Regression
SI 422

Topic:

Prediction of Miles/Gallon on Car dataset using Regression Analysis

Submitted by:
Anjali Kanojia (22N0046)
Piyush Kumar (22N0061)
Tushar Khatri (22N0066)

Supervised by:
Prof. Siuli Mukhopadhyay

Contents

Acknowledgement

I would like to thank my project supervisor **Prof Siuli Mukhopadhyay** for their unwavering support and expert guidance throughout this project. Their advice, feedback, and constructive criticism have been instrumental in shaping the direction of this project and improving its quality. I am truly grateful for their generosity with their time and expertise.

Furthermore, I would like to express my immense gratitude to everyone who has supported me throughout this project. Your encouragement, guidance, and assistance have been invaluable and have enabled me to complete this project successfully

Lastly, I would like to thank all the resources and references that have been instrumental in shaping the ideas and concepts presented in this project. The knowledge and insights gained from these sources have been invaluable and have contributed significantly to the success of this project.

Thank you all once again for your support, guidance, and assistance. This project would not have been possible without you all. I would like to express my immense gratitude to everyone who has supported me throughout this project. Your encouragement, guidance, and assistance have been invaluable and have enabled me to complete this project successfully

Anjali Kanojia (22N0046)
Piyush Kumar (22N0061)
Tushar Khatri (22N0066)

Date: 25/04/2023

Chapter-1

Objective

The objective of this regression prediction project using the mtcars dataset is to develop a model that can predict the fuel efficiency (measured in miles per gallon, or mpg) of a car based on its various characteristics, such as its engine size, horsepower, and weight. The model is trained using historical data from the mtcars dataset, which includes information on 32 different cars and their corresponding mpg values and characteristics.

The goal is to develop a model that accurately predicts the mpg of new cars based on their characteristics, with the ultimate objective of helping car manufacturers and consumers make more informed decisions about fuel efficiency.

The success of the project is measured by the model's ability to accurately predict mpg values on new data that was not used during the training process**.** Finally at the end we obtain all necessary results, conclusions along with interpretation.

Chapter-2

Data Description

The description of the data is as follows:

Feature	Description	Data Type
y	Response Variable	Float
cyl	Number of cylinders	Integer
disp	Displacement (cu.in.)	Float
hp	Gross horsepower	Integer
drat	Rear axle ratio	Float
wt	Weight (1000 lbs.)	Float
qsec	¼ mile time	Float
am	Transmission (0 = automatic, 1 = manual)	Integer
gear	Number of forward gears	Integer
carb	Number of carburetors	Integer

1. **cyl** - A cylinder is a vital part of the engine. It's a chamber where fuel is combusted and power is generated. The cylinder consists of a piston and two valves at the top; an inlet and exhaust valves. Most cars have a 4-, 6-, or 8-cylinder engine. they will either be laid out in a straight line, V-shaped or in a flat arrangement. They will either be laid out in a straight line, V-shaped or in a flat arrangement.

2. **disp** - This metric gives a good proxy for the total amount of power the engine can generate.

3. **hp** - When installed in a car, exhaust systems, carburetor, alternator, power systems, etc all influence the power that actually gets to the drive train

4. **drat** - The rear axle gear ratio indicates the number of turns of the drive shaft for every one rotation of the wheel axle. A vehicle with a high ratio would provide more torque and thus more towing capability, for example. Transmission configuration can often influence a manufacturer's gearing ratio.

5. **wt** - The overall weight of the vehicle per 1000 lbs (half US ton).

6. **qsec** - A performance measure, primarily of acceleration. Fastest time to travel 1/4 mile from standstill (in seconds).

7. **vs** - Binary variable signaling the engine cylinder configuration a V-shape (vs=0) or Straight Line (vs=1). V==0 and S==1. The configuration offers trade-offs in power/torque, design usage in terms of space/size of engine and performance or center of gravity of the vehicle. The geometry and placement of the engine, as influenced by its cylinder head, can have numerous knock-on influences on the vehicle beyond the technical engineering considerations of the cylinder angle.

8. **am** - A binary variable signaling whether the vehicle has automatic (am=0) or manual (am=1) transmission configuration.

9. **gear** - Number of gears in the transmission. Manual transmissions have either 4 or 5 forward gears; Automatic either 3 or 4.

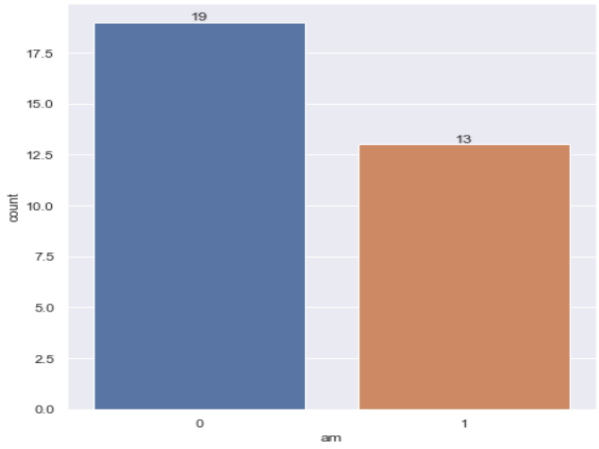
10. **carb** - The number of carburetor barrels. Engines with higher displacement typically have higher barrel configurations to accommodate the increased airflow rate of the larger engine; in other words, more capacity is available for an engine when it may need it versus constraining power output with limited barrels. A vehicle may have multiple physical carburetors, but it's less common; this metric is the sum of the number of carburetors and the number of barrels inside the carburetor.

Chapter-3

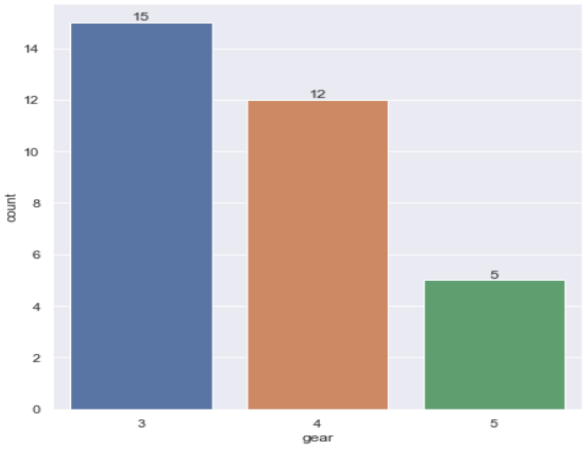
Exploratory Data Analysis

3.1 Univariate Analysis of Categorical Variables

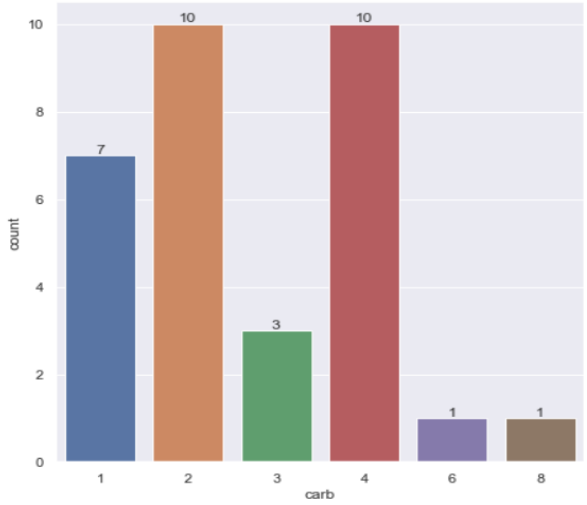
Plot 1



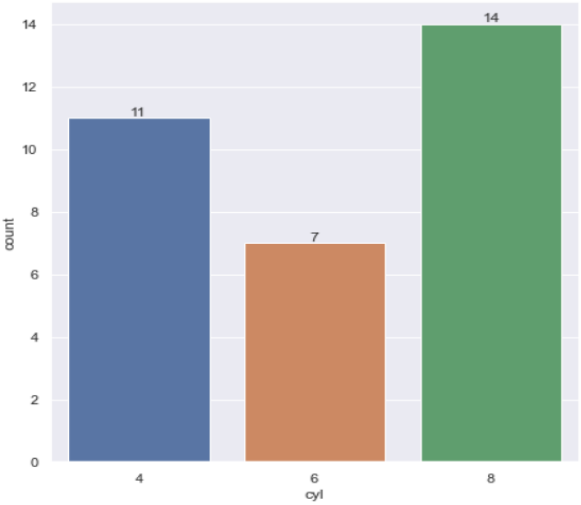
Plot 2



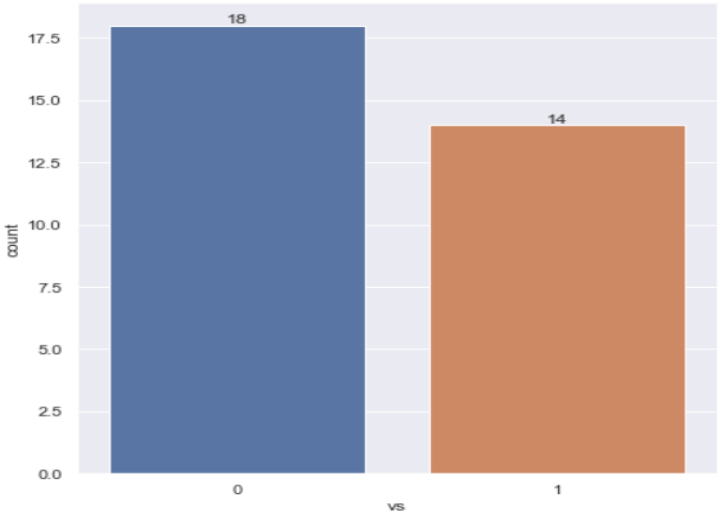
Plot 3



Plot 4



Plot 5



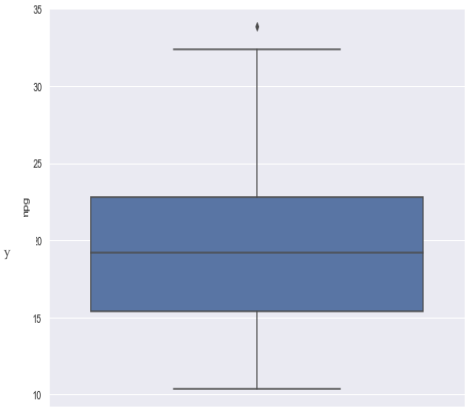
Observations from the above plots:

- From plot 1, it is clearly visible that 19 vehicles have automatic transmissions while 13 have manual one. Most of the cars have automatic transmissions.
- Plot 2 shows that the cars with 3 gears are maximum in number followed by cars with 4 gears and cars with 5 gears.
- Coming to Plot 3, cars with 2 and 4 carbureted barrels are higher in number followed by the cars with 1, 3, 6 and 8 carbureted barrels respectively.
- Plot 4 displays the number of cylinders in the cars. Cars with 8 cylinders are higher in number than cars with 4 cylinders followed by cars with 6 cylinders.
- The last plot (Plot 5) shows the count of engine cylinder configurations. i.e. there are 18 V-shape cylinders and 14 straight-line cylinders.

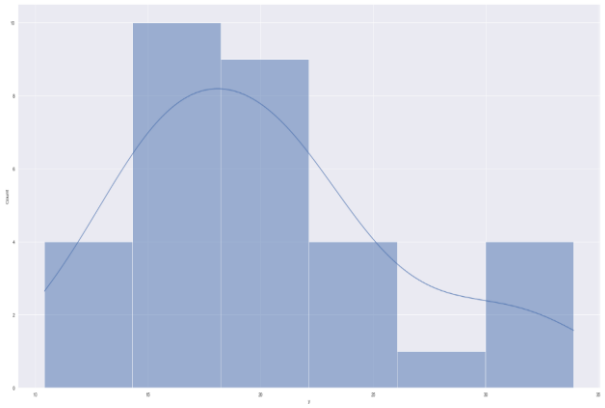
3.2 Univariate Analysis of Continuous Variable

- **Analysis of response variable**

Boxplot for y

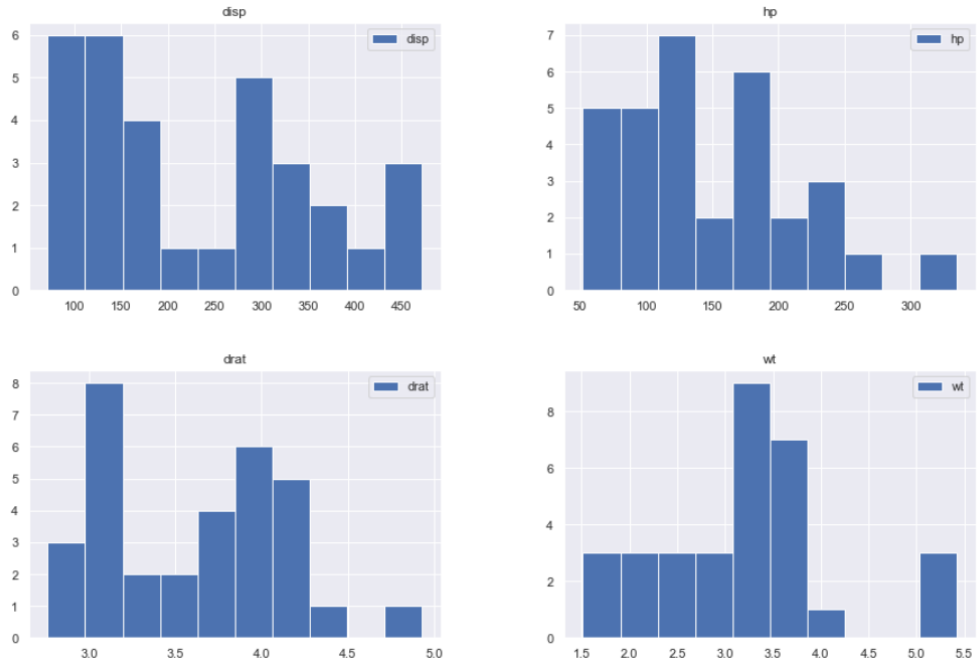


Histogram of y (with frequency polygon)



- The response variable seems to have one “off-the-trend” observation (We will examine this in more detail in upcoming sections). The first, second and third quartiles are 15.5, 19 and 23 us gallons respectively.
- Histogram suggests that there is slight departure from normality, but still the plot is very close the “bell-shape” curve.

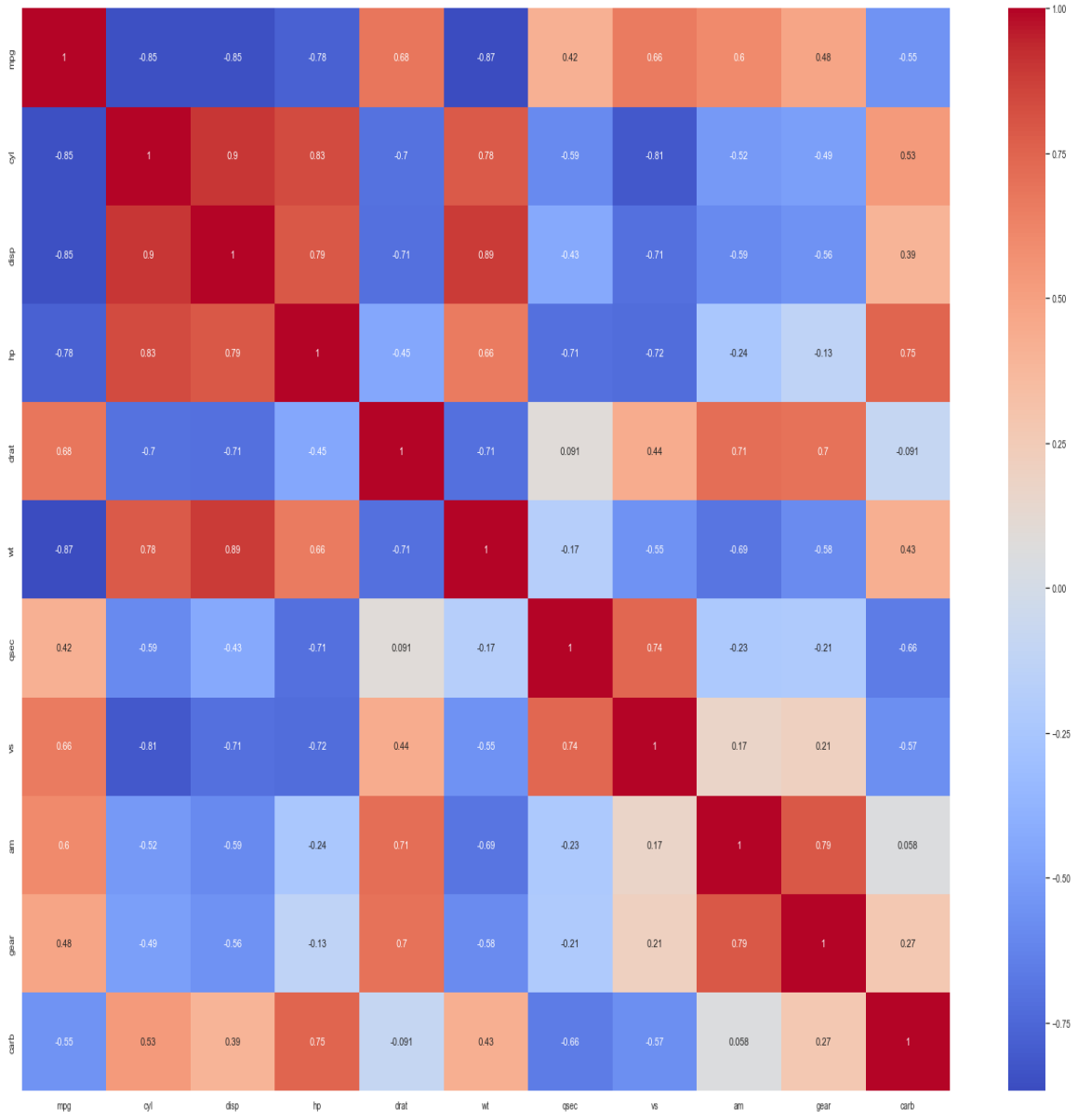
- **Histogram of other continuous variables**



3.3 Bivariate Analysis

- The Correlation Matrix + Heat Map shows that most of the variates are moderately or strongly correlated to each other.
- There is a negative correlation between mpg and weight, indicating that heavier cars tend to have lower fuel efficiency. There is also a negative correlation between mpg and horsepower, indicating that cars with higher horsepower tend to have lower fuel efficiency.
- Dataset includes several categorical variables that may also be related to the response variable. For example, the number of cylinders variable shows a clear relationship with mpg, with cars with fewer cylinders generally having higher fuel efficiency. The transmission type variable also shows a relationship with mpg, with manual cars generally having higher fuel efficiency than automatic cars.

Overall, these relationships suggest that there are several variables in the dataset that may be important predictors for the response.


mpg 1 -0.85 -0.85 -0.78 0.68 -0.87 0.42 0.66 0.6 0.48 -0.55
cyl -0.85 1 0.9 0.83 -0.7 0.78 -0.59 -0.81 -0.52 -0.49 0.53
disp -0.85 0.9 1 0.79 -0.71 0.89 -0.43 -0.71 -0.59 -0.56 0.39
hp -0.78 0.83 0.79 1 -0.45 0.66 -0.71 -0.72 -0.24 -0.13 0.75
drat 0.68 -0.7 -0.71 -0.45 1 -0.71 0.091 0.44 0.71 0.7 -0.091
wt -0.87 0.78 0.89 0.66 -0.71 1 -0.17 -0.55 -0.69 -0.58 0.43
qsec 0.42 -0.59 -0.43 -0.71 0.091 -0.17 1 0.74 -0.23 -0.21 -0.66
vs 0.66 -0.81 -0.71 -0.72 0.44 -0.55 0.74 1 0.17 0.21 -0.57
am 0.6 -0.52 -0.59 -0.24 0.71 -0.69 -0.23 0.17 1 0.79 0.058
gear 0.48 -0.49 -0.56 -0.13 0.7 -0.58 -0.21 0.21 0.79 1 0.27
carb -0.55 0.53 0.39 0.75 -0.091 0.43 -0.66 -0.57 0.058 0.27 1
mpg cyl disp hp drat wt qsec vs am gear carb
1.00 0.75 0.50 0.25 0.00 -0.25 -0.50 -0.75

Chapter-4

Univariate Linear Regression

In order to analyze the relationship between predictors and a response variable, univariate models are fitted and several plots like scatterplot, normal probability plot, density plot, etc. are generated.

These plots are used to assess the functional form of predictors that could improve the model. By fitting these univariate predictor-transformed models, we can determine which transformations are most effective and use them in our final models.

Some significant results of these linear regression models include:

Displacement (disp):

- The estimated coefficient for disp is -0.0412, with a p-value of less than 0.001. This suggests that for each additional cubic inch of engine displacement, the fuel efficiency (mpg) decreases by about 0.0412.
- As suggested by the plots generated, log and inverse disp are fitted and they prove to be beneficial with higher R and adjusted R-squared value.

Horsepower (hp):

- The estimated coefficient for hp is -0.0682, with a p-value of less than 0.001. This suggests that for each additional unit of horsepower, the fuel efficiency (mpg) decreases by about 0.0682.
- Here also, log and inverse hp are fitted, among which inverse hp has the highest R and adjusted R squared value.

Rear Axle Ratio (drat):

- The estimated coefficient for drat is 7.6782, with a p-value of 0.000.
- Here, as anticipated by the plots, squared drt shows slight improvement in terms of R and adjusted R squared when regressed against the response.

Weight (wt):

- The estimated coefficient for wt is -5.3445, with a p-value of less than 0.001. This suggests that for each additional pound of weight, the fuel efficiency (mpg) decreases by about 5.3445.
- Log and inverse wt are regressed against the response, among which log transformed seemed to be the best univariate model.

Acceleration (qsec):

- The estimated coefficient for qsec is 1.4121, with a p-value of 0.001. This suggests that for each additional second in the quarter mile time, the fuel efficiency (mpg) increases by about 1.4121.

Cylinder (cyl):

- The estimated coefficient for cyl is -2.8758, with a p-value of less than 0.001. This suggests that cars with more cylinders tend to have lower fuel efficiency (mpg).

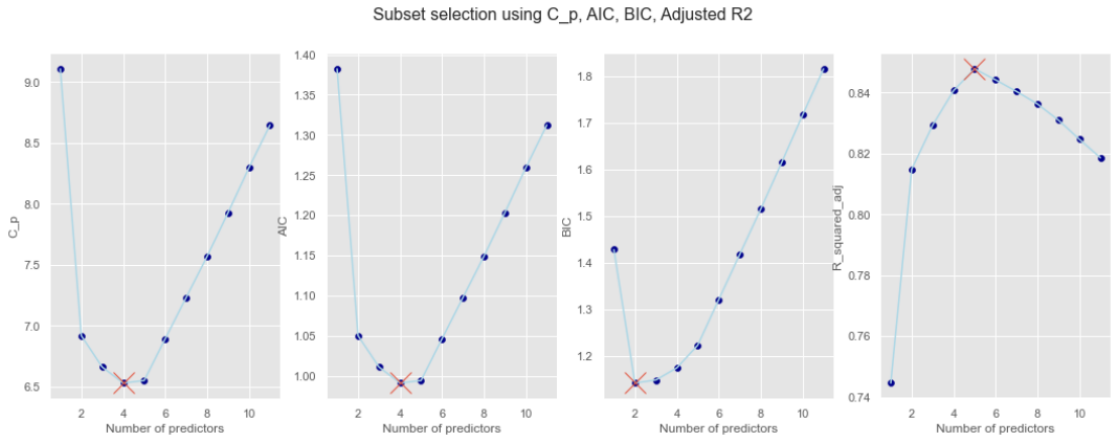
Chapter-5

Model Selection

5.1 Forward Selection

	features	RSS	R_squared	numb_features	C_p	AIC	BIC	R_squared_adj
1	[wt]	278.321938	0.752833	1	9.109392	1.382450	1.428254	0.744594
2	[wt, hp]	195.047755	0.826785	2	6.918906	1.050020	1.141628	0.814840
3	[wt, hp, cyl_6]	173.607207	0.845826	3	6.660721	1.010837	1.148250	0.829307
4	[wt, hp, cyl_6, am_1]	156.237221	0.861252	4	6.529741	0.990960	1.174177	0.840696
5	[wt, hp, cyl_6, am_1, vs_1]	143.728930	0.872360	5	6.550689	0.994139	1.223160	0.847814
6	[wt, hp, cyl_6, am_1, vs_1, gear_5]	141.500414	0.874339	6	6.892879	1.046070	1.320895	0.844180
7	[wt, hp, cyl_6, am_1, vs_1, gear_5, qsec]	139.165580	0.876412	7	7.231748	1.097497	1.418127	0.840366
8	[wt, hp, cyl_6, am_1, vs_1, gear_5, qsec, cyl_8]	136.811129	0.878503	8	7.570003	1.148831	1.515265	0.836243
9	[wt, hp, cyl_6, am_1, vs_1, gear_5, qsec, cyl_...	135.098173	0.880024	9	7.928305	1.203207	1.615445	0.830943
10	[wt, hp, cyl_6, am_1, vs_1, gear_5, qsec, cyl_...	133.757538	0.881215	10	8.298242	1.259349	1.717392	0.824651
11	[wt, hp, cyl_6, am_1, vs_1, gear_5, qsec, cyl_...	131.786214	0.882966	11	8.648470	1.312500	1.816347	0.818597

This table displays the results of a forward selection procedure where additional predictors are added to a model one at a time based on their ability to improve the model's predictive ability. The table shows the features that were added at each step, along with the residual sum of squares (RSS), R-squared value, number of features in the model, and various model selection criteria such as the Mallow C, AIC, BIC, and adjusted R-squared.



The red crosses in above plots represent where these criterion's values are optimum.

The goal of the analysis is to find the subset of features that provides the best balance of predictive power and parsimony (i.e., simplicity). In this case, the model with 5 features (i.e., wt, hp, cyl_6, am_1, vs_1) has the highest adjusted R-squared value (0.847814), which suggests that it explains a large proportion of the variance in the response variable while controlling for the number of predictors in the model. The table also shows that adding additional predictors beyond this point results in diminishing returns in terms of increasing the R-squared value, while increasing the complexity of the model.

Looking at all the criterions from the table obtained, the 5-feature model may be the best choice based on Forward Selection Procedure.

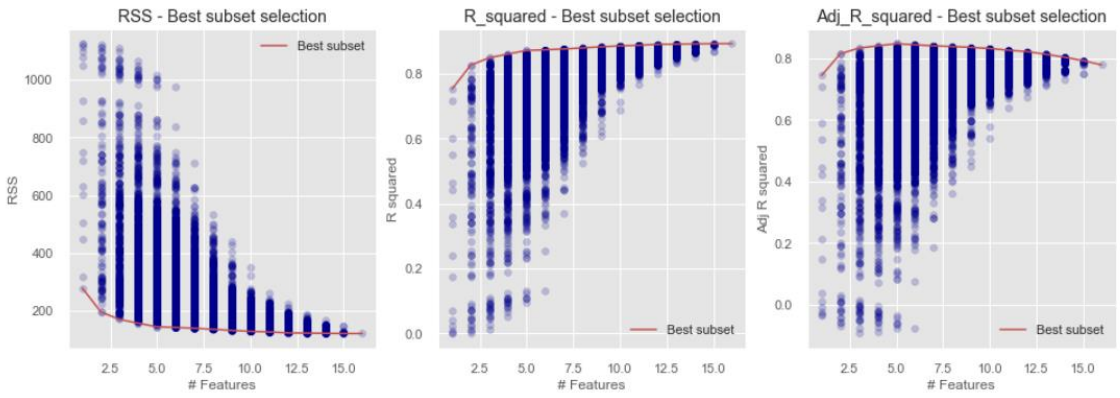
“cyl” has three categories. Here only cyl_6 is included, but we will also take cyl_8 into account as taking cyl variable as a whole makes sense.

Selected Model is: y ~ wt + hp + cyl_6 + cyl_8 + am_1 + vs_1

5.2 Best Subset Selection

Here, we have analyzed all the models by looping over all the possible combinations of predictors in the model and the best is chosen for each sample size based on the RSS, R-squared and adjusted R-squared value.

The following graph is an illustration of all possible models plotted for their RSS, R squared and adjusted R squared value:



The following table gives the models that are considered best by Best Subset Selection:

	numb_features	RSS	R_squared	Ad_R_squared	features
3	1	278.321938	0.752833	0.744594	(wt,)
32	2	195.047755	0.826785	0.814840	(hp, wt)
412	3	169.285930	0.849664	0.833556	(wt, qsec, am_1)
1241	4	156.237221	0.861252	0.840696	(hp, wt, cyl_6, am_1)
4238	5	143.728930	0.872360	0.847814	(hp, wt, cyl_6, am_1, vs_1)
10815	6	141.500414	0.874339	0.844180	(hp, wt, cyl_6, am_1, gear_5, vs_1)
21316	7	138.772565	0.876761	0.840817	(hp, wt, qsec, cyl_8, am_1, gear_5, vs_1)
28092	8	134.806953	0.880283	0.838642	(disp, hp, wt, cyl_6, gear_4, gear_5, carb_2, ...
41582	9	131.216490	0.883472	0.835801	(disp, hp, wt, cyl_6, gear_4, gear_5, carb_2, ...
51306	10	127.973154	0.886352	0.832234	(disp, hp, drat, wt, cyl_6, gear_4, gear_5, ca...
60382	11	125.959402	0.888140	0.826617	(disp, hp, wt, cyl_6, gear_4, gear_5, carb_2, ...
63463	12	122.694089	0.891040	0.822223	(disp, hp, drat, wt, cyl_6, gear_4, gear_5, ca...
64948	13	121.888146	0.891756	0.813579	(disp, hp, drat, wt, qsec, cyl_6, am_1, gear_5...
65437	14	120.960016	0.892580	0.804116	(disp, hp, drat, wt, qsec, cyl_6, am_1, gear_4...
65527	15	120.420368	0.893059	0.792802	(disp, hp, drat, wt, qsec, cyl_6, am_1, gear_4...
65534	16	120.402672	0.893075	0.779022	(disp, hp, drat, wt, qsec, cyl_6, cyl_8, am_1,...

The following table gives the best models based on different criterions:

	Parameter	Value	p	Features
0	R_squared	0.893075	16	(disp, hp, drat, wt, qsec, cyl_6, cyl_8, am_1,...
1	Adj_R_squared	0.847814	5	(hp, wt, cyl_6, am_1, vs_1)
2	AIC	150.882248	5	(hp, wt, cyl_6, am_1, vs_1)
3	BIC	157.982314	3	(wt, qsec, am_1)
4	Cp	9.037120	5	(hp, wt, cyl_6, am_1, vs_1)

Comparing all the criterions mentioned in above table, Best Subset gives the same model as best (same model proved best in Forward Selection)

Chapter-6

Model Diagnostic

The following is the model summary of our selected model:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.872
Model:	OLS	Adj. R-squared:	0.842
Method:	Least Squares	F-statistic:	28.49
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	5.06e-10
Time:	15:47:00	Log-Likelihood:	-69.436
No. Observations:	32	AIC:	152.9
Df Residuals:	25	BIC:	163.1
Df Model:	6		
Covariance Type:	nonrobust		

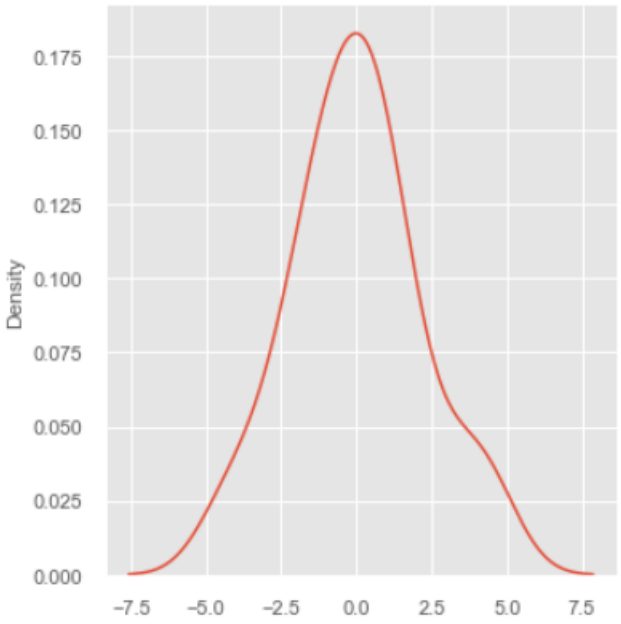
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.1846	3.420	9.118	0.000	24.141	38.228
wt	-2.3734	0.888	-2.674	0.013	-4.201	-0.545
hp	-0.0348	0.014	-2.515	0.019	-0.063	-0.006
cyl_6	-2.0901	1.629	-1.283	0.211	-5.444	1.264
cyl_8	0.2910	3.143	0.093	0.927	-6.182	6.763
am_1	2.7038	1.599	1.691	0.103	-0.588	5.996
vs_1	1.9900	1.760	1.131	0.269	-1.635	5.615

Omnibus:	0.318	Durbin-Watson:	1.889
Prob(Omnibus):	0.853	Jarque-Bera (JB):	0.157
Skew:	0.164	Prob(JB):	0.925
Kurtosis:	2.899	Cond. No.	1.67e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.67e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

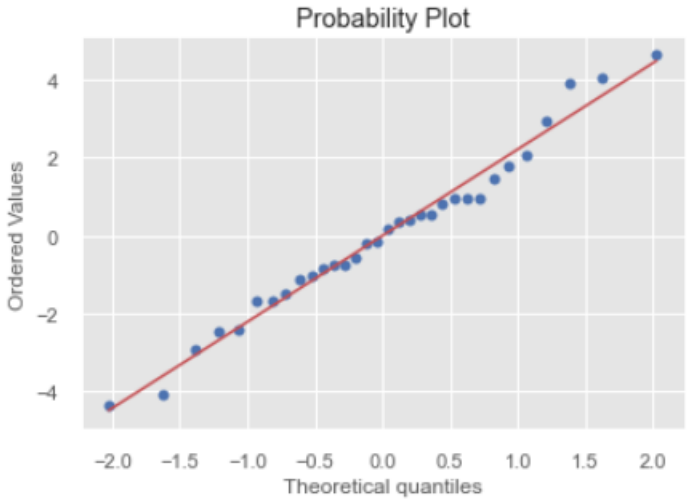
6.1 Residual Analysis

6.1.1 Residual Density Curve



One can easily observe that the distribution is more or less “bell-shaped” and symmetric about 0.

6.1.2 Normal Probability Plot



From the above plot, we can see there is not much departure from normality except at the two extremes. Combining this observation to the one found in 6.1.2, we can say that residuals are almost normally distributed with mean zero.

Now, let's look into the spread of the residuals.

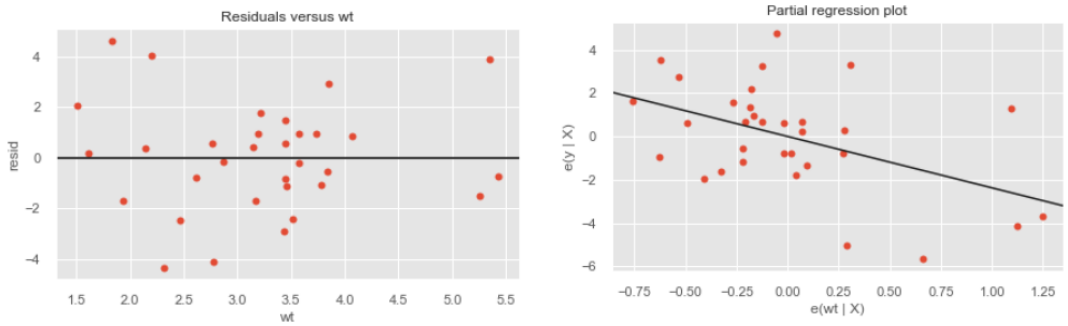
6.1.3 Residual vs Fitted Plot

.

Here, the spread seems to be a bit increasing as Fitted Values are increasing. We will take care of this while analyzing our final model.

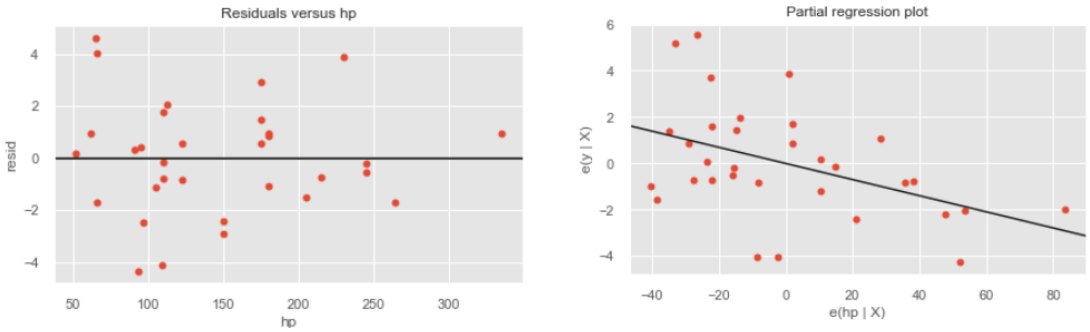
6.2 Added Variable Plots

6.2.1 For "wt"



Clearly, the above plots support the inclusion of "wt" predictor. (except for few "off-the-trend" datapoints)

6.2.2 For “hp”



Observing the above plots, inclusion of “hp” predictor seems reasonable.

6.3 Outlier Detection

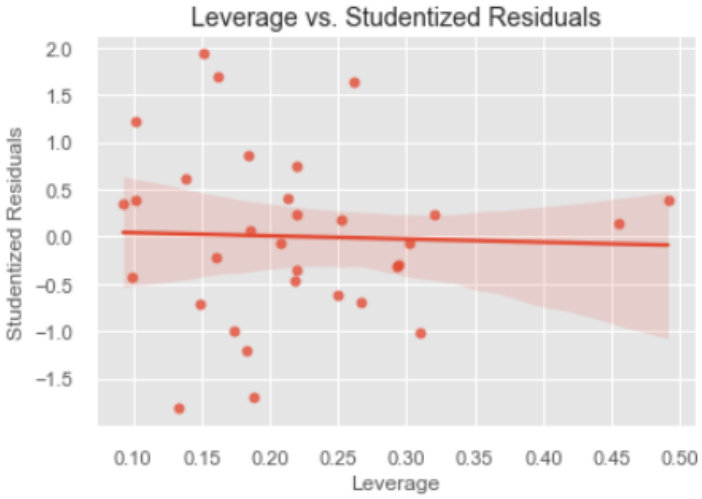
6.3.1 Outliers in X variable

We calculated the leverage values from the hat matrix. The leverage cutoff can be calculated as (2k+2)/n where k is the number of predictors and n is the sample size. So, the leverage cut of 0.4375.

We have identified two outliers, their indices are 23 and 30

6.3.2 Outliers in Y variable

We obtained Studentized residuals to better visualize the outlying observations.

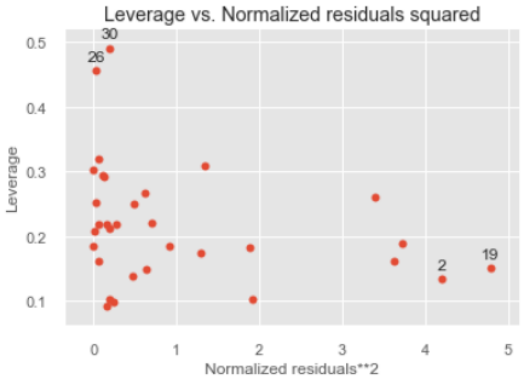


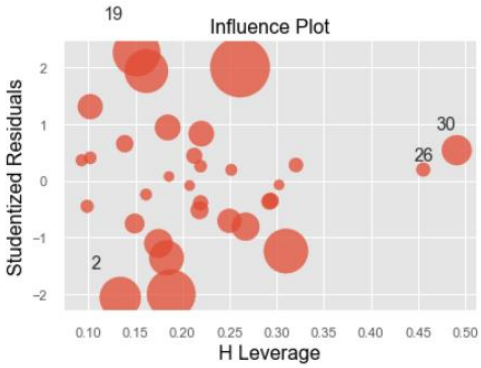
We have used the cutoff of as 2, opposed to 3 in some textbooks. We get four observations to be the outlier. Their indices are 2, 16, 19 and 31.

6.4 Identifying Influential Observations

6.4.1 Identification using Plots

Considering observations with large residuals and high leverage as influential. We have plotted the following plots to visualize the influential observations.

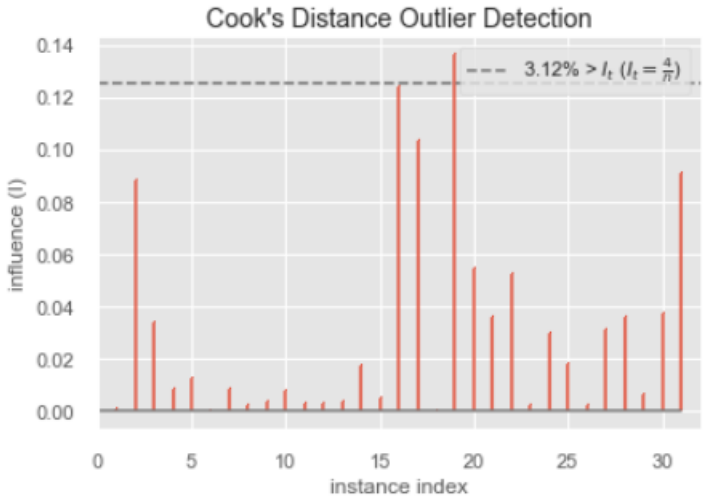




Here, observations having indices 2, 19, 26 and 30 seems to be influential.

6.4.2 Cook's Distance

We have used the thumb rule that Cook's Distance of more than 3 times the mean is a possible outlier. Based on the rule we have decided, observations with indices 16,19 and 31 seems to be influential.
The following plot helps to better visualize the scenario:



6.5 Model Summary after dropping the influential observations

Following is the model summary obtained for the model fitted on the data after removing the influential observations:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.886
Model:                            OLS   Adj. R-squared:                  0.853
Method:                 Least Squares   F-statistic:                     27.21
Date:                Tue, 25 Apr 2023   Prob (F-statistic):           7.20e-09
Time:                        15:47:09   Log-Likelihood:                -60.167
No. Observations:                  28   AIC:                             134.3
Df Residuals:                      21   BIC:                             143.7
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         30.6116      3.590      8.528      0.000      23.146      38.077
wt            -2.2111      0.943     -2.344      0.029      -4.173      -0.250
hp            -0.0369      0.015     -2.510      0.020      -0.068      -0.006
cyl_6         -2.3615      1.700     -1.389      0.179      -5.897       1.174
cyl_8          0.6614      3.195      0.207      0.838      -5.982       7.305
am_1           3.3233      1.730      1.921      0.068      -0.275       6.922
vs_1           2.4135      1.834      1.316      0.202      -1.400       6.227
==============================================================================
Omnibus:                        0.009   Durbin-Watson:                   1.884
Prob(Omnibus):                  0.995   Jarque-Bera (JB):                0.102
Skew:                           0.007   Prob(JB):                        0.950
Kurtosis:                       2.705   Cond. No.                     1.62e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.62e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Comparing this summary with the one at the starting of the Chapter 6, i.e., with the model fitted on the whole data (without removal of the influential observations)

One can see substantial improvement in the model but we are scarce on the data and the number datapoints to be removed accounts for about 9.4% of the data. So, we won't drop any of the observations.

6.6 Detection of Multicollinearity

6.6.1 Analysis after removal of severe collinear variables

We will be using VIF, i.e., Variance Inflation Factor for the detection of possible multicollinearity in the model.

We obtain the following values of the VIF:

	Var	Vif
1	hp	27.35
0	wt	22.62
3	cyl_8	21.76
5	vs_1	4.33
2	cyl_6	3.02
4	am_1	2.61

Clearly here VIF of "hp", "wt" and "cyl_8" exceeds 10.

Now, we will analyze the VIF after removal of variates having high VIF. After analyzing several combinations of removed variates, we get the following results that reduces the VIF the most:

After removal of wt and cyl_8:

	Var	Vif
2	am_1	1.61
3	vs_1	1.51
0	hp	1.43
1	cyl_6	1.27

After removal of hp and cyl_8:

	Var	Vif
3	vs_1	1.64
0	wt	1.57
2	am_1	1.47
1	cyl_6	1.31

After removal of severe collinear variates, we now get satisfactory values of the VIF.

6.6.2 Analysis of the models obtained

MODEL 1:

The following is the model summary for the model having wt and cyl_8 removed:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.836
Model:	OLS	Adj. R-squared:	0.812
Method:	Least Squares	F-statistic:	34.38
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	3.12e-10
Time:	15:47:09	Log-Likelihood:	-73.464
No. Observations:	32	AIC:	156.9
Df Residuals:	27	BIC:	164.3
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	24.3386	2.137	11.387	0.000	19.953	28.724
am_1	5.2475	0.971	5.403	0.000	3.255	7.240
vs_1	2.6780	1.350	1.984	0.058	-0.092	5.448
hp	-0.0477	0.010	-4.688	0.000	-0.069	-0.027
cyl_6	-2.5436	1.140	-2.231	0.034	-4.883	-0.204

Omnibus:	1.507	Durbin-Watson:	1.815
Prob(Omnibus):	0.471	Jarque-Bera (JB):	1.004
Skew:	-0.434	Prob(JB):	0.605
Kurtosis:	2.974	Cond. No.	849.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

MODEL 2:

The following is the model summary for the model having hp and cyl_8 removed:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.820
Model:	OLS	Adj. R-squared:	0.794
Method:	Least Squares	F-statistic:	30.85
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	1.03e-09
Time:	15:47:09	Log-Likelihood:	-74.897
No. Observations:	32	AIC:	159.8
Df Residuals:	27	BIC:	167.1
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	30.1350	3.690	8.167	0.000	22.564	37.706
am_1	1.5549	1.464	1.062	0.298	-1.449	4.558
vs_1	3.8550	1.268	3.040	0.005	1.253	6.457
wt	-3.7319	0.885	-4.217	0.000	-5.548	-1.916
cyl_6	-1.6286	1.183	-1.376	0.180	-4.056	0.799

Omnibus:	0.616	Durbin-Watson:	1.776
Prob(Omnibus):	0.735	Jarque-Bera (JB):	0.720
Skew:	0.246	Prob(JB):	0.698
Kurtosis:	2.455	Cond. No.	29.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Chapter-7

Experimenting and Analyzing Different Models

7.1 Main Effect Models

MODEL 1: (y ~ hp + am_1 + vs_1 + cyl_6 + cyl_8)

The following is the model summary of the model:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.836
Model:	OLS	Adj. R-squared:	0.804
Method:	Least Squares	F-statistic:	26.49
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	1.97e-09
Time:	19:25:17	Log-Likelihood:	-73.460
No. Observations:	32	AIC:	158.9
Df Residuals:	26	BIC:	167.7
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	24.4472	2.571	9.507	0.000	19.162	29.733
am_1	5.1629	1.454	3.551	0.001	2.174	8.151
vs_1	2.5690	1.942	1.323	0.197	-1.424	6.562
hp	-0.0469	0.015	-3.230	0.003	-0.077	-0.017
cyl_6	-2.6525	1.796	-1.477	0.152	-6.344	1.039
cyl_8	-0.2771	3.487	-0.079	0.937	-7.444	6.890

Omnibus:	1.456	Durbin-Watson:	1.814
Prob(Omnibus):	0.483	Jarque-Bera (JB):	0.943
Skew:	-0.420	Prob(JB):	0.624
Kurtosis:	2.992	Cond. No.	1.56e+03

Here, the model seems to be good fit but it has severe multicollinearity.

MODEL 2: (y ~ wt + am_1 + vs_1 + cyl_6 + cyl_8)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.809
Method:	Least Squares	F-statistic:	27.33
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	1.42e-09
Time:	19:25:22	Log-Likelihood:	-73.044
No. Observations:	32	AIC:	158.1
Df Residuals:	26	BIC:	166.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	32.1811	3.729	8.631	0.000	24.517	39.845
am_1	0.6232	1.501	0.415	0.681	-2.463	3.709
vs_1	1.2418	1.904	0.652	0.520	-2.672	5.156
wt	-3.1061	0.920	-3.375	0.002	-4.998	-1.214
cyl_6	-3.7329	1.638	-2.280	0.031	-7.099	-0.367
cyl_8	-4.7483	2.657	-1.787	0.086	-10.210	0.714

Omnibus:	1.627	Durbin-Watson:	1.824
Prob(Omnibus):	0.443	Jarque-Bera (JB):	1.018
Skew:	0.436	Prob(JB):	0.601
Kurtosis:	3.047	Cond. No.	33.9

Here, the model shows a bit improvement over the previous one with less severe.

7.2 Predictor Transformed Model

The following are the models fitted with transformed predictors using the idea from Chapter 5: Univariate Linear Regression.

MODEL 3: (y ~ Inv_hp + am_1 + vs_1 + cyl_6 + cyl_8)
The following is the model summary of the model:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.838
Model:	OLS	Adj. R-squared:	0.807
Method:	Least Squares	F-statistic:	26.95
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	1.64e-09
Time:	19:25:29	Log-Likelihood:	-73.231
No. Observations:	32	AIC:	158.5
Df Residuals:	26	BIC:	167.3
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	13.9407	3.667	3.802	0.001	6.404	21.477
am_1	3.2514	1.307	2.488	0.020	0.566	5.937
vs_1	0.8733	1.927	0.453	0.654	-3.087	4.834
Inv_hp	742.7431	224.395	3.310	0.003	281.492	1203.994
cyl_6	-2.3263	1.824	-1.275	0.214	-6.076	1.424
cyl_8	-3.0317	2.917	-1.039	0.308	-9.028	2.965

Omnibus:	1.416	Durbin-Watson:	2.111
Prob(Omnibus):	0.493	Jarque-Bera (JB):	1.106
Skew:	0.446	Prob(JB):	0.575
Kurtosis:	2.817	Cond. No.	610.

Here, the model shows slight improvement over Model 1.

MODEL 4: (y ~ Inv_wt + am_1 + vs_1 + cyl_6 + cyl_8)
The following is the model summary of the model:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.861
Model:	OLS	Adj. R-squared:	0.835
Method:	Least Squares	F-statistic:	32.27
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	2.33e-10
Time:	19:26:00	Log-Likelihood:	-70.781
No. Observations:	32	AIC:	153.6
Df Residuals:	26	BIC:	162.4
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.0361	3.489	3.450	0.002	4.864	19.208
am_1	-0.2970	1.472	-0.202	0.842	-3.322	2.728
vs_1	0.5213	1.793	0.291	0.774	-3.164	4.206
Inv_wt	31.0476	7.514	4.132	0.000	15.602	46.494
cyl_6	-2.5409	1.611	-1.577	0.127	-5.852	0.770
cyl_8	-4.8826	2.423	-2.015	0.054	-9.863	0.098

Omnibus:	5.189	Durbin-Watson:	2.288
Prob(Omnibus):	0.075	Jarque-Bera (JB):	3.888
Skew:	0.834	Prob(JB):	0.143
Kurtosis:	3.367	Cond. No.	24.8

Here, the model performance is improved. So, this model is preferred over Model 2.

MODEL 5: (y ~ log_wt + am_1 + vs_1 + cyl_6 + cyl_8)
The following is the model summary of the model:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.839
Model:                            OLS   Adj. R-squared:                  0.808
Method:                 Least Squares   F-statistic:                     27.03
Date:                Tue, 25 Apr 2023   Prob (F-statistic):           1.59e-09
Time:                        19:26:13   Log-Likelihood:                -73.191
No. Observations:                  32   AIC:                             158.4
Df Residuals:                      26   BIC:                             167.2
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     17.1486      3.025      5.669      0.000      10.931      23.366
am_1           0.7000      1.501      0.466      0.645      -2.385       3.785
vs_1           0.3584      1.951      0.184      0.856      -3.651       4.368
Log_wt         6.2592      1.883      3.324      0.003       2.388      10.130
cyl_6         -3.4969      1.670     -2.094      0.046      -6.929      -0.065
cyl_8         -6.7777      2.504     -2.707      0.012     -11.924      -1.632
==============================================================================
Omnibus:                        4.525   Durbin-Watson:                   2.139
Prob(Omnibus):                  0.104   Jarque-Bera (JB):                3.187
Skew:                           0.745   Prob(JB):                        0.203
Kurtosis:                       3.412   Cond. No.                         15.3
==============================================================================
```

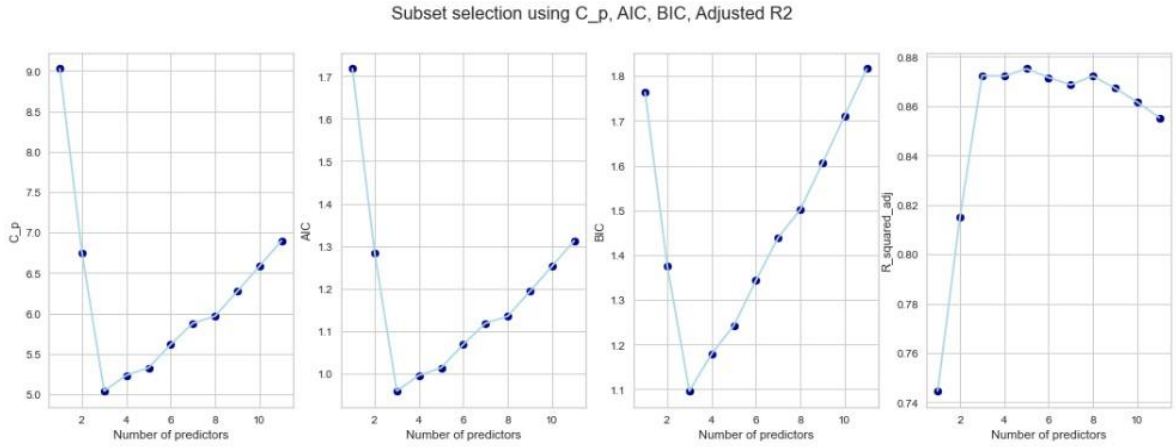
Here, the model performance dips a bit. So, Model 2 is preferred.

7.3 First-Order Interaction Model

Here, considering the base model as "y ~ wt + hp + cyl_6 + cyl_8 + am_1 + vs_1", using forward selection procedure to analyze all the possible first-order interactions.

	features	RSS	R_squared	numb_features	C_p	AIC	BIC	R_squared_adj
1	[wt]	278.321938	0.752833	1	9.026028	1.717453	1.763257	0.744594
2	[wt, hp]	195.047755	0.826785	2	6.752177	1.284789	1.376398	0.814840
3	[wt, hp, wt*hp]	129.761498	0.884764	3	5.040448	0.959085	1.096498	0.872417
4	[wt, hp, wt*hp, vs_1]	125.437079	0.888604	4	5.233777	0.995872	1.179089	0.872101
5	[wt, hp, wt*hp, vs_1, hp*vs_1]	117.888558	0.895308	5	5.326353	1.013487	1.242508	0.875174
6	[wt, hp, wt*hp, vs_1, hp*vs_1, hp*am_1]	116.621933	0.896432	6	5.615238	1.068455	1.343281	0.871576
7	[wt, hp, wt*hp, vs_1, hp*vs_1, hp*am_1, wt*am_1]	114.494922	0.898321	7	5.877236	1.118307	1.438937	0.868665
8	[wt, hp, wt*hp, vs_1, hp*vs_1, hp*am_1, wt*am_...	106.792431	0.905162	8	5.965001	1.135007	1.501441	0.872174
9	[wt, hp, wt*hp, vs_1, hp*vs_1, hp*am_1, wt*am_...	106.049632	0.905821	9	6.270255	1.193090	1.605328	0.867294
10	[wt, hp, wt*hp, vs_1, hp*vs_1, hp*am_1, wt*am_...	105.530095	0.906283	10	6.582487	1.252501	1.710543	0.861655
11	[wt, hp, wt*hp, vs_1, hp*vs_1, hp*am_1, wt*am_...	105.109489	0.906656	11	6.897810	1.312500	1.816347	0.855317

The following plots help to better visualize the best model according to different evaluation criterion:



Observing the table and plots, we finalize the following model:

<u>MODEL 6</u>: (y ~ wt + hp + vs_1 + wt*hp + hp* vs_1)
The following is the model summary of the model:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.895
Model:	OLS	Adj. R-squared:	0.875
Method:	Least Squares	F-statistic:	44.47
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	6.30e-12
Time:	19:41:47	Log-Likelihood:	-66.270
No. Observations:	32	AIC:	144.5
Df Residuals:	26	BIC:	153.3
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	43.7551	5.211	8.397	0.000	33.044	54.466
vs_1	5.4387	3.551	1.531	0.138	-1.861	12.738
wt	-6.8475	1.578	-4.340	0.000	-10.091	-3.604
hp	-0.0886	0.032	-2.779	0.010	-0.154	-0.023
wt:hp	0.0207	0.009	2.293	0.030	0.002	0.039
hp:vs_1	-0.0413	0.032	-1.290	0.208	-0.107	0.025

Omnibus:	1.910	Durbin-Watson:	2.391
Prob(Omnibus):	0.385	Jarque-Bera (JB):	1.265
Skew:	0.209	Prob(JB):	0.531
Kurtosis:	2.121	Cond. No.	1.03e+04

This interaction model sees substantial increment in R and Adjusted R squared value over other main effect and transformed model. Though here the multicollinearity may be problematic.

VIFs are as follows:

	Var	Vif
2	wt*hp	36.60
1	hp	25.91
0	wt	22.22
4	vs_1	18.72
3	hp*vs_1	16.30

Here the multicollinearity is severe, we will employ ridge regression now to alleviate the multicollinearity present here.

MODEL 7: (y ~ wt + hp + vs_1 + wt*hp + hp* vs_) – Using Ridge Regression

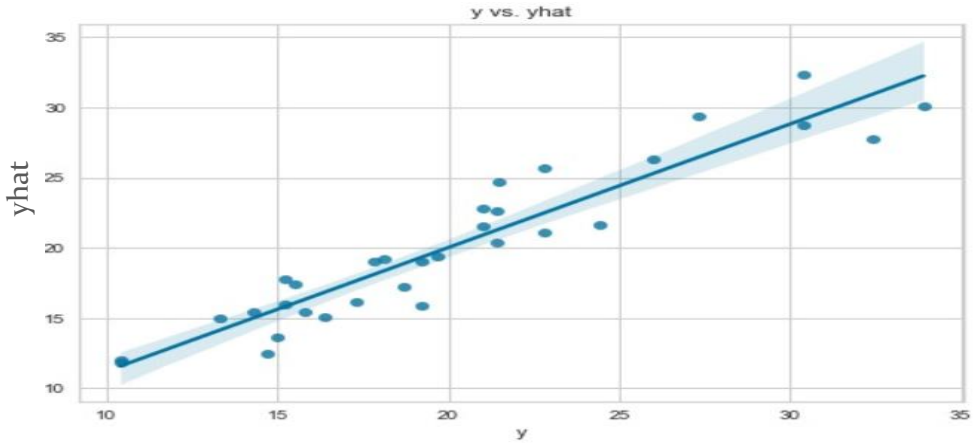
Following are the coefficients estimates for different values of alpha:

	wt	hp	vs_1	wt*hp	hp*vs_1	
coefficent estimates	[-6.59446571	-5.97719308	6.95082679	-1.98059015	5.43868115]	alpha 0.00
coefficent estimates	[-6.50543178	-5.852012	6.75634108	-1.99370904	5.46428942]	alpha 0.01
coefficent estimates	[-6.42244587	-5.73566359	6.57454895	-2.00267907	5.48022497]	alpha 0.02
coefficent estimates	[-6.34475877	-5.62704412	6.40391417	-2.00818015	5.48809327]	alpha 0.03
coefficent estimates	[-6.27174949	-5.52523921	6.24316145	-2.01076692	5.48920307]	alpha 0.04
coefficent estimates	[-6.20289833	-5.42948415	6.09122181	-2.01089554	5.48463006]	alpha 0.05
coefficent estimates	[-6.13776659	-5.33913372	5.94719123	-2.00894397	5.47526484]	alpha 0.06
coefficent estimates	[-6.07598091	-5.2536392	5.81029896	-2.0052275	5.46184968]	alpha 0.07
coefficent estimates	[-6.01722119	-5.17253045	5.67988294	-2.0000107	5.44500691]	alpha 0.08
coefficent estimates	[-5.96121112	-5.0954019	5.55537052	-1.99351682	5.42526114]	alpha 0.09
coefficent estimates	[-5.90771064	-5.0219015	5.43626317	-1.98593519	5.40305673]	alpha 0.10
coefficent estimates	[-5.85650998	-4.95172181	5.32212436	-1.97742706	5.37877178]	alpha 0.11
coefficent estimates	[-5.80742484	-4.88459292	5.21256968	-1.96813037	5.35272926]	alpha 0.12
coefficent estimates	[-5.76029242	-4.82027662	5.10725885	-1.95816352	5.32520603]	alpha 0.13
coefficent estimates	[-5.71496827	-4.75856171	5.00588927	-1.94762852	5.29644026]	alpha 0.14
coefficent estimates	[-5.67132362	-4.69926007	4.90819054	-1.93661349	5.26663737]	alpha 0.15
coefficent estimates	[-5.62924322	-4.64220345	4.81392009	-1.92519478	5.23597502]	alpha 0.16
coefficent estimates	[-5.58862346	-4.58724077	4.72285941	-1.91343871	5.20460724]	alpha 0.17
coefficent estimates	[-5.54937092	-4.53423588	4.63481093	-1.901403	5.17266782]	alpha 0.18
coefficent estimates	[-5.51140102	-4.48306564	4.5495954	-1.889138	5.14027319]	alpha 0.19
coefficent estimates	[-5.47463696	-4.43361831	4.46704965	-1.8766877	5.10752483]	alpha 0.20
coefficent estimates	[-5.43900876	-4.38579219	4.38702468	-1.86409057	5.07451129]	alpha 0.21
coefficent estimates	[-5.40445249	-4.33949445	4.30938406	-1.85138034	5.04130993]	alpha 0.22
coefficent estimates	[-5.3709096	-4.29464009	4.23400247	-1.83858656	5.00798838]	alpha 0.23
coefficent estimates	[-5.33832626	-4.25115113	4.16076455	-1.8257352	4.9746058]	alpha 0.24

Following table displays VIFs corresponding to coefficient estimates:

wt	hp	vs_1	wt*hp	hp*vs_1	
[16.29698119,	32.65191623,	64.87518419,	16.62942979,	21.90395837]	alpha 0.00
[4.63045863,	7.78991568,	14.22348385,	7.28805042,	9.3977994]	alpha 0.01
[2.86625491,	4.13117449,	6.62724111,	4.91637917,	6.21375125]	alpha 0.02
[2.23447766,	2.88278132,	4.01915474,	3.7099543 ,	4.59578498]	alpha 0.03
[1.9099485 ,	2.2806591 ,	2.78095032,	2.96131737,	3.5948456]	alpha 0.04
[1.70595536,	1.92743577,	2.08159533,	2.45141014,	2.91603213]	alpha 0.05
[1.56068882,	1.69245504,	1.64162697,	2.0837511 ,	2.42916549]	alpha 0.06
[1.44854289,	1.52209492,	1.34378768,	1.80778568,	2.06594627]	alpha 0.07
[1.35717361,	1.39082097,	1.13116146,	1.59417811,	1.78671595]	alpha 0.08
[1.27993081,	1.2850869 ,	0.97311626,	1.42470413,	1.56683207]	alpha 0.09
[1.21290995,	1.19708569,	0.85184701,	1.28746391,	1.39020557]	alpha 0.10
[1.15365456,	1.12201088,	0.75636933,	1.17438086,	1.24592042]	alpha 0.11
[1.10052874,	1.0567387 ,	0.67957743,	1.07979893,	1.12633652]	alpha 0.12
[1.05239001,	0.99914481,	0.61669313,	0.99965244,	1.02596627]	alpha 0.13
[1.00840782,	0.94772784,	0.56439832,	0.93095314,	0.9407802]	alpha 0.14
[0.96795747,	0.90139102,	0.52032514,	0.87146197,	0.86776251]	alpha 0.15
[0.93055531,	0.85930991,	0.482744 ,	0.81947254,	0.80461796]	alpha 0.16
[0.89581759,	0.82084922,	0.45036565,	0.77366479,	0.74957362]	alpha 0.17
[0.86343332,	0.78550869,	0.42221183,	0.73300369,	0.7012417]	alpha 0.18
[0.83314598,	0.75288689,	0.39752855,	0.69666779,	0.65852287]	alpha 0.19
[0.80474068,	0.72265629,	0.37572646,	0.66399798,	0.62053682]	alpha 0.20
[0.77803507,	0.6945458 ,	0.35633905,	0.63446003,	0.58657166]	alpha 0.21
[0.75287263,	0.66832814,	0.33899289,	0.60761695,	0.55604646]	alpha 0.22
[0.72911774,	0.64381061,	0.32338587,	0.58310825,	0.52848323]	alpha 0.23
[0.70665183,	0.62082823,	0.30927136,	0.56063421,	0.50348564]	alpha 0.24

Analyzing the above two tables, we choose alpha to be 0.09. Following the model summary:




```
Ridge Regression Model
----------------------
Coefficients: [-5.96121112 -5.0954019   5.55537052 -1.99351682  5.42526114]
Intercept: 17.71707325160746
R-squared: 0.8940234281076307
Adj_R-squared: 0.873643318128329
```

Chapter-8

Model Validation and Final Model Selection

Based on the observations in Chapter-7, we select 3 models as contenders for final model, i.e., Model 3, Model 4 and Model 7.

Evaluating the models on PRESS criterion. (Since data here is too small to split)
Following are the PRESS values of the stated models:

Models	PRESS value
Model 3	4543.973190735056
Model 4	3826.1793061772732
Model 7	1435.3627193971438

Here, **MODEL 7**: (y ~ wt + hp + vs_1 + wt*hp + hp* vs_1) – Using Ridge Regression is the best amongst all.

So, this is our final model.

Conclusion:

1) Performed Exploratory data analysis on the Car dataset consisting of 32 rows and 12 columns
2) Detected multicollinearity using Variance Inflation factor and addressed it using Ridge Regression.
3) Detection and Removal of influential points using proper metrics such as Cooks Distance, DFFITS.
4) Implemented Stepwise Feature Selection technique to extract significant set of predictor variables.
5) Determined the optimal model based on criteria Adjusted R^2, AIC, BIC and Mallows C.
6) Fitted Multiple Regression Model and validated the regression assumptions after doing EDA.
7) Achieved Adjusted R^2 value of 0.8736 and PRESS value 1435.36 in the final model.

Future Scope:

Additional info about data can help us more in the analysis. We can use further techniques to address issues of multicollinearity and that can improve model predictive performance.

----------END----------