Regression Analysis of *Auto-MPG* Dataset

Author: Abhishek Deshpande[1]

Affiliations:

[1]School of Computing, Informatics, and decision systems engineering, Arizona State University

Abstract: This report contains data analysis and interpretation of the Auto-mpg dataset. We find the relation between the output miles per gallons and the other parameters that are affecting the output mileage of different car models of different year. Data Analysis is done using SPSS 25 software. Interpretations of the results are summarized in conclusions.

Introduction:

Since ages, improving the mileage of the cars has been the main aim of the manufacturers. There had been studies on what different parameters that affect the mileage of the car they manufacture and how to reduce the effects from the parameters that reduce them. For instance, engine size, car weight, etc. can have significant effect on the performance of the car as the friction power of the car increases and hence the efficiency of the engine performance reduces. The researchers have also found that model year can also affect the mileage of the car. Any small variation in the engine parameters lead to the change in the output mileage which signifies there is a relation between mpg and the affecting parameters. Hence, we need to perform a linear regression to check which parameters have significant effect on the dependent variable mpg.

Dataset Description:

Source: UCI, This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.

Following table gives the ranges of the data and the type of variables that are used in the dataset.

Variables	Range	Type	Subscript
mpg	9-46.6	Continuous	mpg
cylinders	3,4,5,6,8	Multi-valued discrete	cylinders
displacement	68-455	Continuous	displacement
horsepower	46-230	Continuous	horsepower
weight	1613-5140	Continuous	weight
acceleration	8-24.8	Continuous	acceleration
model year	70-82	Multi-valued discrete	modelyear
origin	1-3	Categorical	origin
car name	Multiple values	Nominal (ID variable)	carname

Total Number of readings are 398 and total number of variables are 9 (one is ID variable).
The variable horsepower had 6 missing values which were replaced by the series mean into the new variable horsepower_1.
There were no abnormal cases found on visual inspection of the dataset. Maximum and minimum values of the variables have chances of being potential outlier.

Explanation of variables

Mpg: Mileage (Miles per gallons).
Cylinders: Number of cylinders in the engine.
Displacement: Piston displacement inside the cylinder.
Horsepower: Brake Power of engine described in horsepower.
Weight: Weight of the car.
Acceleration: Average acceleration of the car
Model Year: The year in which the model was produced.
Origin: The type of origin when the car was manufactured.
Car Name: Name of the car.

Data Cleaning

The data obtained was in the form of csv file separated by width. Needed to import the data into the SPSS software and the provide variable names as described.

Data Preprocessing

There are instances of missing values in the horsepower variable. I chose to replace the missing values by series mean and store the new values in the horsepower_1 variable. (Ref Figure 1.1)

The variable origin can be taken as categorical variable where it can be transformed into 3 dummy variables origin_1, origin_2 and origin_3 using ‘recode into different variables’ in SPSS. (Ref Fig.1.2)

Procedure

Model 1:
The method of entering the variables is using Stepwise method as other methods have less R adjusted than this method along with F_0 value.
F(in)=0.05 is the probability of entering a variable and F(out)=0.1 to drop a variable already considered in a model.

Variables entered are cylinders, displacement, horsepower, weight, acceleration, modelyear, origin_1 and origin_2. Origin_3 is taken as a reference. Origin_1 has given value 1 when origin takes value of 1, origin_2 takes value of 1 when origin has value of 2 and origin_3 takes value of 1 when origin takes value of 3, 0 otherwise in all the cases.
Dependent variable is mpg.
Descriptive Statistics are shown in figure 1.3 as shown.
Correlation coefficients observed are given in figure 1.4.
Displacement, cylinders, weight and horsepower are negatively correlated with mpg and have values greater than 0.7 which suggest significant relation between those.

Variable to variable correlation that were greater than 0.7 were observed are given in pairs as follows,
Displacement-Cylinder: 0.951, displacement-weight: 0.933, displacement-hp1: 0.894, cylinder-weight: 0.896, cylinder-horsepower: 0.839
These show a hint for these variables have multicollinearity.

The R squared value observed here is 0.716 and adjusted R squared value obtained is 0.713, which is a quiet good score and initial guess is that model is quiet adequate. There is need to check assumptions to check if model is adequate by testing assumptions for this model.

Fitted Model:

Mpg= 45.575 -0.05 weight -0.049 horsepower_1 -2.699 origin_1 -1.471 origin_2 (Ref Fig.1.5)

Displacement, acceleration and cylinders variables are excluded in this model.
Weight and horsepower have correlation coefficient -0.816 which is greater than 0.7 signifying possible multicollinearity.

P-P plot obtained is almost linear but is slightly heavy in between 0.5 and 0.8 cumulative probability. (Fig 1.9)

Scatter plot between Studentized Residual value and Standardized Residual value suggests that scatter form particular funnel in the right side.(Fig 1.10)

Histogram obtained looks quite normal. We proceed to assumption testing. (Fig 1.8)

Assumption Testing

Assumption 1: Normality
i) The P-P plot is almost linear but is slightly bulging out in between 0.5 and 0.8
ii) Kurtosis and Skewness values are observed below 2 in all cases.
Normality assumption is not violated.

Assumption 2: Linearity
i) P-P plot although is close to linear is not completely linear. This is an issue to address upon.
ii) Mean value of residuals is 0 for each independent variable w.r.t dependent variable which suggests no-nonlinear relation.
Linearity assumption is not violated.

Assumption 3: Homoscedasticity
i) Scatter plot between Studentized Residual value and Standardized Residual value suggests that scatter form particular funnel in the right side. This denotes heteroscedasticity.
Homoscedasticity assumption is violated.

Assumption 4: Independence (Autocorrelation)
i) Durbin-Watson Test: d statistic observed is 0.883 (fig.1.5)
The lower and upper limits of d statistic are dl=1.8183 and du=1.8493
Here, d < dl, there is probably an evidence of positive autocorrelation.
Independence assumption is violated.

ii) Consider the time sequence no firm statement can be said about auto-correlation. But, overall pattern can be said to be in the form of positive autocorrelation.
Independence assumption is violated.

Assumption 5: Multicollinearity
i) Correlation between weight and horsepower_1 is observed 0.816>0.7 indicating possible multicollinearity.
ii) VIF scores for weight and horsepower_1 are approximately near 4 which indicates possible multicollinearity. (Fig 1.12)
iii) Tolerance scores are approximately near 0.2 which again indicates possible multicollinearity. (Fig.1.12)
Multicollinearity is present.

Assumption 6: Influential Cases
i) Cook's distance D was found using formula 4/(n-k-1) obtained to be 0.0105, so flagged cases were, 72,109,112,117,155,156,325,326,327,328,329,330,331,334,335,336,345,355,361,388,395.
Out of these large influence is made by, 326,327,328,330,331,395.
ii) Residual plots also suggest, presence of outliers outside +/- 2 so influential cases are present.
Cases are present those influence model.

Comments
- Here we can say that model is not adequate as some of the assumptions are violated.
- The model observed can be said to be linear and normal but fails the homoscedasticity assumption.
- Significant variables are weight, horsepower_1, origin_1, origin_2 which are suggested by t-values and standardized beta values.(Fig.1.12)
- Both origin_1 and origin_2 have lower preference than origin_3 as can be observed from the t-value in fig1.12, for both of the variables, negative values are obtained.
- Influential Cases exists that affect the betas and eventually the R squared adjusted value.
- Multicollinearity exists in the model as horsepower_1 and weight variables are correlated.
- Autocorrelation is present because of the dependencies within the data.
- Transformation of data isn't needed in the model as kurtosis and skewness is below 2 for all the variables.

Remedies
- Check the model after removing the influential cases, as those cases maybe special cases taken in error or under special conditions.
- To deal with multicollinearity, try removing one of the variable (horsepower_1 or weight) that has VIF around 4 and correlation 0.8 as one of it is a redundant variable.

Model 2

Deleting all the influential cases from the model, instead of putting both horsepower_1 and weight, this time only weight along with rest other variables is introduced.

Assumptions:
Most of the assumptions except for the multicollinearity and autocorrelation were violated, indicating that the variables weight, displacement, horsepower are correlated with each other.
But there had been improvements in the scatter plots as there had been reduction in the influential cases.

Output

- R squared adjusted that is obtained is 0.771 and R squared is 0.774 which is much better than the earlier model.
- Also, the P-P plot obtained from this model is much better as compared to the earlier model. (Ref fig.1.17)
- The model obtained is as follows:
  Mpg= 43.737 -0.05 weight -0.17 displacement -3.38 origin_2 -2.314 origin_1 (Ref Fig.1.15)

Model	Constant	weight	Horsepower_1	displacement	Origin_1	Origin_2
1	45.575	-0.05	-0.049	0	-2.699	-1.471
2	43.737	-0.05	0	-0.17	-2.314	-3.38

Conclusion
In conclusion, the first model failed to qualify all the assumptions. The second model made improvements in the assumptions. The R squared adjusted value of the second model was better along with a better P-P plot. This provides evidence that the second model is better and that there is also room of improvement in the second model by reduction of multicollinearity by further reduction of variables such as displacement and cylinders. As these make the model over specified but at the same time a check on the R square adjusted value is important for a conclusive result.

References:

1. 'Introduction to Linear Regression' by Douglas.C.Montgomery, Elizabeth.A.Peck, G. Geoffrey Vining
2. https://archive.ics.uci.edu/ml/datasets/auto+mpg. (Data Source)
3. http://web.stanford.edu/~clint/bench/dw05c.htm. (Durbin-Watson Tables)
4. Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning. In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann.

Appendix:

Figure 1.1

Replace Missing Values

```
[DataSet2] E:\578 project\Auto_mpg.sav
```

Result Variables

	Result Variable	N of Replaced Missing Values	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
			First	Last		
1	horsepower_1	6	1	398	398	SMEAN (horsepower)

Figure 1.2

12	origin_1	Numeric	8	2		None	None	8	Right	Nominal	Input
13	origin_2	Numeric	8	2		None	None	8	Right	Nominal	Input
14	origin_3	Numeric	8	2		None	None	8	Right	Nominal	Input
15	origin	Numeric	1	0		None	None	12	Right	Nominal	Input

Figure 1.3

Descriptive Statistics

	Mean	Std. Deviation	N
mpg	23.51457286	7.815984313	398
displacement	193.426	104.2698	398
acceleration	15.56809045	2.757688930	398
cylinders	5.45	1.701	398
origin_1	.6256	.48457	398
origin_2	.1759	.38120	398
weight	2970.42	846.842	398
SMEAN(horsepower)	104.469	38.1992	398

Figure 1.4

Correlations

		mpg	displacement	acceleration	cylinders	origin_1	origin_2	weight	SMEAN (horsepower)
Pearson Correlation	mpg	1.000	-.804	.420	-.775	-.568	.259	-.832	-.771
	displacement	-.804	1.000	-.544	.951	.651	-.374	.933	.894
	acceleration	.420	-.544	1.000	-.505	-.251	.204	-.417	-.684
	cylinders	-.775	.951	-.505	1.000	.604	-.353	.896	.839
	origin_1	-.568	.651	-.251	.604	1.000	-.597	.598	.486
	origin_2	.259	-.374	.204	-.353	-.597	1.000	-.299	-.281
	weight	-.832	.933	-.417	.896	.598	-.299	1.000	.861
	SMEAN(horsepower)	-.771	.894	-.684	.839	.486	-.281	.861	1.000

Figure 1.5

Model Summary[e]

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.832[a]	.692	.691	4.344628058	.692	888.851	1	396	.000	
2	.839[b]	.704	.702	4.265004255	.012	15.924	1	395	.000	
3	.844[c]	.713	.711	4.203547675	.009	12.634	1	394	.000	
4	.846[d]	.716	.713	4.185029068	.003	4.495	1	393	.035	.883

a. Predictors: (Constant), weight

b. Predictors: (Constant), weight, SMEAN(horsepower)

c. Predictors: (Constant), weight, SMEAN(horsepower), origin_1

d. Predictors: (Constant), weight, SMEAN(horsepower), origin_1, origin_2

e. Dependent Variable: mpg

Figure 1.6

4	(Constant)	45.575
	weight	-.005
	SMEAN(horsepower)	-.049
	origin_1	-2.699
	origin_2	-1.471

a. Dependent Variable: mpg

Figure 1.7

4	Correlations	weight	1.000	-.816	-.411	-.139
		SMEAN(horsepower)	-.816	1.000	.117	.106
		origin_1	-.411	.117	1.000	.553
		origin_2	-.139	.106	.553	1.000

Figure 1.8

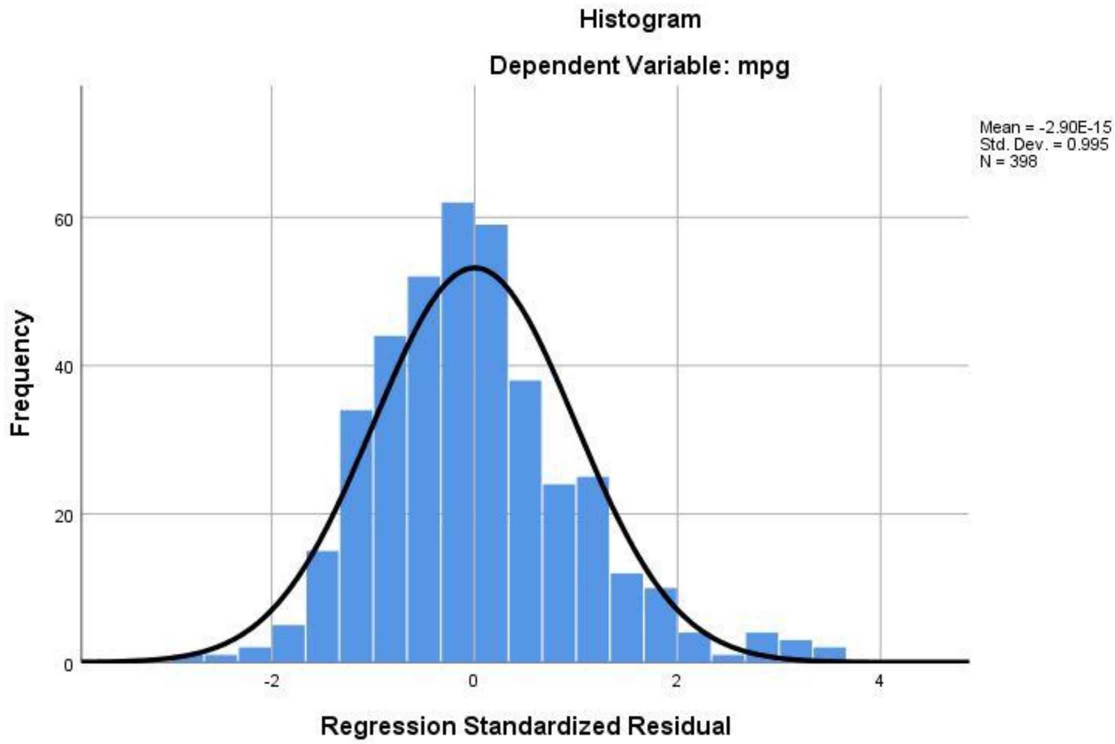


Figure 1.9

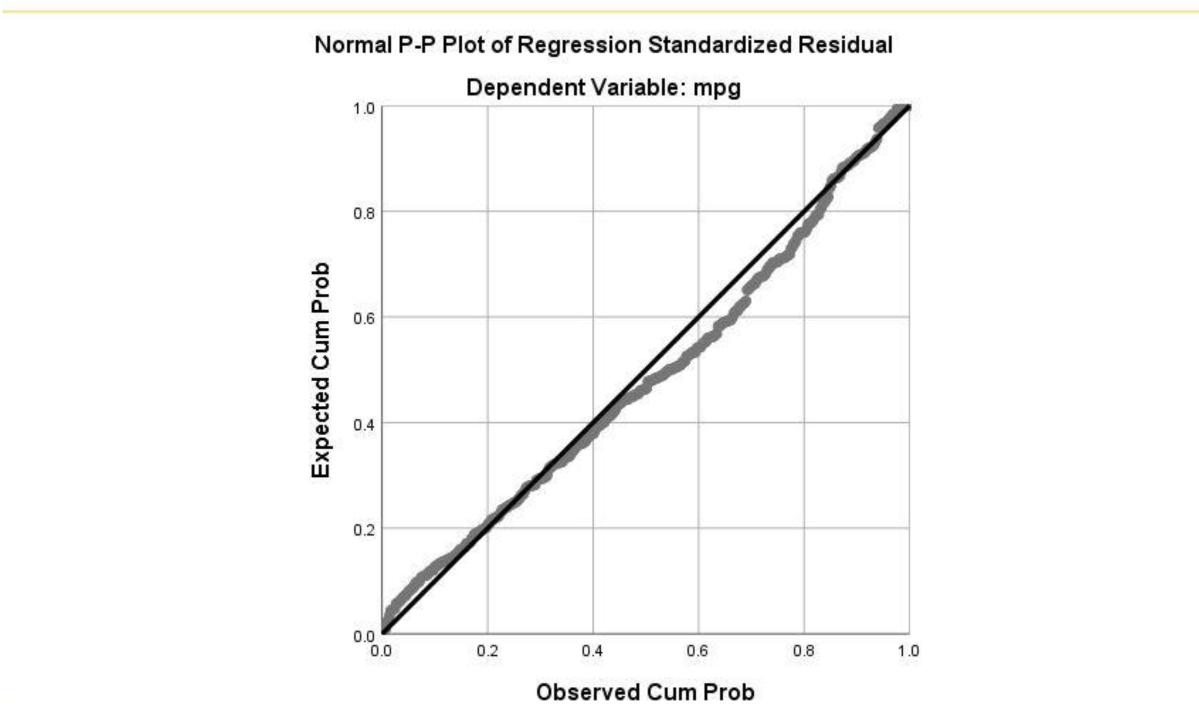


Figure 1.10

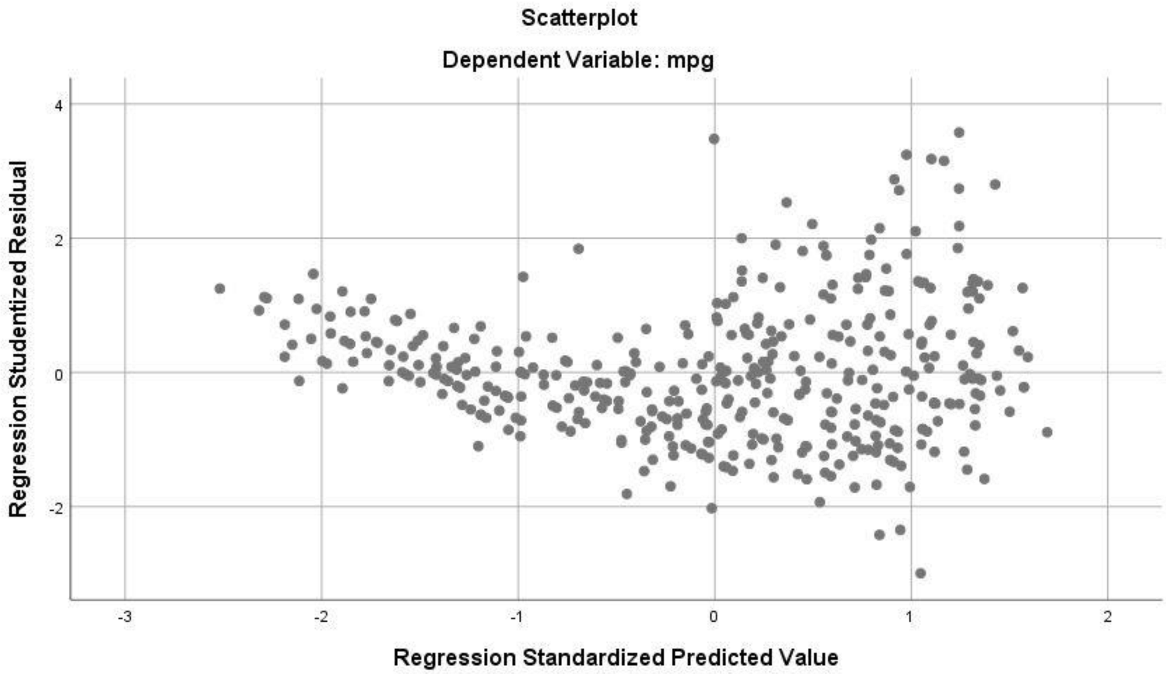


Figure 1.11

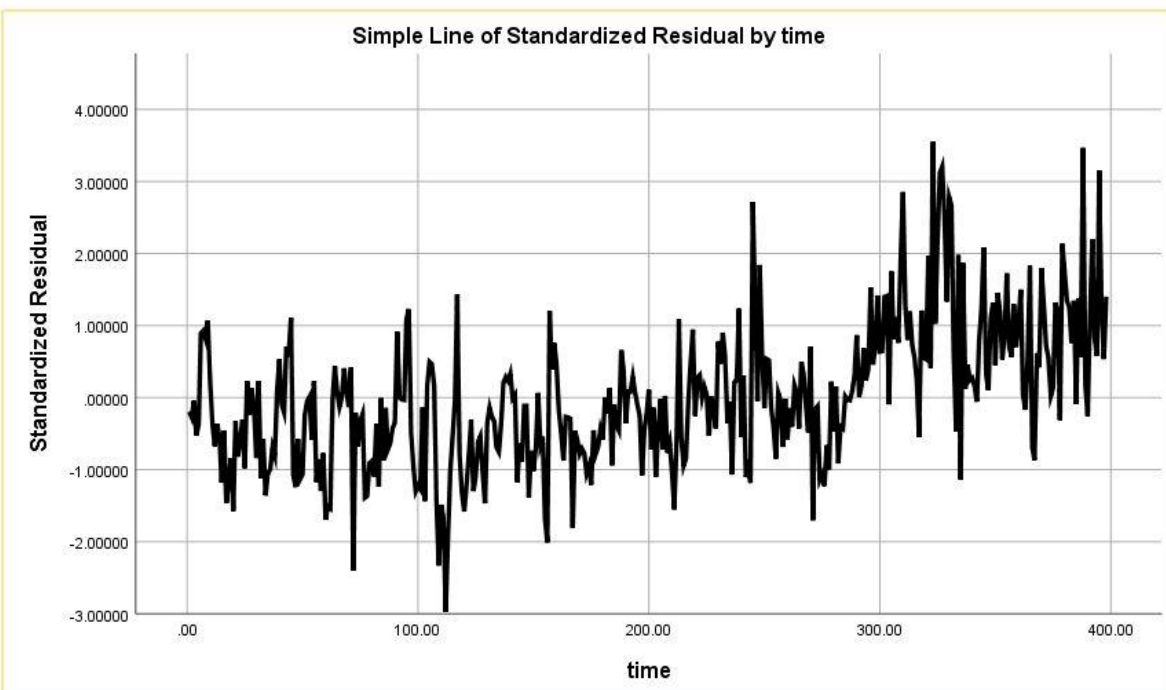


Figure 1.12

Model		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
4	(Constant)	45.575	.842		54.102	.000					
	weight	-.005	.001	-.547	-9.383	.000	-.832	-.428	-.252	.213	4.700
	SMEAN(horsepower)	-.049	.011	-.240	-4.509	.000	-.771	-.222	-.121	.255	3.918
	origin_1	-2.699	.651	-.167	-4.147	.000	-.568	-.205	-.111	.444	2.255
	origin_2	-1.471	.694	-.072	-2.120	.035	.259	-.106	-.057	.631	1.585

a. Dependent Variable: mpg

Figure 1.13 Chunk of data (example)

mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	Car Name	Column
37	4	91	68.00	2025	18.2	82	3	mazda glc custom l	
31	4	91	68.00	1970	17.6	82	3	mazda glc custom	
34	4	108	70.00	2245	16.9	82	3	toyota corolla	
38	4	91	67.00	1995	16.2	82	3	datsun 310 gx	
32	4	91	67.00	1965	15.7	82	3	honda civic (auto)	
38	4	91	67.00	1965	15	82	3	honda civic	
36	4	107	75.00	2205	14.5	82	3	honda accord	
36	4	120	88.00	2160	14.5	82	3	nissan stanza xe	
32	4	144	96.00	2665	13.9	82	3	toyota celica gt	
37	4	85	65.00	1975	19.4	81	3	datsun 210 mpg	
31.6	4	120	74.00	2635	18.3	81	3	mazda 626	
32.3	4	97	67.00	2065	17.8	81	3	subaru	
37.7	4	89	62.00	2050	17.3	81	3	toyota tercel	
39.1	4	79	58.00	1755	16.9	81	3	toyota starlet	
32.4	4	108	75.00	2350	16.8	81	3	toyota corolla	
35.1	4	81	60.00	1760	16.1	81	3	honda civic 1300	
34.1	4	91	68.00	1985	16	81	3	mazda glc 4	
32.9	4	119	100.0	2615	14.8	81	3	datsun 200sx	
33.7	4	107	75.00	2210	14.4	81	3	honda prelude	
24.2	6	146	120.0	2930	13.8	81	3	datsun 810 maxima	
25.4	6	168	116.0	2900	12.6	81	3	toyota cressida	

Figure 1.14

Descriptive Statistics

	Mean	Std. Deviation	N
mpg	23.02207447	7.345599459	376
displacement	196.783	104.4808	376
acceleration	15.46702128	2.635213370	376
weight	2993.99	852.923	376
cylinders	5.51	1.710	376
origin_1	.6489	.47794	376
origin_2	.1596	.36670	376

Figure 1.15

4	(Constant)	43.731	.985
	weight	-.005	.001
	displacement	-.017	.005
	origin_2	-3.338	.626
	origin_1	-2.314	.575

a. Dependent Variable: mpg

Figure 1.16

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.863[a]	.745	.744	3.714441008	.745	1092.557	1	374	.000	
2	.868[b]	.753	.752	3.658367614	.008	12.553	1	373	.000	
3	.873[c]	.761	.760	3.602063777	.008	12.752	1	372	.000	
4	.879[d]	.774	.771	3.514700168	.012	19.723	1	371	.000	.975

a. Predictors: (Constant), weight

b. Predictors: (Constant), weight, cylinders

c. Predictors: (Constant), weight, cylinders, origin_2

d. Predictors: (Constant), weight, cylinders, origin_2, origin_1

e. Dependent Variable: mpg

Figure 1.17

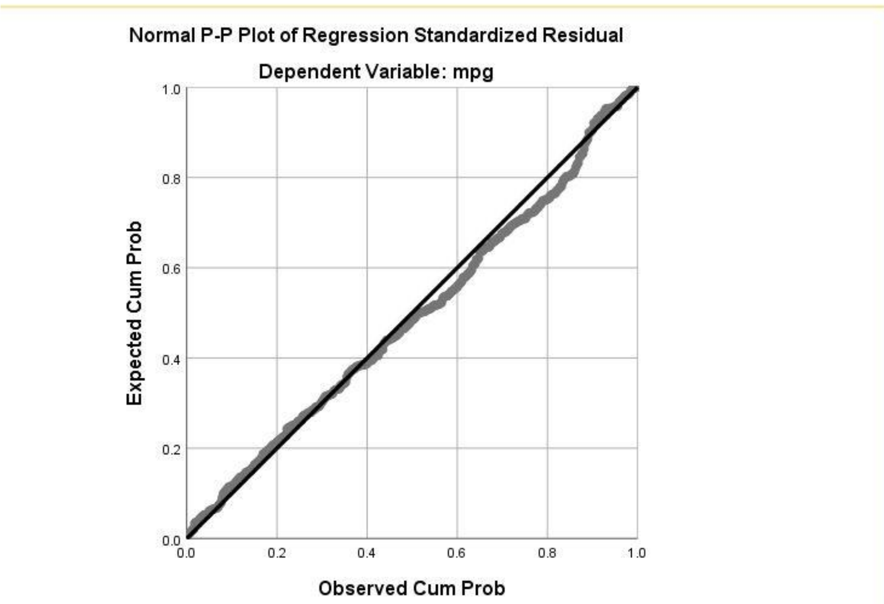
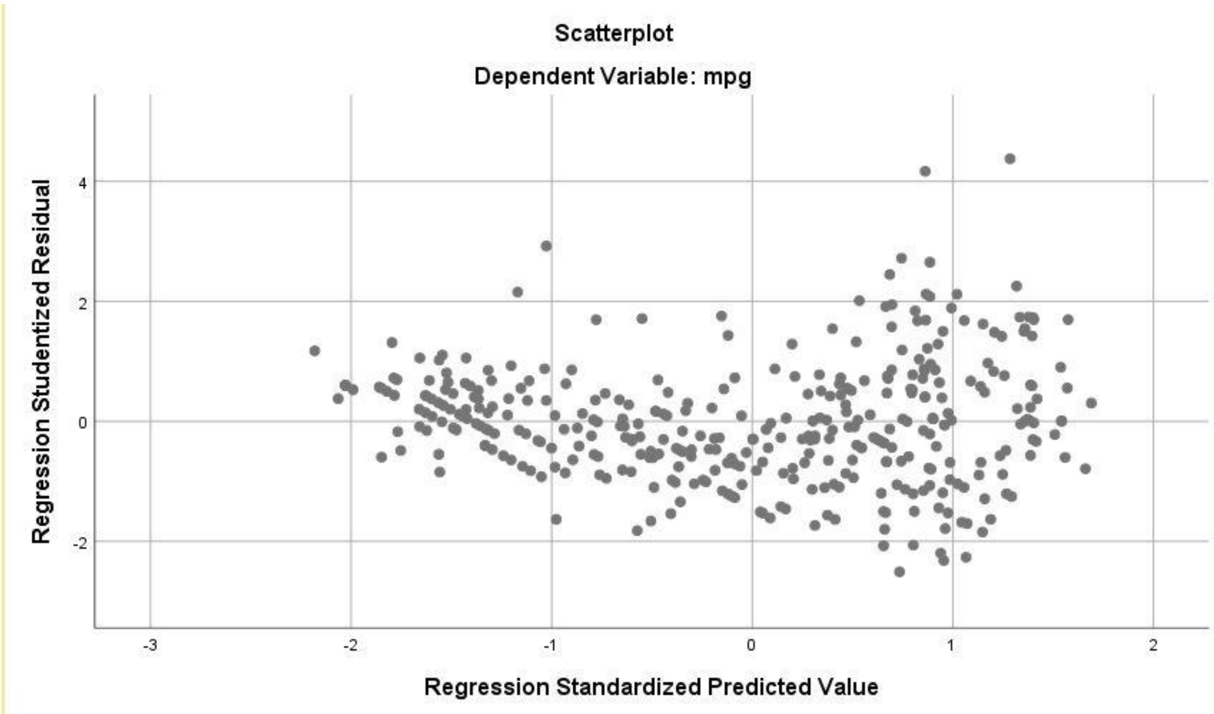


Figure 1.18



Scatterplot
Dependent Variable: mpg
Regression Studentized Residual
4
2
0
-2
-3
-2
-1
0
1
2
Regression Standardized Predicted Value