Statistical Modeling and Analysis Results for Car Prices
MSDS 6372

Submitted to:
Dr. Martin Selzer

Report prepared by:
Tanvi Arora
Rebecca Holsapple
Anjli Solsi

June 20, 2018

Introduction

This report summarizes the statistical modeling and analysis results for the data set on car specifications. Analysis of the data is limited to knowledge and techniques learned in MSDS 6371 and 6372. The purpose of this report is to document the detailed analysis of the proposed questions regarding the data set.

For the Toyota VP Analysis Question 1, an explanation is made to quantify the relationship between horsepower of a car and miles per gallon in the city, considering the dependence of that association on fuel type. This model also accounts for the weight of the car.

For the Toyota VP Analysis Question 2, a description is given to determine the association between body style, excluding convertibles, of a car and mean miles per gallon that can be achieved in the city. This model is used to estimate the difference between mean city miles per gallon between the averages of the various body styles.

For the Predictive Analysis Question, this focuses on predicting the sales prices of cars for the test dataset by creating models with known techniques from the training dataset. The variable selection methods used are: LASSO, LAR, and OLS-stepwise selection. These models are compared using the Adjusted R^2 values, Cross Validation MSPE, and AIC statistics.

The final section includes an Appendix with SAS code for each analysis question and screenshots with additional details.

Note: *Italicized font* represents supplemental tables and figures found in the Appendix.

Data Description

This dataset was created from the "1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook." It has been adapted and provided to us in this format by the SMU MSDS program.

With 24 explanatory variables, describing numerous aspects of model import car and truck specifications; there is a training dataset with 179 observations and the resulting sale price. The test dataset of 24 records is provided with the exact same 24 explanatory variables, but no price is detailed.

Name	Type / Levels	Range	Description
symboling	Int	-3, -2, -1, 0, 1, 2, 3	degree to which the car is more risky than price indicates
make	Factor, 21 levels	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo	brand of vehicle
fuel-type	Factor, 3 levels	diesel, gas	fuel used to provide power

aspiration	Factor, 2 levels	std, turbo	type of internal combustion engine
num-of-doors	Factor, 3 levels	four, two	number of doors of car
body-style	Factor, 5 levels	hardtop, wagon, sedan, hatchback, convertible	shape of car
drive-wheels	Factor, 3 levels	4wd, fwd, rwd	traction of car
engine-location	Factor, 2 levels	front, rear	placement of engine
wheel-base	Num	continuous 86.6 – 120.9	distance between centers of front and rear wheels
length	Num	continuous 141.1 – 208.1	length of car from front to back
width	Num	continuous 60.3 – 72.3	width of car from driver side to passenger side
height	Num	continuous 47.8 – 59.8	length of car roof from road
curb-weight	Factor, 154 levels	continuous 1488 – 4066	total weight of vehicle with standard equipment
engine-typ	Factor, 6 levels	dohc, dohcv, I, ohc, ohcf, ohcv, rotor	type of engine, assembly, design
num-of-cylineders	Factor, 7 levels	eight, five, four, six, three, twelve, two	number of cylinders related to the engine, dictates power
engine-size	Int	continuous 61 – 326	volume of all cylinders in an engine
fuel-system	Factor, 8 levels	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi	how fuel is stored and supplied to the engine
bore	Factor, 38 levels	continuous 2.54 – 3.94	size (diameter) of cylinder in which piston travels
stroke	Factor, 35 levels	continuous 2.07 – 4.17	distance travelled by piston in each cycle of the engine
compression-ratio	Num	continuous 7 – 23	ratio of the volume of the cylinder and combustion chamber when the piston is at the bottom/top
horsepower	Factor, 58 levels	continuous 48 – 288	power an engine produces
peak-rpm	Int	continuous 4150 – 6600	revolutions per minute, how fast vehicle's crankshaft is spinning as car accelerates/decelerates
city-mpg	Int	continuous 13 – 49	miles per gallon in the city

highway-mpg	Int	continuous 16 – 54	miles per gallon on the highway
price	Int	continuous 5118 - 45400	price of the car

Table 1 Description of the Explanatory Variables in the Dataset

Files used:
Train_Auto.csv – the training dataset
Test_Auto.csv – the test dataset

Data Cleaning

The original data in Train_Auto.csv contained a few syntax errors that were not accepted by SAS. First, the column names needed to be changed in order to be read in by SAS. The 15 explanatory variable names that were separated by dashes were changed to underscores; i.e. num-of-doors being changed to num_of_doors. The data was checked to ensure all categorical variables had consistent values.

Exploratory Analysis

To create a model suited to the dataset, the data was examined and transformed in various ways. The training dataset contained 17 missing values. Those missing values were distributed across the following variables: fuel-type (3), num-of-doors (1), curb-weight (3), bore (4), stroke (4), and horsepower (2). Imputation was performed to assign values to those that were missing; the process will be detailed in the next section.

The adequacy of the fit was examined using an ANOVA, which yielded an F-statistic of 64.03 and p-value less than 0.0001 as shown in Table 2. This suggested there was not enough evidence to support that a difference did not exist in the price model of the various cars. Next, the parameter estimates, and plots were reviewed to determine whether a correlation was present between variables.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	9847000014	703357144	64.03	<.0001
Error	163	1790583287	10985174		
Corrected Total	177	11637583301			

Root MSE	3314.38886	R-Square	0.8461
Dependent Mean	13366	Adj R-Sq	0.8329
Coeff Var	24.79790		

Table 2 ANOVA to determine Adequacy of the Fit

Based on the VIF values in Table 3 and correlation values in *Table 4*, city_mpg and highway_mpg are highly correlated, as well as curb_weight and length. Although signs of multicollinearity were present, none of the variables were removed, but rather left to the model to validate. Once validated by the model, there was no correlation present. The left plot in *Figure 1.1* demonstrates a linear relationship between curb_weight and wheel_base, length, and width. Since the relationships are mainly linear, no transformations were applied to those variables. Based on the right plot in *Figure 1.1*, a log transformation was applied to peak_rpm, but that did not significantly change the plot, so the original data was used. Due to the curved

relation of horsepower, a log transform was applied to the variable. No transformations were performed on the variables in *Figure 1.2*.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-55924	17502	-3.20	0.0017	0
symboling	1	316.98140	276.78910	1.15	0.2538	1.91750
wheel_base	1	165.31804	121.42471	1.36	0.1752	9.07050
length	1	-72.24967	65.54863	-1.10	0.2720	10.68392
width	1	566.27215	283.84718	1.99	0.0477	5.81719
height	1	188.14827	154.68243	1.22	0.2256	2.29894
curb_weight	1	0.83030	1.86366	0.45	0.6565	14.96856
engine_size	1	117.43392	15.32124	7.66	<.0001	6.94601
bore	1	-1548.14270	1447.81255	-1.07	0.2865	2.33659
stroke	1	-3348.13518	936.95907	-3.57	0.0005	1.29898
compression_ratio	1	277.01187	89.66330	3.09	0.0024	2.05544
horsepower	1	40.32334	19.26843	2.09	0.0379	9.00915
peak_rpm	1	2.09576	0.74800	2.80	0.0057	2.05065
city_mpg	1	-395.84509	196.24665	-2.02	0.0453	27.31338
highway_mpg	1	294.17984	176.18548	1.67	0.0969	24.71534

Table 3 Parameter Estimates and VIF values

Table 2 and Figure 2.1 represent the model of the original data. Looking at the same variables as before, some estimates, noted in *Table 5*, have statistically significant p-values, and some do not. The residual plot in Figure 2.1 has clustered observations and does not resemble a random spread of data. This does not meet the constant variance assumption; therefore, a log transform will be applied to price. The RStudent plot shows a couple observations that could be outliers, while the QQ-plot shows two distinct outliers. The leverage plot displays a couple observations, but none seem to be large enough to consider. The Cook's D plot shows a single observation with high leverage, but the actual value of the data point is not large.

Figure 2.2 represents the model with a log transformation performed on price. The residual plot displays a random spread, so the constant variance assumption is met. The RStudent plot still shows a few outliers, which will be explored. The QQ-plot is nearly straight supporting the assumption of normality. The leverage and Cook's D plot are similar to the original data and will be explored further. Table 6 displays the statistics and *Table 7* the parameter estimates with the log transformation, demonstrating the statistically significant variables in the model.

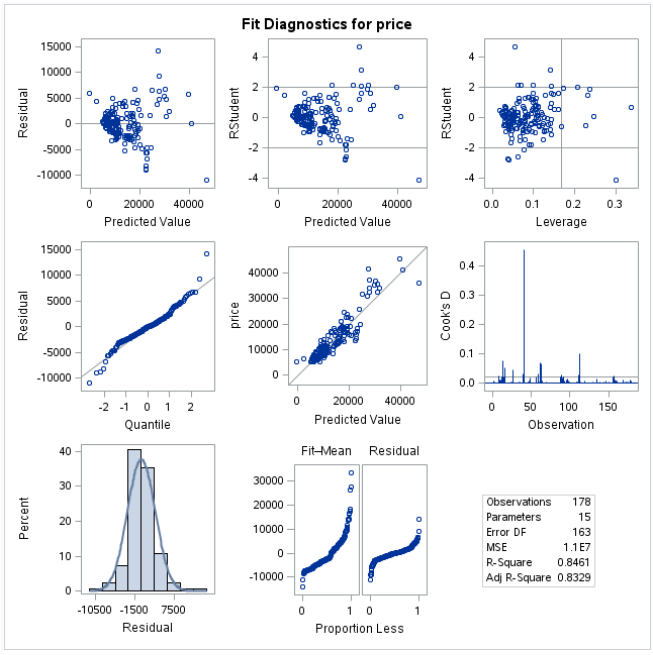

Figure 2.1 Fit Diagnostic Plots for Original Data

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	40.08723	2.86337	87.35	<.0001
Error	163	5.34344	0.03278		
Corrected Total	177	45.43067			

Root MSE	0.18106	R-Square	0.8824
Dependent Mean	9.35992	Adj R-Sq	0.8723
Coeff Var	1.93439		

Table 6 Statistics for Log Transformed Data

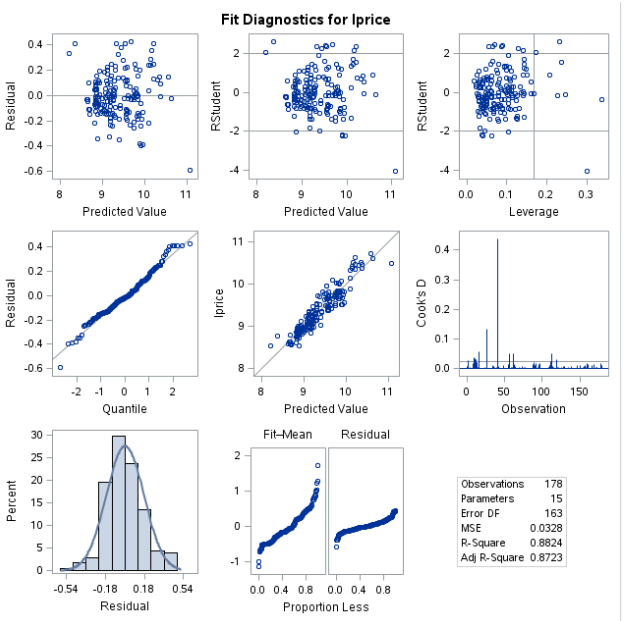

Figure 2.2 Build Model with Log Transformation on price

Based on the studentized residual plot in Figure 3, observations 134 and 41 demonstrate the largest leverages of the data. The Cook's D plot in Figure 3 shows observation 41 having the largest value. Since the value of observation 41 is influential and unique from all the other cars by having 12 cylinders, models will be created without this value.

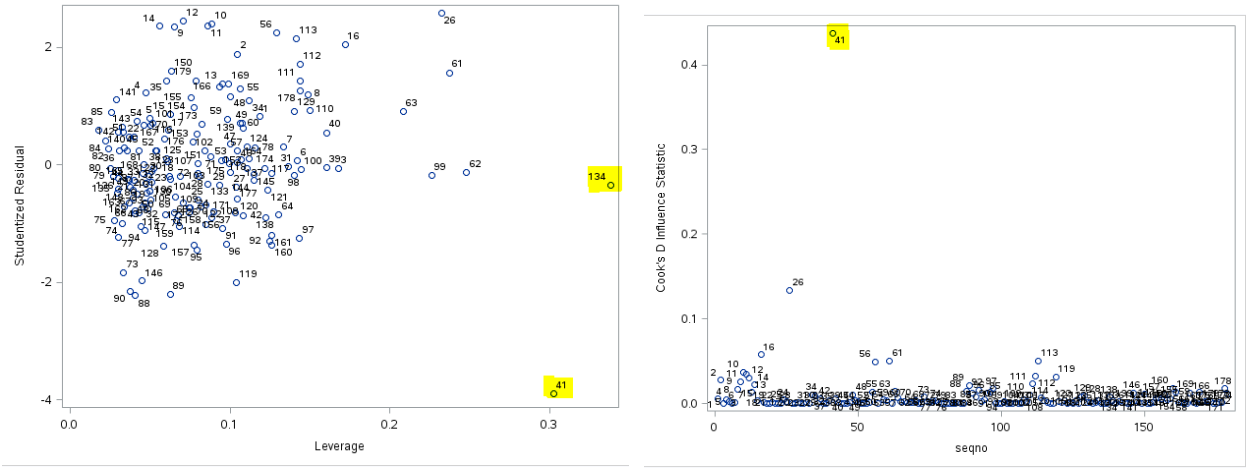

Figure 3 Studentized Residual and Cook's D Plot for Log Transformed Data

Models and Fit

The models required imputation to assign values to the missing data. Categorical variables were imputed with the mode of the respective variable. Quantitative (continuous) variables were imputed with the mean of a single or grouping of variables. Table 8 below identifies the imputed values.

Dataset		Train_Auto.csv	
Field	**Type**	**Formula**	**Final value imputed**
Fuel-type	categorical	Mode of fuel-type	*gas* (Obs 9,58,65)
Num-of-doors	categorical	Mode of num-of-doors	*four* (Obs 53)
Curb-weight	continuous	curb weight is the weight of car with standard equipment, so we take mean of all cars grouped by make,aspiration and body-style	2843 for make=audi (Obs 6) 2968 for make=bmw (Obs 13) 2266 for make=mazda (Obs 46)
bore	continuous	Mean of cars grouped by make	3.21 for make-mazda (Obs 46,47,48,49)
stroke	continuous	Mean of cars grouped by make	3.27 for make-mazda (Obs 46,47,48,49)
horsepower	continuous	horsepower is a property of engine type and engine size (proportional to number of cylinders)	90 for make = isuzu (Obs 38) 118 for make = mercedes-benz (Obs 58)
Dataset	Test_Auto.csv		
Field	**Type**	**Formula**	**Final value imputed**
Curb-weight	continuous	curb weight is the weight of car with standard equipment, so we take mean of all cars grouped	2478 (Obs 6)

		by make,aspiration and body-style	
make	categorical	Not changed. This is treated as *missing at random* and will be handled as an exception during prediction modelling.	

Table 8 Data Imputations performed on Missing Values

Train_Auto.csv was used to generate the model variables using three model selection techniques incorporating cross validation. The code used to create the following models creates a training data set with 75 percent of the data and allocates another 25 percent to the test data. Table 9 below summarizes the models by providing parameter estimates and statistics on the three model selection techniques used.

LASSO	LAR	OLS (Stepwise)
Parameters: width, curb_weight, horsepower	Parameters: width, curb_weight, horsepower	Parameters: wheel_base, height, curb_weight, horsepower, make, aspiration, body_style

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 3.

Effects:	Intercept width curb_weight horsepower

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	28.64395	9.54798	177.26
Error	128	6.89467	0.05386	
Corrected Total	131	35.53862		

Root MSE	0.23209
Dependent Mean	9.34774
R-Square	0.8060
Adj R-Sq	0.8014
AIC	-247.67100
AICC	-247.19481
SBC	-370.13980
ASE (Train)	0.05223
ASE (Test)	0.05322
CV PRESS	4.59387

Cross Validation Details

Index	Observations Fitted	Observations Left Out	CV PRESS
1	106	26	1.1961
2	105	27	0.5467
3	98	34	1.2807
4	110	22	0.7261
5	109	23	0.8443
Total			4.5939

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 3.

Effects: Intercept width curb_weight horsepower

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	28.64395	9.54798	177.26
Error	128	6.89467	0.05386	
Corrected Total	131	35.53862		

Root MSE	0.23209
Dependent Mean	9.34774
R-Square	0.8060
Adj R-Sq	0.8014
AIC	-247.67100
AICC	-247.19481
SBC	-370.13980
ASE (Train)	0.05223
ASE (Test)	0.05322
CV PRESS	4.59387

Cross Validation Details

Index	Observations Fitted	Observations Left Out	CV PRESS
1	106	26	1.1961
2	105	27	0.5467
3	98	34	1.2807
4	110	22	0.7261
5	109	23	0.8443
Total			4.5939

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 7.

Effects:	Intercept wheel_base height curb_weight horsepower make aspiration body_style

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	28	34.30707	1.22525	102.47
Error	103	1.23155	0.01196	
Corrected Total	131	35.53862		

Root MSE	0.10935
Dependent Mean	9.34774
R-Square	0.9653
Adj R-Sq	0.9559
AIC	-425.03794
AICC	-406.62210
SBC	-475.43669
ASE (Train)	0.00933
ASE (Test)	0.02281
CV PRESS	2.07252

Cross Validation Details

Index	Observations Fitted	Observations Left Out	CV PRESS
1	106	26	0.4424
2	105	27	0.4645
3	98	34	0.5929
4	110	22	0.2534
5	109	23	0.3194
Total			2.0725

Parameter Estimates in *Figure 4*.

Parameter Estimates Parameter / DF / Estimate Intercept 1 6.694032 width 1 0.020146 curb_weight 1 0.000376 horsepower 1 0.003545	Parameter Estimates Parameter / DF / Estimate Intercept 1 6.694032 width 1 0.020146 curb_weight 1 0.000376 horsepower 1 0.003545	

Table 9 Model Selection Techniques Implemented with Statistical Output

Toyota VP Questions

Analysis 1

The first question focuses on quantifying how horsepower is related to the car's miles per gallon in the city. The linear regression model is: $\log(city\_mpg) = \beta_0 + \beta_1 \log(horsepower) + \beta_2(curb\_weight)$.

The assumptions are verified using the plots in Figure 5. Judging the scatter plot, QQ-plot, and histogram of residuals, there is a slight right skew, but not strong enough evidence against the assumption of normality. Examining the pairwise scatterplots, there is a strong linear trend between the log of horsepower and the log of city_mpg. Based on the residual plot in Figure 5, there is little evidence of heteroscedasticity. We will assume independence within and between observations. There are two noticeable observations of concern and high leverage, observations 26 and 41. The general trend seems to follow that as horsepower increases, the miles per gallon in the city decreases.

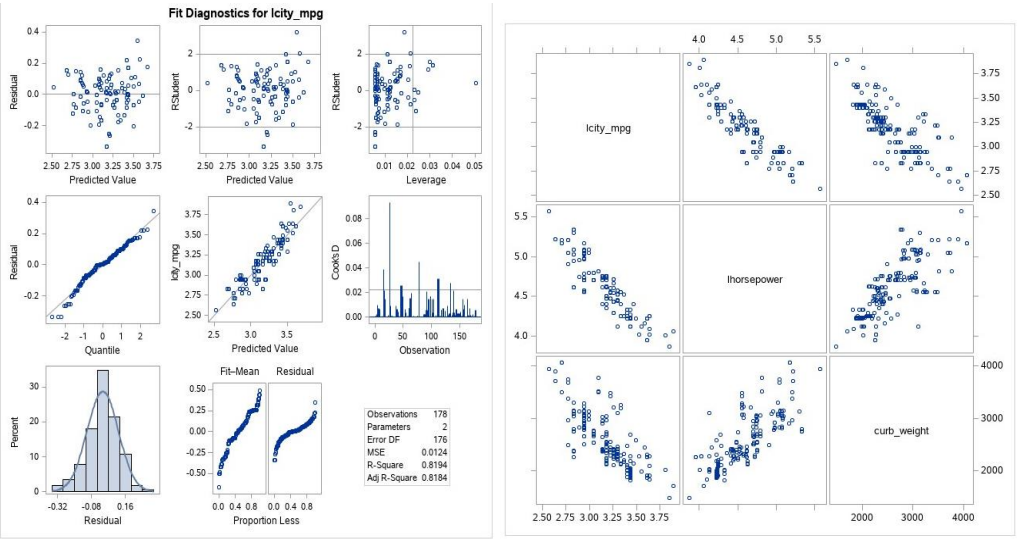

Figure 5 Fit Diagnostic and Matrix Plots for Analysis 1 Assumptions

Based on the developed model, one would reject the null hypothesis. Holding the curb weight constant, a doubling of the vehicle's horsepower equates to a multiplicative change of $2^{-0.560} = 0.678$ in the median of the distribution vehicle's miles per gallon in the city. In other words, a doubling of horsepower decreases the estimated median of city miles per gallon 67.8%. A 95% confidence interval for β_1 is (-0.632,-0.488). Therefore a 95% confidence interval for the median decrease after a doubling of horsepower is:$\left(2^{-0.632}, 2^{-0.488}\right) = (28.7\%, 35.5\%)$. The interaction between the horsepower and fuel type are not statistically significant (p-value = 0.9072). *Table 10* provides the necessary statistics to calculate the following values.

Analysis 2

The second question focuses on analyzing how the body style, excluding convertibles, of the car is associated with the mean miles per gallon achieved in the city. Out of the 179 observations, excluding the convertible cars, the model uses 173 remaining observations. The model has a very low F-statistic and is not significant, according to the left side of Table 11. At a p-value of 0.1590 for the α=0.05 level of significance, we fail to reject the null hypothesis, i.e. there is not sufficient evidence to show that the difference in mean miles per gallon in the city is different for cars with different body styles, excluding the convertibles.

For the results of the two-way ANOVA on the right of Table 11, the model improves by including fuel_type. The F-statistic rises enough to lower the p-value to 0.0001, making the model significant. As per the model statistics, at a p-value of 0.1220, body style is still not significant enough to explain the mpg in the city. Fuel type is a highly significant variable in this model with a p-value to 0.0001. The interaction between body style and fuel type is not significant enough to say that mean miles per gallon in the city of a car is dependent on body style.

Dependent Variable: city_mpg

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	229.268050	76.422683	1.75	0.1590
Error	169	7386.177037	43.705190		
Corrected Total	172	7615.445087			

R-Square	Coeff Var	Root MSE	city_mpg Mean
0.030106	26.25577	6.610990	25.17919

Source	DF	Type I SS	Mean Square	F Value	Pr > F
body_style	3	229.2680497	76.4226832	1.75	0.1590

Source	DF	Type III SS	Mean Square	F Value	Pr > F
body_style	3	229.2680497	76.4226832	1.75	0.1590

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	23.71428571	B	1.44263630	16.44	<.0001
body_style hardtop	-2.42857143	B	2.88527261	-0.84	0.4011
body_style hatchback	2.60774818	B	1.67987066	1.55	0.1224
body_style sedan	1.35548173	B	1.60916108	0.84	0.4008
body_style wagon	0.00000000	B	.	.	.

Analyze effects for Body_style and fuel_type on city_mpg

The GLM Procedure

Dependent Variable: city_mpg

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1142.015478	190.335913	4.88	0.0001
Error	166	6473.429609	38.996564		
Corrected Total	172	7615.445087			

R-Square	Coeff Var	Root MSE	city_mpg Mean
0.149960	24.80113	6.244723	25.17919

Source	DF	Type I SS	Mean Square	F Value	Pr > F
body_style	3	229.2680497	76.4226832	1.96	0.1220
fuel_type	1	822.7173405	822.7173405	21.10	<.0001
body_style*fuel_type	2	90.0300880	45.0150440	1.15	0.3178

Source	DF	Type III SS	Mean Square	F Value	Pr > F
body_style	3	232.9932896	77.6644299	1.99	0.1172
fuel_type	1	276.8752772	276.8752772	7.10	0.0085
body_style*fuel_type	2	90.0300880	45.0150440	1.15	0.3178

Table 11 Parameter Estimates and Related Statistics for Analysis 2

Figure 6 shows the interaction plot for mpg in the city for different body styles and fuel type. Hardtop cars have low mileage per gallon in the city. There are no samples in the dataset for hardtop cars using diesel. Hatchback cars have high mileage per gallon in the city for some cars, but the mean mpg in the city is lower for cars using gas than cars using diesel. Sedans using diesel have lower mean mpg in the city as compared to the hatchback, but better mean mpg as compared to sedans using gas. Mean mpg in the city for wagons is almost the same for cars using gas or diesel. While there seems to be some relation from the interaction plot, it is not significant and does not provide sufficient evidence to suggest that body styles affect the mean mpg in the city.

Applying a Bonferroni adjustment to the differences of mean mileage per gallon in the city for different body styles, the difference between the hatchback and wagon is the only one that is slightly significant at a p-value of 0.0486. Considering the effect of fuel type,

the differences between hardtop models using gas are moderately significant compared to hatchbacks using diesel with a p-value of 0.0133, and slightly significant with hatchbacks using gas at a p-value of 0.0547. The differences between hardtop models using gas are significant with sedan models using diesel, demonstrated by a p-value of 0.0003. The difference between the hatchback diesel model is moderately significant when compared to wagons using gas, having a p-value of 0.0257. However, the differences between hatchback models using either gas or diesel are both significant when compared to sedans using either gas (p-value of 0.0329) or diesel (p-value of 0.0024). The differences between sedans using diesel and sedans using gas are highly significant, having a p-value less than 0.0001. Another moderately significant difference is between sedans using diesel and wagons using gas with a p-value of 0.0002. This output is seen in Table 12 and demonstrated differently in *Table 13*.

Least Squares Means

Effect	body_style	fuel_type	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
body_style	hardtop		Non-est	.	.	.	.	.	.	.
body_style	hatchback		32.0603	3.1492	166	10.18	<.0001	0.05	25.8428	38.2779
body_style	sedan		27.8036	0.9120	166	30.49	<.0001	0.05	26.0029	29.6042
body_style	wagon		24.2895	2.3211	166	10.46	<.0001	0.05	19.7067	28.8722
fuel_type		die	Non-est	.	.	.	.	.	.	.
fuel_type		gas	23.6838	0.7432	166	31.87	<.0001	0.05	22.2165	25.1512
body_style*fuel_type	hardtop	gas	21.2857	2.3603	166	9.02	<.0001	0.05	16.6257	25.9458
body_style*fuel_type	hatchback	die	38.0000	6.2447	166	6.09	<.0001	0.05	25.6707	50.3293
body_style*fuel_type	hatchback	gas	26.1207	0.8200	166	31.86	<.0001	0.05	24.5018	27.7396
body_style*fuel_type	sedan	die	31.8571	1.6690	166	19.09	<.0001	0.05	28.5620	35.1523
body_style*fuel_type	sedan	gas	23.7500	0.7359	166	32.27	<.0001	0.05	22.2970	25.2030
body_style*fuel_type	wagon	die	25.0000	4.4157	166	5.66	<.0001	0.05	16.2819	33.7181
body_style*fuel_type	wagon	gas	23.5789	1.4326	166	16.46	<.0001	0.05	20.7504	26.4075

Table 12 Bonferroni Adjustment applied to Analysis 2 – Least Squares Means

The second part of this analysis focuses on estimating the difference between mean city miles per gallon between the average of hardtop, hatchback, sedans, and wagons. Based on Table 14, the difference between body style and mean city miles per gallon is evident. The most statistically significant contrast is that between the hatchback and the average of the hardtop, sedan, and wagon, with a p-value of 0.0059. The least significant is the comparison of sedan cars to the rest of the body styles.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
hatchback	1	304.3910944	304.3910944	7.79	0.0059
hardtop	1	62.3973988	62.3973988	1.60	0.2081
sedan	1	3.8514535	3.8514535	0.10	0.7539
wagon	1	8.4124010	8.4124010	0.22	0.6432

Table 14 Contrast of difference of city_mpg between body_style

Predictive Models

In order to apply the selection methods detailed below, the dataset had to be verified for fit, assumptions, and outliers identified. This part of the analysis was performed in the exploratory phase of the data. Table 2 and Figure 2.1 represent the model of the original data. The residual plot in Figure 2.1 has clustered observations and does not resemble a random spread of data. This does not meet the constant variance assumption; therefore, a log transform will be applied to price. The RStudent plot shows a couple observations that could be outliers, while the QQ-plot shows two distinct outliers. The leverage plot displays a couple observations, but none seem to be large enough to consider. The Cook's D plot shows a single observation with high leverage, but the actual value of the data point is not large.

Based on the studentized residual plot in Figure 8, observation 41 demonstrates the largest leverage of the data. This is also supported by the Cook's D plot in Figure 3. Since the value of observation 41 is influential and unique from all the other cars by having 12 cylinders, models will be created without this value.

Having performed a log transformation on price and removing observation 41, the assumptions for the model can be verified. Based on the plots in Figure 2.1 and 7, the Cook's D value decreases from 0.4 to 0.15 by implementing these changes, as well as the leverage decreasing. Figure 7 represents the model with a log transformation performed on price and observation 41 removed. The residual plot displays a random spread of data, so constant variance will be assumed. The RStudent plot shows a few outliers; however, since the observations lie between -2 and +2, this is acceptable. The histogram and nearly straight QQ-plot support the assumption of normality. Independence will be assumed within and between data.

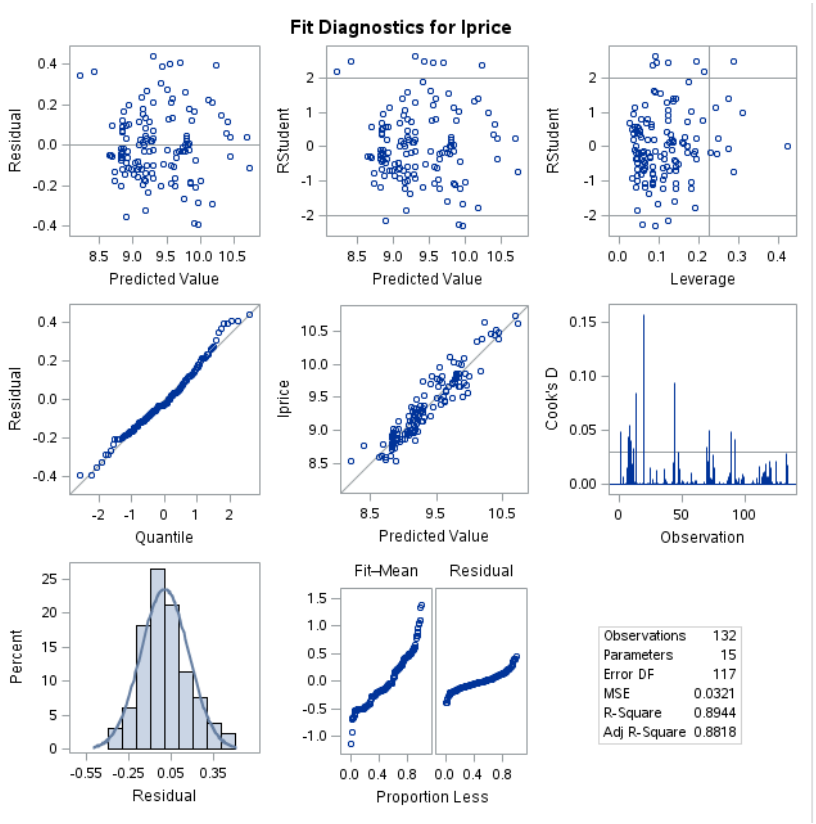

Figure 7 Fit Diagnostic Plots for Data used for Selection Techniques

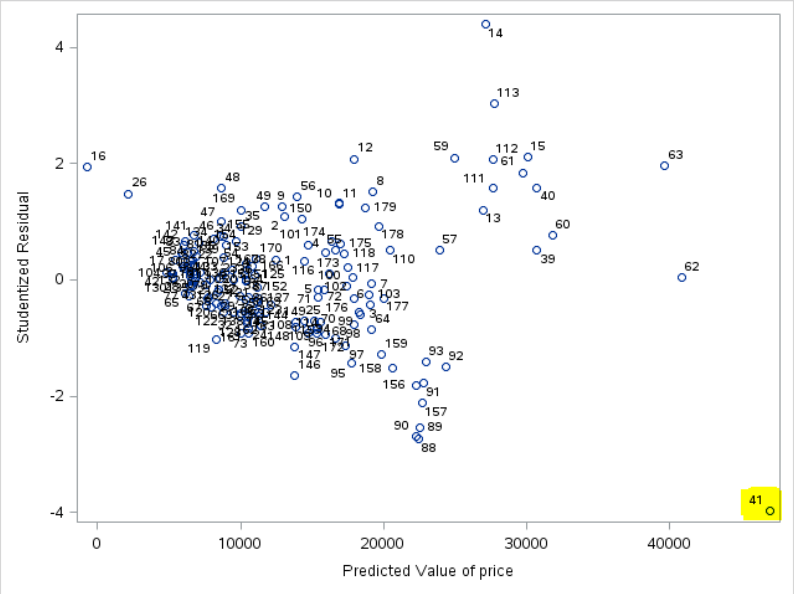

Figure 8 Studentized Residuals of Data Used for Selection Techniques

The three model selection techniques performed for this predictive analysis were the LASSO method, LAR method, ad OLS (Stepwise) selection. Table 15 shows the statistics used to determine the final model out of the competing models. The statistics used were external cross validation, the adjusted R^2 value, and the AIC. It should be noted that the models and statistics were identical for the LASSO and LAR models.

Predictive Models	Adjusted R^2	MSE	Cross Validation MSPE	AIC
Model 1 - LASSO	0.8014	0.23209	0.05322	-247.671
Model 2 – LAR	0.8014	0.23209	0.05322	-247.671
Model 3 - OLS	0.9559	0.10935	0.02281	-425.03794

Table 15 Predictive Modeling Selection Techniques and Relevant Statistics

Based upon the analysis, the selection method chosen to predict the price of cars was the OLS-stepwise selection method. With an adjusted R^2 value of 0.9559, the model fits the data and appears as a better representative of the dataset. The cross validation mean squared prediction error is 0.02281, indicating the expected squared distance between the predicted price of a car and the actual price of a car. Since this value is significantly smaller than the LASSO and LAR models, this supports using the OLS method for prediction. Although the AIC seems to favor the LASSO and LAR models, it does not outweigh the other statistics.

Conclusion

This report summarizes the statistical modeling and analysis results for the data set on car specifications. This involved modeling groups of variables as well as conducting a predictive analysis to determine the best method to forecast the prices of cars.

For the Toyota VP Analysis Question 1 it was found that doubling of the vehicle's horsepower equates to a multiplicative change of 0.678 in the median of the distribution vehicle's miles per gallon in the city. This analysis held the curb weight constant while accounting for the dependence on fuel type and weight of the car.

For the Toyota VP Analysis Question 2, a description of statistics demonstrates that there is not sufficient evidence to show that the difference in mean miles per gallon in the city is different for cars with different body styles, excluding convertibles. The 2-way ANOVA shows an improvement in the model with the addition of the variable fuel type, however, it is still not significant in demonstrating a dependence of mean miles per gallon in the city on body style. After performing an adjustment and examining individual relationships, the two most significant existed between sedans using diesel and sedans using gas, and sedans using diesel ad wagons using gas. The adjusted values provide greater clarity into the significance of differences that exist.

For the Predictive Analysis Question, out of the three models that were created, LASSO, LAR, and OLS-stepwise, the OLS technique was chosen to complete the predictive analysis due to a large adjusted R^2 value and small mean square prediction error. The variables selected in the models for LASSO and LAR are not what is expected, considering only width, curb weight, and horsepower. The strength of the OLS method comes from the model considering a greater breadth of vehicle specifications, including variables like horsepower, make, aspiration, and body style. These models demonstrate the types of factors that may be more heavily weighted than popular belief in the car buying process. It would be interesting to continue the study by applying different groupings of variables and developing/comparing models to identify factors that influence the prices of cars.

Appendix

Figures/Tables

Correlation															
Variable	symboling	wheel_base	length	width	height	curb_weight	engine_size	bore	stroke	compression_ratio	horsepower	peak_rpm	city_mpg	highway_mpg	price
symboling	1.0000	-0.5307	-0.3703	-0.2164	-0.5427	-0.2081	-0.0871	-0.1111	0.0338	-0.1562	0.0967	0.2930	-0.0409	0.0221	-0.0512
wheel_base	-0.5307	1.0000	0.8815	0.8061	0.5939	0.7699	0.5509	0.4837	0.1789	0.2010	0.3513	-0.3538	-0.4607	-0.5293	0.5571
length	-0.3703	0.8815	1.0000	0.8583	0.4843	0.8831	0.6760	0.6160	0.1401	0.1229	0.5725	-0.2668	-0.6574	-0.6918	0.6853
width	-0.2164	0.8061	0.8583	1.0000	0.2877	0.8590	0.7147	0.5384	0.1844	0.1234	0.6101	-0.2116	-0.6322	-0.6765	0.7397
height	-0.5427	0.5939	0.4843	0.2877	1.0000	0.2765	0.0495	0.2038	-0.0498	0.2351	-0.1128	-0.3108	-0.0251	-0.0759	0.0960
curb_weight	-0.2081	0.7699	0.8831	0.8590	0.2765	1.0000	0.8432	0.6523	0.1680	0.0811	0.7584	-0.2489	-0.7544	-0.7949	0.8226
engine_size	-0.0871	0.5509	0.6760	0.7147	0.0495	0.8432	1.0000	0.5801	0.2038	-0.0261	0.8211	-0.2349	-0.6454	-0.6710	0.8733
bore	-0.1111	0.4837	0.6160	0.5384	0.2038	0.6523	0.5801	1.0000	-0.1424	-0.0550	0.5763	-0.2537	-0.5911	-0.5954	0.5527
stroke	0.0338	0.1789	0.1401	0.1844	-0.0498	0.1680	0.2038	-0.1424	1.0000	0.1693	0.1066	-0.0447	-0.0425	-0.0443	0.0911
compression_ratio	-0.1562	0.2010	0.1229	0.1234	0.2351	0.0811	-0.0261	-0.0550	0.1693	1.0000	-0.2565	-0.4060	0.3765	0.3226	0.0022
horsepower	0.0967	0.3513	0.5725	0.6101	-0.1128	0.7584	0.8211	0.5763	0.1066	-0.2565	1.0000	0.1330	-0.8211	-0.8007	0.8131
peak_rpm	0.2930	-0.3538	-0.2668	-0.2116	-0.3108	-0.2489	-0.2349	-0.2537	-0.0447	-0.4060	0.1330	1.0000	-0.1404	-0.0851	-0.0728
city_mpg	-0.0409	-0.4607	-0.6574	-0.6322	-0.0251	-0.7544	-0.6454	-0.5911	-0.0425	0.3765	-0.8211	-0.1404	1.0000	0.9729	-0.6879
highway_mpg	0.0221	-0.5293	-0.6918	-0.6765	-0.0759	-0.7949	-0.6710	-0.5954	-0.0443	0.3226	-0.8007	-0.0851	0.9729	1.0000	-0.6976
price	-0.0512	0.5571	0.6853	0.7397	0.0960	0.8226	0.8733	0.5527	0.0911	0.0022	0.8131	-0.0728	-0.6879	-0.6976	1.0000

Table 3 Correlations in the training dataset

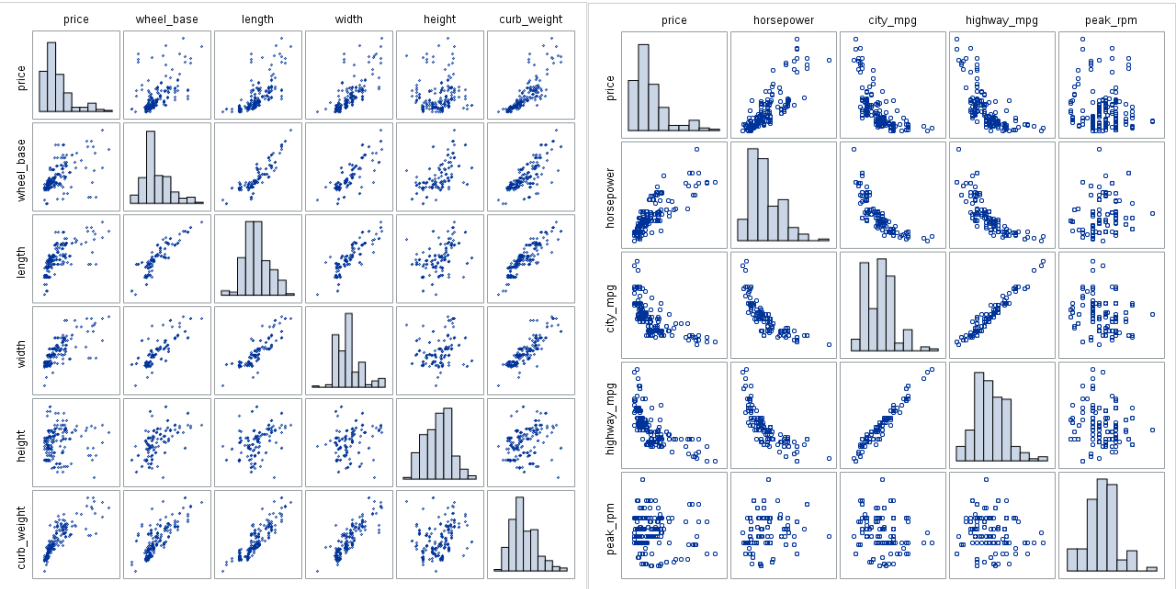

Figure 1.1 Matrix Plots among explanatory variables

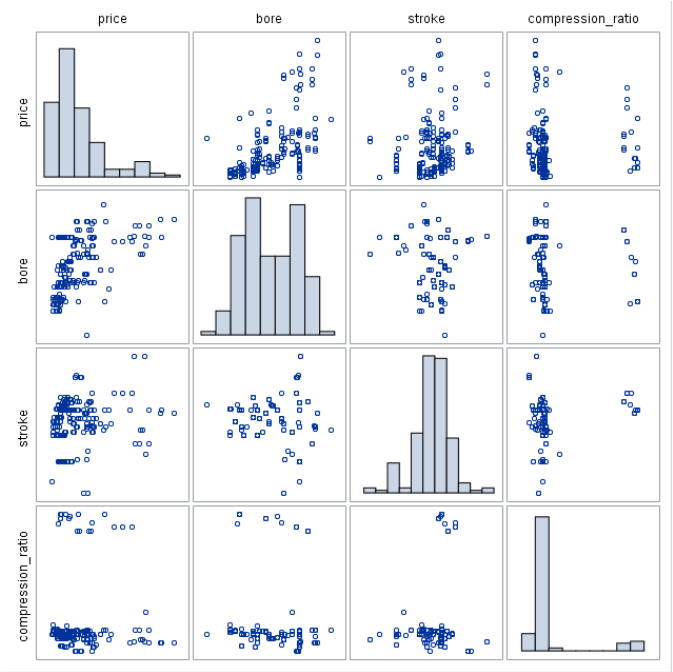

Figure 1.2 Matrix Plots among explanatory variables

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-55924	17502	-3.20	0.0017
symboling	1	316.98140	276.78910	1.15	0.2538
wheel_base	1	165.31804	121.42471	1.36	0.1752
length	1	-72.24967	65.54863	-1.10	0.2720
width	1	566.27215	283.84718	1.99	0.0477
height	1	188.14827	154.68243	1.22	0.2256
curb_weight	1	0.83030	1.86366	0.45	0.6565
engine_size	1	117.43392	15.32124	7.66	<.0001
bore	1	-1548.14270	1447.81255	-1.07	0.2865
stroke	1	-3348.13518	936.95907	-3.57	0.0005
compression_ratio	1	277.01187	89.66330	3.09	0.0024
horsepower	1	40.32334	19.26843	2.09	0.0379
peak_rpm	1	2.09576	0.74800	2.80	0.0057
city_mpg	1	-395.84509	196.24665	-2.02	0.0453
highway_mpg	1	294.17984	176.18548	1.67	0.0969

Table 5 Build Model with Original Data

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.29202	0.95610	5.53	<.0001
symboling	1	0.00933	0.01512	0.62	0.5380
wheel_base	1	0.00171	0.00663	0.26	0.7969
length	1	0.00053571	0.00358	0.15	0.8813
width	1	0.03371	0.01551	2.17	0.0311
height	1	0.00720	0.00845	0.85	0.3953
curb_weight	1	0.00023209	0.00010181	2.28	0.0239
engine_size	1	0.00237	0.00083696	2.84	0.0051
bore	1	-0.01833	0.07909	-0.23	0.8170
stroke	1	-0.09870	0.05118	-1.93	0.0555
compression_ratio	1	0.02216	0.00490	4.53	<.0001
horsepower	1	0.00372	0.00105	3.53	0.0005
peak_rpm	1	0.00006116	0.00004086	1.50	0.1364
city_mpg	1	-0.04086	0.01072	-3.81	0.0002
highway_mpg	1	0.02528	0.00962	2.63	0.0094

Table 7 Build Model with Log Transformation on price

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	7.987502	0.498003	16.04
wheel_base	1	0.019042	0.004396	4.33
height	1	-0.038738	0.008096	-4.79
curb_weight	1	0.000541	0.000066076	8.19
horsepower	1	0.002082	0.000673	3.10
make alfa-romero	1	0.147534	0.104624	1.41
make audi	1	0.232826	0.070559	3.30
make bmw	1	0.472110	0.068352	6.91
make chevrolet	1	-0.011309	0.103167	-0.11
make dodge	1	-0.167589	0.081639	-2.05
make honda	1	0.011597	0.076137	0.15
make isuzu	1	-0.288939	0.125462	-2.30
make jaguar	1	-0.203827	0.107031	-1.90
make mazda	1	0.093261	0.062387	1.49
make mercedes-be	1	0.101458	0.086091	1.18
make mitsubishi	1	-0.240115	0.067967	-3.53
make nissan	1	-0.051227	0.064843	-0.79
make peugot	1	-0.175755	0.063069	-2.79
make plymouth	1	-0.181934	0.072622	-2.51
make porsche	1	0.499753	0.113581	4.40
make saab	1	0.212522	0.095765	2.22
make subaru	1	-0.105027	0.066360	-1.58
make toyota	1	-0.080670	0.057899	-1.39
make volkswagen	1	0.045089	0.065621	0.69
make volvo	0	0	.	.
aspiration std	1	-0.095387	0.032821	-2.91
aspiration turbo	0	0	.	.
body_style convertible	1	0.303792	0.080615	3.77
body_style hardtop	1	0.307589	0.086014	3.58
body_style hatchback	1	0.016110	0.046606	0.35
body_style sedan	1	0.069707	0.038992	1.79
body_style wagon	0	0	.	.

Figure 4 Parameter Estimates for OLS (Stepwise) Selection Model

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	6.025161756	0.12507891	48.17	<.0001	5.778304470	6.272019043
lhorsepower	-0.559888220	0.03635826	-15.40	<.0001	-0.631645336	-0.488131105
curb_weight	-0.000104811	0.00002441	-4.29	<.0001	-0.000152980	-0.000056641

Source	DF	Type I SS	Mean Square	F Value	Pr > F
lhorsepower	1	9.89007101	9.89007101	1106.59	<.0001
fuel_type	1	0.08265977	0.08265977	9.25	0.0027
lhorsepowe*fuel_type	1	0.00012181	0.00012181	0.01	0.9072
curb_weight	1	0.55102216	0.55102216	61.65	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
lhorsepower	1	0.31928941	0.31928941	35.73	<.0001
fuel_type	1	0.00880717	0.00880717	0.99	0.3222
lhorsepowe*fuel_type	1	0.02134277	0.02134277	2.39	0.1241
curb_weight	1	0.55102216	0.55102216	61.65	<.0001

Table 10 Parameter Estimates and Related Statistics for Toyota VP Analysis 1

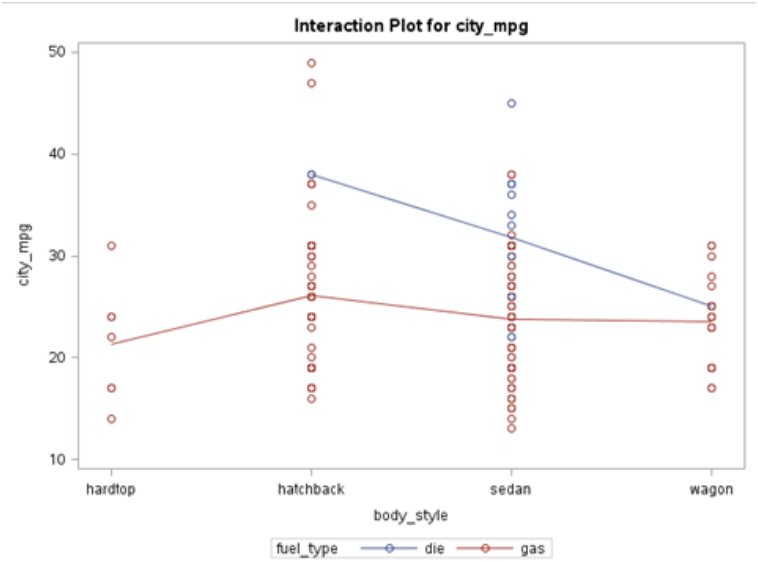

Figure 6 Interaction Plot for mpg in the city for different body style and fuel type

Differences of Least Squares Means																	
Effect	body_style	fuel_type	_body_style	_fuel_type	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
body_style	hardtop		hatchback		Non-est	.	.	.	.	Bonferroni	.	.	.	.	.	.
body_style	hardtop		sedan		Non-est	.	.	.	.	Bonferroni	.	.	.	.	.	.
body_style	hardtop		wagon		Non-est	.	.	.	.	Bonferroni	.	.	.	.	.	.
body_style	hatchback		sedan		4.2568	3.2786	166	1.30	0.1960	Bonferroni	0.5879	0.05	-2.2163	10.7298	-3.6724	12.1859
body_style	hatchback		wagon		7.7709	3.9121	166	1.99	0.0486	Bonferroni	0.1459	0.05	0.04689	15.4949	-1.6906	17.2323
body_style	sedan		wagon		3.5141	2.4939	166	1.41	0.1607	Bonferroni	0.4820	0.05	-1.4097	8.4379	-2.5173	9.5455
fuel_type		die		gas	Non-est	.	.	.	.	Bonferroni	.	.	.	.	.	.
body_style*fuel_type	hardtop	gas	hatchback	die	-16.7143	6.6759	166	-2.50	0.0133	Bonferroni	0.2784	0.05	-29.8949	-3.5337	-37.3131	3.8845
body_style*fuel_type	hardtop	gas	hatchback	gas	-4.8350	2.4987	166	-1.94	0.0547	Bonferroni	1.0000	0.05	-9.7682	0.09827	-12.5447	2.8748
body_style*fuel_type	hardtop	gas	sedan	die	-10.5714	2.8907	166	-3.66	0.0003	Bonferroni	0.0072	0.05	-16.2788	-4.8641	-19.4910	-1.6519
body_style*fuel_type	hardtop	gas	sedan	gas	-2.4643	2.4724	166	-1.00	0.3203	Bonferroni	1.0000	0.05	-7.3456	2.4170	-10.0929	5.1643
body_style*fuel_type	hardtop	gas	wagon	die	-3.7143	5.0069	166	-0.74	0.4592	Bonferroni	1.0000	0.05	-13.5997	6.1712	-19.1634	11.7348
body_style*fuel_type	hardtop	gas	wagon	gas	-2.2932	2.7610	166	-0.83	0.4074	Bonferroni	1.0000	0.05	-7.7445	3.1581	-10.8126	6.2261
body_style*fuel_type	hatchback	die	hatchback	gas	11.8793	6.2983	166	1.89	0.0610	Bonferroni	1.0000	0.05	-0.5558	24.3145	-7.5545	31.3131
body_style*fuel_type	hatchback	die	sedan	die	6.1429	6.4639	166	0.95	0.3433	Bonferroni	1.0000	0.05	-6.6192	18.9049	-13.8018	26.0875
body_style*fuel_type	hatchback	die	sedan	gas	14.2500	6.2879	166	2.27	0.0247	Bonferroni	0.5193	0.05	1.8354	26.6646	-5.1517	33.6517
body_style*fuel_type	hatchback	die	wagon	die	13.0000	7.6482	166	1.70	0.0911	Bonferroni	1.0000	0.05	-2.1003	28.1003	-10.5989	36.5989
body_style*fuel_type	hatchback	die	wagon	gas	14.4211	6.4070	166	2.25	0.0257	Bonferroni	0.5399	0.05	1.7714	27.0707	-5.3479	34.1900
body_style*fuel_type	hatchback	gas	sedan	die	-5.7365	1.8595	166	-3.08	0.0024	Bonferroni	0.0501	0.05	-9.4078	-2.0651	-11.4741	0.001195
body_style*fuel_type	hatchback	gas	sedan	gas	2.3707	1.1018	166	2.15	0.0329	Bonferroni	0.6903	0.05	0.1953	4.5460	-1.0290	5.7704
body_style*fuel_type	hatchback	gas	wagon	die	1.1207	4.4912	166	0.25	0.8033	Bonferroni	1.0000	0.05	-7.7465	9.9879	-12.7370	14.9784
body_style*fuel_type	hatchback	gas	wagon	gas	2.5417	1.6507	166	1.54	0.1255	Bonferroni	1.0000	0.05	-0.7173	5.8008	-2.5516	7.6351
body_style*fuel_type	sedan	die	sedan	gas	8.1071	1.8240	166	4.44	<.0001	Bonferroni	0.0003	0.05	4.5059	11.7084	2.4790	13.7353
body_style*fuel_type	sedan	die	wagon	die	6.8571	4.7206	166	1.45	0.1482	Bonferroni	1.0000	0.05	-2.4629	16.1772	-7.7084	21.4227
body_style*fuel_type	sedan	die	wagon	gas	8.2782	2.1995	166	3.76	0.0002	Bonferroni	0.0049	0.05	3.9355	12.6208	1.4914	15.0649
body_style*fuel_type	sedan	gas	wagon	die	-1.2500	4.4766	166	-0.28	0.7804	Bonferroni	1.0000	0.05	-10.0884	7.5884	-15.0628	12.5628
body_style*fuel_type	sedan	gas	wagon	gas	0.1711	1.6106	166	0.11	0.9155	Bonferroni	1.0000	0.05	-3.0089	3.3510	-4.7986	5.1407
body_style*fuel_type	wagon	die	wagon	gas	1.4211	4.6423	166	0.31	0.7599	Bonferroni	1.0000	0.05	-7.7445	10.5866	-12.9029	15.7450

Table 13 Bonferroni Adjustment applied to Analysis 2 – Differences of Least Squares Means

SAS Code

```
**load training dataset;
options VALIDVARNAME=ANY;
%web_drop_table(WORK.train_auto);
FILENAME REFFILE '/home/tanvia0/MSDS 6372/Data/Train_Auto.csv';
PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=WORK.train_auto;
        GETNAMES=YES;
        DATAROW=2;
RUN;
PROC CONTENTS DATA=WORK.train_auto; RUN;

proc datasets lib=work;
modify train_auto;
        rename 'fuel-type'n = fuel_type;
        rename 'num-of-doors'n = num_of_doors;
        rename 'body-style'n = body_style;
        rename 'drive-wheels'n = drive_wheels;
        rename 'engine-location'n = engine_location;
        rename 'wheel-base'n = wheel_base;
        rename 'curb-weight'n = curb_weight;
        rename 'engine-type'n = engine_type;
        rename 'num-of-cylinders'n = num_of_cylinders;
```

```
        rename 'engine-size'n = engine_size;
        rename 'fuel-system'n = fuel_system;
        rename 'compression-ratio'n = compression_ratio;
        rename 'peak-rpm'n = peak_rpm;
        rename 'city-mpg'n = city_mpg;
        rename 'highway-mpg'n = highway_mpg;
quit;
proc print data=train_auto;
run;


%let vcurb_weight_mean; ** variable to calculate mean of curbweight;
proc univariate data=train_auto;
class make,aspiration,body_style;
var curb_weight;
run;


**first imputation of null values**;
data train_auto_imp;
set train_auto;
if fuel_type eq '?' then fuel_type='gas';   **mode of fuel_type;
if num_of_doors eq '?' then num_of_doors='four';     **mode of num_of_doors;
** curb weight is the weight of car with standard equipment, so we consider what matters most. for null
values, we take mean of make,aspiration and body_style;
if notdigit(curb_weight) && make eq 'audi' then curb_weight=2843;
if notdigit(curb_weight) && make eq 'bmw' then curb_weight=2968;
if notdigit(curb_weight) && make eq 'mazda' then curb_weight=2266;
if bore eq '' then bore=3.21;    ** mean of make;
if stroke eq '' then stroke=3.27;     **mean of make;
** curb_weight is loaded into SAS as a character due to null values, so convert to a numeric type;
vcurb_weight = input(curb_weight,8.);
drop curb_weight;
rename vcurb_weight = curb_weight;
**horsepower is a property of engine type and engine size. Since we do not have any other data for
engine_size=119,
to take average for, we will consider number_of_cylinders which is proportional to engine_size;
if horsepower eq '?' && make = 'isuzu' then horsepower=90;
if horsepower eq '?' && make = 'mercedes-benz' then horsepower=118;
** log of price;
lprice = log(price);
**lcurb_weight = log(curb_weight);
**lpeak_rpm = log(peak_rpm);
**lhorsepower = log(horsepower);
seqno = _N_;
run;
```

EXPLORATORY

```
** check  VIF for quantitative variables;
proc reg data=train_auto_imp;
model price = symboling wheel_base length width height curb_weight engine_size bore stroke
compression_ratio horsepower peak_rpm city_mpg highway_mpg / VIF;
run;

** check  VIF for quantitative variables;
**Correlation;
proc reg data=train_auto_imp;
model price = symboling wheel_base length width height curb_weight engine_size bore stroke
compression_ratio horsepower peak_rpm city_mpg highway_mpg / VIF;

**Plot Data;
proc sgscatter data=train_auto_imp;
matrix price wheel_base length width height curb_weight / diagonal=(histogram) ;*datalabel=seqno;
run;
proc sgscatter data=train_auto_imp;
matrix price horsepower city_mpg highway_mpg peak_rpm/ diagonal=(histogram) ;*datalabel=seqno;
run;
proc sgscatter data=train_auto_imp;
matrix price bore stroke compression_ratio/ diagonal=(histogram) ;*datalabel=seqno;
run;

**Build models and check residual plots;
**influential Observations and fit;
proc reg data=train_auto_imp plots=all;
model price = symboling wheel_base length width height curb_weight engine_size bore stroke
compression_ratio horsepower peak_rpm city_mpg highway_mpg;
output out = t student=res cookd = cookd h = lev p = yhat;
run;
quit;

** try log transformation for price, since the residual plot is not a random cloud;
proc reg data=train_auto_imp plots=all;
model lprice = symboling wheel_base length width height curb_weight engine_size bore stroke
compression_ratio horsepower peak_rpm city_mpg highway_mpg;
output out = t student=res cookd = cookd h = lev p = yhat;
run;
quit;

**Identify Outliers;
**get labels on res vs leverage;
proc sgplot data = t;
scatter y = res  x = lev / datalabel = seqno; run;
**Get labels on Cook's D vs. subject;
proc sgplot data=t;
scatter y = cookd  x = seqno / datalabel = seqno; run;
quit;
```

TOYOTA VP ANALYSIS 1

```
proc sgscatter data = train_auto_imp;
matrix city_mpg horsepower curb_weight;
run;

proc sgscatter data = train_auto_imp;
matrix lcity_mpg lhorsepower curb_weight;
run;

proc reg data  = train_auto_imp;
model city_mpg = horsepower;
run;

proc reg data  = train_auto_imp;
model lcity_mpg = lhorsepower;
run;

**Linear model how horsepower is related to city mpg holding curb weight constant;
proc glm data = train_auto_imp plots=all;
model city_mpg = horsepower curb_weight / solution clparm;
run;

**Linear model how horsepower and fuel type are related to city mpg holding curb weight constant;
proc glm data = train_auto_imp plots=all;
class fuel_type (ref='gas');
model city_mpg = horsepower fuel_type curb_weight / solution clparm;
run;

**Linear model how lhorsepower is related to lcity mpg holding curb weight constant;
proc glm data = train_auto_imp plots=all;
model lcity_mpg = lhorsepower curb_weight / solution clparm;
run;

**Linear model how lhorsepower and fuel type are related to lcity mpg holding curb weight constant;
proc glm data = train_auto_imp plots=all;
class fuel_type (ref='gas');
model lcity_mpg = lhorsepower | fuel_type curb_weight / solution clparm;
run;

** try log transformation for price, since the residual plot is not a random cloud;
proc reg data=train_auto_imp plots=all;
model lcity_mpg = lhorsepower;
output out = t student=res cookd = cookd h = lev p = yhat;
run;
quit;
```

```
proc sgplot data = t;
scatter y = res  x = lev / datalabel = seqno;
run;

**Get labels on Cook's D vs. subject;
proc sgplot data = t;
scatter y = cookd  x = seqno / datalabel = seqno;
run;
quit;
```

TOYOTA VP ANALYSIS 2

```
** use imputed training dataset ;
** filter out convertibles as per requirement ;
data train_auto_citympg;
set train_auto_imp;
if body_style = 'convertible' then delete;
lcity_mpg = log(city_mpg);
run;

** response variable - mpg per city(city_mpg)
** factor variables -
** var 1 - body_style ( excluding convertibles);
proc glm data=train_auto_citympg plots=(DIAGNOSTICS RESIDUALS);
class body_style ;
model city_mpg = body_style /solution cli clm ;
run;
quit;

** 2 way ANOVA
** response variable - mpg per city(city_mpg)
** factor variables -
** var 1 - body_style ( excluding convertibles)
** var 2 - fuel_type;
** Bonferroni adjustment;
title 'Analyze effects for Body_style and fuel_type on city_mpg';
proc mixed data=train_auto_citympg plots= (ResidualPanel);
class body_style fuel_type;
model city_mpg = body_style |fuel_type / solution;
lsmeans body_style |fuel_type / CL adjust = bon;
run;

quit;

title 'Contrast for the different body styles';
proc glm data=train_auto_citympg plots=(DIAGNOSTICS RESIDUALS);
class body_style fuel_type  ;
model city_mpg = body_style fuel_type ;
```

```
*contrast order : hardtop hatchback sedan wagon;
contrast 'hatchback' body_style 0.34 -1 0.33 0.33/e;
contrast 'hardtop' body_style -1 0.34 0.33 0.33/e;
contrast 'sedan' body_style 0.34 0.33 -1 0.33/e;
contrast 'wagon' body_style 0.34 0.33 0.33 -1/e;
run;
```

SELECT MODEL FOR PREDICTIONS

```
** load test dataset **;
%web_drop_table(WORK.test_auto);
FILENAME REFFILE '/home/tanvia0/MSDS 6372/Data/Test_Auto.csv';
PROC IMPORT DATAFILE=REFFILE
DBMS=CSV
OUT=WORK.test_auto;
GETNAMES=YES;
DATAROW=2;
RUN;

PROC CONTENTS DATA=WORK.test_auto; RUN;

%web_open_table(WORK.test_auto);

**rename fields  in testds **;
proc datasets lib=work;
modify test_auto;
rename 'fuel-type'n = fuel_type;
rename 'num-of-doors'n = num_of_doors;
rename 'body-style'n = body_style;
rename 'drive-wheels'n = drive_wheels;
rename 'engine-location'n = engine_location;
rename 'wheel-base'n = wheel_base;
rename 'curb-weight'n = curb_weight;
rename 'engine-type'n = engine_type;
rename 'num-of-cylinders'n = num_of_cylinders;
rename 'engine-size'n = engine_size;
rename 'fuel-system'n = fuel_system;
rename 'compression-ratio'n = compression_ratio;
rename 'peak-rpm'n = peak_rpm;
rename 'city-mpg'n = city_mpg;
rename 'highway-mpg'n = highway_mpg;
quit;

data test_auto_imp;
set test_auto;
** curb weight is the weight of car with standard equipment, so we consider what matters most. for null
values, we take mean of
```

```
make,aspiration and body_style. for the test dataset, due to insufficient data for the same make, we will
consider mean of same aspiration
and body style cars;
if notdigit(curb_weight) then curb_weight = '2478';
** curb_weight is loaded into SAS as a character due to null values, so convert to a numeric type;
vcurb_weight = input(curb_weight,8.);
drop curb_weight;
rename vcurb_weight = curb_weight;
lprice=.;
seqno = 180 + _N_;
run;

PROC CONTENTS DATA=WORK.test_auto_imp; RUN;

** without Obs 41 **
** add a new variable ,that generates a random number to group our data later;
data train_auto_imp_ms_wo41;
 set train_auto_imp;
 if seqno = 41 then delete;
 RandNumber = ranuni(11);
run;

** generate train data set where RandNumber > 0.25;

data trainds_wo41;
 set train_auto_imp_ms_wo41;
 if RandNumber <= 1/4 then delete;
run;

* generate test data set where RandNumber < 0.25;

data testds_wo41;
 set train_auto_imp_ms_wo41;
 if RandNumber > 1/4 then delete;
run;

** run proc glm with LASSO and Cross validation selection;
ods graphics on;
proc glmselect data=trainds_wo41 testdata = testds_wo41
          seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot CRITERIONPANEL);
 class symboling make fuel_type aspiration num_of_doors body_style drive_wheels engine_location
engine_type engine_size fuel_system;
 model lprice = symboling wheel_base length width height curb_weight engine_size bore stroke
compression_ratio horsepower peak_rpm city_mpg
 highway_mpg make fuel_type aspiration num_of_doors body_style drive_wheels engine_location
engine_type engine_size fuel_system
 / selection=LASSO( choose=CV stop=CV) CVdetails  ;
run; quit;
```

```
ods graphics off;

ods graphics on;
proc glmselect data=trainds_wo41 testdata = testds_wo41
        seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot CRITERIONPANEL);
 class symboling make fuel_type aspiration num_of_doors body_style drive_wheels engine_location
engine_type engine_size fuel_system;
 model lprice = symboling wheel_base length width height curb_weight engine_size bore stroke
compression_ratio horsepower peak_rpm city_mpg
 highway_mpg make fuel_type aspiration num_of_doors body_style drive_wheels engine_location
engine_type engine_size fuel_system
 / selection=stepwise( choose=CV stop=CV) CVdetails  ;
run; quit;
ods graphics off;

ods graphics on;
proc glmselect data=trainds_wo41 testdata = testds_wo41
        seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot CRITERIONPANEL);
 class symboling make fuel_type aspiration num_of_doors body_style drive_wheels engine_location
engine_type engine_size fuel_system;
 model lprice = symboling wheel_base length width height curb_weight engine_size bore stroke
compression_ratio horsepower peak_rpm city_mpg
 highway_mpg make fuel_type aspiration num_of_doors body_style drive_wheels engine_location
engine_type engine_size fuel_system
 / selection=LAR( choose=CV stop=CV) CVdetails  ;
run; quit;
ods graphics off;


**prediction dataset **;
data auto_all;
set  train_auto_imp_ms_wo41 test_auto_imp;
run;

** use variables returned from stepwise model **;

proc glm data=auto_all plots=all;
class make aspiration body_style ;
model lprice =wheel_base height curb_weight horsepower make aspiration body_style  /solution cli clm
clparm;
output out = results_model_swm p = Predict;
run;

** stepwise model has a variable make, one of the test dataset does not value for its make, so re-run
predictions with same variables from stepwise but the make **;
proc glm data=auto_all plots=all;
class  aspiration body_style ;
```

```
model lprice =wheel_base height curb_weight horsepower  aspiration body_style  /solution cli clm
clparm;
output out = results_model_sw p = Predict;
run;

data results_swm;
set results_model_swm;
price = exp(Predict);
keep seqno price;
where seqno > 180;
run;

proc print data=results_swm;
run;

data results_sw;
set results_model_sw;
price = exp(Predict);
keep seqno price;
where seqno > 180;
run;

proc print data=results_sw;
run;
```