Factors Affecting Miles Per Gallon of Fuel

Contents

1. Objective statement 2

2. Methods 2

2.1 Data description 2

2.2 Analysis 4

2.3 Sample size requirement 4

2.4 Variate(s)/model description 5

2.5 Assumptions in multiple regression analysis 5

2.6 Multiple regression model estimation and assessment of overall model fit 6

2.7 Results validation 6

3. Results and Interpretations 6

3.1 Assumptions testing 7

3.1.1 Missing data analysis 7

3.1.2 Homoscedasticity test 8

3.1.3 Normality test 10

3.1.4 Correlation test 13

3.2 Multiple regression model 13

3.3 Results validation 23

4. Conclusion 28

1. Objective statement

The goal of the analysis in this report is to explore the relationship between miles per gallon (MPG) and other relevant variables by performing multiple regression analyses on the data set of "auto-mpg.sav". We chose MPG as the dependent variable with the aim of developing a model to predict vehicle fuel efficiency. By deeply analyzing the relationship between each independent variable and the dependent variable, we hope to identify factors that significantly influence MPG to provide a deeper understanding of vehicle design and performance.

2. Methods

2.1 Data description

For the purpose of this technical report, secondary data was collected from the "Auto MPG Data Set" used in the 1983 American Statistical Association Exposition, which is currently archived in the StatLib Library maintained at the Carnegie Mellon University. The data contains 398 observations of 9 variables, which are mpg, cylinders, displacement, horsepower, weight, acceleration, modelyear, origin and carname.

Fuel efficiency is represented by the variable *mpg*. It measures how far a car can travel on a specific amount of fuel. The lesser the fuel used, the higher the fuel efficiency.

The number of cylinders is represented by the variable *cylinders*. The cylinder is the power unit of a car's engine, where fuel is burned and converted into mechanical energy to power the vehicle.

The engine displacement is represented by the variable *displacement*. Engine size refers to the volume of the cylinders of the engine. Generally, larger engine sizes can take in more fuel and air, resulting in higher power.

The horsepower is represented by the variable *horsepower*. Horsepower measures how fast force is produced from a car engine, and how well it can accomplish it.

The vehicle weight is represented by the variable *weight*. Car weight typically increases fuel consumption because of the additional power needed to move the vehicle.

The time to accelerate from 0 to 60 mph is represented by the variable *acceleration*. Theoretically, if the time to accelerate from 0 to 60 mph is shorter, more fuel will be consumed.

The *modelyear* variable represents the year in which each car model was launched;

The *origin* variable identifies the car's country of origin ("1" is American car, "2" is European car, and "3" is Japanese car)，later, we intervers them into dummy variable in section 3.

The carname is a nominal variable, which describes the car brands.

There are a variables table to clearly summary the situation as follow:

Order	Name	Label	Measure
1	mpg	miles per gallon	Scale
2	cylinders	number of cylinders	Scale
3	displacement	engine displacement	Scale
4	horsepower	horsepower	Scale
5	weight	vehicle weight (lbs.)	Scale
6	acceleration	time to accelerate from 0 to 60 mph (sec.)	Scale
7	modelyear	model year (modulo 100)	Nominal
8	origin	origin of car	Nominal
9	carname	car labels (type)	Nominal

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
v1_miles per gallon	398	9.0	46.6	23.515	7.8160
v2_number of cylinders	398	3	8	5.45	1.701
v3_engine displacement	398	68.0	455.0	193.426	104.2698
v4_horsepower	398	46	230	104.47	38.199
v5_vehicle weight (lbs.)	398	1613	5140	2970.42	846.842
v6_time to accelerate from 0 to 60 mph (sec.)	398	8.0	24.8	15.568	2.7577
v7_model year (modulo 100)	398	70	82	76.01	3.698
v8_origin of car	398	1	3	1.57	.802
Valid N (listwise)	398				

2.2 Analysis

Multiple regression analysis is a statistical technique that can be used to analyze the relationship between a single dependent(criterion)variable and several independent (predictor)variables.

The objective of multiple regression analysis is to use the independent variables whose values are known to predict the single dependent value selected by the researcher. And we will perform multiple regression analyses to explore the relationship between miles per gallon (MPG) and other relevant variables.

2.3 Sample size requirement

Rule: The ratio of sample size to variables should never fall below 5:1.

Since we have 9 variables, the overall sample size of the data set we use for analysis needs to be at least 5*9=45.

For the data set, we will use 70% as the Estimation sample and the other 30% as the Validation sample.

For Estimation sample :

398*70%=278.6,

278.6>45,

For Validation sample :

398*30%=119.4,

119.4>45,

So this data set meets the sample size requirement for all analysis.

2.4 Variate(s)/model description

Multiple regression model:

$Y=bo+b_1X_1+b_2X_2+\cdots\cdots+bnXn+e$ where:

bo =constant number of MPG independent of Xn

Xn = the n th independent variable bn =

coefficient of the n th independent variable e

=predication error(residual)

2.5 Assumptions in multiple regression analysis

Initially, all the assumption testing procedures will be conducted using the overall sample to test for data validity, which include descriptive statistics (in Section 2.1), Missing data analysis and replace(in Section 3.1.1), normality test (in Section 3.1.3), correlation test (in Section 3.1.4) and homogeneity test (in Section 3.1.2).

1. Linearity Assessment: Scrutinizing the linear relationship between dependent and independent variables through scatterplot analysis.

2. Homoscedasticity Evaluation: Verifying the consistency of variance in residuals (homoscedasticity) across the regression model.

3. Normality Check: Analyzing the distribution of residuals for normality using plots and statistical tests.

4. Identification of Influential Points: Detecting and mitigating the impact of outliers or leverage points through diagnostic plots.

2.6 Multiple regression model estimation and assessment of overall model fit

1. Model Estimation: This includes selecting relevant independent variables and employing SPSS to compute the regression coefficients. The process involves fitting the model to the data, ensuring that the chosen variables are relevant and contribute significantly to the model.

2. Assessment of Model Fit: This is conducted through the evaluation of various statistical measures such as R-squared, Adjusted R-squared, and F-statistics. These metrics help in determining how well the model explains the variability in the dependent variable and the overall effectiveness of the model.

2.7 Results validation

1. Residual Analysis: This includes checking residuals for randomness and absence of patterns, indicating a well-fitting model.

2. Cross-Validation: Implementing techniques like k-fold cross-validation to assess the model's performance on unseen data.

3. Comparative Analysis: Comparing the developed model with other models using statistical tests to ensure its superiority and appropriateness.

3. Results and Interpretations

Based on the dataset, we can find there is a total 9 variables with independent variable and dependent variable. For the variable of "v8_origin of car", we would like to transform it to the dummy variables "American", "European" and "Japanese" to analyze and for the "v9_car labels (type)", because it is the string variable and there are so many kinds of labels in this variable, so we would not to use this variable and exclude it to the independent variable when we make analysis. In this way, except for the v9 and our dependent variable, we totally have 9 variables in our database (v8 to 3 dummy variables.)

3.1 Assumptions testing

3.1.1 Missing data analysis

At the beginning of the study, we would like to test if there is missing value in the dataset.

Case Processing Summary[a]

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
v1_miles per gallon	100	100.0%	0	0.0%	100	100.0%
v2_number of cylinders	100	100.0%	0	0.0%	100	100.0%
v3_engine displacement	100	100.0%	0	0.0%	100	100.0%
v4_horsepower	94	94.0%	6	6.0%	100	100.0%
v5_vehicle weight (lbs.)	100	100.0%	0	0.0%	100	100.0%
v6_time to accelerate from 0 to 60 mph (sec.)	100	100.0%	0	0.0%	100	100.0%
v7_model year (modulo 100)	100	100.0%	0	0.0%	100	100.0%
v8_origin of car	100	100.0%	0	0.0%	100	100.0%
v9_car labels (type)	100	100.0%	0	0.0%	100	100.0%

a. Limited to first 100 cases.

Based on the above result, we find there is missing value in the dataset for the v4_horsepower, but only 6%. And we want to check if it is MCAR or MAR. Based on this, we transform the missing value to 0 and the others to 1, using the compare mean and we get

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
v8_origin of car	Equal variances assumed	3.552	.060	-.737	396	.462	-.243	.330	-.892	.406
	Equal variances not assumed			-1.133	5.379	.305	-.243	.215	-.784	.297

Because the sig>0.05, so we think this is MCAR and we want to use the mean value to replace the missing data and it will good to make analysis.

3.1.2 Homoscedasticity test

Before we build the model, we would like to test the Constant variance (homoscedasticity).

For this one, we will use one-way ANOVA to test it, make the scale independent variables into the dependent list opinion and make one of the nominal independent variables into the factor and then we can see the test result.

At here, we put v2-v6 into the dependent list and choice v7 (model year (modulo 100) into the factor and we get the result.

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
v2_number of cylinders	Based on Mean	16.344	12	385	.000
	Based on Median	5.909	12	385	.000
	Based on Median and with adjusted df	5.909	12	233.676	.000
	Based on trimmed mean	17.113	12	385	.000
v3_engine displacement	Based on Mean	15.347	12	385	.000
	Based on Median	7.739	12	385	.000
	Based on Median and with adjusted df	7.739	12	254.423	.000
	Based on trimmed mean	15.166	12	385	.000
v4_horsepower	Based on Mean	10.151	12	385	.000
	Based on Median	8.008	12	385	.000
	Based on Median and with adjusted df	8.008	12	307.762	.000
	Based on trimmed mean	9.880	12	385	.000
v5_vehicle weight (lbs.)	Based on Mean	8.874	12	385	.000
	Based on Median	6.467	12	385	.000
	Based on Median and with adjusted df	6.467	12	295.017	.000
	Based on trimmed mean	8.739	12	385	.000
v6_time to accelerate from 0 to 60 mph (sec.)	Based on Mean	1.826	12	385	.042
	Based on Median	1.509	12	385	.118
	Based on Median and with adjusted df	1.509	12	342.992	.119
	Based on trimmed mean	1.736	12	385	.057

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
v2_number of cylinders	Between Groups	214.364	12	17.864	7.361	.000
	Within Groups	934.322	385	2.427		
	Total	1148.686	397			
v3_engine displacement	Between Groups	843936.528	12	70328.044	7.798	.000
	Within Groups	3472326.536	385	9019.030		
	Total	4316263.063	397			
v4_horsepower	Between Groups	144167.012	12	11805.091	10.385	.000
	Within Groups	435126.621	385	1136.708		
	Total	579293.633	397			
v5_vehicle weight (lbs.)	Between Groups	39676146.954	12	3306345.579	5.195	.000
	Within Groups	245028826.285	385	636438.510		
	Total	284704973.239	397			
v6_time to accelerate from 0 to 60 mph (sec.)	Between Groups	405.045	12	33.754	4.971	.000
	Within Groups	2614.080	385	6.790		
	Total	3019.125	397			

In the same way, we will try again by using the v8 (origin of the car) into the factor and we get the result.

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
v2_number of cylinders	Based on Mean	128.106	2	395	.000
	Based on Median	110.437	2	395	.000
	Based on Median and with adjusted df	110.437	2	334.952	.000
	Based on trimmed mean	140.496	2	395	.000
v3_engine displacement	Based on Mean	109.388	2	395	.000
	Based on Median	109.468	2	395	.000
	Based on Median and with adjusted df	109.468	2	273.177	.000
	Based on trimmed mean	110.011	2	395	.000
v4_horsepower	Based on Mean	49.438	2	395	.000
	Based on Median	26.527	2	395	.000
	Based on Median and with adjusted df	26.527	2	302.650	.000
	Based on trimmed mean	44.604	2	389	.000
v5_vehicle weight (lbs.)	Based on Mean	44.543	2	395	.000
	Based on Median	43.602	2	395	.000
	Based on Median and with adjusted df	43.602	2	344.967	.000
	Based on trimmed mean	44.795	2	395	.000
v6_time to accelerate from 0 to 60 mph (sec.)	Based on Mean	6.239	2	395	.002
	Based on Median	4.226	2	395	.015
	Based on Median and with adjusted df	4.226	2	334.774	.015
	Based on trimmed mean	5.713	2	395	.004

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
v2_number of cylinders	Between Groups	419.662	2	209.831	113.691	.000
	Within Groups	729.024	395	1.846		
	Total	1148.686	397			
v3_engine displacement	Between Groups	1833062.349	2	916531.174	145.792	.000
	Within Groups	2483200.715	395	6286.584		
	Total	4316263.063	397			
v4_horsepower	Between Groups	138894.595	2	69447.297	61.342	.000
	Within Groups	440399.038	389	1132.131		
	Total	579293.633	391			
v5_vehicle weight (lbs.)	Between Groups	103462718.801	2	51731359.400	112.744	.000
	Within Groups	181242254.438	395	458841.150		
	Total	284704973.239	397			
v6_time to accelerate from 0 to 60 mph (sec.)	Between Groups	203.951	2	101.975	14.308	.000
	Within Groups	2815.174	395	7.127		
	Total	3019.125	397			

Based on the above result, we can find the sig<0.05, we reject the null hypotheses, and we think the data is not very good.

3.1.3 Normality test

After that, we would like to test the normality for the independent variables, which helps us to find that whether the dataset is suitable to conduct further exploration latter.

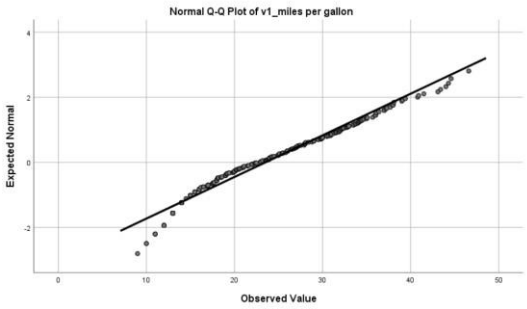
Descriptives

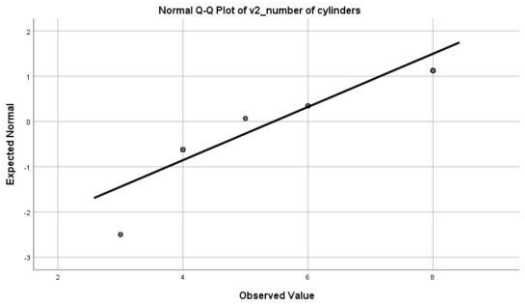
	Statistic													Std. Error		
	Mean	95% Confidence Interval for Mean		5% Trimmed Mean	Median	Variance	Std. Deviation	Minimum	Maximum	Range	Interquartile Range	Skewness	Kurtosis	Mean	Skewness	Kurtosis
		Lower Bound	Upper Bound													
v2_number of cylinders	5.45	5.29	5.62	5.41	4.00	2.893	1.701	3	8	5	4	.527	-1.377	.085	.122	.244
v3_engine displacement	193.426	183.151	203.701	187.359	148.500	10872.199	104.2698	68.0	455.0	387.0	160.0	.720	-.747	5.2266	.122	.244
v4_horsepower	104.47	100.71	108.23	101.75	95.00	1459.178	38.199	46	230	184	49	1.096	.754	1.915	.122	.244
v5_vehicle weight (lbs.)	2970.42	2886.97	3053.88	2937.31	2803.50	717140.991	846.842	1613	5140	3527	1388	.531	-.786	42.448	.122	.244
v6_time to accelerate from 0 to 60 mph (sec.)	15.568	15.296	15.840	15.520	15.500	7.605	2.7577	8.0	24.8	16.8	3.4	.279	.419	.1382	.122	.244
v7_model year (modulo 100)	76.01	75.65	76.37	76.01	76.00	13.672	3.698	70	82	12	6	.012	-1.181	.185	.122	.244
v8_origin of car	1.57	1.49	1.65	1.53	1.00	.643	.802	1	3	2	1	.924	-.818	.040	.122	.244
v1_miles per gallon	23.515	22.744	24.285	23.224	23.000	61.090	7.8160	9.0	46.6	37.6	11.6	.457	-.511	.3918	.122	.244

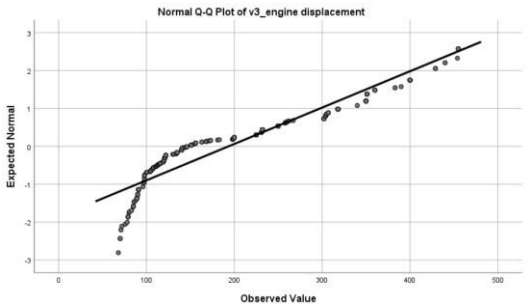
Tests of Normality

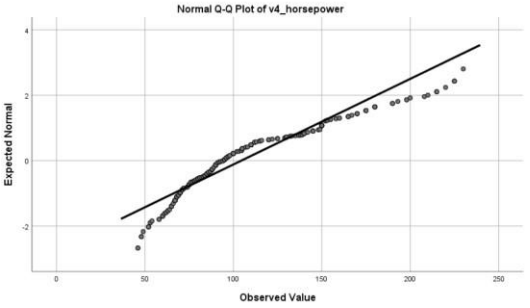
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
v2_number of cylinders	.326	398	.000	.749	398	.000
v3_engine displacement	.183	398	.000	.880	398	.000
v4_horsepower	.155	398	.000	.905	398	.000
v5_vehicle weight (lbs.)	.093	398	.000	.941	398	.000
v6_time to accelerate from 0 to 60 mph (sec.)	.051	398	.015	.992	398	.040
v7_model year (modulo 100)	.106	398	.000	.946	398	.000
v8_origin of car	.388	398	.000	.674	398	.000
v1_miles per gallon	.079	398	.000	.968	398	.000

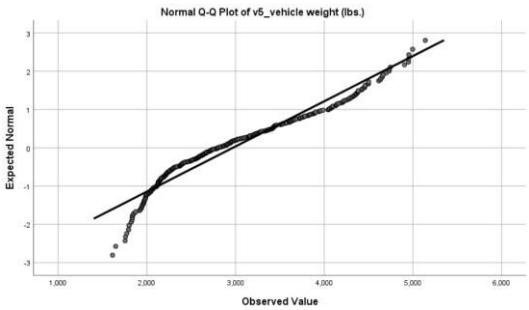
a. Lilliefors Significance Correction

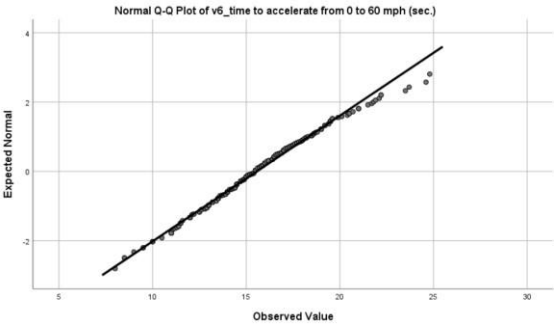


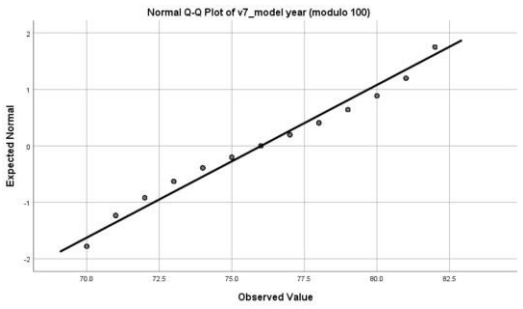












Unfortunately, we find the data are not the normally distribution, so we try to use the log to change the data (we exclude v7 and v8 cause they are nominal data).

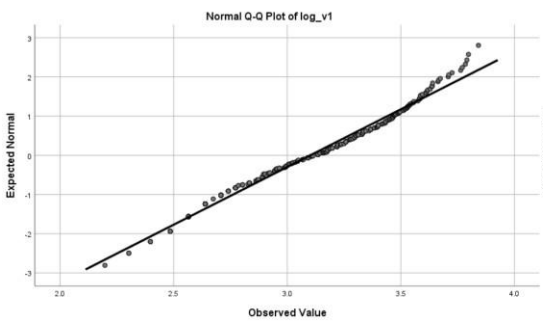
Descriptives

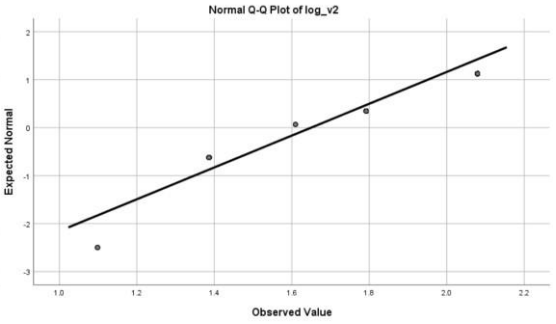
	Statistic													Std. Error		
	Mean	95% Confidence Interval for Mean		5% Trimmed Mean	Median	Variance	Std. Deviation	Minimum	Maximum	Range	Interquartile Range	Skewness	Kurtosis	Mean	Skewness	Kurtosis
		Lower Bound	Upper Bound													
log_v1	3.1014	3.0679	3.1349	3.1037	3.1355	.115	.33966	2.20	3.84	1.64	.51	-.136	-.803	.01703	.122	.244
log_v2	1.6500	1.6203	1.6798	1.6441	1.3863	.091	.30151	1.10	2.08	.98	.69	.368	-1.509	.01511	.122	.244
log_v3	5.1232	5.0708	5.1755	5.1161	5.0004	.282	.53128	4.22	6.12	1.90	.94	.226	-1.342	.02663	.122	.244
log_v4	4.5889	4.5553	4.6224	4.5825	4.5539	.116	.34086	3.83	5.44	1.61	.50	.365	-.388	.01709	.122	.244
log_v5	7.9569	7.9292	7.9845	7.9544	7.9386	.079	.28066	7.39	8.54	1.16	.49	.156	-1.088	.01407	.122	.244
log_v6	2.7293	2.7114	2.7471	2.7328	2.7408	.033	.18073	2.08	3.21	1.13	.22	-.360	.630	.00906	.122	.244

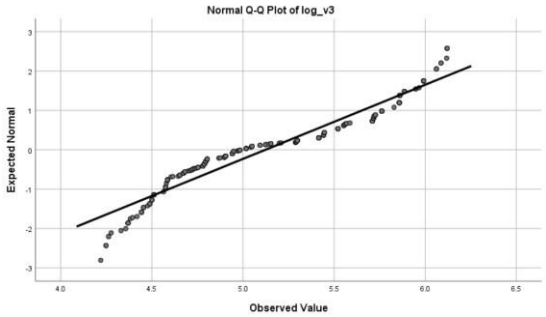
Tests of Normality

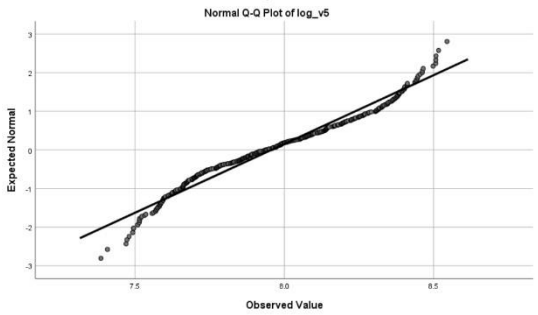
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
log_v1	.060	398	.002	.982	398	.000
log_v2	.332	398	.000	.758	398	.000
log_v3	.141	398	.000	.926	398	.000
log_v4	.089	398	.000	.976	398	.000
log_v5	.081	398	.000	.964	398	.000
log_v6	.056	398	.004	.989	398	.005

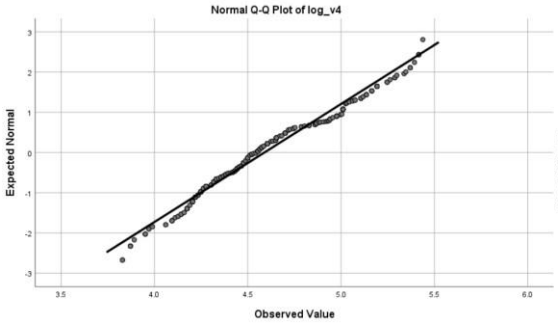
a. Lilliefors Significance Correction

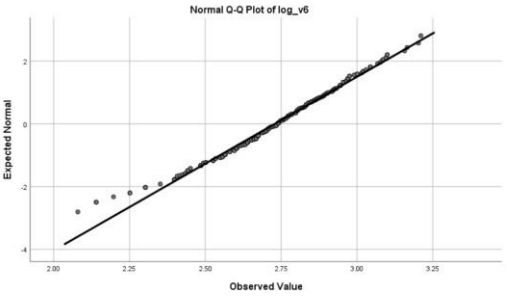












Although the P value for the data is still not good, when we see the normal Q-Q plot, we find it better compared with the previous.

And when we try to use the square and inverse, we find similar results to the log.

So, we will try to use the original data and log data to build the model and make a comparison.

3.1.4 Correlation test

After that, we want to see the Linearity between independent variables.

Correlation Matrix

		v2_number of cylinders	v3_engine displacement	v4_horsepower	v5_vehicle weight (lbs.)	v6_time to accelerate from 0 to 60 mph (sec.)	v7_model year (modulo 100)
Correlation	v2_number of cylinders	1.000	.951	.839	.896	-.505	-.349
	v3_engine displacement	.951	1.000	.894	.933	-.544	-.370
	v4_horsepower	.839	.894	1.000	.861	-.684	-.412
	v5_vehicle weight (lbs.)	.896	.933	.861	1.000	-.417	-.307
	v6_time to accelerate from 0 to 60 mph (sec.)	-.505	-.544	-.684	-.417	1.000	.288
	v7_model year (modulo 100)	-.349	-.370	-.412	-.307	.288	1.000

Based on the above result, we think there might be the linear relationship between variables, and we want to see for the log variable group.

Correlation Matrix

		log_v2	log_v3	log_v4	log_v5	log_v6	v7_model year (modulo 100)
Correlation	log_v2	1.000	.946	.820	.881	-.505	-.340
	log_v3	.946	1.000	.866	.943	-.524	-.330
	log_v4	.820	.866	1.000	.867	-.710	-.391
	log_v5	.881	.943	.867	1.000	-.426	-.284
	log_v6	-.505	-.524	-.710	-.426	1.000	.311
	v7_model year (modulo 100)	-.340	-.330	-.391	-.284	.311	1.000

Based on the above result, we think there is a linear relationship between variables, but seems better compare with the original dataset.

After that, we will use both the original dataset and the log dataset to build up the model and make a comparison.

3.2 Multiple regression model

When we use the **original data** v1_miles per gallon as the **Dependent Variable**; v2_number of cylinders, v3_engine displacement, v4_horsepower, v5_vehicle weight

(lbs.), v6_time to accelerate from 0 to 60 mph (sec.), v7_model year (modulo 100), v10_American, v11_European, v12_Japanese as the **Independent Variables** to build the linear regression model, SPSS will help us choice the suitable variables in the model and we get the results as below.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.832[a]	.692	.691	4.3446
2	.899[b]	.808	.807	3.4347
3	.905[c]	.819	.818	3.3338
4	.906[d]	.822	.820	3.3189
5	.907[e]	.823	.821	3.3060

a. Predictors: (Constant), v5_vehicle weight (lbs.)
b. Predictors: (Constant), v5_vehicle weight (lbs.), v7_model year (modulo 100)
c. Predictors: (Constant), v5_vehicle weight (lbs.), v7_model year (modulo 100), v10_American
d. Predictors: (Constant), v5_vehicle weight (lbs.), v7_model year (modulo 100), v10_American, v3_engine displacement
e. Predictors: (Constant), v5_vehicle weight (lbs.), v7_model year (modulo 100), v10_American, v3_engine displacement, v6_time to accelerate from 0 to 60 mph (sec.)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	46.317	.795		58.243	.000
	v5_vehicle weight (lbs.)	-.008	.000	-.832	-29.814	.000
2	(Constant)	-14.198	3.968		-3.578	.000
	v5_vehicle weight (lbs.)	-.007	.000	-.722	-31.161	.000
	v7_model year (modulo 100)	.757	.049	.358	15.447	.000
3	(Constant)	-16.141	3.870		-4.170	.000
	v5_vehicle weight (lbs.)	-.006	.000	-.640	-22.972	.000
	v7_model year (modulo 100)	.770	.048	.364	16.176	.000
	v10_American	-2.170	.432	-.135	-5.027	.000
4	(Constant)	-17.089	3.879		-4.406	.000
	v5_vehicle weight (lbs.)	-.007	.001	-.752	-12.615	.000
	v7_model year (modulo 100)	.800	.049	.378	16.197	.000
	v10_American	-2.510	.458	-.156	-5.476	.000
	v3_engine displacement	.010	.005	.139	2.133	.034
5	(Constant)	-18.956	3.973		-4.772	.000
	v5_vehicle weight (lbs.)	-.007	.001	-.790	-12.689	.000
	v7_model year (modulo 100)	.794	.049	.376	16.122	.000
	v10_American	-2.670	.463	-.166	-5.761	.000
	v3_engine displacement	.016	.006	.209	2.841	.005
	v6_time to accelerate from 0 to 60 mph (sec.)	.154	.076	.054	2.019	.044

a. Dependent Variable: v1_miles per gallon

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	v2_number of cylinders	-.153[b]	-2.449	.015	-.122	.197
	v3_engine displacement	-.218[b]	-2.844	.005	-.142	.130
	v4_horsepower	-.215[b]	-3.990	.000	-.197	.259
	v6_time to accelerate from 0 to 60 mph (sec.)	.088[b]	2.909	.004	.145	.826
	v7_model year (modulo 100)	.358[b]	15.447	.000	.614	.906
	v10	-.110[b]	-3.189	.002	-.158	.642
	v11	.011[b]	.393	.695	.020	.911
	v12	.094[b]	3.048	.002	.152	.806
2	v2_number of cylinders	-.019[c]	-.377	.707	-.019	.191
	v3_engine displacement	.015[c]	.233	.816	.012	.122
	v4_horsepower	-.012[c]	-.257	.797	-.013	.235
	v6_time to accelerate from 0 to 60 mph (sec.)	.020[c]	.799	.425	.040	.797
	v10	-.135[c]	-5.027	.000	-.245	.640
	v11	.058[c]	2.509	.013	.125	.896
	v12	.068[c]	2.798	.005	.140	.802
3	v2_number of cylinders	.033[d]	.656	.512	.033	.183
	v3_engine displacement	.139[d]	2.133	.034	.107	.107
	v4_horsepower	-.024[d]	-.547	.584	-.028	.235
	v6_time to accelerate from 0 to 60 mph (sec.)	.018[d]	.762	.446	.038	.797
	v11	-.006[d]	-.206	.837	-.010	.630
	v12	.006[d]	.206	.837	.010	.574
4	v2_number of cylinders	-.078[e]	-1.136	.257	-.057	.095
	v4_horsepower	-.097[e]	-1.900	.058	-.096	.174
	v6_time to accelerate from 0 to 60 mph (sec.)	.054[e]	2.019	.044	.101	.620
	v11	.004[e]	.149	.882	.008	.612
	v12	-.004[e]	-.149	.882	-.008	.558
5	v2_number of cylinders	-.083[f]	-1.205	.229	-.061	.095
	v4_horsepower	-.055[f]	-.870	.385	-.044	.112
	v11	.001[f]	.021	.983	.001	.610
	v12	-.001[f]	-.021	.983	-.001	.555

a. Dependent Variable: v1_miles per gallon

b. Predictors in the Model: (Constant), v5_vehicle weight (lbs.)

c. Predictors in the Model: (Constant), v5_vehicle weight (lbs.), v7_model year (modulo 100)

d. Predictors in the Model: (Constant), v5_vehicle weight (lbs.), v7_model year (modulo 100), v10

e. Predictors in the Model: (Constant), v5_vehicle weight (lbs.), v7_model year (modulo 100), v10, v3_engine displacement

f. Predictors in the Model: (Constant), v5_vehicle weight (lbs.), v7_model year (modulo 100), v10, v3_engine displacement, v6_time to accelerate from 0 to 60 mph (sec.)

According to the Exclusive variables table we can see that V11V12(European and Japanese) are not significant so we can delete them and aren't conduct further exploration and modeling process, thus we just analyze American cars' relationship between mpg and independent variables. We can find the best model which can explain the largest value with the significant independent variables (sig<0.05) is the model 5. The function for model 5 is:

$$V1 = -18.956 - 0.007V5 + 0.794V7 - 2.670V10 + 0.016V3 + 0.154V6$$

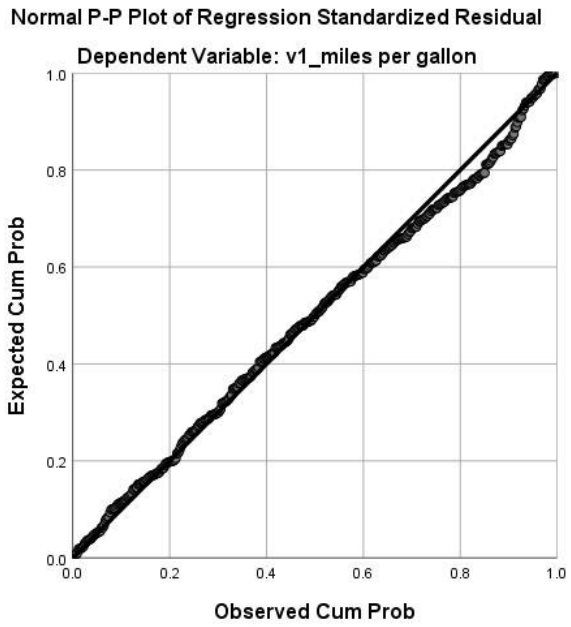
This model can explain **82.1%** data and with the vehicle weight increase 1 lbs, the miles per gallon will decrease 0.07; when the model year increase model 100, the miles per gallon will increase 0.794; if the vehicle is American car, the miles per gallon will decrease 2.670; if the engine displacement increase 1, the miles per gallon will increase 0.016; when time to accelerate from 0 to 60 mph (sec) increase 1 second, the miles per gallon will increase 0.154. The other independent variables will not be used in the function because they are not significant as shown above (Excluded Variables).

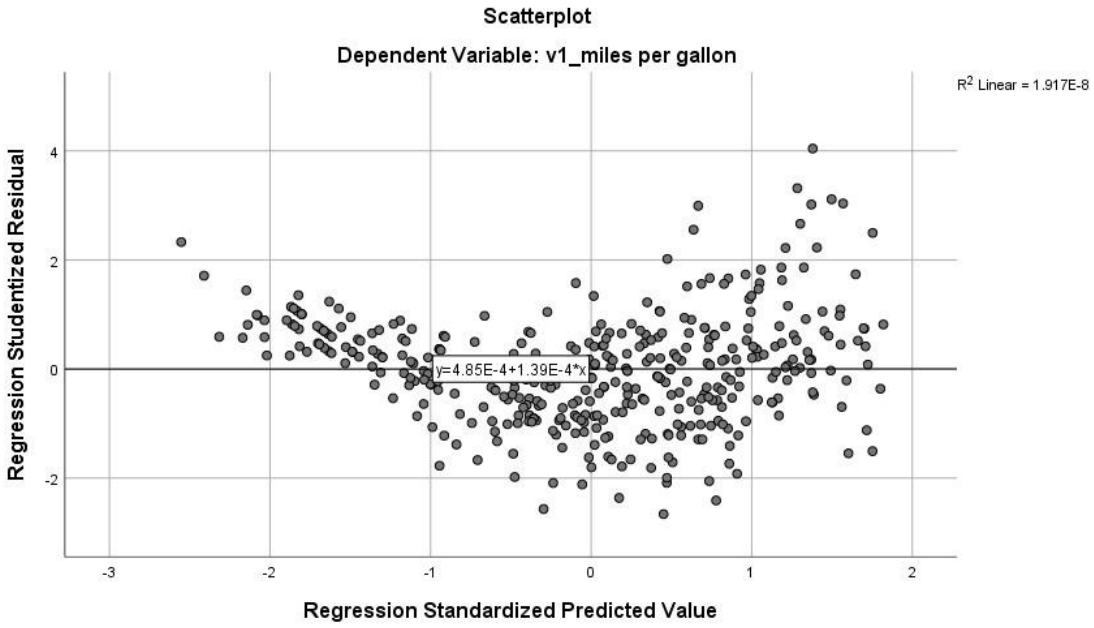
After that, we want to see the residuals for the model

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5.408	36.419	23.515	7.0921	398
Std. Predicted Value	-2.553	1.820	.000	1.000	398
Standard Error of Predicted Value	.214	1.153	.395	.095	398
Adjusted Predicted Value	5.203	36.379	23.511	7.0915	398
Residual	-8.7117	13.2980	.0000	3.2851	398
Std. Residual	-2.635	4.022	.000	.994	398
Stud. Residual	-2.660	4.042	.000	1.003	398
Deleted Residual	-8.8775	13.4279	.0032	3.3457	398
Stud. Deleted Residual	-2.681	4.124	.001	1.007	398
Mahal. Distance	.672	47.318	4.987	3.495	398
Cook's Distance	.000	.091	.003	.008	398
Centered Leverage Value	.002	.119	.013	.009	398

a. Dependent Variable: v1_miles per gallon





Based on the residence, we find the residence P-P plot is a little bit deviate to the downside, and same for the scatterplot. But in general, it is still suitable for the normal distribution and concentrate in the scope of [-2,2]

After that, we want to use the **log data** to build the model and see the result.

When we use the original data log_v1_miles per gallon as the **Dependent Variable**; log_v2_number of cylinders, log_v3_engine displacement, log_v4_horsepower, log_v5_vehicle weight (lbs.), log_v6_time to accelerate from 0 to 60 mph (sec.), v7_model year (modulo 100), v10_American, v11_European, v12_Japanese as the **Independent Variables** to build the linear regression model and we get the results as below.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.874[a]	.765	.764	.16497
2	.939[b]	.882	.881	.11712
3	.942[c]	.886	.886	.11488
4	.943[d]	.889	.888	.11355
5	.944[e]	.891	.889	.11294

a. Predictors: (Constant), log_v5
b. Predictors: (Constant), log_v5, v7_model year (modulo 100)
c. Predictors: (Constant), log_v5, v7_model year (modulo 100), log_v4
d. Predictors: (Constant), log_v5, v7_model year (modulo 100), log_v4, v10_American
e. Predictors: (Constant), log_v5, v7_model year (modulo 100), log_v4, v10_American, log_v6

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	11.522	.235		49.056	.000
	log_v5	-1.058	.029	-.874	-35.874	.000
2	(Constant)	8.055	.242		33.288	.000
	log_v5	-.936	.022	-.773	-42.837	.000
	v7_model year (modulo 100)	.033	.002	.357	19.767	.000
3	(Constant)	7.728	.251		30.832	.000
	log_v5	-.791	.042	-.653	-19.021	.000
	v7_model year (modulo 100)	.031	.002	.334	17.970	.000
	log_v4	-.145	.036	-.146	-4.067	.000
4	(Constant)	7.313	.279		26.172	.000
	log_v5	-.734	.045	-.607	-16.424	.000
	v7_model year (modulo 100)	.031	.002	.336	18.280	.000
	log_v4	-.149	.035	-.150	-4.229	.000
	v10_American	-.048	.015	-.069	-3.206	.001
5	(Constant)	7.566	.299		25.315	.000
	log_v5	-.672	.052	-.555	-12.897	.000
	v7_model year (modulo 100)	.031	.002	.335	18.342	.000
	log_v4	-.239	.052	-.239	-4.553	.000
	v10_American	-.050	.015	-.071	-3.331	.001
	log_v6	-.122	.053	-.065	-2.297	.022

a. Dependent Variable: log_v1

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	log_v2	-.223[b]	-4.351	.000	-.215	.220
	log_v3	-.320[b]	-4.448	.000	-.220	.111
	log_v4	-.363[b]	-7.711	.000	-.364	.236
	log_v6	.114[b]	4.306	.000	.213	.819
	v7_model year (modulo 100)	.356[b]	19.450	.000	.702	.918
	v10	-.039[b]	-1.248	.213	-.063	.614
	v11	.002[b]	.075	.940	.004	.913
	v12	.035[b]	1.260	.208	.064	.778
2	log_v2	-.091[c]	-2.420	.016	-.122	.212
	log_v3	-.136[c]	-2.555	.011	-.129	.107
	log_v4	-.159[c]	-4.284	.000	-.213	.213
	log_v6	.033[c]	1.667	.096	.084	.780
	v10	-.064[c]	-2.876	.004	-.144	.612
	v11	.051[c]	2.757	.006	.139	.897
	v12	.006[c]	.281	.779	.014	.773
3	log_v2	-.063[d]	-1.661	.098	-.084	.204
	log_v3	-.084[d]	-1.550	.122	-.079	.100
	log_v6	-.077[d]	-2.544	.011	-.128	.315
	v10	-.068[d]	-3.127	.002	-.157	.611
	v11	.039[d]	2.157	.032	.109	.874
	v12	.024[d]	1.188	.236	.060	.741
4	log_v2	-.038[e]	-.994	.321	-.051	.194
	log_v3	.005[e]	.080	.936	.004	.073
	log_v6	-.085[e]	-2.825	.005	-.142	.313
	v11	.011[e]	.516	.606	.026	.603
	v12	-.012[e]	-.516	.606	-.026	.537
5	log_v2	-.060[f]	-1.559	.120	-.079	.187
	log_v3	-.055[f]	-.840	.402	-.043	.066
	v11	.008[f]	.355	.723	.018	.601
	v12	-.008[f]	-.355	.723	-.018	.535

a. Dependent Variable: log_v1

b. Predictors in the Model: (Constant), log_v5

c. Predictors in the Model: (Constant), log_v5, v7_model year (modulo 100)

d. Predictors in the Model: (Constant), log_v5, v7_model year (modulo 100), log_v4

e. Predictors in the Model: (Constant), log_v5, v7_model year (modulo 100), log_v4, v10

f. Predictors in the Model: (Constant), log_v5, v7_model year (modulo 100), log_v4, v10, log_v6

According Exclusive variables table we can should delete logv2 logv3 v11 v12. Based on the above result, we can find the best model which can explain the largest value with

the significant independent variables (sig<0.05) is the model 5. The function for model 5 is:

$$logV1 = 7.566 - 0.672logV5 + 0.031V7 - 0.239logV4 - 0.05V10 - 0.122logV6$$

This model can explain **88.9%** data and with the vehicle weight increase 1% lbs, the miles per gallon will decrease 0.672%; when the model year increase modeulo 100, the miles per gallon will increase 3.1%; if the vehicle is American car, the miles per gallon will decrease 5%; if the horsepower increase 1%, the miles per gallon will decrease 0.239%; when time to accelerate from 0 to 60 mph (sec) increase 1%, the miles per gallon will decrease 0.122%.The other independent variables will not be used in the function because they are not significant as shown above (Excluded Variables).

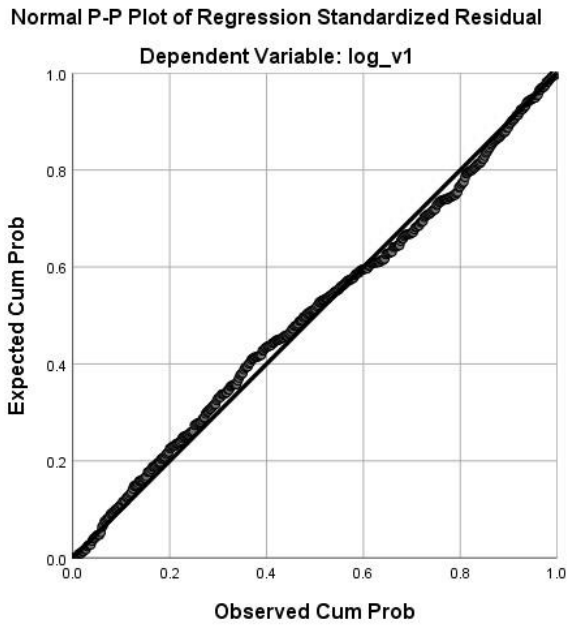
Compared with original dataset, log data has obviously increase. (from 82.1% to 88.9%)

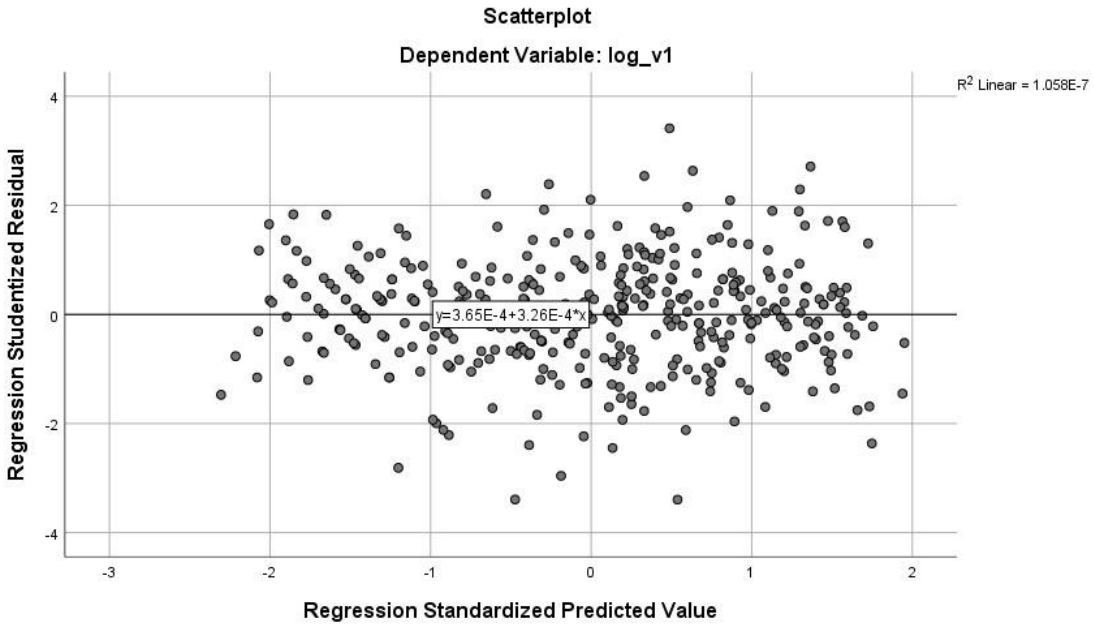
After that, we want to see the residuals for the model

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.3584	3.7246	3.0983	.32101	392
Std. Predicted Value	-2.305	1.951	.000	1.000	392
Standard Error of Predicted Value	.008	.029	.014	.003	392
Adjusted Predicted Value	2.3687	3.7254	3.0982	.32094	392
Residual	-.38132	.38258	.00000	.11217	392
Std. Residual	-3.378	3.389	.000	.994	392
Stud. Residual	-3.394	3.412	.000	1.002	392
Deleted Residual	-.38578	.38771	.00008	.11412	392
Stud. Deleted Residual	-3.442	3.460	.000	1.006	392
Mahal. Distance	1.012	24.364	4.987	3.138	392
Cook's Distance	.000	.035	.003	.005	392
Centered Leverage Value	.003	.062	.013	.008	392

a. Dependent Variable: log_v1





Based on the residence, we find the residence P-P plot and the scatterplot are better compare with the original dataset.

And we think using the log dataset will better to show the result.

3.3 Results validation

Based on the model building we think the log dataset will build the better model, so we will create Estimation sample and Validation sample based on log model.

For the data, we will use 70% as the Estimation sample and the other 30% as the Validation sample.

For Estimation sample we get

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.881[a]	.777	.776	.16323
2	.940[b]	.883	.882	.11847
3	.941[c]	.886	.884	.11735
4	.942[d]	.887	.886	.11661

a. Predictors: (Constant), log_v5
b. Predictors: (Constant), log_v5, v7_model year (modulo 100)
c. Predictors: (Constant), log_v5, v7_model year (modulo 100), v11
d. Predictors: (Constant), log_v5, v7_model year (modulo 100), v11, log_v4
e. Dependent Variable: log_v1

ANOVA[a]

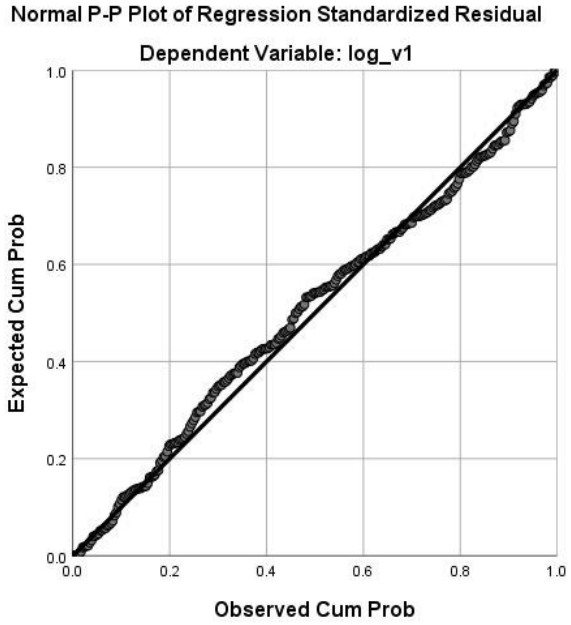
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25.350	1	25.350	951.446	.000[b]
	Residual	7.274	273	.027		
	Total	32.623	274			
2	Regression	28.806	2	14.403	1026.177	.000[c]
	Residual	3.818	272	.014		
	Total	32.623	274			
3	Regression	28.891	3	9.630	699.280	.000[d]
	Residual	3.732	271	.014		
	Total	32.623	274			
4	Regression	28.952	4	7.238	532.248	.000[e]
	Residual	3.672	270	.014		
	Total	32.623	274			

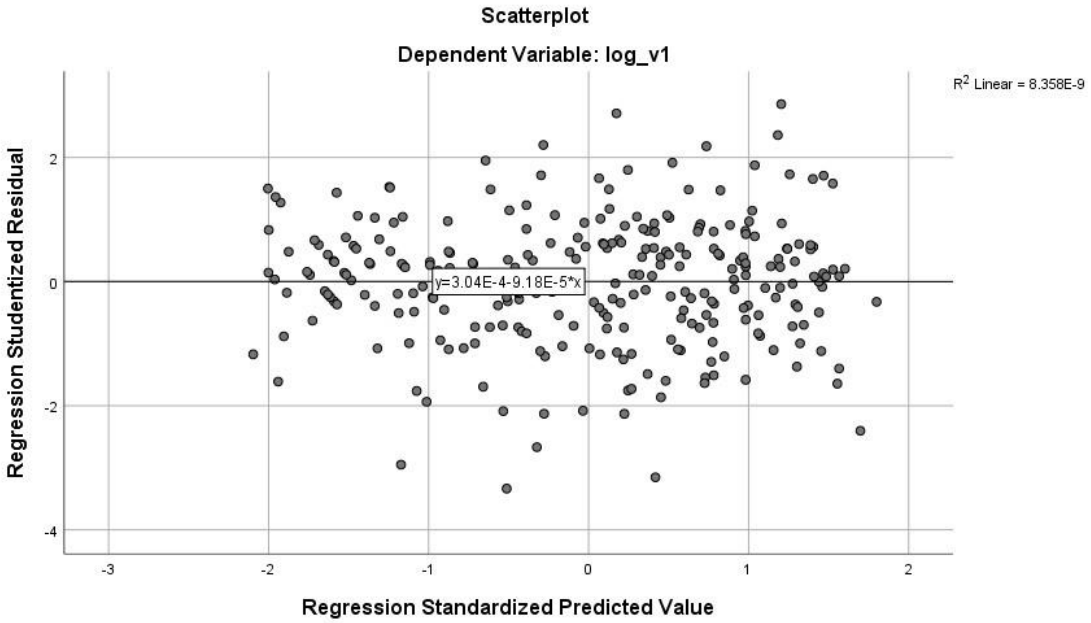
a. Dependent Variable: log_v1
b. Predictors: (Constant), log_v5
c. Predictors: (Constant), log_v5, v7_model year (modulo 100)
d. Predictors: (Constant), log_v5, v7_model year (modulo 100), v11
e. Predictors: (Constant), log_v5, v7_model year (modulo 100), v11, log_v4

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	11.493	.272		42.306	.000
	log_v5	-1.054	.034	-.881	-30.846	.000
2	(Constant)	8.204	.288		28.507	.000
	log_v5	-.943	.026	-.788	-36.547	.000
	v7_model year (modulo 100)	.032	.002	.339	15.692	.000
3	(Constant)	7.942	.304		26.137	.000
	log_v5	-.919	.027	-.769	-33.751	.000
	v7_model year (modulo 100)	.032	.002	.348	16.033	.000
	v11	.050	.020	.055	2.491	.013
4	(Constant)	7.722	.319		24.187	.000
	log_v5	-.820	.054	-.686	-15.151	.000
	v7_model year (modulo 100)	.031	.002	.333	14.644	.000
	v11	.046	.020	.050	2.266	.024
	log_v4	-.100	.048	-.099	-2.109	.036

a. Dependent Variable: log_v1





The function is

$$logV1 = 7.722 - 0.82logV5 + 0.031V7 + 0.046V11 - 0.1logV4$$

The adjust R square is 0.886

For the Validation Sample

Model Summary[e]

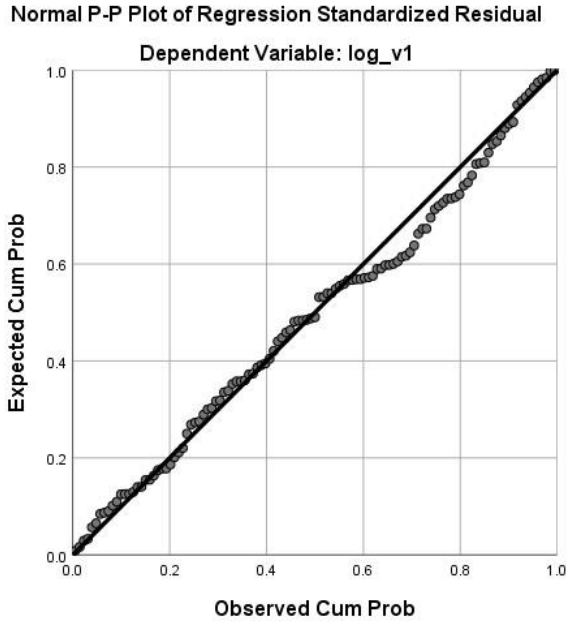
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.852[a]	.726	.723	.17055
2	.934[b]	.873	.871	.11650
3	.943[c]	.889	.886	.10927
4	.946[d]	.895	.891	.10694

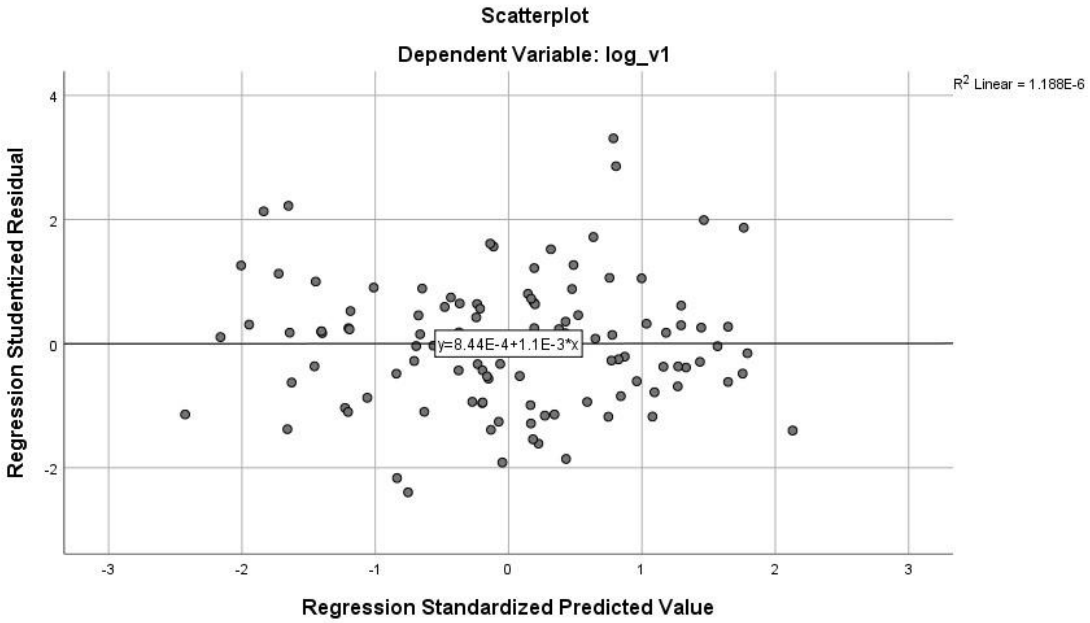
a. Predictors: (Constant), log_v5
b. Predictors: (Constant), log_v5, v7_model year (modulo 100)
c. Predictors: (Constant), log_v5, v7_model year (modulo 100), log_v4
d. Predictors: (Constant), log_v5, v7_model year (modulo 100), log_v4, log_v6
e. Dependent Variable: log_v1

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	11.523	.486		23.714	.000
	log_v5	-1.060	.061	-.852	-17.446	.000
2	(Constant)	7.561	.478		15.813	.000
	log_v5	-.903	.044	-.726	-20.685	.000
	v7_model year (modulo 100)	.036	.003	.404	11.510	.000
3	(Constant)	7.059	.465		15.176	.000
	log_v5	-.665	.071	-.535	-9.334	.000
	v7_model year (modulo 100)	.031	.003	.355	10.144	.000
	log_v4	-.232	.057	-.247	-4.073	.000
4	(Constant)	7.421	.479		15.503	.000
	log_v5	-.541	.086	-.435	-6.275	.000
	v7_model year (modulo 100)	.033	.003	.367	10.605	.000
	log_v4	-.404	.090	-.430	-4.499	.000
	log_v6	-.235	.096	-.141	-2.444	.016

a. Dependent Variable: log_v1





The function is

$$logV1 = 7.421 - 0.541logV5 + 0.033V7 + 0.235logV6 - 0.404logV4$$

The adjust R square is 0.891 and with the vehicle weight increase 1% lbs, the miles per gallon will decrease 0.541%; when the model year increase modeulo 100, the miles per gallon will increase 3.3%; if the horsepower increase 1%, the miles per gallon will decrease 0.404%; when time to accelerate from 0 to 60 mph (sec) increase 1%, the miles per gallon will increase 0.235%

There is some of difference between Validation Sample, Estimation Sample and the original dataset, which because based on the different data we can always get the different results. But for most of the variable are similar and get the close coefficients.

4. Conclusion

In conclusion, because the data can't follow the normal distribution, and we use inverse data to log, and then we find the normality increase even if it's also not follow the normal distribution, we compare the model of original data and log data, finding the log data model can explain more data, which can cover 88.9% data, the influence factors are model year and American cars, vehicle weight, horsepower and the time to accelerate from 0 to 60 mph (sec.). The vehicle weight (lbs) and horsepower are the most significantly influence factors for the mpg. Because vehicle weight increases will make inertia and resistance increase, mechanical efficiency decreases braking energy loss, tire and rolling resistance, and power transmission efficiency decreases. The highhorsepower engine designs may pursue more powerful power output, these design changes may cause fuel to burn insufficiently and thus affect fuel efficiency, high RPM operation may cause engine fuel efficiency to decrease in a certain speed range due to possible reduction in mechanical and thermal efficiency, as well as additional mechanical and friction losses, all of which can cause fuel efficiency to decrease. Thus, these variables will cause to mpg decrease.