Automobiles Data Analysis

Richard Murad

Introduction

The automobile dataset consists of 398 vehicles manufactured from the 1970s to the early 2000s. This dataset includes the following specifications:

Variable Name	Description
mpg	fuel efficiency measured in miles per gallon
cylinders	number of cylinders in engine
displacement	total volume of engine cylinders (in cubic inches)
horsepower	engine power
weight	vehicle weight in pounds
acceleration	time (in seconds) to accelerate from 0 to 60 mph
model_year	model year
origin	country origin of vehicle
car_name	car name

This data set was provided by Dr.Matheny and I do not remember exactly where it came from, but I found it to be interesting because we can see the different specifications of vehicles and analyze how these specifications correlate with the performance of vehicles.

Methods/Expectations

Throughout this analysis, I will be utilizing three very important procedures. A one-way ANOVA test will help us find any significant differences for the average acceleration time between three groups of vehicles: 4-cylinder, 6-cylinder, and 8-cylinder vehicles. It is important to keep in mind that acceleration in this dataset is measured in seconds to accelerate from 0 to 60 miles per hour, so a lower time would mean faster acceleration. I expect all the groups to be significantly different from each other in terms of average acceleration time; 8-cylinder vehicles having the lowest acceleration time, 6-cylinder having the second highest, and 4-cylinder having the highest amount of acceleration time. A two-cylinder difference sounds like it would be significant in the acceleration of a vehicle because more cylinders generate more power. I will also be using multiple linear regression to test if acceleration time has a relationship with fuel efficiency (miles per gallon) and horsepower of a vehicle. I expect acceleration time to increase (slower acceleration) as fuel efficiency increases since acceleration power must be sacrificed for better fuel efficiency. As for horsepower, I expect an inversely proportional relationship with acceleration time because of the increase in vehicle power decreasing acceleration time. Lastly, I will be using a Wilcoxon rank-sum test to see if there is a significant difference in the distribution of fuel efficiency (mpg) between American (1) vehicles and Japanese (3) vehicles. For this method, I do not know what to expect. I do know that Honda and Toyota brands come from Japan and both of those are known for being fuel efficient, but this dataset was collected about 20 years ago so I cannot tell for certain (only very few Hondas and Toyotas are found in this dataset).

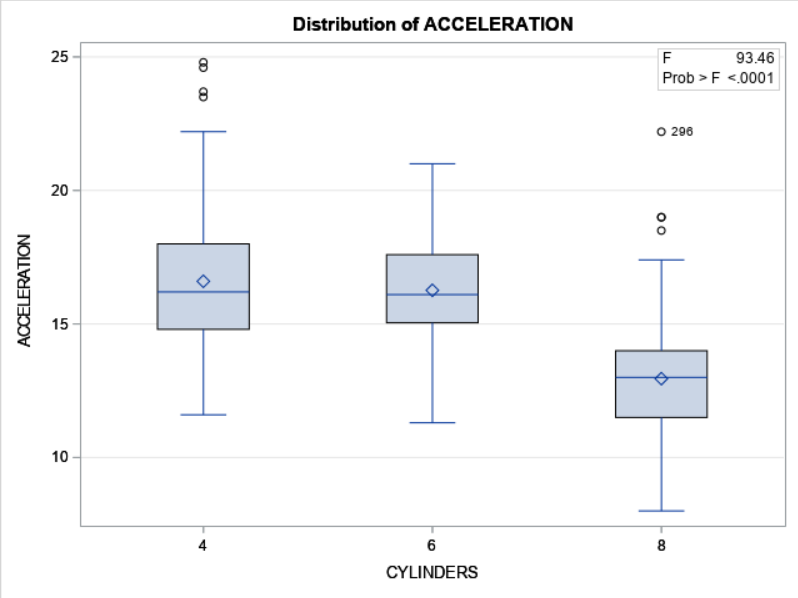
1. Data overview for ANOVA test

Dependent Variable: ACCELERATION ACCELERATION

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	961.547035	480.773517	93.46	<.0001
Error	388	1995.919717	5.144123		
Corrected Total	390	2957.466752			

R-Square	Coeff Var	Root MSE	ACCELERATION Mean
0.325125	14.56850	2.268066	15.56829

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CYLINDERS	2	961.5470347	480.7735174	93.46	<.0001



Level of CYLINDERS	N	ACCELERATION	
		Mean	Std Dev
4	204	16.6014706	2.38220988
6	84	16.2630952	2.02114007
8	103	12.9553398	2.22475943

Comparisons significant at the 0.05 level are indicated by ***.

CYLINDERS Comparison	Difference Between Means	95% Confidence Limits		
4 - 6	0.3384	-0.2397	0.9165	
4 - 8	3.6461	3.1071	4.1851	***
6 - 4	-0.3384	-0.9165	0.2397	
6 - 8	3.3078	2.6522	3.9633	***
8 - 4	-3.6461	-4.1851	-3.1071	***
8 - 6	-3.3078	-3.9633	-2.6522	***

Fisher's Test ^

Comparisons significant at the 0.05 level are indicated by ***.

CYLINDERS Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
4 - 6	0.3384	-0.3534	1.0302	
4 - 8	3.6461	3.0011	4.2911	***
6 - 4	-0.3384	-1.0302	0.3534	
6 - 8	3.3078	2.5233	4.0923	***
8 - 4	-3.6461	-4.2911	-3.0011	***
8 - 6	-3.3078	-4.0923	-2.5233	***

Tukey's Test ^

Assumptions for ANOVA test

It is important to note that the 3- and 5-cylinder vehicles were omitted due to infrequency. All assumptions were satisfied:

Independence: there is no overlap in data observations.

Homogeneity of Variance: as you can see in the output of group standard deviations, the lowest standard deviation multiplied by two is greater than the highest standard deviation (2.02114*2 = 4.04228 > 2.38221).

Normally Distributed Groups: all boxplots show that the mean and median are very close to each other, which is an indicator that all the groups are normal. There is a slight skew in these boxplots, but not enough to violate normality. There are also a few outliers in 4-cylinder and 8-cylinder vehicles according to SAS, but since the sample sizes are above 200 and above 100 for those respected groups, the outliers don't have too much of an effect on these distributions.

Results for ANOVA test

Null Hypothesis: All the vehicle acceleration means are the same for every group of cylinders (4,6, and 8).

Alternative Hypothesis: At least one vehicle acceleration mean group is different than another.

Our result p-value is <0.0001 which results in significant findings (p-value<0.05). This means we reject the null and conclude that 8-cylinder vehicles have a significantly lower acceleration time (in seconds to accelerate from 0 to 60 mph) than 4-cylinder and 6-cylinder vehicles.

Conclusion for ANOVA test

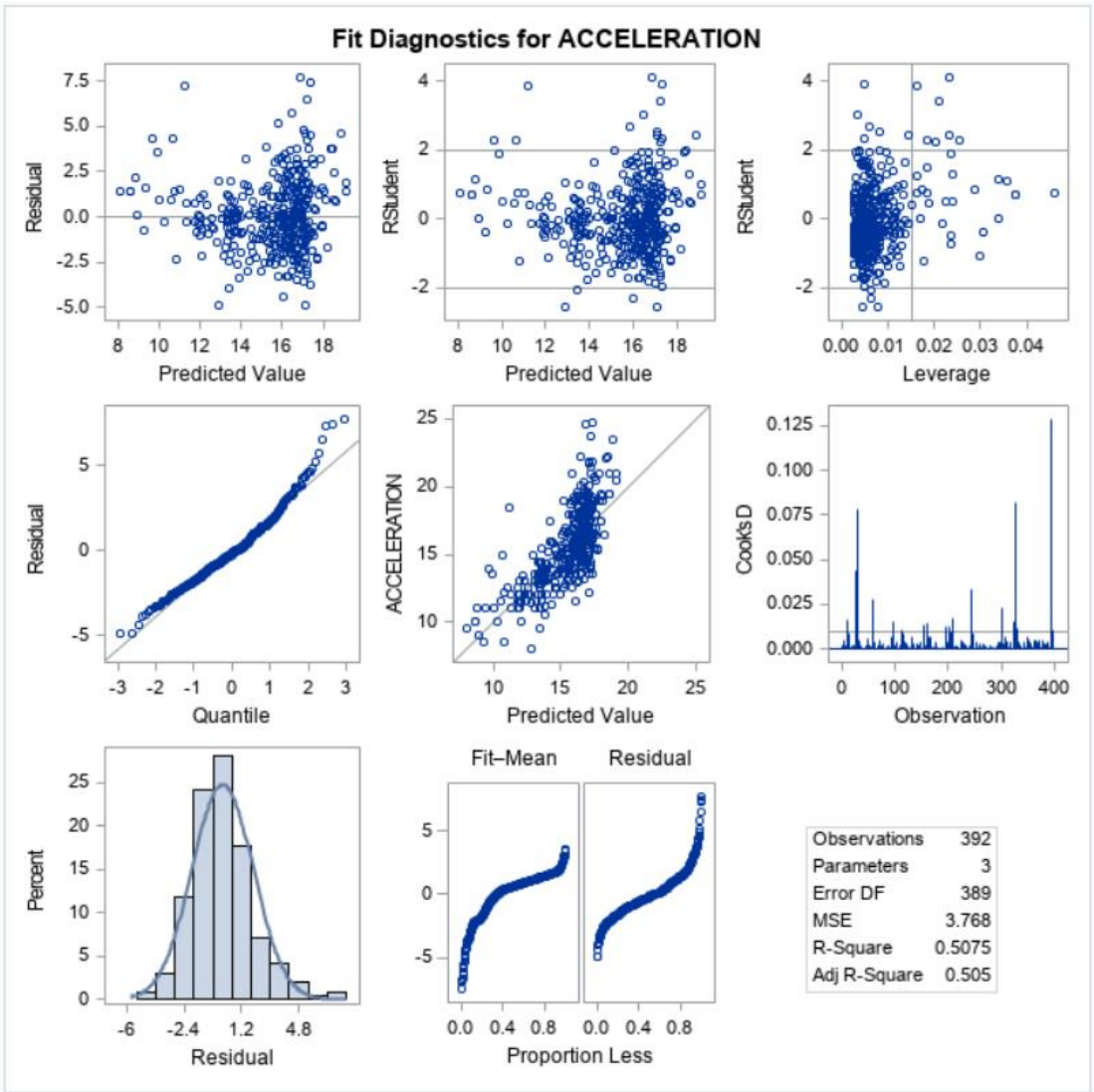
Our results make complete sense. Lower acceleration time from 0 to 60 MPH means faster acceleration, so we would expect 8-cylinder vehicles to be superior to other lower cylinder vehicles. This is because each cylinder has its own power stroke, so the more cylinders you have, the more power your vehicle is generating. Based off my findings, I was a bit surprised that the average 6-cylinder vehicle acceleration time wasn't significantly lower than 4-cylinder vehicles. Since we found a significant difference in 6- and 8-cylinder acceleration mean times, we would expect to find one in 4- and 6-cylinder vehicles as well. This result motivates further research to see why that is the case.

2. Data overview for Multiple Linear Regression test

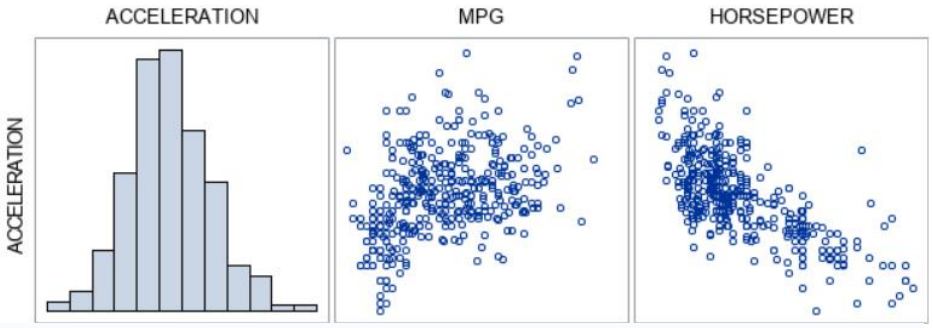
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1510.29527	755.14763	200.41	<.0001
Error	389	1465.73524	3.76796		
Corrected Total	391	2976.03051			

Root MSE	1.94112	R-Square	0.5075
Dependent Mean	15.54133	Adj R-Sq	0.5050
Coeff Var	12.49007		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	24.75570	0.84903	29.16	<.0001	0
MPG	MPG	1	-0.10151	0.02004	-5.07	<.0001	2.53774
HORSEPOWER	HORSEPOWER	1	-0.06542	0.00406	-16.10	<.0001	2.53774

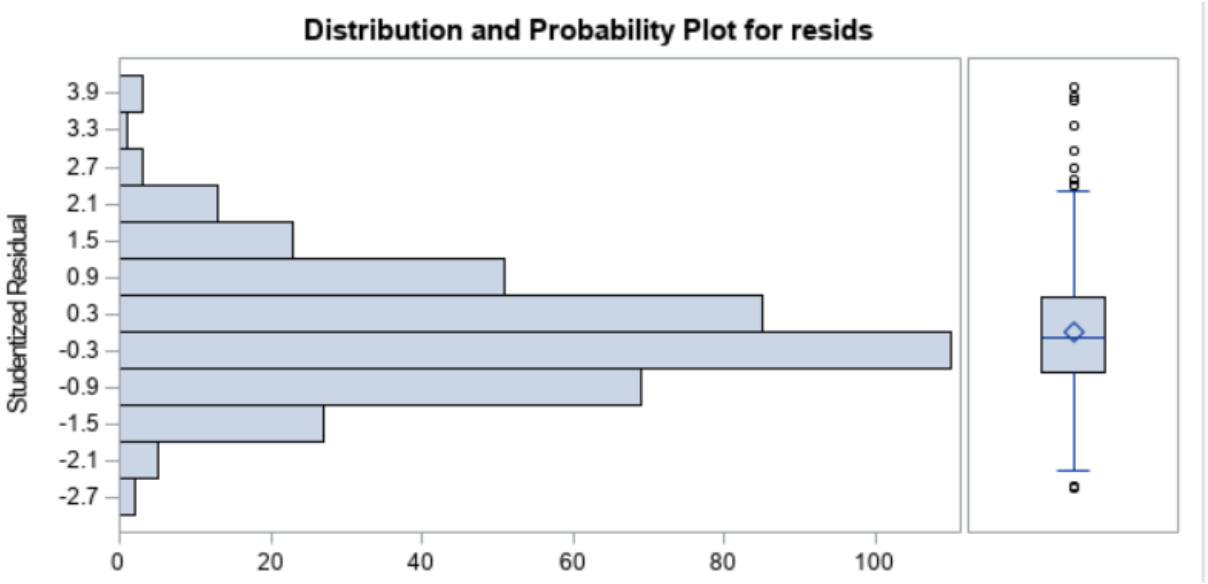


	ACCELERATION	MPG	HORSEPOWER
ACCELERATION	1.00000	0.42029	-0.68920
ACCELERATION		<.0001	<.0001
	398	398	392



Moments			
N	392	Sum Weights	392
Mean	0.00077453	Sum Observations	0.30361489
Std Deviation	1.00215322	Variance	1.00431107
Skewness	0.75850318	Kurtosis	1.47235778
Uncorrected SS	392.685863	Corrected SS	392.685628
Coeff Variation	129388.93	Std Error Mean	0.05061638

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.968426	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.076583	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.437524	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.621617	Pr > A-Sq	<0.0050



Assumptions for Multiple Linear Regression Test

Linearity: This assumption meets the requirements since the r value for the mpg variable shows a moderate positive linear relationship (0.42029) and the horsepower variable shows a stronger linear relationship (-0.68920) but this time it's negative. Logically this makes sense because the more horsepower your vehicle has, the faster it will accelerate from 0 to 60 mph, resulting in less time. As far as MPG (fuel efficiency), we notice that acceleration time typically increases as fuel efficiency increases. This also makes sense since there is typically a trade-off between fuel-efficiency and acceleration.

Normality of Residuals: Skewness (0.7585) and Kurtosis (1.4724) value not too extreme. Not much skewness in histogram and boxplot; mean and median are close together. Looking at the studentized residual plot, there are only a couple of outliers greater than or less than 3 standard deviations away from the mean. Sample size is very large (392) so the outliers don't have much effect. Q-Q plot has about 2-4% deviation in the left tail and 10-15% deviation in the right tail. Tests of normality violate the assumption of normality, but these tests are sensitive to sample size, and we have a large sample size. Based off this evidence, it is safe to say that the assumption of residual normality is met.

Residual Equality of Variance: This assumption is completely violated since the residual plot shows a "funnel shape" distribution of variances, meaning that the variances are not equal throughout the range of predicted values.

Residual Independence: There is obvious clustering on the right side of the residual plot, so this violates the assumption of residual independence.

Collinearity: both VIF values are 2.53774 and these values are well below 10, so the collinearity assumption is not violated.

Results for Multiple Linear Regression Test

Null Hypothesis: There are no significant predictors of acceleration.

Alternative Hypothesis: There is at least one significant predictor of acceleration.

All p-values are <.0001, which means we reject null: the overall model is significant and both mpg and horsepower are significant predictors of acceleration. More specifically:

50.75% of the variation in a vehicle's acceleration time (from 0 to 60 mph) is explained by fuel efficiency (in mpg) and horsepower. (r^2=0.5075).

For every mile per gallon increase in a vehicle, there is on average a 0.10 decrease in acceleration time from 0 to 60 mph while all other variables are held constant.

For every horsepower increase in a vehicle, there is on average a 0.065 decrease in acceleration time from 0 to 60 mph while all other variables are held constant.

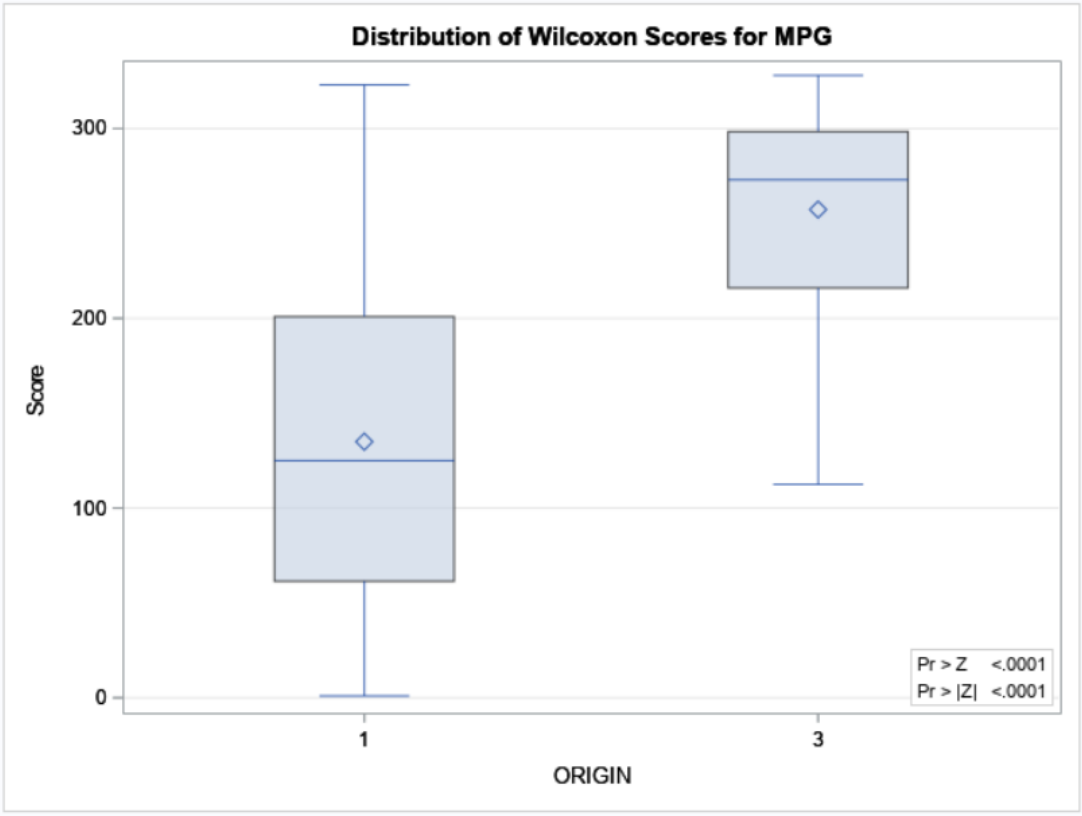
Conclusion for Multiple Linear Regression Test

This is an inaccurate result because logically you would expect longer acceleration time as miles per gallon increases. Anytime you accelerate a vehicle, the engine will work harder to burn more fuel. Of course, fuel efficient cars will oppose that type of fuel burning, otherwise it wouldn't be as efficient. Here we see the opposite; our acceleration time is a tenth of a second less (faster acceleration) for every mile per gallon increase. Even the scatterplot matrix (r=0.42) shows that this can't be true. The reason why we got these values is the fact that two assumptions were violated before the model was generated; this will lead to an unreliable and inaccurate model. The alternative approach of this analysis is to use other methods when an assumption violation is encountered; for example, the weighted least squares regression technique can be used to satisfy the violation of homoscedasticity.

3. Data overview for the Wilcoxon rank-sum test

Wilcoxon Scores (Rank Sums) for Variable MPG Classified by Variable ORIGIN					
ORIGIN	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	249	33624.50	40960.50	734.058017	135.038153
3	79	20331.50	12995.50	734.058017	257.360759
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	20331.5000
Normal Approximation	
Z	9.9931
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
t Approximation	
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
Z includes a continuity correction of 0.5.	



Assumptions for the Wilcoxon rank-sum test

There are no assumptions for non-parametric testing, but it is important to note that since there is some obvious skewing in the distribution of Japanese vehicles (3) for MPG, it might be a better idea to use the median since it's less affected by skewness compared to the mean.

Results for the Wilcoxon rank-sum test

Null Hypothesis: The median fuel efficiency (mpg) of vehicles originating from America (1) and Japan (3) are the same.

Alternative Hypothesis: The median fuel efficiency (mpg) of vehicles originating from America (1) and Japan (3) are not the same.

Our result p-value is <0.0001 which results in significant findings (p-value<0.05). This means we reject the null and conclude that the cars originating from Japan have a significantly higher median fuel efficiency (mpg) than cars from America (from the 1970s to 2000s).

Conclusion for the Wilcoxon rank-sum test

Our result influences research as to why Japan had significantly higher median fuel efficiency than American cars. There can be many factors that explain this like government regulations, fuel prices, etc. I wanted to do a Kruskal-Wallis test and add European cars in the mix, but I haven't figured out the correct code to do post hoc testing. One suggestion for further analysis of this data is to test the statistical power of this procedure.

SAS CODE APPENDIX

```
data A_subset_cylinders;
     set Automobiles;
     if cylinders in (4,6,8);
     run;

data A_subset_origin;
     set Automobiles;
     if origin in (1,3);
     run;

proc anova data= A_subset_cylinders;
     class cylinders;
     model acceleration=cylinders;
     means cylinders / tukey lsd cldiff;
     run;

PROC CORR data=Automobiles pearson plots=matrix(histogram);
     var acceleration mpg horsepower;
     RUN;

proc reg data=Automobiles;
     model acceleration = mpg horsepower / vif;
     output out=residuals p = predict student=resids;
     run;

PROC UNIVARIATE data=residuals normal plot;
     var resids;
     RUN;

proc npar1way data=A_subset_origin wilcoxon ;
     var mpg;
     class origin;
     run;
```