuel economy is mechanically correlated with weight and horsepower and it has often proven difficult to separately identify preferences for these different characteristics. Furthermore, fuel economy is highly negatively correlated with price in the cross-section, suggesting that larger vehicles have more observed and unobserved amenities.

Fuel Economy, expressed in miles per gallon (MPG), is an index of the overall “effectiveness” achieved with a motor vehicle which consumes fuel. It measures what you get (miles travelled) versus what you put in (gallons of fuel). It is related to engine power load, vehicle speed and engine efficiency. For a given speed and engine efficiency, fuel economy is high for low power requirements and decreases as power goes up.

Over the last few decades, the average weight of a vehicle sold in the U.S. climbed steadily after surviving the oil embargoes of the 1970s. Today, however, auto companies are putting a lot of effort into reducing weight – [Lotus](https://www.autoblog.com/lotus/) set up an entire [lightweight structures division](http://green.autoblog.com/2008/05/18/lotus-creates-lightweight-structures-division/), [BMW](https://www.autoblog.com/bmw/) is [investing millions in the production of carbon fiber](http://green.autoblog.com/2009/10/29/megacity-will-be-bmws-first-to-use-carbon-fiber-on-a-large-sca/) and [Jaguar loves aluminum](http://green.autoblog.com/2009/06/28/video-jaguar-extols-the-glory-of-aluminum/) – because every ounce you take out of a car improves the vehicle's performance and [fuel economy](https://www.autoblog.com/tag/fuel+economy/). Options for weight savings that automakers are investigating include installing things like plastic fuel tanks and using carbon fiber instead of steel. Today, one of the main reasons automakers want to reduce weight is because it's a great way to increase MPG numbers. The [Environmental Protection Agency(EPA) says](http://www.fueleconomy.gov/feg/driveHabits.shtml) that for every 100 pounds taken out of the vehicle, the [fuel economy](https://www.autoblog.com/tag/fuel+economy/) is increased by 1-2 percent. Based on a gallon of gasoline costing $2.58, this translates to savings of between $0.03-$0.05 a gallon.

Horsepower is an important factor in an automobile’s fuel consumption. Horsepower is a means of measuring power. Power itself is a measure of how much work can be done in a specific time period.  The two common measurements of output that are applied to cars and their engines are torque and horsepower. On average, today’s vehicles offer about twice the horsepower of their counterparts from the early 1980s. How much horsepower do you actually need to satisfy your everyday driving needs? Progressive drivers are assessing their vehicle needs based on fuel consumption, not on top speed and quarter-mile time. So higher horsepower cars may be “slower” than lower horsepower cars. They may get better or worse gas mileage than lower horsepower cars (generally higher horsepower comes from burning more fuel, so get lower mpg). This change in mindset can save fuel, save money and reduce your impact on the environment.

“Speed Kills MPG”. Unfortunately, it's true. Your car's gas mileage decreases once it gets past its optimal speed. For most cars, this is around 55-60 mph. This means that every time you go over this speed, you're essentially wasting gas and money - and creating unnecessary greenhouse gases.  
  
You'd be surprised to learn that a slight decrease in your highway driving speed can significantly reduce your gas consumption, while only adding a few minutes to your travel time.

According to studies backed by the department of energy, the average car will be at its advertised MPG at 55 mph. But as the speed increases:  
  
      - 3% less efficient at 60 mph  
      - 8% less efficient at 65 mph  
      - 17% less efficient at 70 mph  
      - 23% less efficient at 75 mph  
      - 28% less efficient at 80 mph

The above-mentioned factors will impact the MPG of a vehicle as per the analysis made in various studies. We in this analysis are going to make a prediction on how the input variables Weight, Speed, Volume and Horsepower are impacting the performance of a vehicle regarding the output variable MPG.

**METHODOLOGY**

**Exploratory Data Analysis:** Methodology ,conclusion are shown below once we start our model building.

**Check for the Outliers. How spread apart the input variables to the output variable.**

**1)Hypothesis Test for the HP.**

H0: The variable HP have any outliers.

H1: The variable HP does not have any outliers.

**Process**: I am using the BOX PLOT to check for the outliers and the distribution of the data.

Executed the below code to achieve the results:

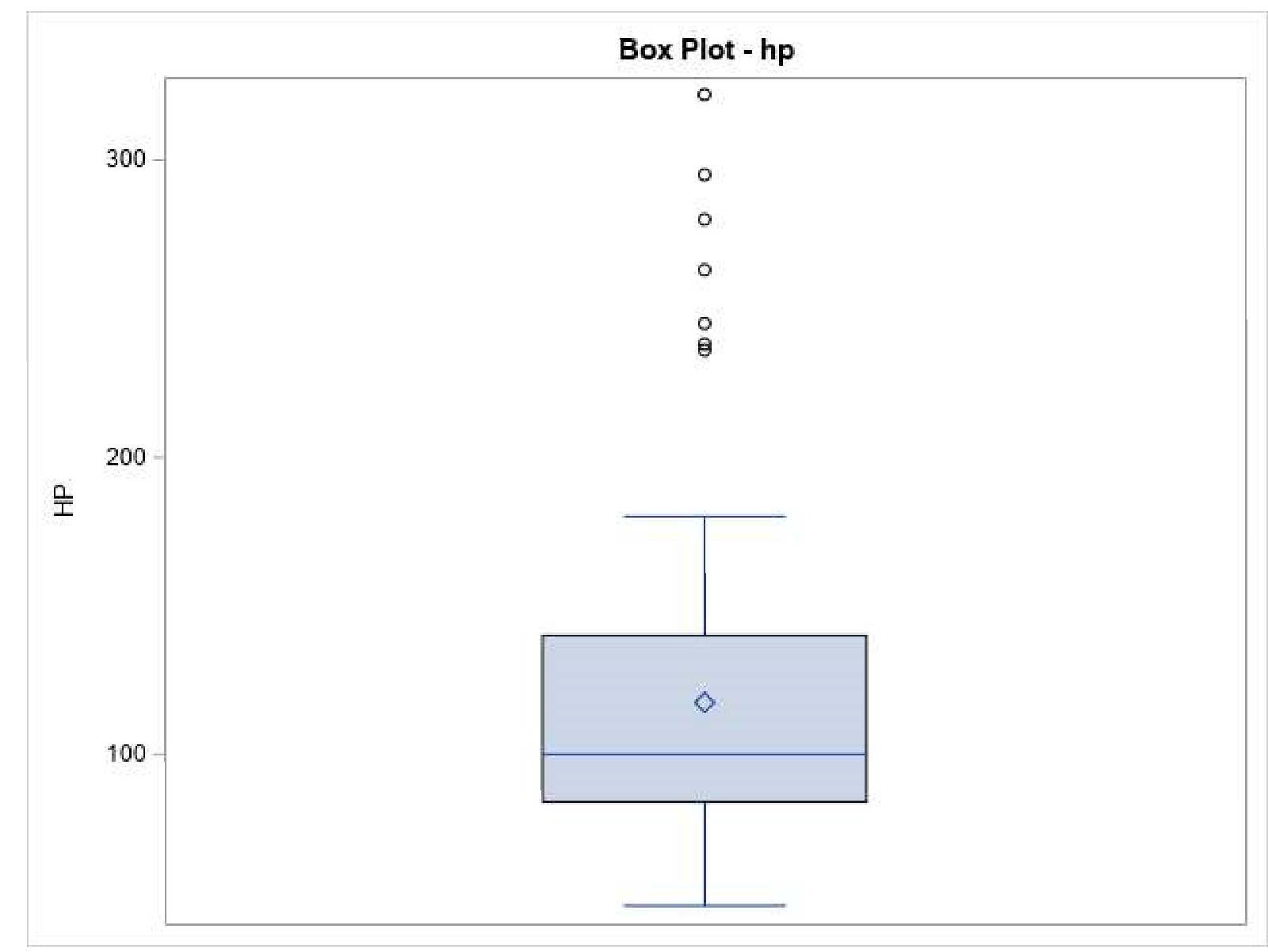
PROC SGPLOT DATA = newdata;

VBOX HP;

title 'Box Plot - hp';

RUN;

**Results:** From I came to know the data had some outliers.



**Conclusion:** As per the above results, we conclude that, we fail Reject Null Hypothesis H0.

**2)Hypothesis Test for the VOL.**

H0: The variable VOL have any outliers.

H1: The variable VOL does not have any outliers.

**Process**: I am using the BOX PLOT to check for the outliers and the distribution of the data.

Executed the below code to achieve the results:

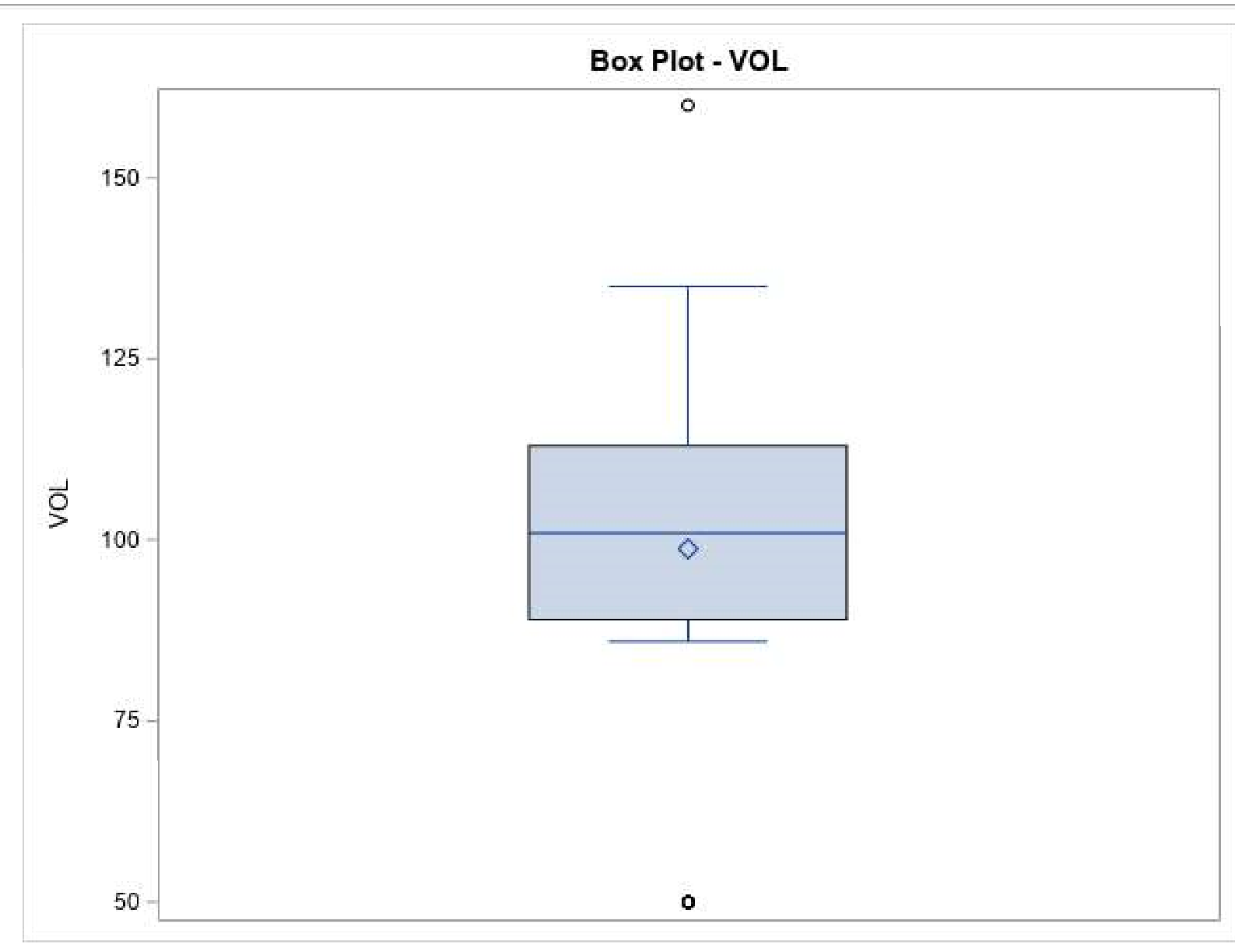
PROC SGPLOT DATA = newdata;

VBOX VOL;

title 'Box Plot - VOL';

RUN;

**Results:** From I came to know the data had some outliers.



**Conclusion:** As per the above results, we conclude that, we fail to Reject Null Hypothesis H0.

**3)Hypothesis Test for the SP.**

H0: The variable SP have any outliers.

H1: The variable SP does not have any outliers.

**Process**: I am using the BOX PLOT to check for the outliers and the distribution of the data.

Executed the below code to achieve the results:

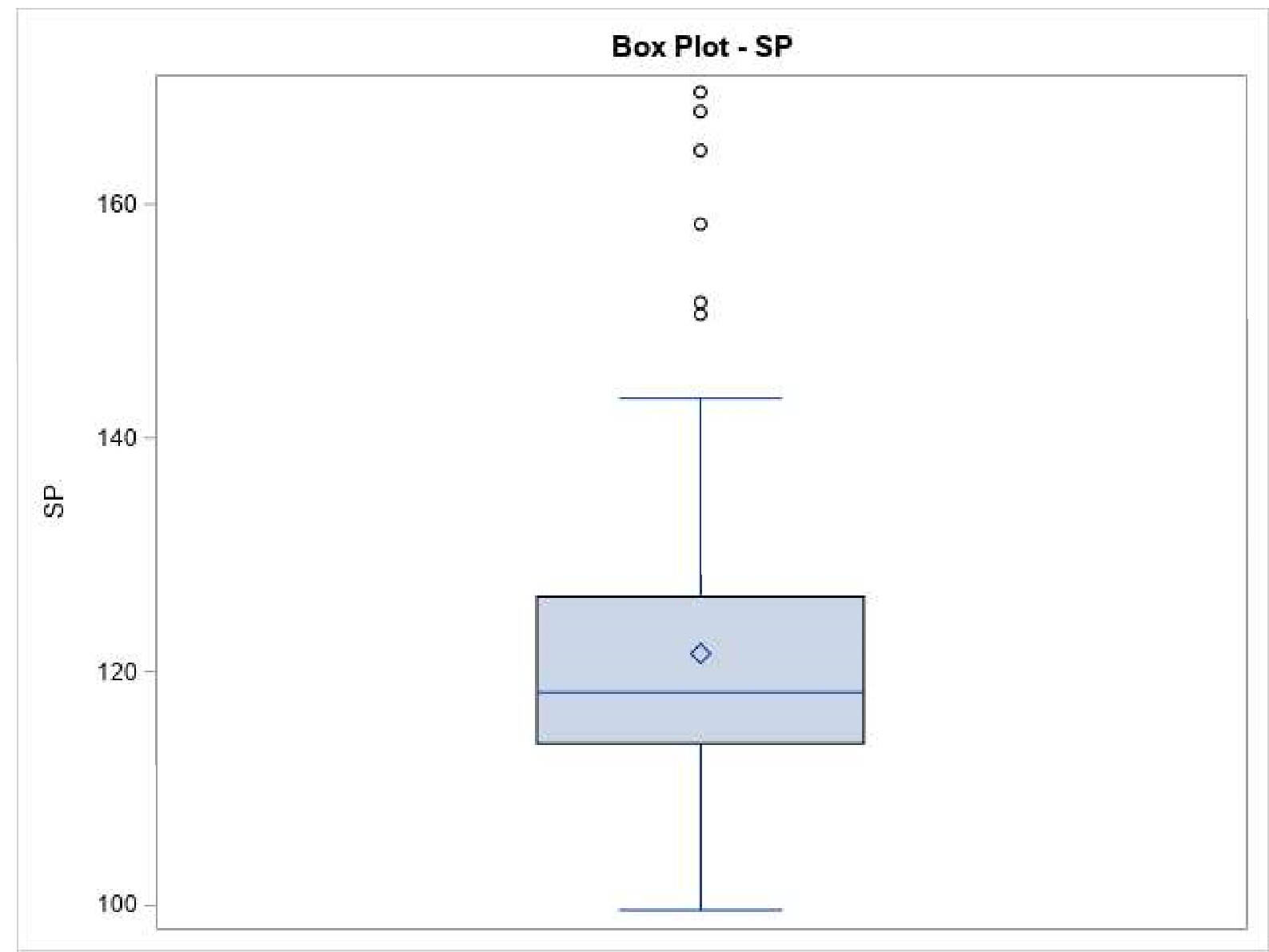
PROC SGPLOT DATA = newdata;

VBOX SP;

title 'Box Plot - SP';

RUN;

**Results:** From I came to know the data had some outliers.



**Conclusion:** As per the above results, we conclude that, we fail to reject Null Hypothesis H0.

**4)Hypothesis Test for the WT.**

H0: The variable WT have any outliers.

H1: The variable WT does not have any outliers.

**Process**: I am using the BOX PLOT to check for the outliers and the distribution of the data.

Executed the below code to achieve the results:

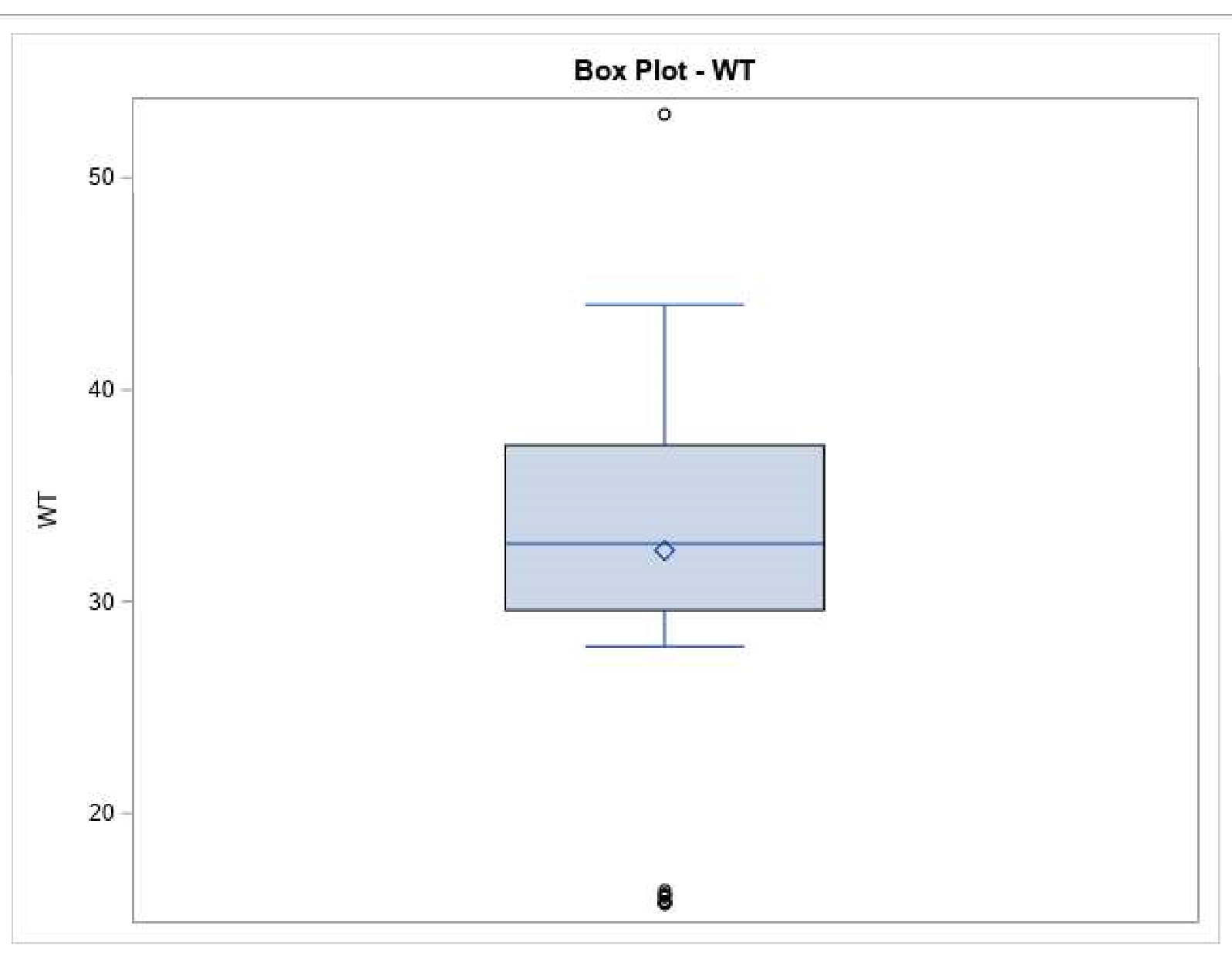
PROC SGPLOT DATA = newdata;

VBOX WT;

title 'Box Plot - WT';

RUN;

**Results:** From I came to know the data had some outliers.



**Conclusion:** As per the above results, we conclude that, we fail to Reject Null Hypothesis H0.

**5)Hypothesis Test for the MPG.**

H0: The variable MPG have any outliers.

H1: The variable MPG does not have any outliers.

**Process**: I am using the BOX PLOT to check for the outliers and the distribution of the data.

Executed the below code to achieve the results:

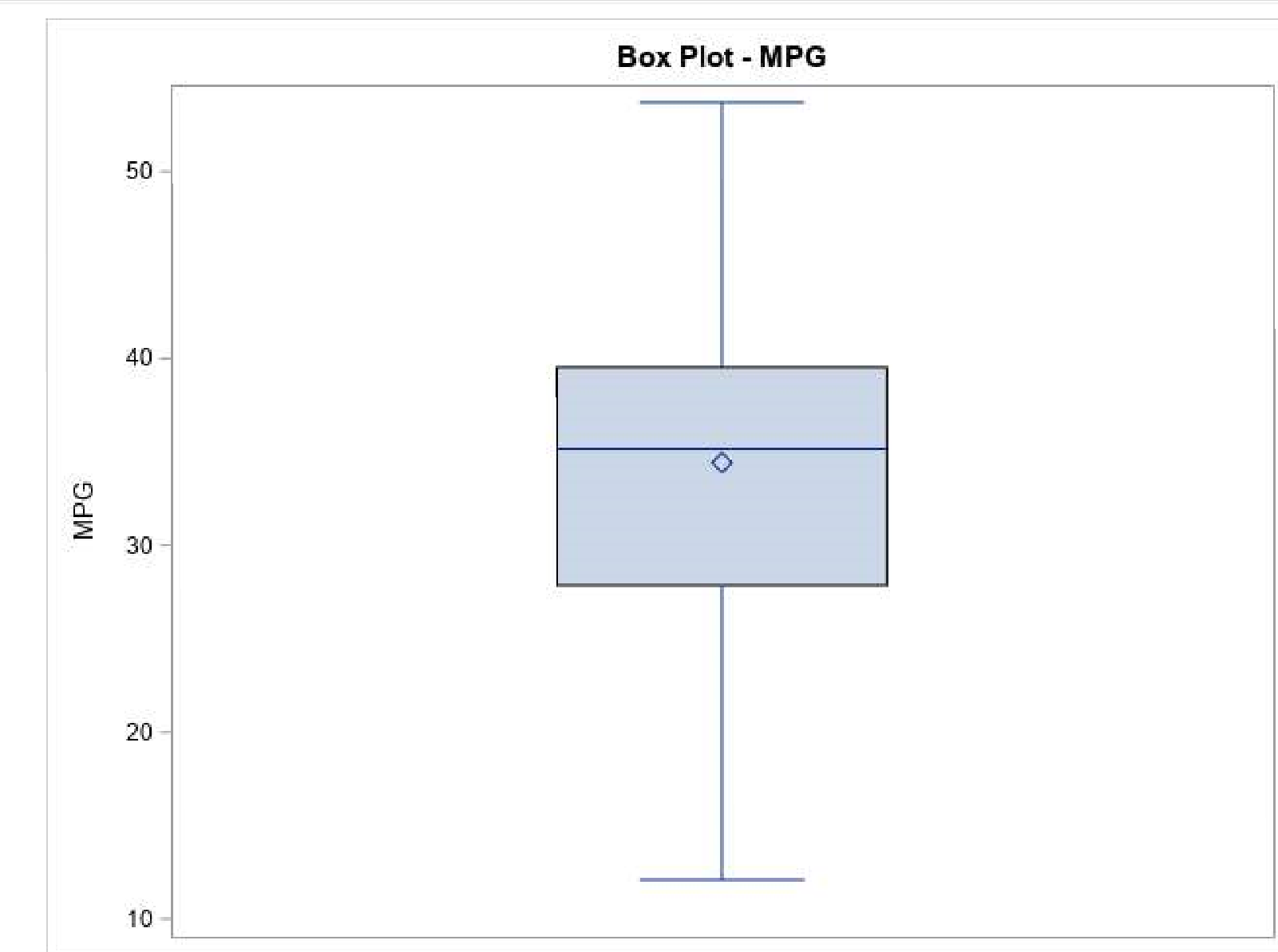
PROC SGPLOT DATA = newdata;

VBOX MPG;

title 'Box Plot - MPG';

RUN;

**Results:** From I came to know the data had NO outliers.



**Conclusion:** As per the above results, we conclude that, we Reject Null Hypothesis H0.

**Scatter plot Analysis: To Find the correlation b/n Output (MPG) & (HP,VOL,SP,WT)**

**1)Hypothesis Test for correlation between HP(x1) and MPG(y).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Process**: I am using the SCATTER PLOT to check for the direction, strength and linearity of the relationship.

Executed the below code to achieve the results:

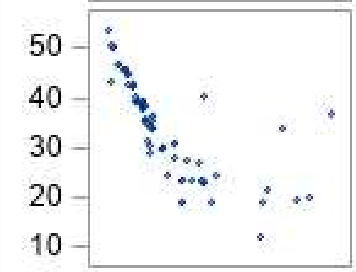
PROC sgscatter DATA = newdata;

matrix HP MPG VOL SP WT;

title 'Scatter Plot: Correlation Graph - t1';

RUN;

**Results:** From the below graph it can be interpreted that direction of the relationship is negative and strength of the relationship is probably moderate. If we draw two parallel lines, majority of the data points are falling between both the lines. So, both the variables can be interpreted to be linear. The variables are independent of each other and they are not running to collinearity problem.



**Conclusion:** As per the above results, we conclude that, we Reject Null Hypothesis H0.

**2)Hypothesis Test for correlation between VOL(x2) and MPG(y).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Process**: I am using the SCATTER PLOT to check for the direction, strength and linearity of the relationship.

Executed the below code to achieve the results:

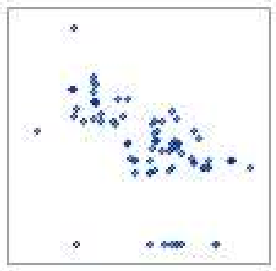
PROC sgscatter DATA = newdata;

matrix HP MPG VOL SP WT;

title 'Scatter Plot: Correlation Graph - t1';

RUN;

**Results:** From the below graph it can be interpreted that direction of the relationship is negative and strength of the relationship is probably moderate. If we draw two parallel lines, majority of the data points are falling between both the parallel lines. So, both the variables can be interpreted to be linear. The variables are independent of each other and they are not running to collinearity problem.



**Conclusion:** As per the above results, we conclude that, we Reject Null Hypothesis H0.

**3)Hypothesis Test for correlation between SPD(x3) and MPG(y).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Process**: I am using the SCATTER PLOT to check for the direction, strength and linearity of the relationship.

Executed the below code to achieve the results:

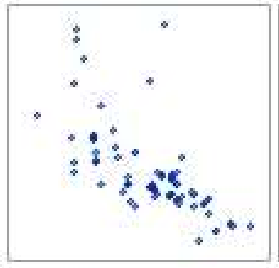
PROC sgscatter DATA = newdata;

matrix HP MPG VOL SP WT;

title 'Scatter Plot: Correlation Graph - t1';

RUN;

**Results:** From the below graph it can be interpreted that direction of the relationship is negative and strength of the relationship is probably moderate. If we draw two parallel lines, majority of the data points are falling between both the parallel lines. So, both the variables can be interpreted to be linear. The variables are independent of each other and they are not running to collinearity problem.



**Conclusion:** As per the above results, we conclude that, we Reject Null Hypothesis H0.

**4)Hypothesis Test for correlation between WT(x4) and MPG(y).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Process**: I am using the SCATTER PLOT to check for the direction, strength and linearity of the relationship.

Executed the below code to achieve the results:

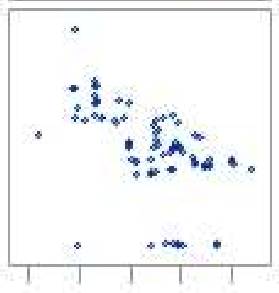
PROC sgscatter DATA = newdata;

matrix HP MPG VOL SP WT;

title 'Scatter Plot: Correlation Graph - t1';

RUN;

**Results:** From the below graph it can be interpreted that direction of the relationship is negative and strength of the relationship is probably moderate. If we draw two parallel lines, majority of the data points are falling between both the parallel lines. So, both the variables can be interpreted to be linear. The variables are independent of each other and they are not running to collinearity problem.



**Conclusion:** As per the above results, we conclude that, we Reject Null Hypothesis H0.

**5)Hypothesis Test for correlation between HP(x1) and VOL(x2).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Process**: I am using the SCATTER PLOT to check for the direction, strength and linearity of the relationship.

Executed the below code to achieve the results:

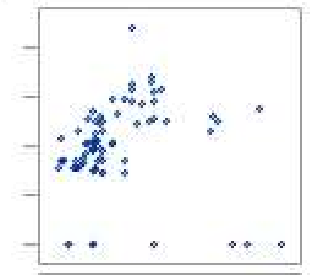
PROC sgscatter DATA = newdata;

matrix HP MPG VOL SP WT;

title 'Scatter Plot: Correlation Graph - t1';

RUN;

**Results:** From the below graph it can be interpreted that direction of the relationship is positive and strength of the relationship is probably moderate. If we draw two parallel lines, majority of the data points are falling between both the parallel lines. So, both the variables can be interpreted to be linear. The variables are independent of each other and they are not running to collinearity problem.



**Conclusion:** As per the above results, we conclude that, we Reject Null Hypothesis H0.

**6)Hypothesis Test for correlation between HP(x1) and SP(x3).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Process**: I am using the SCATTER PLOT to check for the direction, strength and linearity of the relationship.

Executed the below code to achieve the results:

PROC sgscatter DATA = newdata;

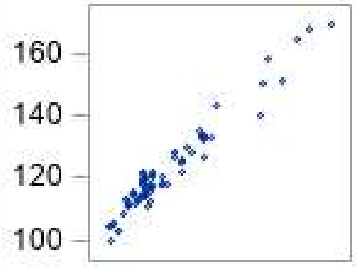
matrix HP MPG VOL SP WT;

title 'Scatter Plot: Correlation Graph - t1';

RUN;

**Results:** From the below graph it can be interpreted that direction of the relationship is positive and strength of the relationship is strong. If we draw two parallel lines, majority of the data points are falling between both the parallel lines. So, both the variables can be interpreted to be linear. The variables are interacting together in the model, they are not independent. So, there is collinearity problem between HP and SP.

Insight: The important insight has come out to be that between Horsepower and speed, there could be a collinearity problem.



**Conclusion:** As per the above results, we conclude that, we fail to Reject Null Hypothesis H0.

**7)Hypothesis Test for correlation between HP(x1) and WT(x4).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Process**: I am using the SCATTER PLOT to check for the direction, strength and linearity of the relationship.

Executed the below code to achieve the results:

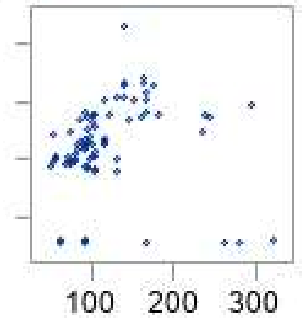
PROC sgscatter DATA = newdata;

matrix HP MPG VOL SP WT;

title 'Scatter Plot: Correlation Graph - t1';

RUN;

**Results:** From the below graph it can be interpreted that direction of the relationship is positive and strength of the relationship is probably moderate. If we draw two parallel lines, majority of the data points are falling between both the parallel lines. So, both the variables can be interpreted to be linear. The variables are independent of each other and they are not running to collinearity problem.



**Conclusion:** As per the above results, we conclude that, we Reject Null Hypothesis H0.

**8)Hypothesis Test for correlation between VOL(x2) and SPD(x3).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Process**: I am using the SCATTER PLOT to check for the direction, strength and linearity of the relationship.

Executed the below code to achieve the results:

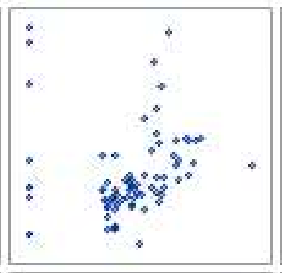
PROC sgscatter DATA = newdata;

matrix HP MPG VOL SP WT;

title 'Scatter Plot: Correlation Graph - t1';

RUN;

**Results:** From the below graph it can be interpreted that direction of the relationship is positive and strength of the relationship is probably moderate. If we draw two parallel lines, majority of the data points are falling between both the parallel lines. So, both the variables can be interpreted to be linear. The variables are independent of each other and they are not running to collinearity problem.



**Conclusion:** As per the above results, we conclude that, we Reject Null Hypothesis H0.

**9)Hypothesis Test for correlation between VOL(x2) and WT(x4).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Process**: I am using the SCATTER PLOT to check for the direction, strength and linearity of the relationship.

Executed the below code to achieve the results:

PROC sgscatter DATA = newdata;

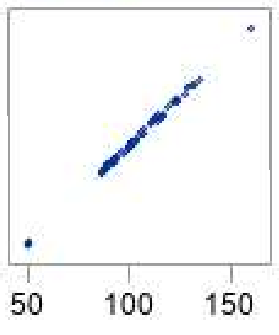
matrix HP MPG VOL SP WT;

title 'Scatter Plot: Correlation Graph - t1';

RUN;

**Results:** From the below graph it can be interpreted that direction of the relationship is positive and strength of the relationship is strong. If we draw two parallel lines, majority of the data points are falling between both the parallel lines. So, both the variables can be interpreted to be linear. The variables are interacting together in the model, they are not independent. So, there is collinearity problem between VOL and WT.

Insight: The important insight has come out to be that between Volume and Weight, there could be a collinearity problem.



**Conclusion:** As per the above results, we conclude that, we fail to Reject Null Hypothesis H0.

**10)Hypothesis Test for correlation between SPD(x3) and WT(x4).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Process**: I am using the SCATTER PLOT to check for the direction, strength and linearity of the relationship.

Executed the below code to achieve the results:

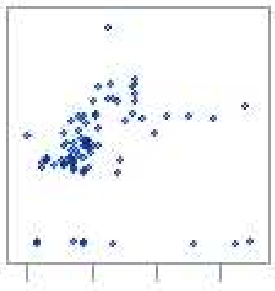
PROC sgscatter DATA = newdata;

matrix HP MPG VOL SP WT;

title 'Scatter Plot: Correlation Graph - t1';

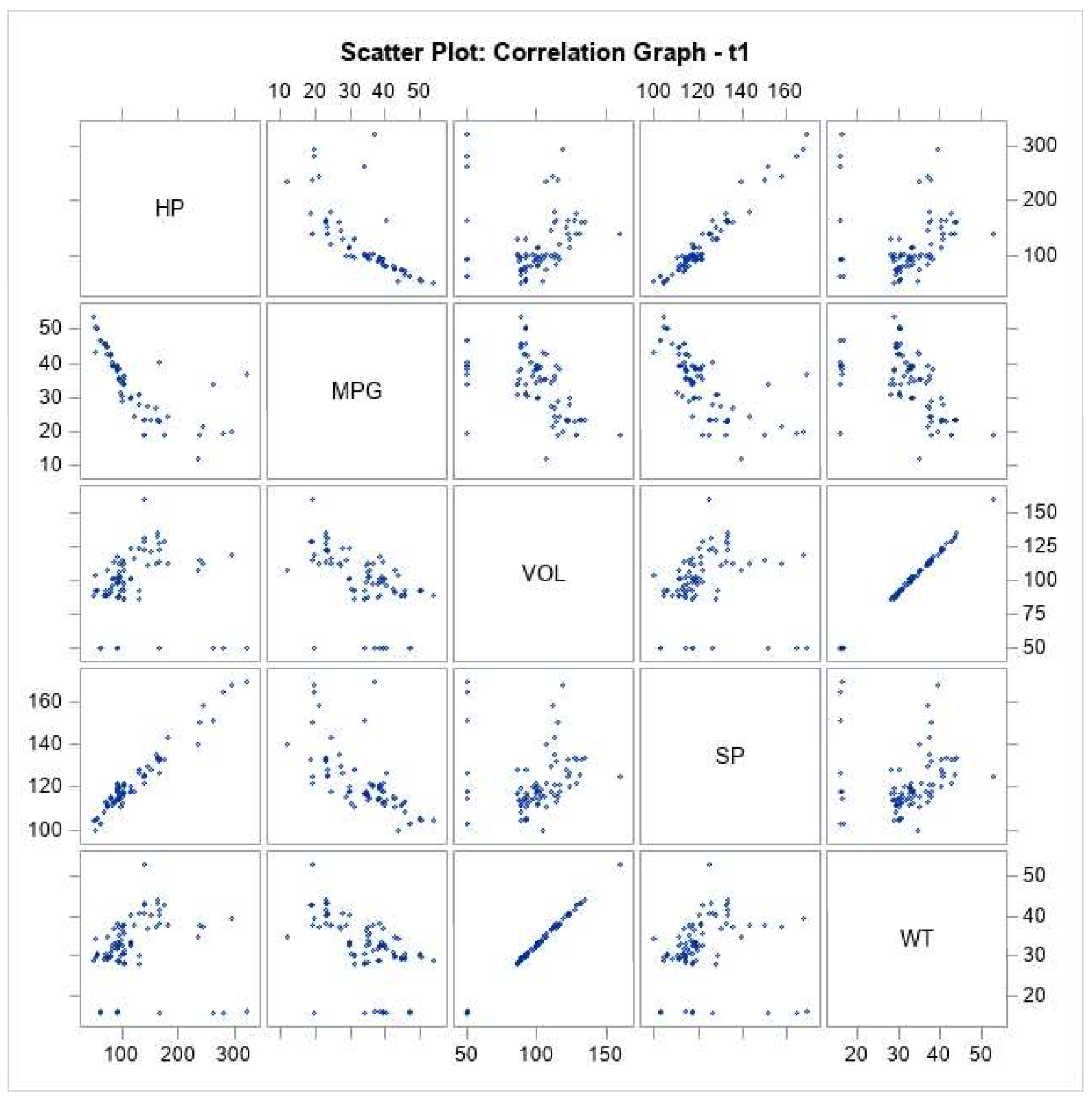
RUN;

**Results:** From the below graph it can be interpreted that direction of the relationship is positive and strength of the relationship is probably moderate. If we draw two parallel lines, majority of the data points are falling between both the parallel lines. So, both the variables can be interpreted to be linear. The variables are independent of each other and they are not running to collinearity problem.



**Conclusion:** As per the above results, we conclude that, we Reject Null Hypothesis H0.

**Consolidated Graph:**



**Pearson Correlation Coefficient Analysis:**

**Process**: I am using the Pearson Correlation Coefficient Matrix to strongly confirm the direction, strength and linearity of the relationship. The correlation coefficient is represented by “r”. If the absolute r value is greater than 0.85, then we say the 2 variables are strongly correlated. If the absolute value is less than 0.85, then it is moderate correlation. Based on the absolute r value, we perform regression analysis and will derive more insights.

Executed the below code to achieve the results:

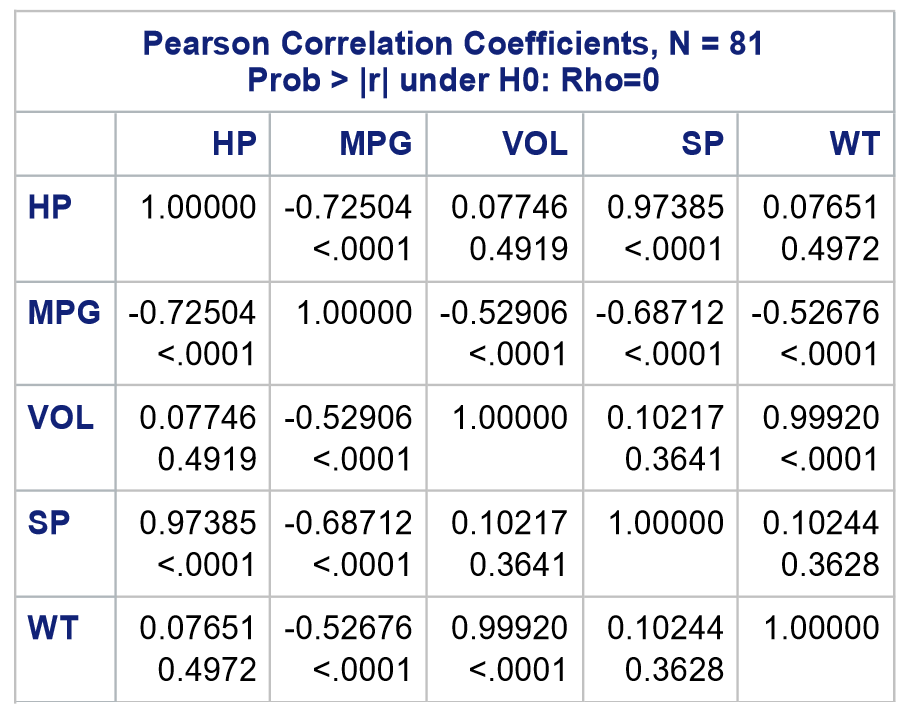
proc corr data = newdata ;

VAR HP MPG VOL SP WT ;

title "Correlation pairs -t1 ";

run;

**Code Results:**



**1)Hypothesis Test for correlation between HP(x1) and MPG(y).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Insights:** Correlation Coefficient: -0.72504, the absolute value of r is less than 0.85. The relationship between 2 variables is negative and the strength of the relationship between 2 variables is moderate. Hence, the 2 variables are independent and they are linear with a value of 0.72504 which ranges between -1 to 1. It is not running into collinearity problem.

**Conclusion:** As per the above results, we conclude that, we reject Null Hypothesis H0.

**2)Hypothesis Test for correlation between VOL(x2) and MPG(y).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Insights:** Correlation Coefficient: -0.52906, the absolute value of r is less than 0.85. The relationship between 2 variables is negative and the strength of the relationship between 2 variables is moderate. Hence, the 2 variables are independent and they are linear with a value of 0.52906 which ranges between -1 to 1. It is not running into collinearity problem.

**Conclusion:** As per the above results, we conclude that, we reject Null Hypothesis H0.

**3)Hypothesis Test for correlation between SPD(x3) and MPG(y).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Insights:** Correlation Coefficient: -0.68712, the absolute value of r is less than 0.85. The relationship between 2 variables is negative and the strength of the relationship between 2 variables is moderate. Hence, the 2 variables are independent and they are linear with a value of 0.68712 which ranges between -1 to 1. It is not running into collinearity problem.

**Conclusion:** As per the above results, we conclude that, we reject Null Hypothesis H0.

**4)Hypothesis Test for correlation between WT(x4) and MPG(y).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Insights:** Correlation Coefficient: -0.52676, the absolute value of r is less than 0.85. The relationship between 2 variables is negative and the strength of the relationship between 2 variables is moderate. Hence, the 2 variables are independent and they are linear with a value of 0.52676 which ranges between -1 to 1. It is not running into collinearity problem.

**Conclusion:** As per the above results, we conclude that, we reject Null Hypothesis H0.

**NOTE:** All the input variables HP, VOL, SPD and WT are negatively correlated to the output variable MPG. So, I now have to decide if I can build a good model with the available data as from the above analysis, it can be interpreted that most of the **relationship between the variables is moderate.**

**5)Hypothesis Test for correlation between HP(x1) and SP(x3).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Insights:** Correlation Coefficient: 0.97385, the absolute value of r is greater than 0.85. The relationship between 2 variables is positive and the strength of the relationship between 2 variables is strong and heavily correlated. It is running into collinearity problem. The 2 variables are independent and they are linear with a value of 0.97385 which ranges between -1 to 1.

**Conclusion:** As per the above results, we conclude that, we fail to reject Null Hypothesis H0.

**6)Hypothesis Test for correlation between VOL(x2) and WT(x4).**

H0: The relationship between 2 variables are correlated.

H1: The relationship between 2 variables are not correlated.

**Insights:** Correlation Coefficient: 0.99920, the absolute value of r is greater than 0.85. The relationship between 2 variables is positive and the strength of the relationship between 2 variables is strong and heavily correlated. It is running into collinearity problem. The 2 variables are independent and they are linear with a value of 0.99920 which ranges between -1 to 1.

**Conclusion:** As per the above results, we conclude that, we fail to reject Null Hypothesis H0.

**NOTE:** The relationship between HP and SPD, VOL and WT will create a problem in our model due to correlation. The correlation coefficient for Weight and Volume is coming out to be too high, so if we remove the impact of other variables on weight and volume to see what the actual correlation is. Because of multiple effects, the values might come up to be so high. So if we remove the effects of other variables on weight and volume, then we can find the actual correlation that we call as a pure correlation or partial correlation between 2 variables.

**Examples of Multiple Effects**: {HP,WT}, {MPG,WT}, {SP,WT} on WT, VOL.

**Pure or Partial Correlation Analysis:**

**Hypothesis Testing:**

H0: Effect of other variables on HP and SP.

H1: No effect of other variables on HP and SP.

**Code Used:**

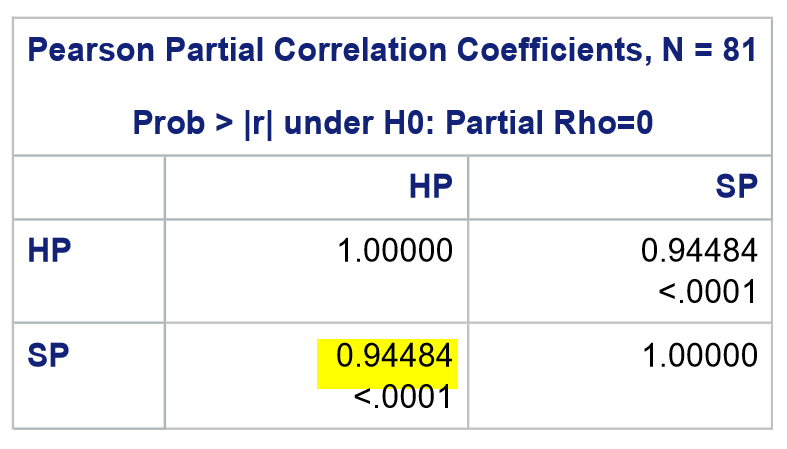
proc corr data=newdata plots=scatter(alpha=.20 .30);

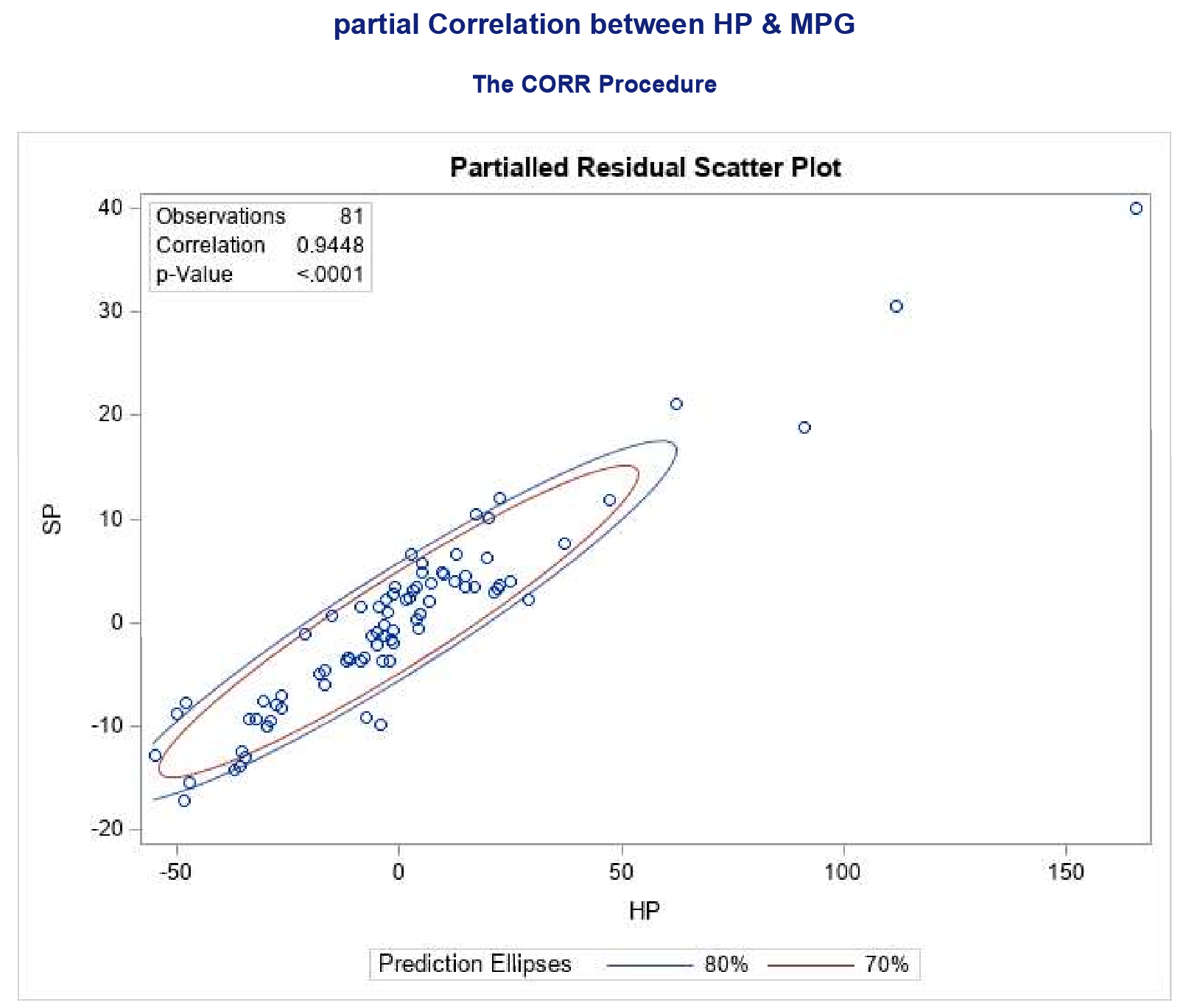
var HP SP;

partial MPG VOL WT;

run;

**Results:** Earlier the HP and SP value is0.97385 which is greater than 0.94484, we were successfully able to remove the affect of other variables on HP and SP.





**Conclusion:** Hence, we fail to reject Null Hypothesis H0.

**Hypothesis Testing:**

H0: Effect of other variables on VOL and WT.

H1: No effect of other variables on VOL and WT.

**Code Used:**

proc corr data=newdata plots=scatter(alpha=.20 .30);

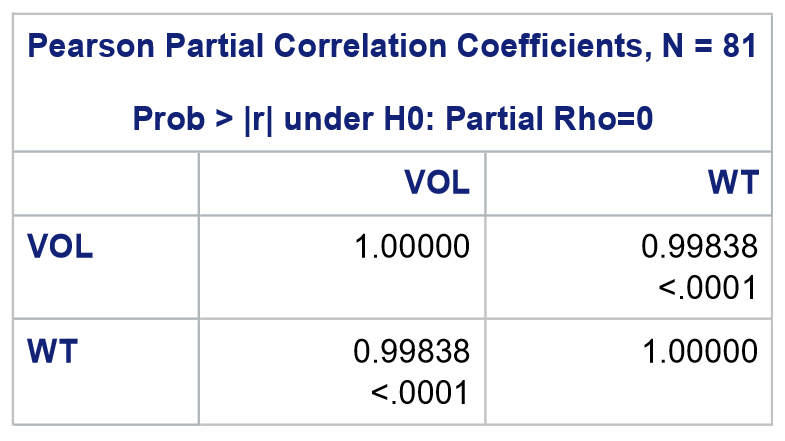
var VOL WT;

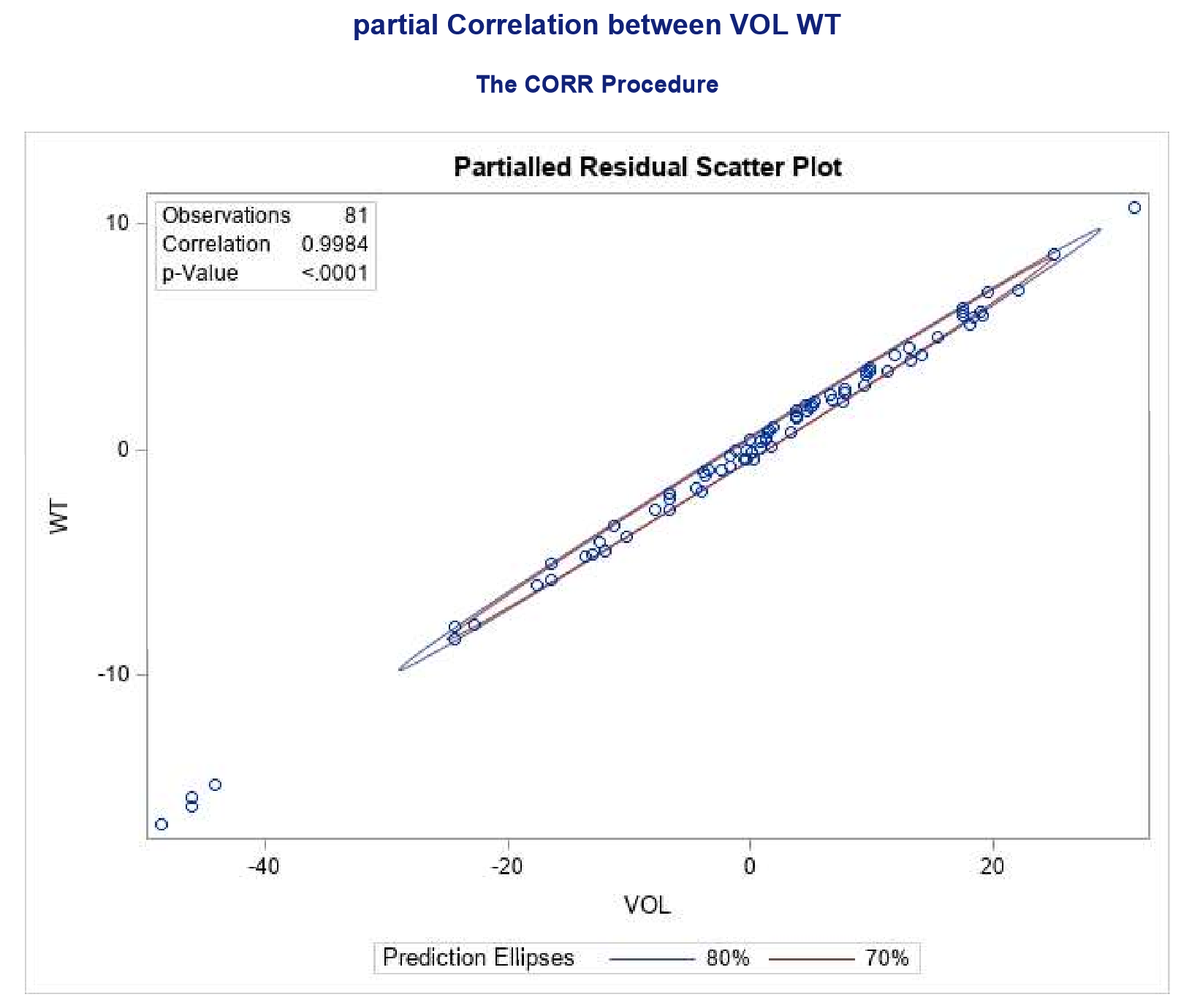
partial HP MPG SP;

title 'partial Correlation between VOL WT';

run;

**Results:** Earlier the VOL and WT value is0.99920 which is greater than 0.99838, we were successfully able to remove the affect of other variables on VOL and WT.





**Conclusion:** Hence, we fail to reject Null Hypothesis H0.

***Methodology ,Code and Conclusions:***

The below all information shows the methodology analysis and its respective code used , conclusion and insights drawn for each and every process mentioned to get detailed understand of the concepts used in order to get the better results in our final model. Note: extra done log transformation to increase the accuracy of the model.

**Creating Multi-Linear Regression Model-01:**

We have observed that the output variable MPG(y) is continuous and input variables VOL, HP, SP, WT are continuous in nature, in such case we are developing the model to perform multiple linear regression.

**Code Used:**

proc reg data = newdata;

model MPG = HP VOL SP WT;

output out=MODEL.CAR;

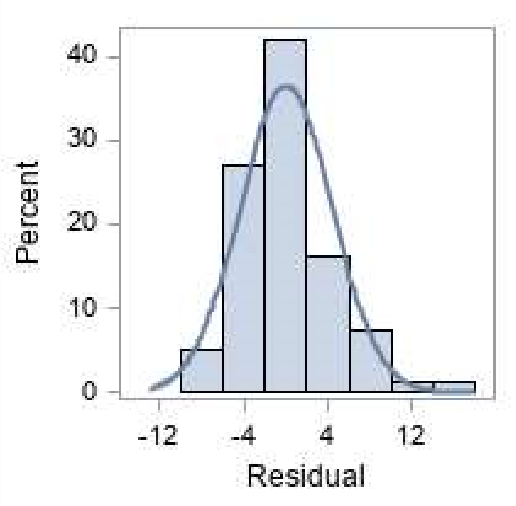
title "Regression with all Required Statistics to perform EDA";

run;

**Hypothesis Testing:**

H0: Residuals are not normally distributed.

H1: Residuals are normally distributed.



**Results:** As we know that, the error should be normally distributed but because of the residual distribution plot shown above, there is some skewness on the right tail and it is not normally distributed but probably it is near to normal distribution as shown in the above graph.

**Conclusion:** We fail to reject null hypothesis.

**Insights:** From this we get to know that the model which we have built is not fitting the data well.

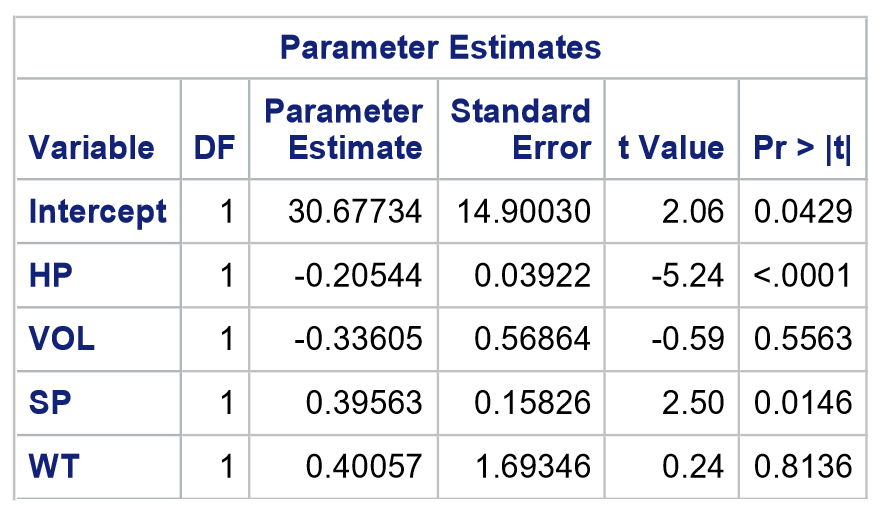
**Coefficient Analysis:** Consideringthe permissible error limit alpha=0.05

**Hypothesis Testing:**

H0: Coefficients are not significant to use in our model

H1: Coefficients are significant to use in our model

**Results:**



* The probability associated with beta(0) = 0.0429 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(0) in our model which means beta(0) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(1) < 0.0001 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(1) in our model which means beta(1) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(2) = 0.5563 which is greater than alpha value 0.05, then we say beta(2) value cannot be used in our model that we have built as it is not significant.
* The probability associated with beta(3) = 0.0146 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(3) in our model which means beta(3) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(4) = 0.8136 which is greater than alpha value 0.05, then we say beta(4) value cannot be used in our model that we have built as it is not significant.

**Conclusion:** We fail to reject Null Hypothesis.

**R-Square Analysis(Coefficient of Determination):** It actually represents percentage variation in Output variable MPG explained by input variables HP, VOL, SP, WT.

The higher the R-Square value the better the model fits the data.

**Hypothesis Testing:**

H0: R-Square value greater than 0.80.

H1: R-Square value not greater than 0.80.

**Results:**



Here, 0.2799 is less than 0.80, hence it is not a good model. The coefficient value of volume and weight are becoming insignificant so this is also pointing towards collinearity issue.

**Conclusion:** We reject Null Hypothesis

Insights: We have already figured out {VOL,WT} and {HP,SP} were collinear, probably their interaction is causing the issue. ToO fix this issue, let’s see what happens if we build the model using only VOL, only WT, or both VOL & WT.

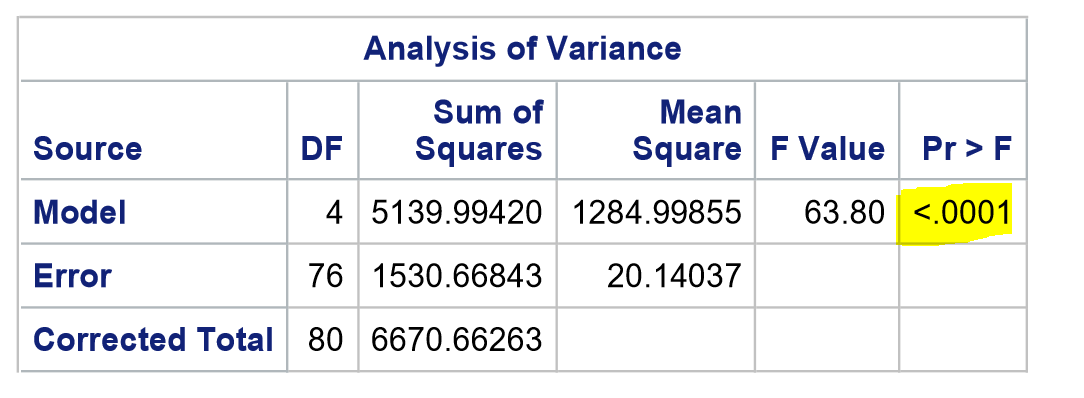
**F-Statistic Analysis:** It is used to check the overall significance of the model.

**Hypothesis Testing:**

H0: Overall significance of the model is not good

H1: Overall significance of the model is good

**Results:** The p-value for F-Statistic is less than 0.0001 which is less than 0.05 so from this we say that the overall significance of the model is good.



**Conclusion:** We reject Null Hypothesis.

**Creating Multi-Linear Regression Model-02:** As per the insights drawn from above model, we are going to develop another model Multi Collinearity Check, Model will be based on VOL.

**Code Used:**

proc reg data = newdata;

model MPG = VOL;

output out=MODEL.CARV;

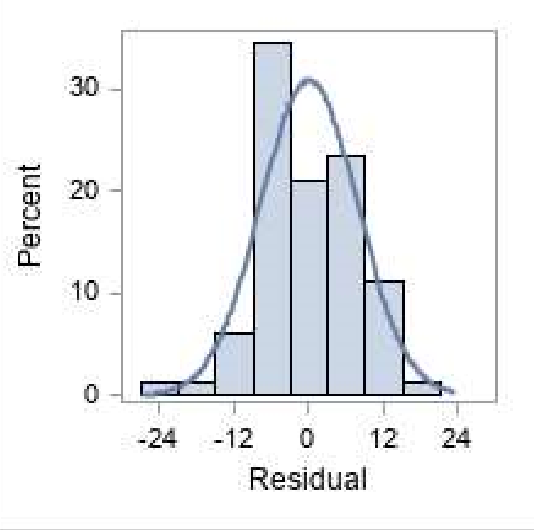
title " volume parameter statistics";

run;

**Hypothesis Testing:**

H0: Residuals are not normally distributed.

H1: Residuals are normally distributed.



**Results:** As we know that, the error should be normally distributed but because of the residual distribution plot shown above, there is some skewness on the left tail and it is not normally distributed but probably it is near to normal distribution as shown in the above graph.

**Conclusion:** We fail to reject null hypothesis.

**Insights:** From this we get to know that the model which we have built is not fitting the data well.

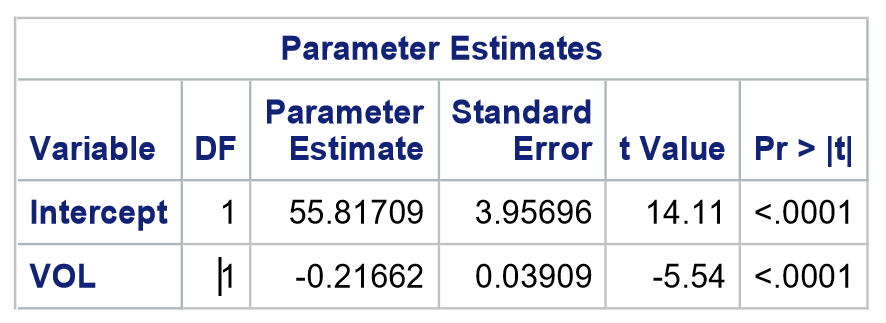
**Coefficient Analysis:** Consideringthe permissible error limit alpha=0.05

**Hypothesis Testing:**

H0: Coefficients are not significant to use in our model

H1: Coefficients are use significant to use in our model

**Results:**



* The probability associated with beta(0) < 0.0001 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(0) in our model which means beta(0) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(1) < 0.0001 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(1) in our model which means beta(1) value is significantly different than 0 and it can be used in our model that we have built.

**Conclusion:** We reject Null Hypothesis.

**R-Square Analysis(Coefficient of Determination):** It actually represents percentage variation in Output variable MPG explained by input variables VOL.

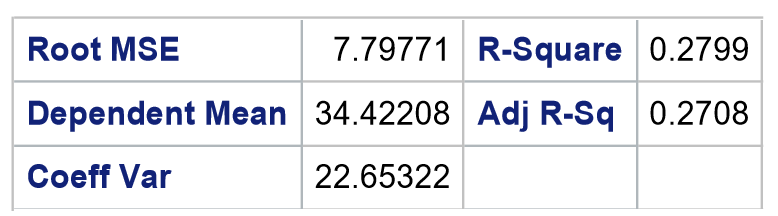
The higher the R-Square value the better the model fits the data.

**Hypothesis Testing:**

H0: R-Square value greater than 0.80.

H1: R-Square value not greater than 0.80.

**Results:**



Here, 0.2708 is less than 0.80, hence it is absolutely bad model when compared to the earlier model. Volume is doing a bad job on explaining the variation in the output variable MPG.

**Conclusion:** We reject Null Hypothesis

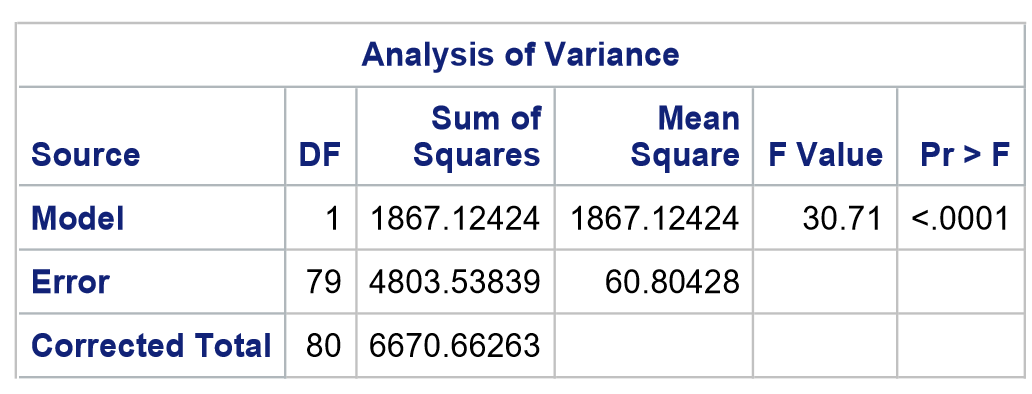
**F-Statistic Analysis:** It is used to check the overall significance of the model.

**Hypothesis Testing:**

H0: Overall significance of the model is not good

H1: Overall significance of the model is good

**Results:** The p-value for F-Statistic is less than 0.0001 which is less than 0.05 so from this we say that the overall significance of the model is good.



**Conclusion:** We reject Null Hypothesis.

**Creating Multi-Linear Regression Model-03:** As per the insights drawn from above model, we are going to develop another model Multi Collinearity Check, Model will be based on WT.

**Code Used:**

proc reg data = newdata;

model MPG = WT;

output out=MODEL.CARW;

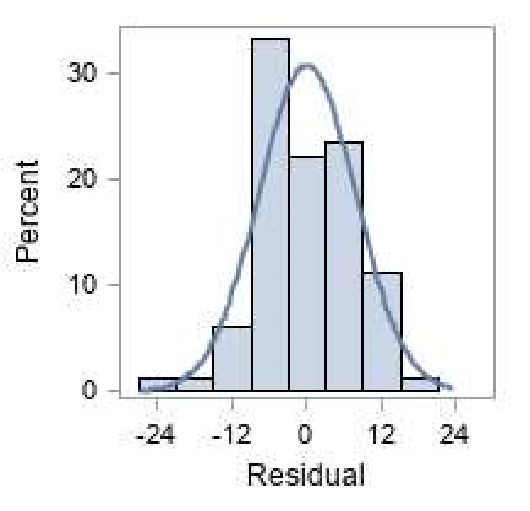
title " weight parameter statistics";

run;

**Hypothesis Testing:**

H0: Residuals are not normally distributed.

H1: Residuals are normally distributed.



**Results:** As we know that, the error should be normally distributed but because of the residual distribution plot shown above, there is some skewness on the left tail and it is not normally distributed but probably it is near to normal distribution as shown in the above graph.

**Conclusion:** We fail to reject null hypothesis.

**Insights:** From this we get to know that the model which we have built is not fitting the data well.

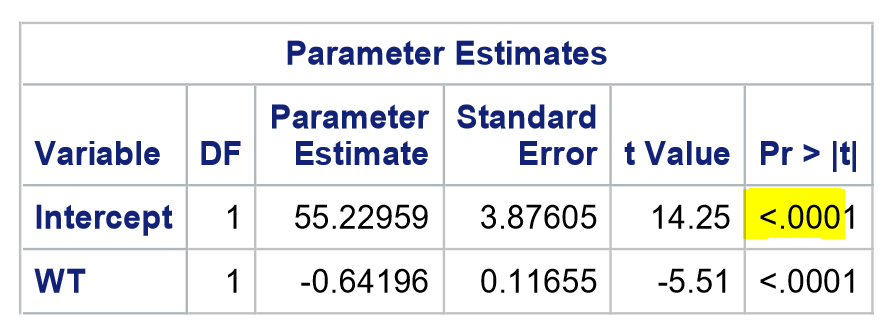
**Coefficient Analysis:** Consideringthe permissible error limit alpha=0.05

**Hypothesis Testing:**

H0: Coefficients are not significant to use in our model

H1: Coefficients are significant to use in our model

**Results:**



* The probability associated with beta(0) < 0.0001 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(0) in our model which means beta(0) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(1) < 0.0001 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(1) in our model which means beta(1) value is significantly different than 0 and it can be used in our model that we have built.

**Conclusion:** We reject Null Hypothesis.

**R-Square Analysis(Coefficient of Determination):** It actually represents percentage variation in Output variable MPG explained by input variables WT.

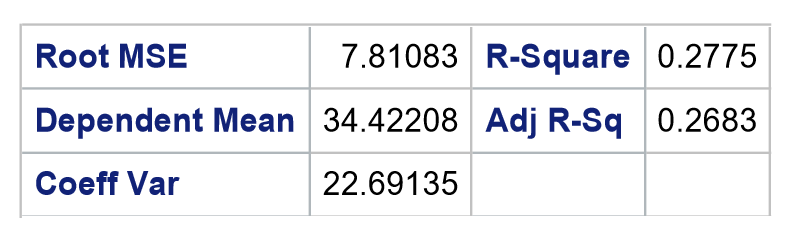
The higher the R-Square value the better the model fits the data.

**Hypothesis Testing:**

H0: R-Square value greater than 0.80.

H1: R-Square value not greater than 0.80.

**Results:**



Here, 0.2775 is less than 0.80, hence it is absolutely bad model when compared to the earlier model. WT is doing a bad job on explaining the variation in the output variable MPG.

**Conclusion:** We reject Null Hypothesis

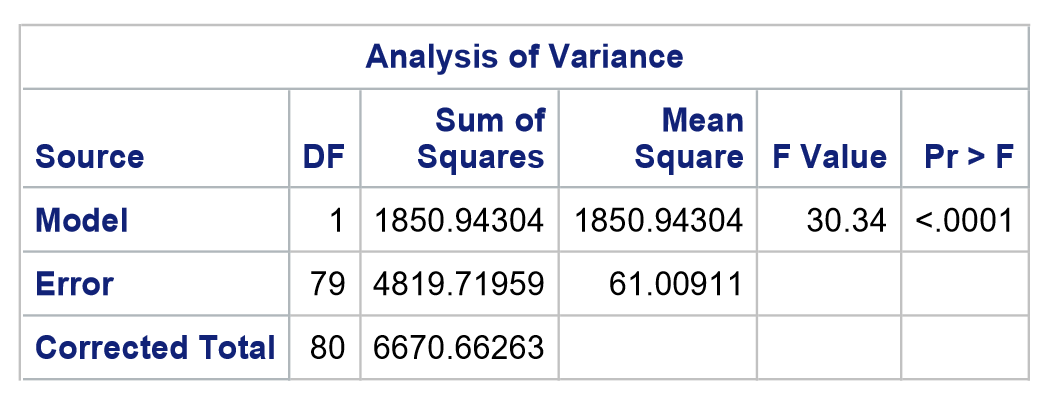
**F-Statistic Analysis:** It is used to check the overall significance of the model.

**Hypothesis Testing:**

H0: Overall significance of the model is not good

H1: Overall significance of the model is good

**Results:** The p-value for F-Statistic is less than 0.0001 which is less than 0.05 so from this we say that the overall significance of the model is good.



**Conclusion:** We reject Null Hypothesis.

**Creating Multi-Linear Regression Model-04:** As per the insights drawn from above model, we are going to develop another model Multi Collinearity Check, Model will be based on VOL & WT.

**Code Used:**

proc reg data = newdata;

model MPG = VOL WT;

output out=MODEL.CARVW;

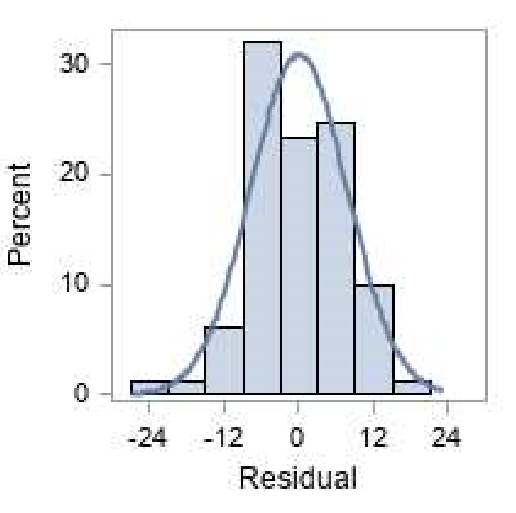
title " Combine volume + weight parameter statistics";

run;

**Hypothesis Testing:**

H0: Residual is not normally distributed.

H1: Residual is normally distributed.



**Results:** As we know that, the error should be normally distributed but because of the residual distribution plot shown above, there is some skewness on the left tail and it is not normally distributed but probably it is near to normal distribution as shown in the above graph.

**Conclusion:** We fail to reject null hypothesis.

**Insights:** From this we get to know that the model which we have built is not fitting the data well.

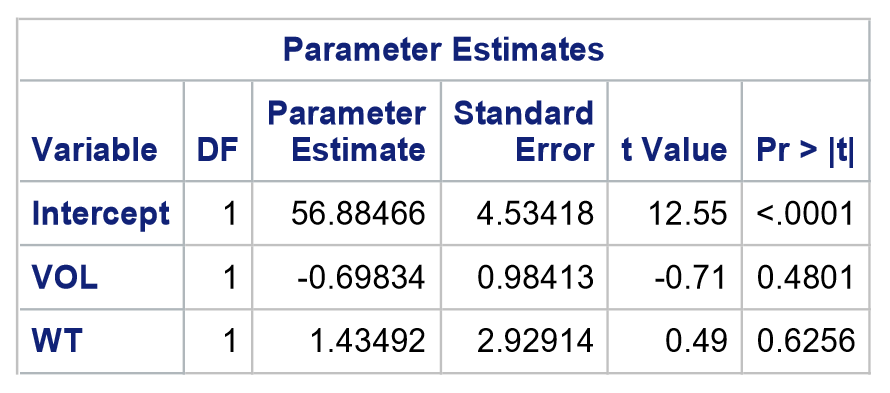
**Coefficient Analysis:** Consideringthe permissible error limit alpha=0.05

**Hypothesis Testing:**

H0: Coefficients are not significant to use in our model

H1: Coefficients are significant to use in our model

**Results:**



* The probability associated with beta(0) < 0.0001 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(0) in our model which means beta(0) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(1) = 0.4801 which is greater than alpha value 0.05, which is not significant to use in our model.
* The probability associated with beta(2) = 0.6256 which is greater than alpha value 0.05, which is not significant to use in our model.

**Conclusion:** We fail to reject Null Hypothesis.

Insights: If we individually use VOL, WT in our models, then they became significant. If we put both together in our model, due to some interaction they are running into some collinearity issue and they became insignificant. Hence, we cannot use them in our model.

**R-Square Analysis(Coefficient of Determination):** It actually represents percentage variation in Output variable MPG explained by input variables VOL & WT.

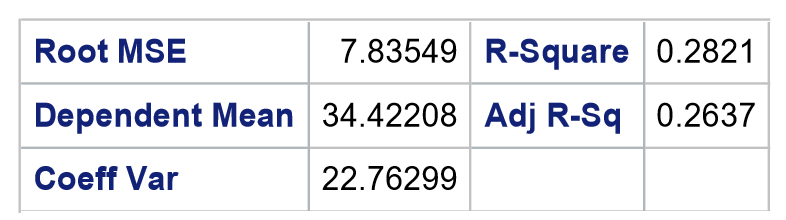
The higher the R-Square value the better the model fits the data.

**Hypothesis Testing:**

H0: R-Square value greater than 0.80.

H1: R-Square value not greater than 0.80.

**Results:**



Here, 0.2821 is less than 0.80, hence it is absolutely bad model. VOL & WT together are doing a bad job on explaining the variation in the output variable MPG.

**Conclusion:** We reject Null Hypothesis

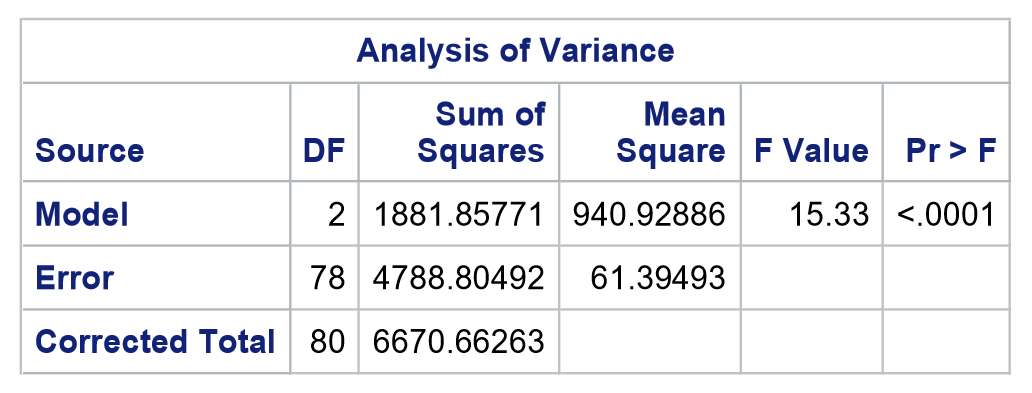
**F-Statistic Analysis:** It is used to check the overall significance of the model.

**Hypothesis Testing:**

H0: Overall significance of the model is not good

H1: Overall significance of the model is good

**Results:** The p-value for F-Statistic is less than 0.0001 which is less than 0.05 so from this we say that the overall significance of the model is good.



**Conclusion:** We reject Null Hypothesis.

Insights: In order to solve collinearity problem, we might have to drop one of the variables either VOL or WT, so we should go with better approach in order to solve collinearity problem. So we will find an approach which is better either than dropping any variable or remove any influential observation. Let’s understand, if we remove one variable we will be losing 81 observations but if we remove one influential observation, we will lose only one row. So in order to achieve this, we will be using influential index plot/deletion diagnostic plot.

**Influential Index Plot Analysis:**

**Code Used:**

proc reg data = newdata plots(only) = (CooksD(label) DFFits(label));

model MPG = HP VOL SP WT/VIF TOL COLLIN;

/\* output out=MODEL.CAR1;\*/

/\*output out=MODEL.CAR1 pred=Pred rstudent=RStudent dffits=DFFits cookd=CooksD; /\* optional: output statistics \*/

output out=MODELCAR r=residual h=hat rstudent=rstudent cookd = cooksd dffits=DFFits pred=Pred;

title "Regression with all Required Statistics to perform EDA - ta";

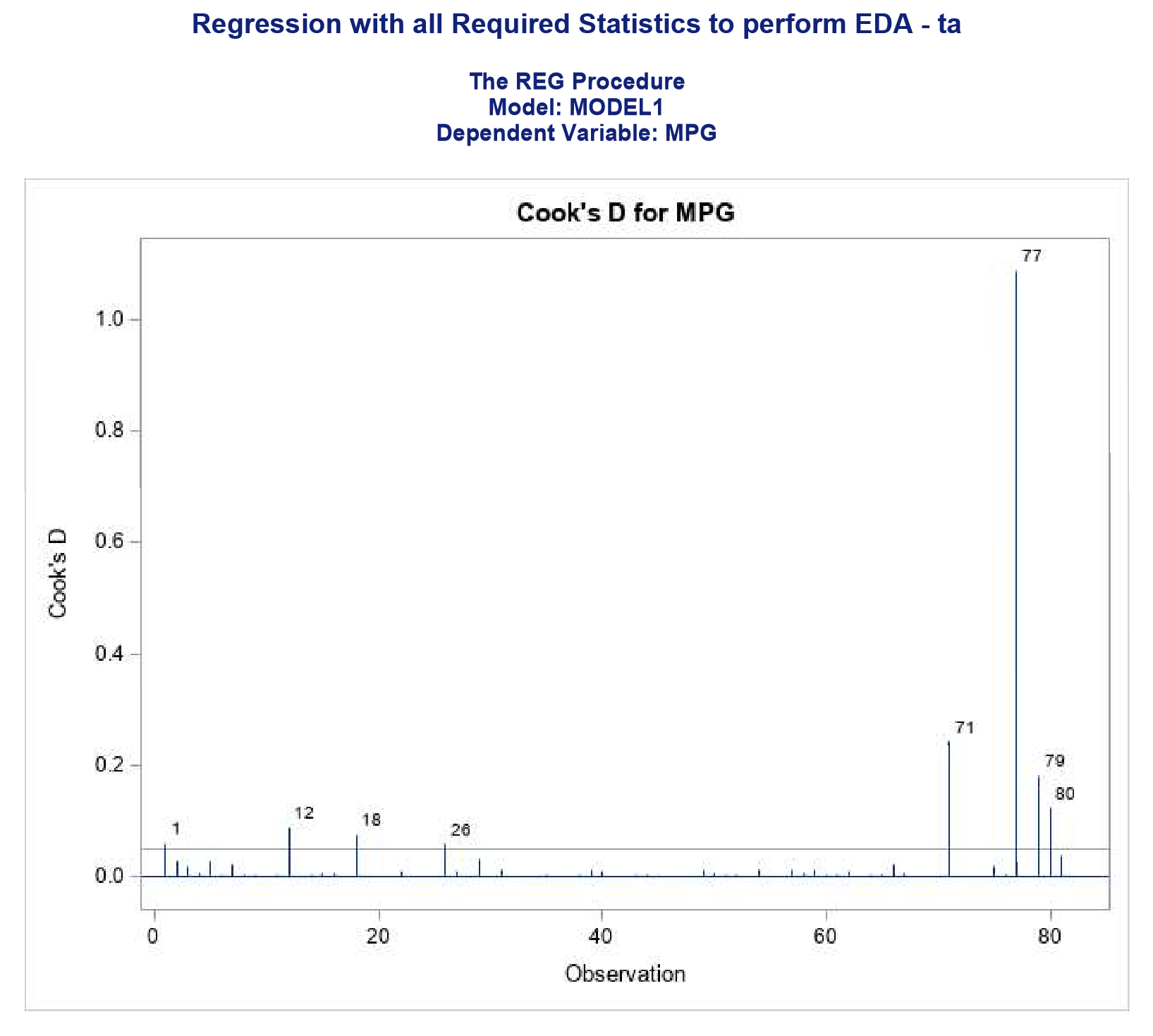
run;

proc print data = MODELCAR; /\*Note that tempfile is the output from the proc reg statements. \*/

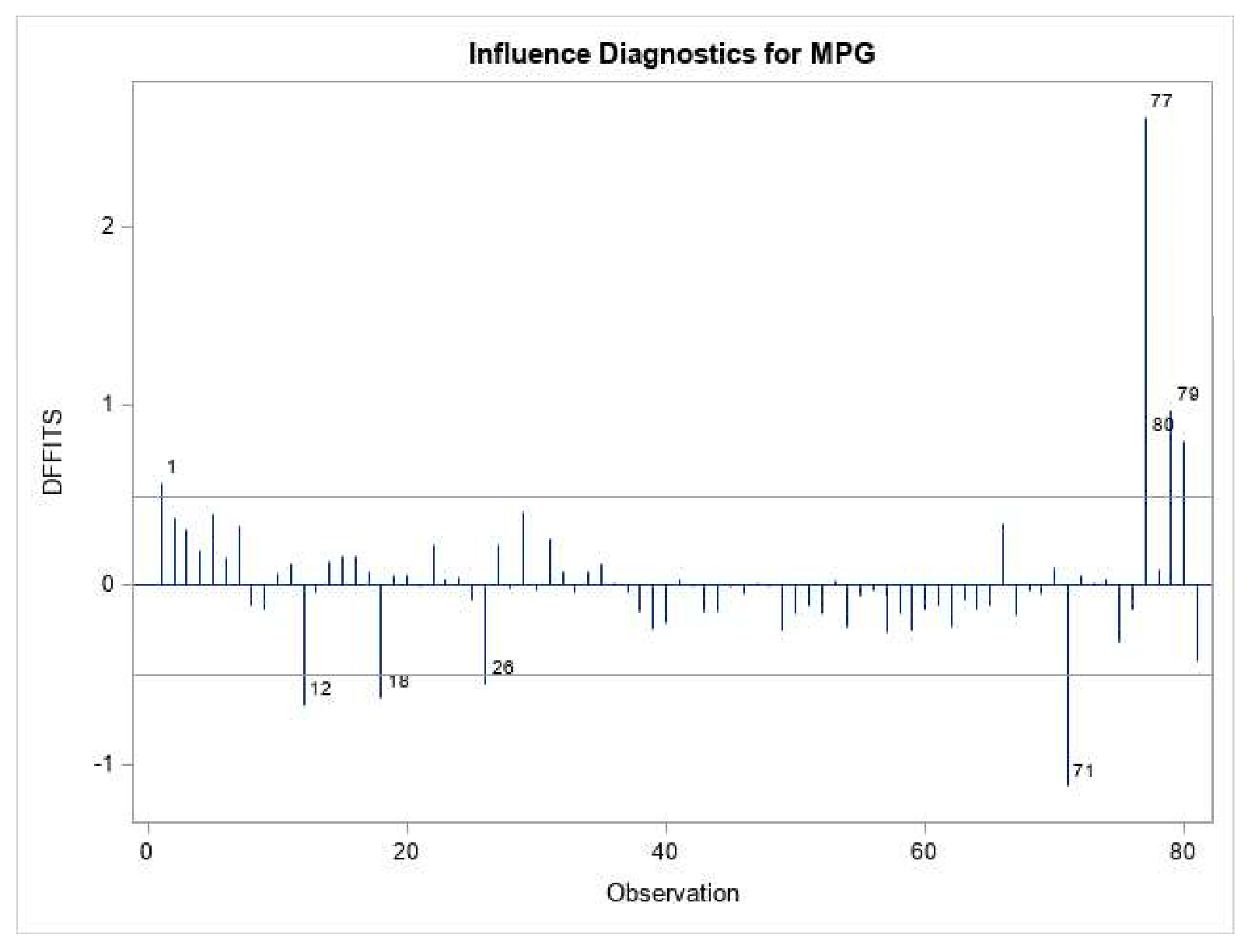
var MPG HP VOL SP WT residual hat rstudent cooksd;

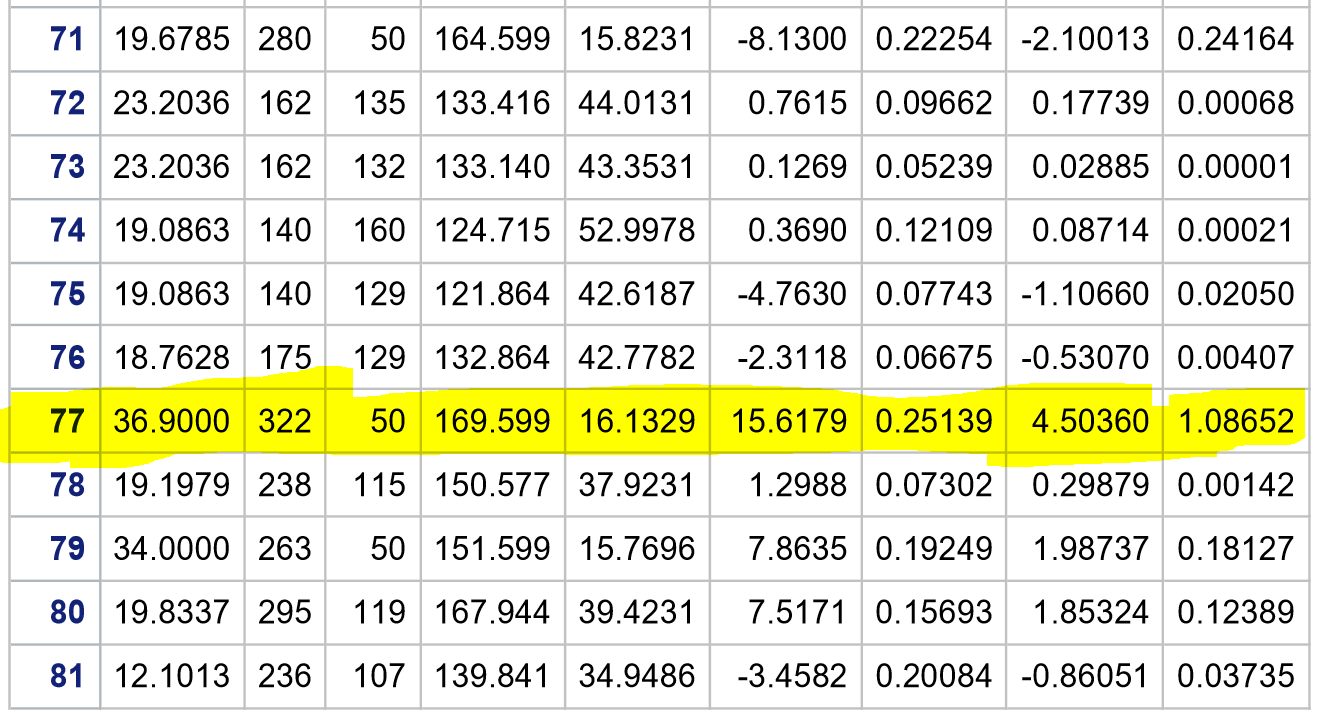
title "Influential Observations Output - Table";

run;



Results: Any observation which is greater than 0.5 will be considered as an influential observation. Probably, 77th row is having some issue.





**Removal of Influential Observation 77:**

**Hypothesis Testing:**

H0: 77th observation is not an Influential Observation

H1: 77th observation is an Influential Observation

**Code used:**

Data newdataWithObs77Removed;

set newdata;

if \_n\_ = 77 then delete; /\*\*\*\_n\_ is a SAS variable which counts observations\*\*/

proc reg data = newdataWithObs77Removed plots(only) = (CooksD(label) DFFits(label));

model MPG = HP VOL SP WT/VIF TOL COLLIN;

output out=tempfile1 r=residual h=hat rstudent=rstudent cookd = cooksd dffits=DFFits pred=Pred;

title "Proc Reg with Obs 77 removed -yh ";

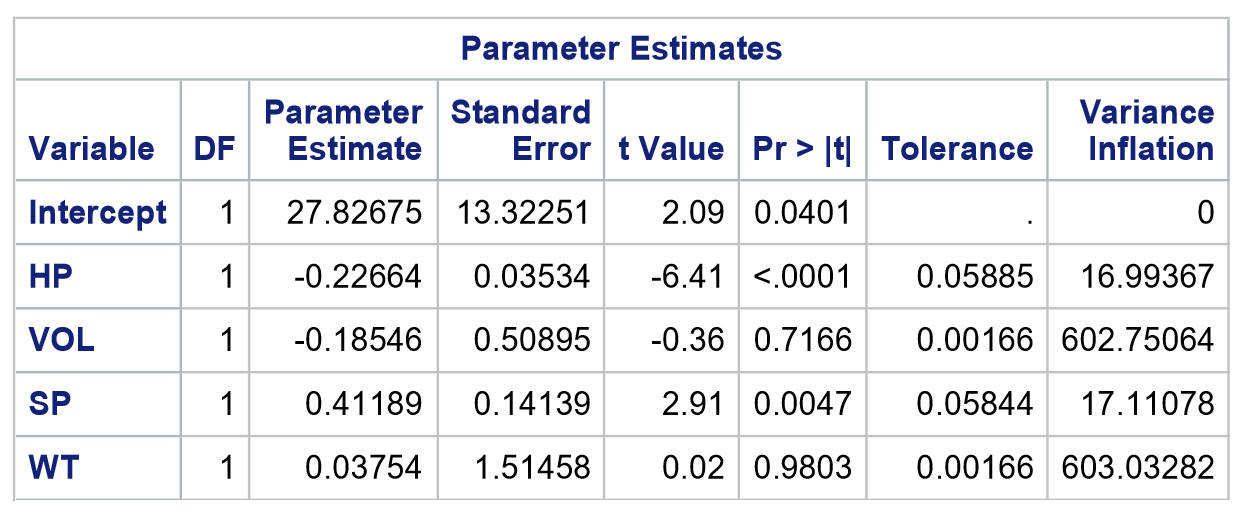
run;

proc print data = tempfile1;

var MPG HP VOL SP WT residual hat rstudent cooksd;

title "Influential Obs without Obs 77";

run;



**Results:**

* The probability associated with beta(0) = 0.0401 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(0) in our model which means beta(0) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(1) < 0.0001 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(1) in our model which means beta(1) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(2) = 0.7166 which is greater than alpha value 0.05, then we say beta(2) value cannot be used in our model that we have built as it is not significant.
* The probability associated with beta(3) = 0.0047 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(3) in our model which means beta(3) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(4) = 0.9803 which is greater than alpha value 0.05, then we say beta(4) value cannot be used in our model that we have built as it is not significant.

Variation Inflation factor for VOL = 602.75064 and WT = 603.03282. As we know that if VIF Value is greater than 10, there is a collinearity problem going on in our model.

**Conclusion:** We reject Null Hypothesis.

Insights: For all the variables VOL, WT, SP, HP, the VIF is greater than 10 which means all input variables are involved in collinearity problem. However, the collinearity value for WT is highest. Probably we can go ahead and remove the input variable WT. It is not that easy to rationalize, we will try to add another level of explanation over here so we are sure that we will go ahead with added variable plots.

**R-Square Analysis(Coefficient of Determination):** It actually represents percentage variation in Output variable MPG explained by input variables VOL, HP & SP.

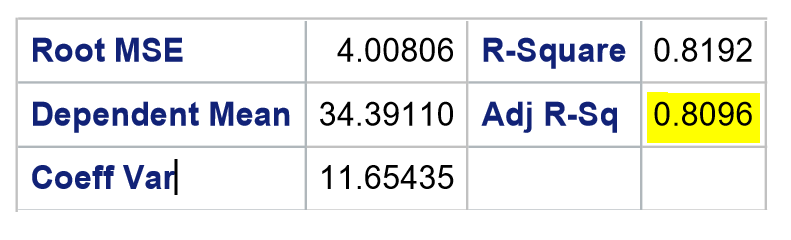
The higher the R-Square value the better the model fits the data.

**Hypothesis Testing:**

H0: R-Square value greater than 0.80.

H1: R-Square value not greater than 0.80.

**Results:**



Here, 0.8096 is greater than 0.80, hence it is a good model.

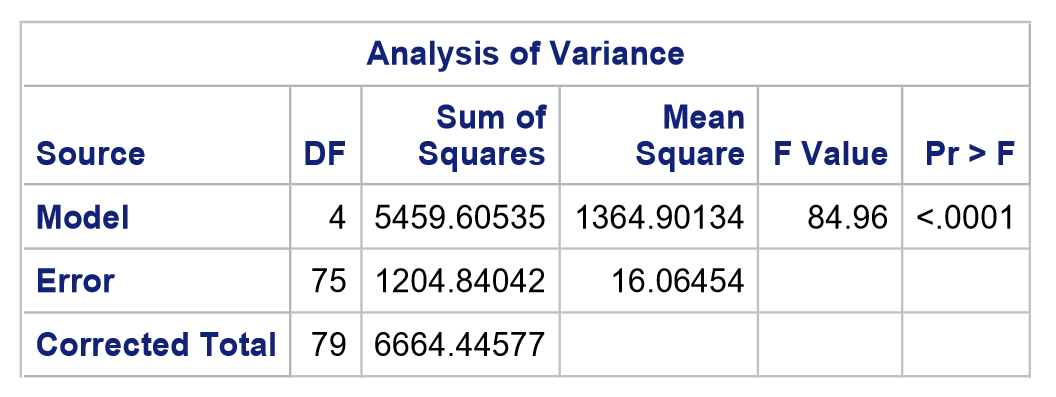
**Conclusion:** We fail to reject Null Hypothesis

**F-Statistic Analysis:** It is used to check the overall significance of the model.

**Hypothesis Testing:**

H0: Overall significance of the model is not good

H1: Overall significance of the model is good



**Results:** The p-value for F-Statistic is less than 0.0001 which is less than 0.05 so from this we say that the overall significance of the model is good.

**Conclusion:** We reject Null Hypothesis.

**Added Variable Plot Analysis:** The Added Variable Plot helps us evaluate the residuals (and coefficients) of the predictor/input variables in a multiple regression while holding the other variables constant. This is important because predictor variables may be correlated with one another (either negatively or positively).

**Code Used:**

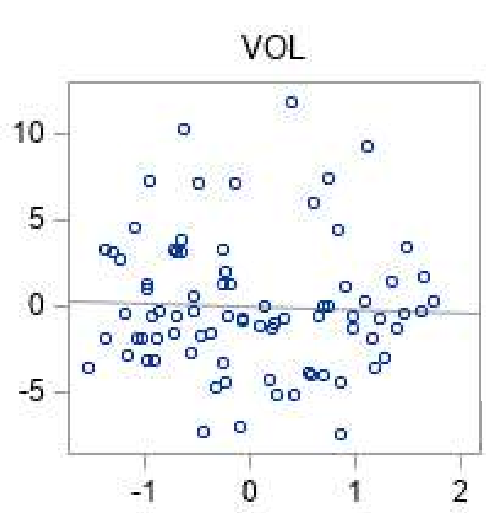
proc reg data = newdataWithObs77Removed;

model MPG = HP VOL SP WT / partial;

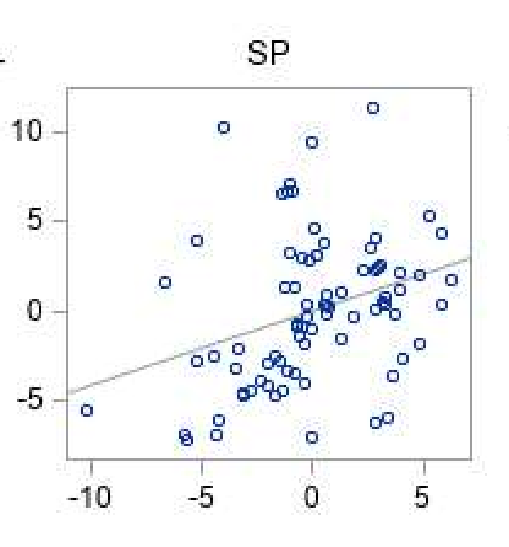
plot r.\*HP;

title "Proc Reg with Obs 78 del,77 removed -yav plot-T ";

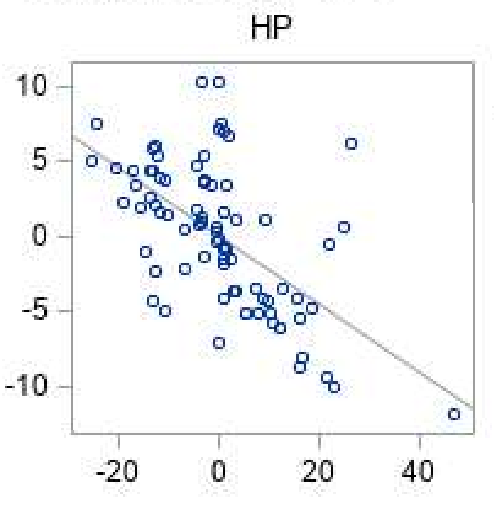
run;



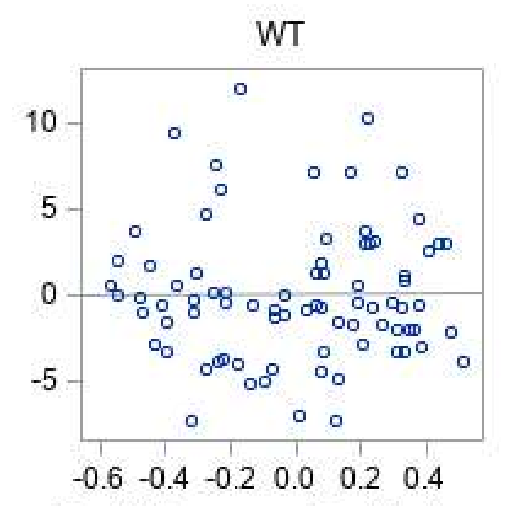
From the figure we can observe that we will have very less impact and less negative correlation.



From the figure we can observe that, it has great impact and positively correlated.



From the figure we can observe that it has great impact and negatively correlated.



We observe from the above figure that, it has least impact and we are not sure about whether it is positively correlated or negatively correlated.

**Insights:** By observing VIF variable values and added variable plots for the variables, VOL and WT contributing very less on explaining the variation of the output variable MPG. So, we are getting a sense that variable WT is contributing less to the Mileage of the car. Probably, we should delete the variable WT for the overall model building as WT has got the highest collinearity value. So, we will build the final model without weight.

**Final Model without WT variable:**

**Code Used:**

proc reg data = newdataWithObs77Removed plots(only) = (CooksD(label) DFFits(label));

model MPG = HP VOL SP /VIF TOL COLLIN;

output out=tempfile3 r=residual h=hat rstudent=rstudent cookd = cooksd dffits=DFFits pred=Pred;

title "Proc Reg with Obs 78 del,77 removed -ya - final results";

run;

proc print data = tempfile3;

var MPG HP VOL SP residual hat rstudent cooksd;

title "Influential Obs without Obs 78del,77 - final model without wt";

run;

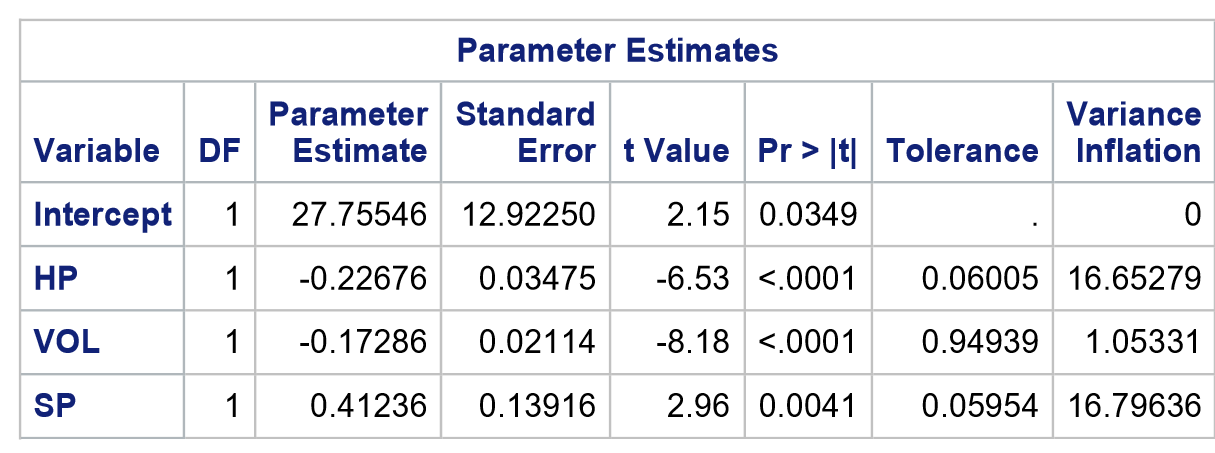
**Coefficient Analysis:** Consideringthe permissible error limit alpha=0.05

**Hypothesis Testing:**

H0: Coefficients are not significant to use in our model

H1: Coefficients are significant to use in our model

**Results:**



* The probability associated with beta(0) = 0.0349 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(0) in our model which means beta(0) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(1) < 0.0001 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(0) in our model which means beta(0) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(2) < 0.0001 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(0) in our model which means beta(0) value is significantly different than 0 and it can be used in our model that we have built.
* The probability associated with beta(3) = 0.0041 which is less than alpha value 0.05, then we say there is only less than 5% chance of going wrong if we use beta(0) in our model which means beta(0) value is significantly different than 0 and it can be used in our model that we have built.

**Conclusion:** We reject Null Hypothesis.

**Insights:** All variables are doing a great job and we can significantly use the variables in our model. Earlier we had collinearity issue with VOL and WT. The moment we have removed WT from the relationship, the variable VOL became significant.

**R-Square Analysis(Coefficient of Determination):** It actually represents percentage variation in Output variable MPG explained by input variables VOL, HP & SP.

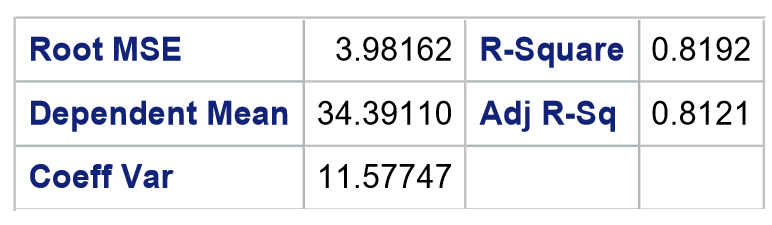
The higher the R-Square value the better the model fits the data.

**Hypothesis Testing:**

H0: R-Square value greater than 0.80.

H1: R-Square value not greater than 0.80.

**Results:**



Here, 0.8121 is greater than 0.80, hence it is a good model. Earlier model adjusted R-square value is 0.8096 which is less than 0.8121. In this model, the R-square value has increased when compared to the previous model. The variation explained also got better in this model.

**Conclusion:** We fail to reject Null Hypothesis

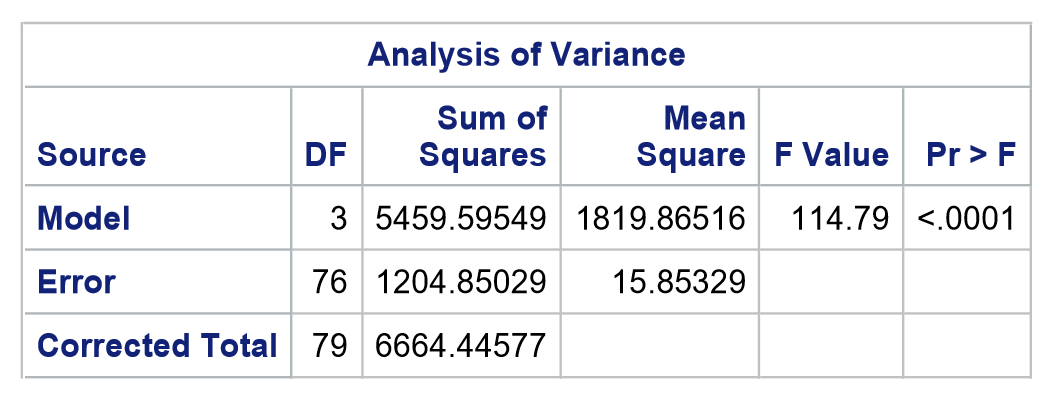
**F-Statistic Analysis:** It is used to check the overall significance of the model.

**Hypothesis Testing:**

H0: Overall significance of the model is not good

H1: Overall significance of the model is good

**Results:** The p-value for F-Statistic is less than 0.0001 which is less than 0.05 so from this we say that the overall significance of the model is good.



**Conclusion:** We reject Null Hypothesis.

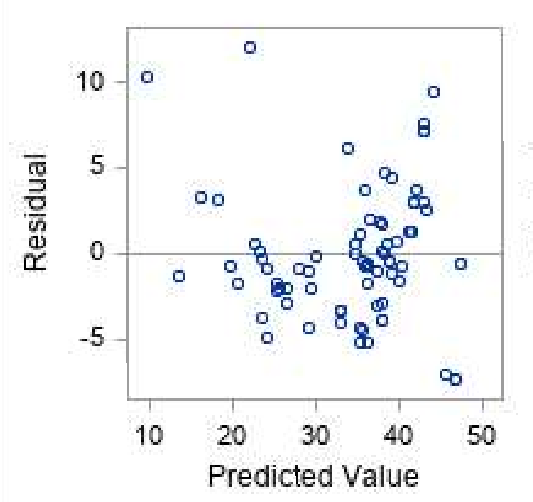
**Final Model:**

**MPG = 27.75546 – 0.22676(HP) – 0.17286(VOL) + 0.41236(SP)**

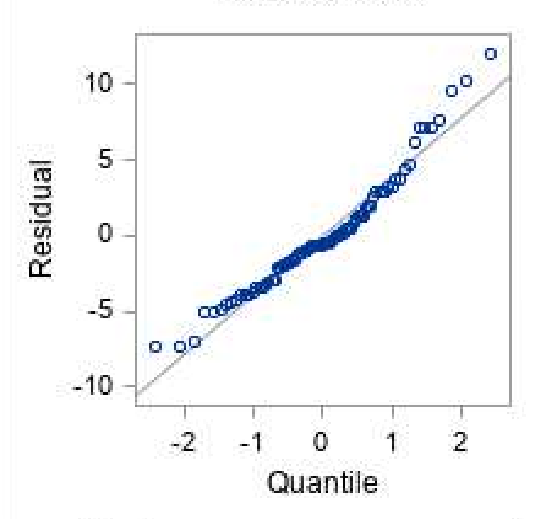
All these 3 variables are doing a great job of explaining the variation in MPG of the car.

**Line Assumptions validation for the Final Model.**

**Step 01: Need to evaluate the errors.(residual VS fitted values).** As on how , the fitted values are increasing, the error values should not increase or decrease. The errors should be constant. From the below graph shows that the errors are constant.



**Speo02: To Check whether the errors are getting normally distributed or not.** From the below graph shown that the, majority of the datapoints are around its line, we say that the datapoints are normally distributed.



Insight: Our final model is good with line assumptions as well.

**Transformations:** Below transformations are used in order to improve the accuracy of the model.

***Log Transformation:*** We have done log transformation for all input and output variables in order to increase the model accuracy.

**Code used:**

data newdata\_logtransformation;

set newdataWithObs77Removed;

HPLOG=log10(HP);

MPGLOG=log10(MPG);

VOLLOG=log10(VOL);

SPLOG=log10(SP);

WTLOG=log10(WT);

run;

Proc print data = newdata\_logtransformation;

Title "My Newdata log transformation";

Run;

proc reg data = newdata\_logtransformation plots(only) = (CooksD(label) DFFits(label));

model MPGLOG = HPLOG VOLLOG SPLOG /VIF TOL COLLIN;

output out=tempfilelog r=residual h=hat rstudent=rstudent cookd = cooksd dffits=DFFits pred=Pred;

title "Model log transformation";

run;

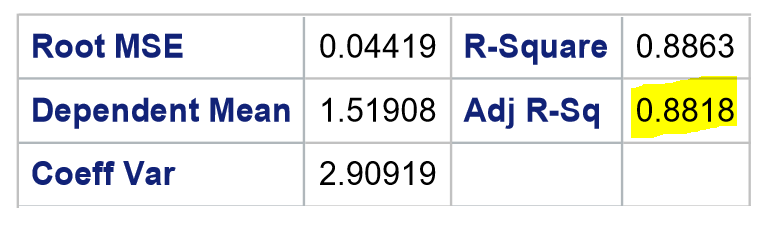
proc print data = tempfilelog;

var MPG HP VOL SP residual hat rstudent cooksd;

title "Model log transformation - final results";

run;

**Results:**



**Insights:** Using the transformation technique,The significance of the model has increased by using log transformation. The adjusted R-Square value is 0.8818.

***Square Transformation:*** We have done square transformation for all input and output variables in order to increase the model accuracy.

**Code used:**

data newdata\_sqrttransformation;

set newdataWithObs77Removed;

HPsqrt=sqrt(HP);

MPGsqrt=sqrt(MPG);

VOLsqrt=sqrt(VOL);

SPsqrt=sqrt(SP);

WTsqrt=sqrt(WT);

run;

Proc print data = newdata\_sqrttransformation;

Title "My Newdata sqrt transformation";

Run;

proc reg data = newdata\_sqrttransformation plots(only) = (CooksD(label) DFFits(label));

model MPGsqrt = HPsqrt VOLsqrt SPsqrt /VIF TOL COLLIN;

output out=tempfilesqrt r=residual h=hat rstudent=rstudent cookd = cooksd dffits=DFFits pred=Pred;

title "Model sqrt transformation";

run;

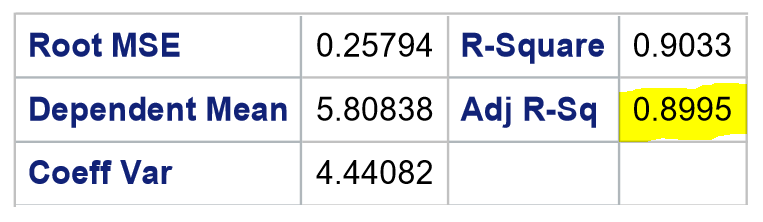
proc print data = tempfilesqrt;

var MPG HP VOL SP residual hat rstudent cooksd;

title "Model sqrt transformation - final results";

run;

**Results:**



**Insights:** Using the transformation techniques**,** The significance of the model has increased by using sqrt transformation. The adjusted R-Square value is 0.8995.

**Future research on the study**:

Apply Ridge and Lasso Regression on this data and check whether it is giving better Adjst.R\_sqrd

and less RMSE values are not