Regression Analysis
Project

The Influence of Multiple Factors on Fuel Economy

Name: Xuejiao Liu
Company: New Era Life Insurance Company
Email: xuejiao@gmail.com

Goal

Mileage is the distance traveled in comparison to the fuel/gas filled in the vehicle. There are a number of variables, such as make, model, year, and engine option that affect a vehicle's gas mileage. The purpose of this project is to determine which variables are significant or important in influencing the Miles per Gallon (MPG) of vehicles. Statistic methods, the scatter plot, ANOVA the residual plot, regression analysis and outliers detection are utilized with SAS software in order to achieve our goal.

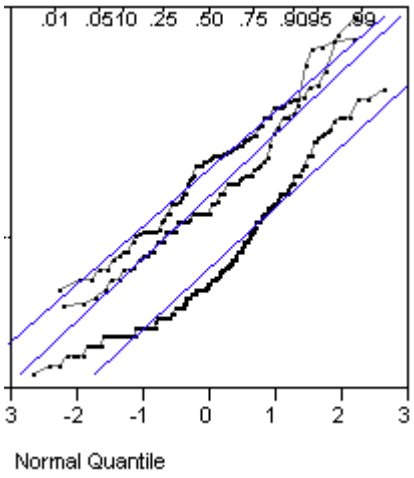
This project illustrates the key techniques using SAS JMP software. Emphasis is on understanding how to go from a practical question to a statistical technique that is readily available in good statistical software packages, and how to interpret and make use of the output of these statistical analysis routines.

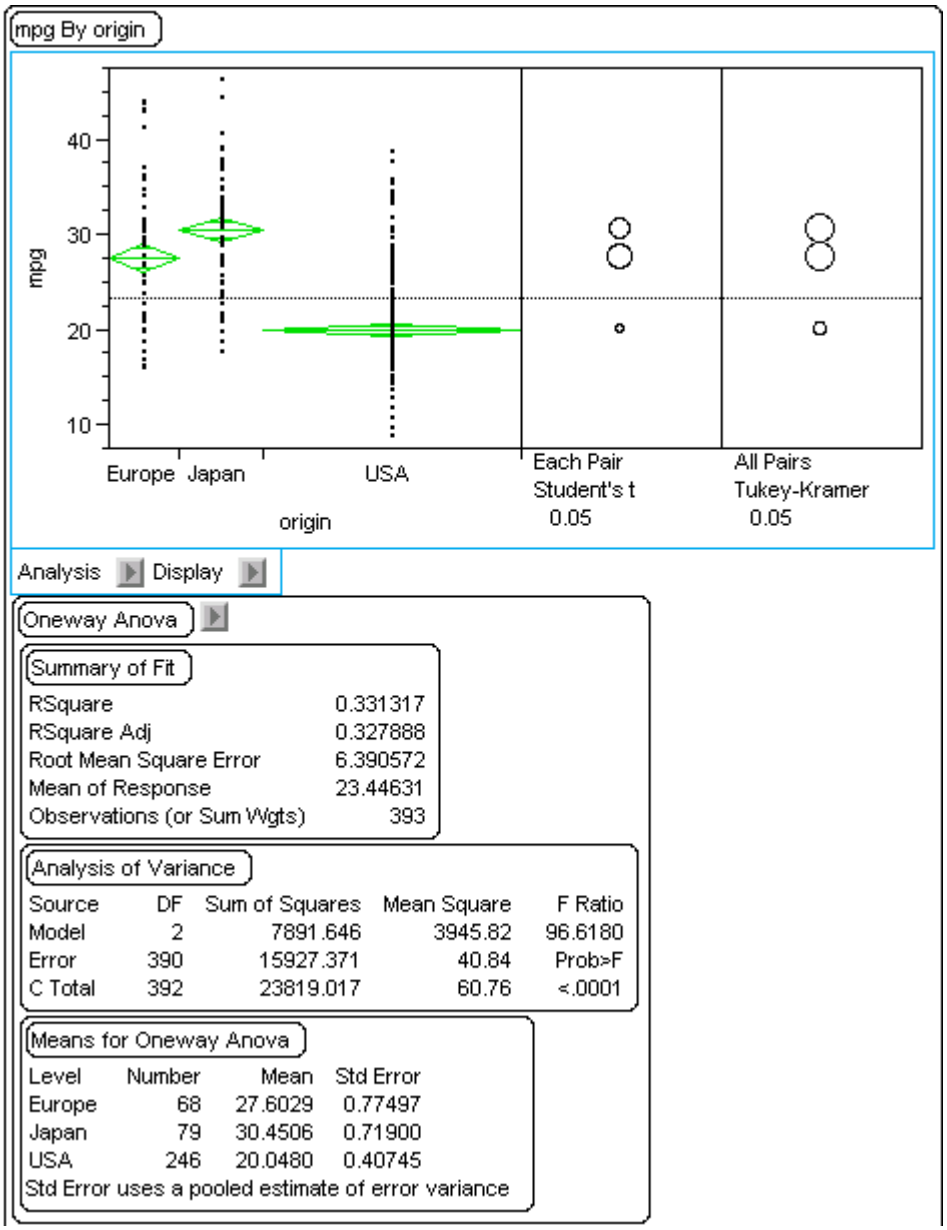
The data is downloaded from the Fuel Economy Database:
(www.fueleconomy.gov/FEG/download.shtml)

Analysis

First, I examine the correlation of the MPG of the cars and origin (American, European, or Japanese).

The following plot shows comparing the test means for three origins of cars:





In the plot's left side, the middle line in the diamond is the response group mean for the group and the vertical endpoints form the 95% confidence interval for the mean. The variances are not significantly unequal.

Means Comparisons

Dif=Mean[i]-Mean[j]	Japan	Europe	USA
Japan	0.0000	2.8477	10.4027
Europe	-2.8477	0.0000	7.5550
USA	-10.4027	-7.5550	0.0000

Alpha= 0.05

Comparisons for each pair using Student's t

t

1.96609

Abs(Dif)-LSD	Japan	Europe	USA
Japan	-1.99915	0.76926	8.77785
Europe	0.76926	-2.15479	5.83355
USA	8.77785	5.83355	-1.13290

Positive values show pairs of means that are significantly different.

Comparisons for all pairs using Tukey-Kramer HSD

q*

2.35270

Abs(Dif)-LSD	Japan	Europe	USA
Japan	-2.39225	0.36057	8.45835
Europe	0.36057	-2.57849	5.49506
USA	8.45835	5.49506	-1.35567

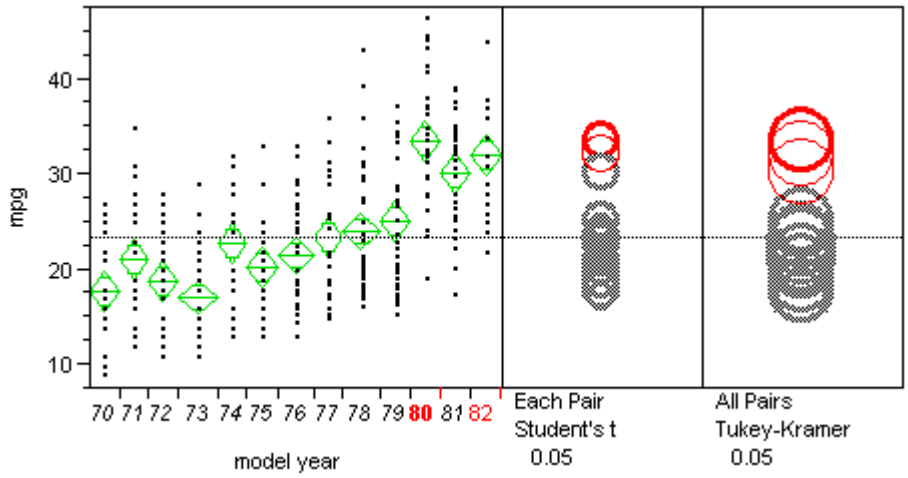
Positive values show pairs of means that are significantly different.

The plot shows that the confidence intervals don't overlap, we can conclude that the means are significantly different. The p-value is less than 0.0001, it conform our conclusion.

From the chart of "mans for one-way Anova", it discovered that the sample size for the three origins are different. So, I applied comparison circles graphical technique, because this graphic works in general with both equal and unequal sample sizes. The plot displays three circles for three groups (origins of cars), which have no intersection. Tukey-Kramer (HSD) is applied to adjust for multiple comparisons to decrease Type I error. In the Tukey table, it still display that the means are significant different.

Conclusion: There is strong evidence that the car's miles per gallon are significant different for these different origins. The Japanese cars have the highest Mpg (mean is 30.45 mpg) and the American cars have the lowest Mpg (mean is 20.05), European cars are in middle of them (mean is bout 27.60).

The model year is also an important factor on the MPG of automobiles. I then examine the correlation of the MPG of the cars and model year. The data contains vehicles MPG information of car made in 1970 – 1982. I use LSD(least significant Different) and Tukey-Kramer HSD to test if it is significantly different between the model years.



Alpha= 0.05

Comparisons for each pair using Student's t

t

1.96626

Abs(Dif)-LSD	80	82	81	79	78	77	74	76	71	75	72	70	73
80	-3.1534	-1.6611	0.1002	5.2201	6.4051	6.9109	7.4566	8.8547	9.1457	10.0722	11.5716	12.6236	13.4320
82	-1.6611	-3.0465	-1.2861	3.8343	5.0221	5.5246	6.0693	7.4710	7.7590	8.6869	10.1853	11.2377	12.0503
81	0.1002	-1.2861	-3.1534	1.9665	3.1516	3.6573	4.2030	5.6011	5.8922	6.8187	8.3180	9.3700	10.1784
79	5.2201	3.8343	1.9665	-3.0985	-1.9121	-1.4080	-0.8628	0.5371	0.8266	1.7538	3.2527	4.3049	5.1155
78	6.4051	5.0221	3.1516	-1.9121	-2.7810	-2.2869	-1.7448	-0.3340	-0.0538	0.8777	2.3738	3.4274	4.2505
77	6.9109	5.5246	3.6573	-1.4080	-2.2869	-3.1534	-2.6077	-1.2096	-0.9186	0.0080	1.5073	2.5593	3.3677
74	7.4566	6.0693	4.2030	-0.8628	-1.7448	-2.6077	-3.2724	-1.8782	-1.5839	-0.6589	0.8415	1.8929	2.6969
76	8.8547	7.4710	5.6011	0.5371	-0.3340	-1.2096	-1.8782	-2.8616	-2.5791	-1.6486	-0.1518	0.9014	1.7213
71	9.1457	7.7590	5.8922	0.8266	-0.0538	-0.9186	-1.5839	-2.5791	-3.2112	-2.2855	-0.7856	0.2661	1.0723
75	10.0722	8.6869	6.8187	1.7538	0.8777	0.0080	-0.6589	-1.6486	-2.2855	-3.0465	-1.5480	-0.4956	0.3170
72	11.5716	10.1853	8.3180	3.2527	2.3738	1.5073	0.8415	-0.1518	-0.7856	-1.5480	-3.1534	-2.1014	-1.2930
70	12.6236	11.2377	9.3700	4.3049	3.4274	2.5593	1.8929	0.9014	0.2661	-0.4956	-2.1014	-3.0985	-2.2880
73	13.4320	12.0503	10.1784	5.1155	4.2505	3.3677	2.6969	1.7213	1.0723	0.3170	-1.2930	-2.2880	-2.6383

Positive values show pairs of means that are significantly different.

Comparisons for all pairs using Tukey-Kramer HSD

q*

3.33408

Abs(Dif)-LSD	80	82	81	79	78	77	74	76	71	75	72	70	73
80	-5.3470	-3.8179	-2.0935	3.0454	4.3369	4.7173	5.2212	6.7601	6.9319	7.9155	9.3780	10.4489	11.4096
82	-3.8179	-5.1657	-3.4429	1.6968	2.9931	3.3678	3.8700	5.4150	5.5816	6.5676	8.0286	9.1003	10.0679
81	-2.0935	-3.4429	-5.3470	-0.2081	1.0834	1.4637	1.9676	3.5065	3.6783	4.6619	6.1244	7.1953	8.1560
79	3.0454	1.6968	-0.2081	-5.2540	-3.9601	-3.5826	-3.0796	-1.5376	-1.3685	-0.3836	1.0781	2.1494	3.1136
78	4.3369	2.9931	1.0834	-3.9601	-4.7156	-4.3551	-3.8573	-2.2969	-2.1435	-1.1514	0.3056	1.3794	2.3649
77	4.7173	3.3678	1.4637	-3.5826	-4.3551	-5.3470	-4.8431	-3.3042	-3.1324	-2.1488	-0.6863	0.3846	1.3453
74	5.2212	3.8700	1.9676	-3.0796	-3.8573	-4.8431	-5.5489	-4.0166	-3.8391	-2.8582	-1.3939	-0.3239	0.6292
76	6.7601	5.4150	3.5065	-1.5376	-2.2969	-3.3042	-4.0166	-4.8524	-4.6949	-3.7046	-2.2464	-1.1733	-0.1933
71	6.9319	5.5816	3.6783	-1.3685	-2.1435	-3.1324	-3.8391	-4.6949	-5.4452	-4.4628	-2.9995	-1.9290	-0.9720
75	7.9155	6.5676	4.6619	-0.3836	-1.1514	-2.1488	-2.8582	-3.7046	-4.4628	-5.1657	-3.7048	-2.6331	-1.6654
72	9.3780	8.0286	6.1244	1.0781	0.3056	-0.6863	-1.3939	-2.2464	-2.9995	-3.7048	-5.3470	-4.2761	-3.3154
70	10.4489	9.1003	7.1953	2.1494	1.3794	0.3846	-0.3239	-1.1733	-1.9290	-2.6331	-4.2761	-5.2540	-4.2898
73	11.4096	10.0679	8.1560	3.1136	2.3649	1.3453	0.6292	-0.1933	-0.9720	-1.6654	-3.3154	-4.2898	-4.4736

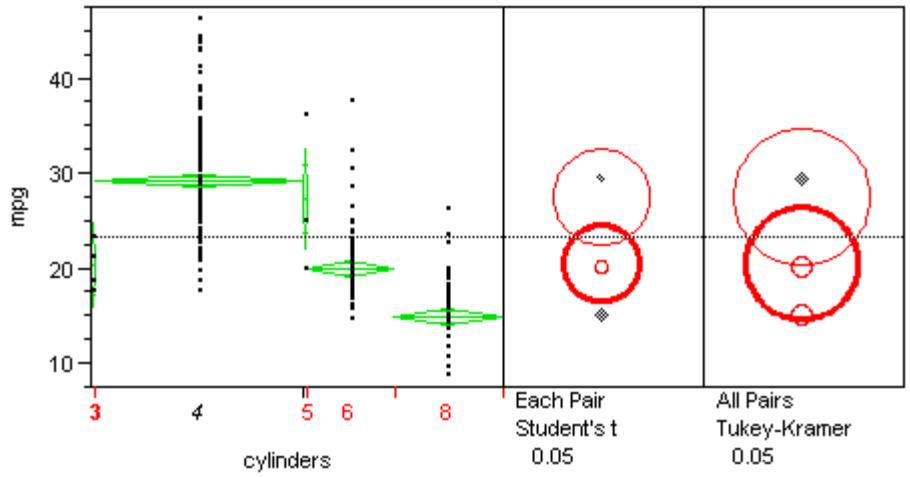
Positive values show pairs of means that are significantly different.

In the plot's left side, the middle line in the diamond is the response group mean for the group and the vertical endpoints form the 95% confidence interval for the mean. The plot shows that the confidence intervals don't overlap for model years 80 and 82 with model years 70's. The elements in the table show the absolute value of the difference in the means, minus the LSD. If the values are positive, it's significant different. The table shows that the means are significantly different.

Conclusion: The mpg of the cars of 80 and 82 model years is significant different from the mpg of cars of 70's.

The third important factor on MPG is the number of engine cylinders. I use LSD and Tukey-Kramer HSD to test if it is significantly different between the different number of

engine cylinders. The sample sizes of the cars with cylinder 3 or 5 are very small, so only 4, 6 and 8 cylinders engine cars are being considered. There are three circles in LSD and Tukey-Kramer HSD graphic. The three circles don't intersect and in the Tukey-Kramer HSD table. We can conclude that the means are significantly different.



Alpha= 0.05
Comparisons for each pair using Student's t

t					
1.96613					
Abs(Dif)-LSD	4	5	3	6	8
4	-0.9244	-3.4880	4.0376	8.0750	13.1713
5	-3.4880	-7.5476	-0.2435	1.9606	6.9894
3	4.0376	-0.2435	-6.5364	-4.1555	0.8761
6	8.0750	1.9606	-4.1555	-1.4349	3.6469
8	13.1713	6.9894	0.8761	3.6469	-1.2881

Positive values show pairs of means that are significantly different.
Comparisons for all pairs using Tukey-Kramer HSD

q*					
2.74066					
Abs(Dif)-LSD	4	5	3	6	8
4	-1.2885	-5.6062	2.1987	7.5996	12.7297
5	-5.6062	-10.5209	-3.0247	-0.1795	4.8566
3	2.1987	-3.0247	-9.1114	-6.0196	-0.9797
6	7.5996	-0.1795	-6.0196	-2.0002	3.1098
8	12.7297	4.8566	-0.9797	3.1098	-1.7955

Positive values show pairs of means that are significantly different.

The plot displays three circles for three groups (number of cylinders). Tukey-Kramer (HSD) is applied to adjust for multiple comparisons to decrease Type I error. In the Tukey table, it display that the means are significant different.

Conclusion: There is strong evidence that the car's miles per gallon are significant different for vehicles with different number of cylinder engines. The 4-cylinder cars have the highest Mpg; the 8-cylinder cars have the lowest Mpg.