Automated Variable Selection and Predictive Modeling of the Gasoline Consumption Dataset

Prepared by Amee Amin

Assignment 5

MSPA Course 410, Summer 2017

Professors Chad R. Bhatti and Tom Miller

Introduction

Fuel economy is one of the most sought-after vehicle qualities. Vehicles that can cover more miles per gallon, and therefore have a higher fuel economy, save consumers money at the gas pump. However, the diversity in vehicle options can make it difficult for consumers to know when they're getting a good fuel economy, given the size and power of the vehicle they want. In this analysis we built a predictive model for vehicle gasoline consumption based on the Gasoline Consumption Case dataset in "Regression Analysis by Example" by Chatterjee and Hadi. The dataset was originally listed in a 1975 Motor Trend magazine. The dataset contains 30 different vehicles and 12 variables, of which 9 are continuous, 2 discrete, and 1 nominal. Our response variable was miles per gallon. Given common knowledge about fuel economy, we expected variables indicating larger vehicle size and power would strongly correlate with less miles per gallon. We performed a log-transformation on the response variable and defined multiple linear regressions for a full model and subset model, developed with an automated variable selection method. We found that the subset model using forward selection had the highest adjusted correlation coefficient. The full model had no significant predictors and displayed characteristics of multicollinearity.

Section 1: Sample Population

The dataset contains 30 observations of 12 variables: mpg, displacement, horsepower, torque, compression ratio, rear axle ratio, number of carburetor barrels and transmission speeds, transmission type, and overall length, width, and height. These variables are described in Table 1 below. Our predictor variable for this analysis was miles per gallon (mpg). There were no missing or erroneous values in the dataset. Due to the small dataset size, we did not remove any values from this analysis.

The variables included in the dataset were chosen by Motor Trend magazine in 1975, which we assume is targeted towards an audience of vehicle enthusiasts. Because that audience is more likely to be familiar with vehicle engineering, the variables we explored in this analysis may not be the most consumer-friendly terms. The authors of "Regression Analysis by Example" created this dataset in order for students to explore collinearity, so we also assumed that variables were specifically included to show collinearity in our models. We noted that vehicle models were not included in the dataset.

Table 1: All Variables

Variable	Description	Type of Variable
mpg	miles/gallon	continuous
displacement	cubic inches	continuous
horsepower	feet/pound	continuous
torque	feet/pound	continuous
compression ratio	ratio of the maximum to minimum volume in the cylinder of an internal combustion engine	continuous
rear axle ratio	ratio of the number of gear teeth on the ring gear of the rear axle and the pinion gear on the driveshaft	continuous
carburetor barrels	number of carburetor barrels	discrete
transmission speeds	number of transmission speeds	discrete
overall length	inches	continuous
width	inches	continuous
weight	pounds	continuous
transmission type	1 = automatic, 0 = manual	nominal

Miles per gallon are the main indicator of fuel economy for consumers. The more miles you can drive per gallon, the higher the fuel economy. Engine displacement, in cubic inches, is an indicator of a vehicle's size and power. Engine displacement is approximately proportional to the volume of fuel-air mixture drawn into an engine's cylinders. More engine displacement means more fuel being used during ignition, which means less fuel economy. Horsepower and torque, both feet/pound, are related to engine power. Torque describes how much work an engine can do, and horsepower describes how quickly that work can be done. Race and sports cars famously have high levels of horsepower, because high horsepower enables fast acceleration. However, this also consumes more fuel, and lowers the overall fuel economy. Additionally, larger engines tend to have higher levels of horsepower because they can naturally move more air per revolution, which also lowers the fuel economy.

The compression ratio represents the ratio of the maximum to minimum volume in the cylinder of a vehicle's engine. The higher the compression ratio, the more compressed air is in the cylinder. Moreover, the more air is compressed, the more powerful the explosion is from the fuel-air mixture. This means that fuel can be used for more miles per gallon, which increases the overall fuel economy. The rear axle ratio represents the ratio of the number of gear teeth on the ring gear of rear axle to the pinion gear on the driveshaft. Rear axle ratios are particularly important for trucks or any vehicles intended for towing. As the numerical ratio increases, a vehicle's towing capacity increases, because there is more low-speed torque. This means that it takes less throttle to accelerate a vehicle and it's load. However, this also lowers the vehicle's fuel economy.

Carburetor barrels also indicate the power of a vehicle. The barrel is a tube-like part of a carburetor that blends fuel and air in the proper ratio for combustion. More powerful vehicles have more carburetors, which means that a vehicle uses more fuel at any given speed. Therefore, as the number of carburetor barrels increases, fuel economy generally decreases. On the other hand, increasing the number of transmission speeds generally increases a vehicle's fuel economy. This is because more gear options enable an engine to operate at a more efficient speed more often. Higher engine efficiency means a vehicle wastes less fuel. For similar efficiency reasons, manual transmissions tend have better fuel economy. Automatic

transmissions require a torque convertor that burns fuel less efficiently than manual transmissions, which use a simple clutch. Although automatic transmission engines today can be just as efficient as, if not more than, manual transmission engines, the dataset in our analysis is from 1975.

Last, vehicle size and weight generally indicate less fuel economy. The larger the length and width of a vehicle, the more it will weigh, and the more fuel it will take to move the vehicle. Varying vehicle materials may create variation in size and weight, but we can normally assume that larger cars weigh more. In the next section we explore each predictor variable's relationship with mpg further and determine whether our assumptions about fuel economy are reflected in the data.

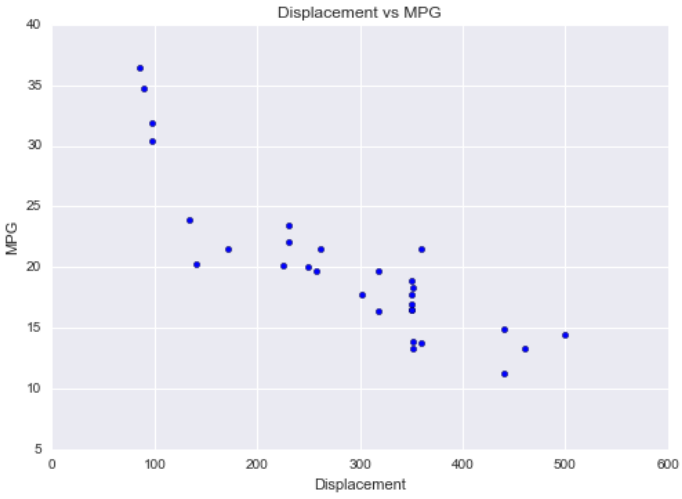
Section 2: Exploratory Data Analysis

We created scatterplots of each of the predictor variables with our response variable mpg to explore the linearity of their relationships, and then develop a simple linear regression model of weight versus mpg.

Section 2.1: Scatterplots of Predictor vs Response Variable

In the scatterplot below, we observed that mpg decreases as engine displacement increases, which is the pattern we expected. Most vehicles in our dataset had an engine displacement between 200 and 400 cubic inches. There are 4 vehicles which have a much higher mpg and an engine displacement below 100. There is a significant decrease of approximately 15 mpg from an engine displacement below 100 to above 100, which may indicate a structural change in the size of the car or engine components.

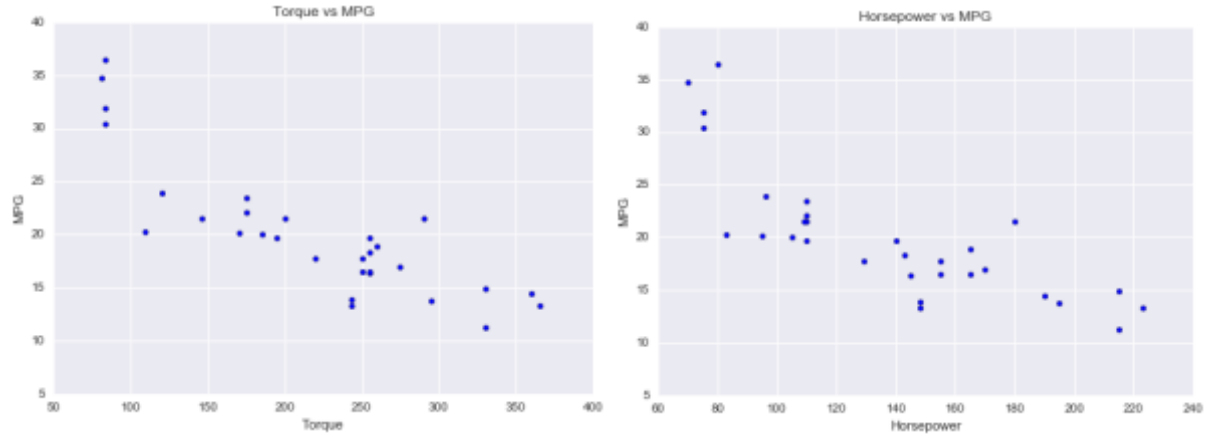
Fig 1: Scatterplot of Displacement vs MPG



As a vehicle's horsepower and torque increase (below), we also observed that mpg tends to decrease. Looking at a side-by-side comparison of torque and horsepower to mpg, the variables have nearly identical scatterplots and we can expect to see some collinearity in our model. Given that torque is in the computational formula for horsepower, the similarity in their

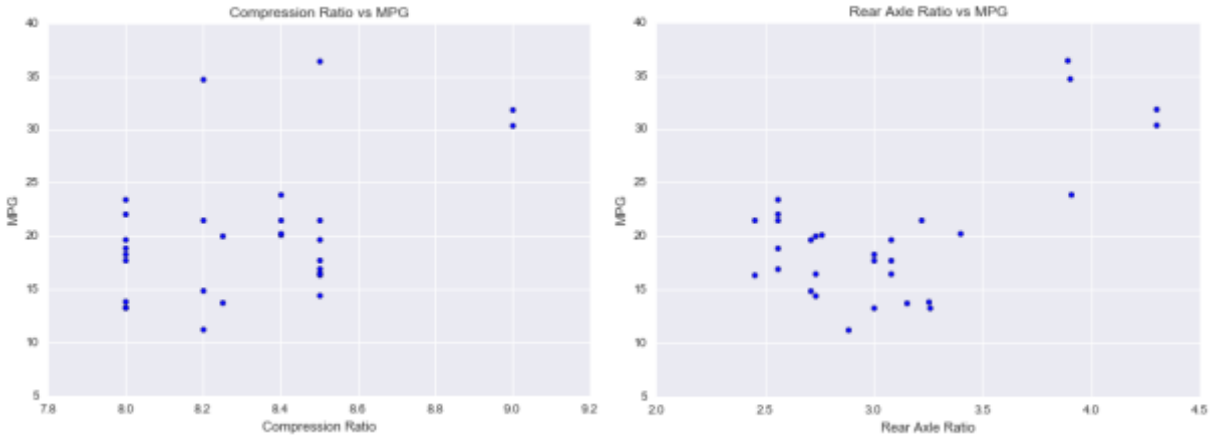
relationship to mpg is expected. Most of the vehicles in our dataset have a torque between 150 and 300, or 100 and 180 horsepower.

Fig 2: Scatterplots of Torque (left) and Horsepower (right) vs MPG



The scatterplots of compression ratio and rear axle ratio below reveal that they may not be the best discriminating features for mpg. Mpg only slightly increases with increased numerical compression ratio, although we expected mpg to increase with a more strongly positive relationship. Mpg tends to stay in the same range for vehicles with a compression ratio between 8.0 and 8.6. There were a few outliers in our dataset where cars have very high fuel economy, but they range from a numerical compression ratio of 8.2 to 9.0. This may indicate that that other engine components influence overall fuel economy more than compression ratio.

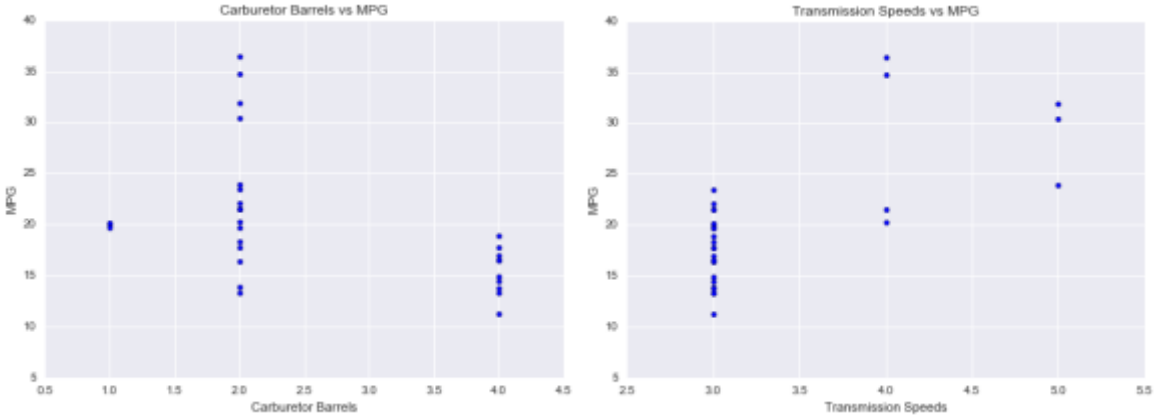
Fig 3: Scatterplots of Compression Ratio (left) and Rear Axle Ratio (right) vs MPG



We observed no clear trend between mpg and rear axle ratio (above, right). Most of the vehicles in our dataset are clustered between a rear axle ratio of 2.5 and 3.5 and an mpg between 10 and 25. We observed a few outliers with high fuel economy with high real axle ratios as well (approximately 4.0 and above). The clustering of rear axle ratios also suggests that other engine components have more influence on overall fuel economy.

Scatterplots of carburetor barrels and transmission speeds with mpg below show the patterns we expected: mpg tends to decrease with an increasing number of carburetor barrels, and mpg tends to increase with an increasing number of transmission speeds. There seems to be a variation of 10-15 mpg for vehicles with the same number of carburetor barrels, which suggests that other engine components can still strongly influence fuel economy given a number of barrels. Most of the vehicles in our dataset have three transmission speeds, which would make it hard to discriminate mpg based on transmission speed alone. There are 7 vehicles with a transmission speed of 4.0 or higher. The variation in mpg for vehicles with the same number of transmission speeds also suggests that other engine components can strongly influence fuel economy given any number of speeds.

Fig 4: Scatterplots of Carburetor Barrels (left) and Transmission Speeds (right) vs MPG

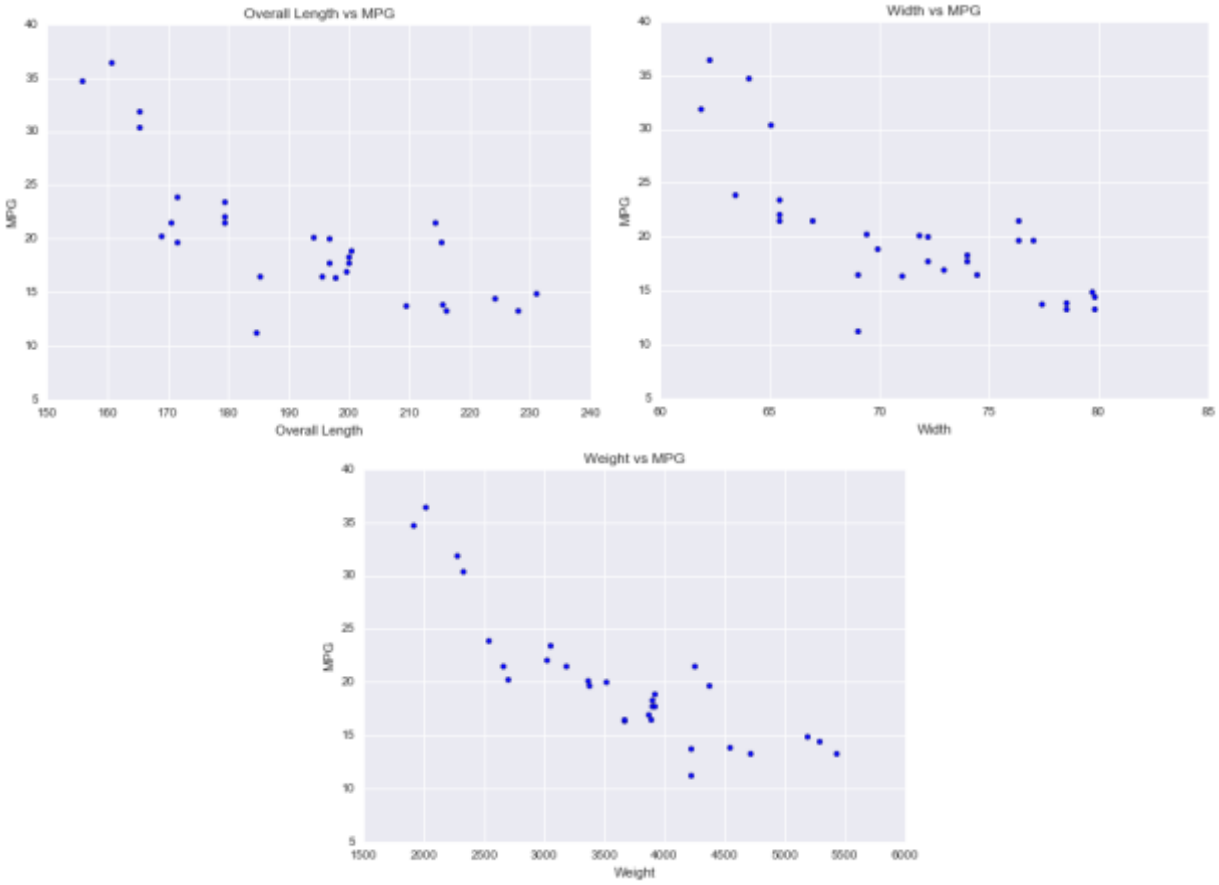


The scatterplot of transmission type versus mpg (below) shows that vehicles with automatic transmission clearly have lower mpg than vehicles with manual transmission. Vehicles with manual transmission have an mpg between 20 and 35, and vehicles with automatic transmission have an mpg between 10 and 25. While more engine features would be needed to further discriminate mpg beyond these ranges, transmission type may be a good predictor for mpg.

Fig 5: Scatterplot of Transmission Type vs MPG

Last, the scatterplots of overall length, width, and weight versus mpg (below) show that mpg tends to decrease as overall length, width, and weight increases. This pattern reflects our assumption about size and mpg. Looking at side-by-side comparisons, the scatterplots below appear nearly identical, as we observed between torque and horsepower above in Fig 2. Most vehicles are within a length of 170 and 220 inches, a width of 65 and 80 inches, and a weight of 2500 and 5500 pounds. Overall length, width, and weight may be good predictors for mpg, but we can expect to see collinearity if more than one is included a model.

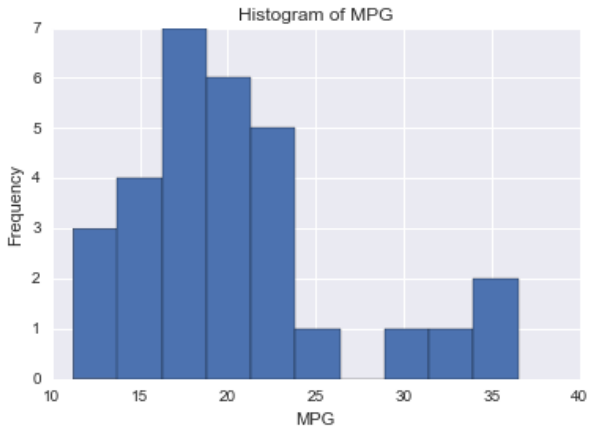
Fig 6: Scatterplots of Overall Length (Left), Width (Right), and Weight (below) vs MPG



Section 2.2: Transformation of the Response Variable

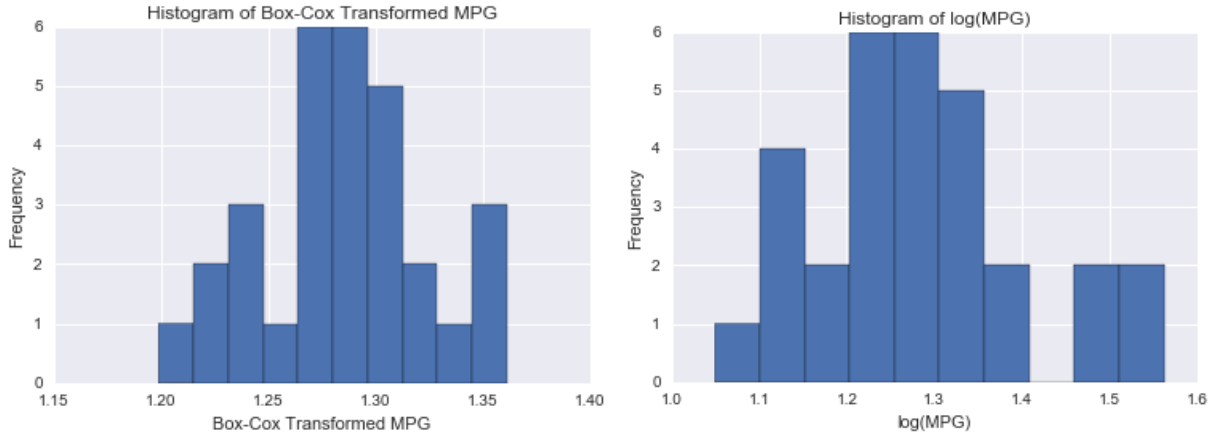
We observed the distribution of the response variable in order to determine whether our dataset validates the assumptions of linear regression. In order for the linear regression method to be accurate, a response variable should have at least a conditional normal distribution. In the histogram of mpg below, we observed that mpg has non-normal distribution. It is positively skewed with a tail to the right, and is leptokurtic, meaning our dataset is more concentrated around the mean than a normal distribution.

Fig 7: Histogram of MPG



In order to make our response variable reflect a more normal distribution, we performed a Box-Cox transformation and Log transformation of mpg and compare them below. The histogram of Box-Cox transformed mpg below (left) show a more normal distribution of mpg than above. There is a more apparent bell shape, although the tails decrease and then increase in frequency. The log transformation of mpg below (right) also shows a more normal distribution than above, however, it is difficult to tell which transformation better improves the distribution of mpg. In the next section, we perform simple regressions of weight versus mpg including a box-cox transformed mpg and log-transformed mpg.

Fig 8: Histogram Comparison of Box-Cox (left) and Log (right) Transformations of MPG



Section 2.3: Simple Linear Regression

We defined a simple linear regression between weight versus mpg. The OLS regression results shown in the table below.

Table 2.3.1: OLS Regression Results for Weight vs MPG

Dep. Variable:	y	**R-squared:**	0.727
Model:	OLS	**Adj. R-squared:**	0.717
Method:	Least Squares	**F-statistic:**	74.52
Date:	Fri, 21 Jul 2017	**Prob (F-statistic):**	2.23e-09
Time:	14:52:04	**Log-Likelihood:**	-77.631
No. Observations:	30	**AIC:**	159.3
Df Residuals:	28	**BIC:**	162.1
Df Model:	1		
Covariance Type:	nonrobust		

As expected, weight has a strong positive correlation with mpg with a correlation coefficient of 0.727 (above). The F-statistic is significant, so we can reject the null hypothesis that the regression coefficients are zero.

Table 2.3.2: OLS Regression Results for Weight vs MPG

Omnibus:	1.774	**Durbin-Watson:**	1.745
Prob(Omnibus):	0.412	**Jarque-Bera (JB):**	1.554
Skew:	0.433	**Prob(JB):**	0.460
Kurtosis:	2.297	**Cond. No.**	1.51e+04

The linear model has a skew of 0.433 and a kurtosis of 2.297. We defined two additional linear models of weight versus mpg using transformations of the response variable.

Table 2.3.3 OLS Regression Comparison of Box-Cox (left) and Log (right) Transformations of MPG

Dep. Variable:	y	**R-squared:**	0.733
Model:	OLS	**Adj. R-squared:**	0.724
Method:	Least Squares	**F-statistic:**	76.98
Date:	Fri, 21 Jul 2017	**Prob (F-statistic):**	1.59e-09
Time:	14:52:04	**Log-Likelihood:**	75.046
No. Observations:	30	**AIC:**	-146.1
Df Residuals:	28	**BIC:**	-143.3
Df Model:	1		
Covariance Type:	nonrobust		

Dep. Variable:	y	**R-squared:**	0.750
Model:	OLS	**Adj. R-squared:**	0.741
Method:	Least Squares	**F-statistic:**	84.08
Date:	Fri, 21 Jul 2017	**Prob (F-statistic):**	6.29e-10
Time:	14:52:04	**Log-Likelihood:**	41.100
No. Observations:	30	**AIC:**	-78.20
Df Residuals:	28	**BIC:**	-75.40
Df Model:	1		
Covariance Type:	nonrobust		

Looking at Table 2.3.3 above and 2.3.4 below, we can compare linear regressions models using a Box-Cox transformed and log-transformed mpg. The Box-Cox transformed mpg (left, above) produced a linear model with a correlation coefficient of 0.733. The log-transformed mpg (right, above) produced a linear model with slightly higher correlation coefficient of 0.750. Both models have significant F-statistics, so we can reject the null hypothesis that the regression coefficients are zero for both models.

Table 2.3.4 OLS Regression Comparison of Box-Cox (left) and Log (right) Transformations of MPG

Omnibus:	9.762	**Durbin-Watson:**	1.759
Prob(Omnibus):	0.008	**Jarque-Bera (JB):**	9.892
Skew:	-0.853	**Prob(JB):**	0.00711
Kurtosis:	5.237	**Cond. No.**	1.51e+04

Omnibus:	1.001	**Durbin-Watson:**	1.710
Prob(Omnibus):	0.606	**Jarque-Bera (JB):**	0.378
Skew:	-0.261	**Prob(JB):**	0.828
Kurtosis:	3.172	**Cond. No.**	1.51e+04

The model of Box-Cox transformed mpg has a skew of -0.853 and kurtosis of 5.237. The model of log-transformed mpg has a skew of -0.261 and kurtosis of 3.172. The skew of these two models is slightly negative, compared to the positively skewed regression of weight and non-transformed mpg. On the other hand, the kurtosis of these two models is slightly higher than the regression of weight and non-transformed mpg. Ultimately, both transformations improved the strength of the correlation coefficient.

Section 3: Multiple Linear Regression – Full Model

We defined a multiple linear regression of mpg using the full model. We used log-transformed mpg as the response variable. The OLS regression results are shown in the table below.

Table 3.1: OLS Regression Results for Multiple Regression – Full Model

Dep. Variable:	mpg	**R-squared:**	0.849
Model:	OLS	**Adj. R-squared:**	0.757
Method:	Least Squares	**F-statistic:**	9.226
Date:	Fri, 21 Jul 2017	**Prob (F-statistic):**	2.53e-05
Time:	14:52:06	**Log-Likelihood:**	48.686
No. Observations:	30	**AIC:**	-73.37
Df Residuals:	18	**BIC:**	-56.56
Df Model:	11		
Covariance Type:	nonrobust		

The linear model has a correlation coefficient of 0.849 and an adjusted correlation coefficient of 0.757 (above), which indicates a strong, positive linear relationship. The F-statistic is 9.226 and significant, so we can reject the null hypothesis that the regression coefficients are zero.

Table 3.2: OLS Regression Results for Multiple Regression – Full Model

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
const	1.1203	0.583	1.922	0.071	-0.104 2.345
displacement	-0.0014	0.001	-1.217	0.239	-0.004 0.001
horsepower	-0.0020	0.002	-1.185	0.251	-0.006 0.002
torque	0.0022	0.002	1.274	0.219	-0.001 0.006
compression_ratio	0.0150	0.059	0.254	0.803	-0.109 0.139
rear_axle_ratio	0.0447	0.060	0.741	0.468	-0.082 0.171
carburetor_barrels	0.0139	0.025	0.564	0.580	-0.038 0.066
transmission_speeds	-0.0085	0.059	-0.143	0.888	-0.133 0.116
overall_length	0.0041	0.002	1.642	0.118	-0.001 0.009
width	-0.0036	0.006	-0.576	0.572	-0.017 0.009
weight	-0.0001	0.000	-1.161	0.261	-0.000 0.000
transmission_type	0.0331	0.058	0.573	0.574	-0.088 0.154

Looking at Table 3.2 above, the p-values for the feature coefficients show that none of the coefficients are significant at a significance level of 0.05. Furthermore, the confidence intervals of all the coefficients include 0. We cannot reject the null hypothesis that the coefficient is zero for any variable. Table 3.3 below shows that the model has a slightly positive skew of 0.082 and positive kurtosis of 2.744. It is difficult to interpret this model, which may be due to the effects of multicollinearity.

Table 3.3: OLS Regression Results for Multiple Regression – Full Model

Omnibus:	0.064	**Durbin-Watson:**	1.892
Prob(Omnibus):	0.969	**Jarque-Bera (JB):**	0.115
Skew:	0.082	**Prob(JB):**	0.944
Kurtosis:	2.744	**Cond. No.**	1.96e+05

The QQ Norm plot of residuals (Fig 9) show that the residuals closely follow a normal distribution, with only a slight skew at the tails. Furthermore, scatterplot of fitted values versus residuals shows that the residuals validate the assumption of linear regression. The values are slightly clustered in the center and show no apparent pattern. In the next section we use an automated variable selection method to see if using a subset of features improves the model.

Fig 9: QQ Norm Plot of Residuals, Multiple Regression – Full Model

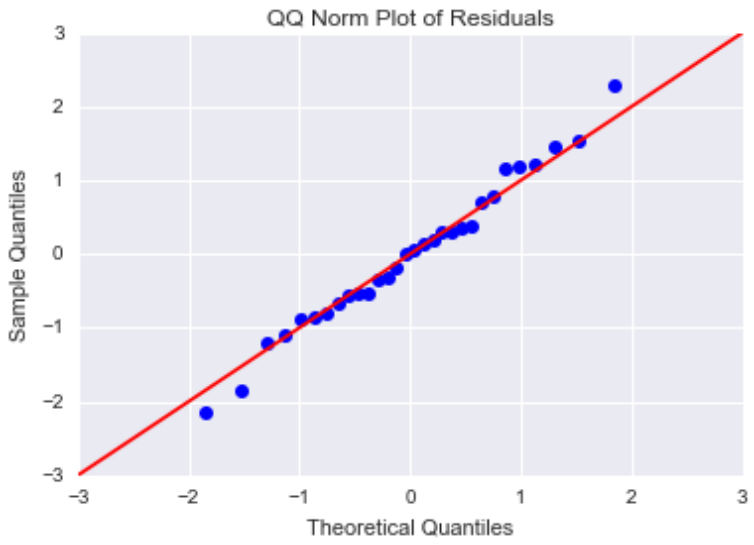
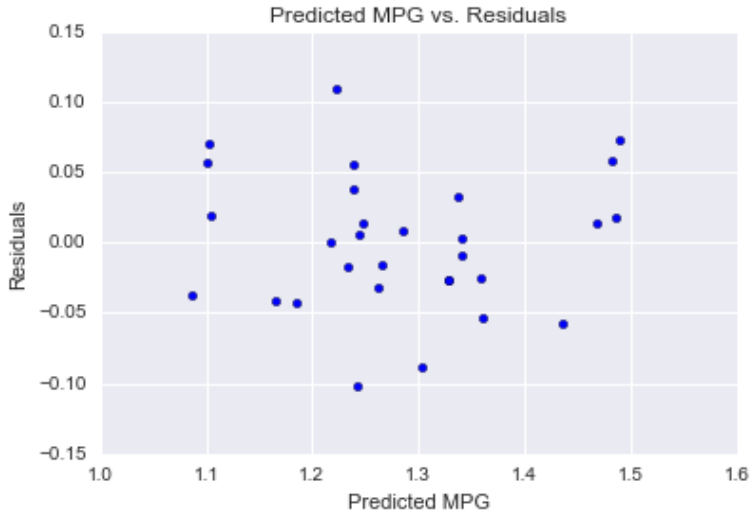


Fig 10: Predicted MPG vs Residuals, Multiple Regression – Full Model



In general, the model is difficult to interpret, as it has a high correlation coefficient and residuals that follow a normal distribution, but no significant feature coefficients.

Section 4: Multiple Linear Regression – Subset Model

We defined a multiple linear regression of mpg using a subset model. We used log-transformed mpg as the response variable and a forward selection method. The forward selection method begins with an empty equation and adds predictors one at a time. Each predictor that is not already in the model is tested for inclusion, based on whether it improves the adjusted r-squared value. The OLS regression results based on the forward selection method are shown in the table below.

Table 4.1: OLS Regression Results for Multiple Regression – Subset Model

Dep. Variable:	mpg	**R-squared:**	0.806
Model:	OLS	**Adj. R-squared:**	0.783
Method:	Least Squares	**F-statistic:**	35.98
Date:	Fri, 21 Jul 2017	**Prob (F-statistic):**	2.10e-09
Time:	14:52:08	**Log-Likelihood:**	44.884
No. Observations:	30	**AIC:**	-81.77
Df Residuals:	26	**BIC:**	-76.16
Df Model:	3		
Covariance Type:	nonrobust		

The linear model has a correlation coefficient of 0.806 and an adjusted correlation coefficient of 0.783 (above), which indicates a strong, positive linear relationship. The F-statistic is 35.98 and significant, so we can reject the null hypothesis that the regression coefficients are zero, similar to the full model.

Table 4.2: OLS Regression Results for Multiple Regression – Subset Model

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	0.9951	0.370	2.686	0.012	0.234 1.757
displacement	-0.0010	0.000	-6.374	0.000	-0.001 -0.001
compression_ratio	0.0667	0.043	1.534	0.137	-0.023 0.156
transmission_type	0.0467	0.044	1.053	0.302	-0.044 0.138

Looking at Table 4.2 above, only three features were selected by our forward selection method to be included in the model: displacement, compression ratio, and transmission type. The p-value for the displacement coefficient is significant, so we can reject the null hypothesis that the coefficient is zero. The coefficient value is negative, which validates our assumption that increasing engine displacement will decrease mpg. The p-values for the coefficients for compression ratio and transmission type are not significant, and include 0 in their confidence intervals. Table 4.3 below shows that the model has a slight positive skew of 0.175 and positive kurtosis of 2.311.

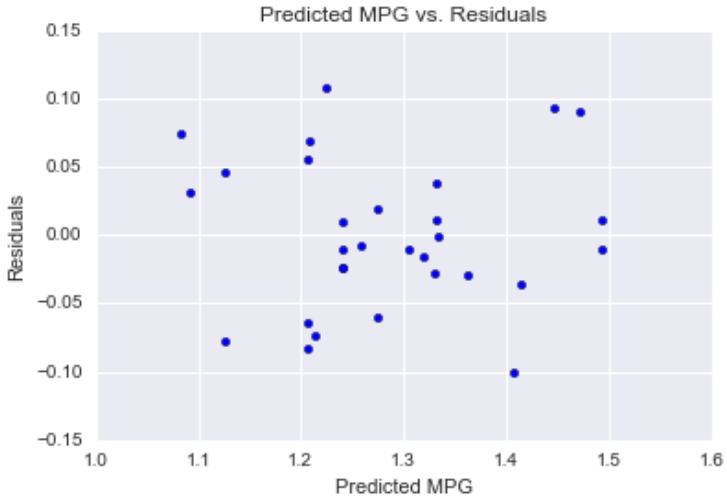
Table 4.3: OLS Regression Results for Multiple Regression – Subset Model

Omnibus:	0.709	**Durbin-Watson:**	1.713
Prob(Omnibus):	0.701	**Jarque-Bera (JB):**	0.746
Skew:	0.175	**Prob(JB):**	0.689
Kurtosis:	2.311	**Cond. No.**	1.08e+04

The QQ-Norm plot of residuals (Fig 11) shows that the residuals closely follow a normal distribution with a slight skew at the tails. However, the residuals seem to have a spiral-like pattern along the normal distribution line. Furthermore, a scatterplot of fitted values versus residuals shows that the residuals validate the assumption of linear regression. The values are slightly clustered in the center and show no apparent pattern, similar to those of the full regression model.

Fig 11: QQ Norm Plot of Residuals, Multiple Regression – Subset Model

Fig 12: Predicted MPG vs Residuals, Multiple Regression – Subset Model



The subset model seems to be a better predictor for mpg due to a higher adjusted correlation coefficient, higher F-statistic, and a significant coefficient. In the next section we compare the full and subset models further.

Section 4: Model Comparison and Recommendation

Overall, we would recommend the subset model over the full model to be used in predicting mpg. The subset model improved the adjusted correlation coefficient value from 0.757 to 0.783, which indicates a stronger positive linear relationship. Interestingly, the F-statistic for the subset model is nearly quadruple that of the full model. A high F-statistic means that our data does not reflect our null hypothesis. Therefore, the higher the F-statistic, the more confidently you can reject the null hypothesis that the regression coefficients are equal to zero. Furthermore, the full model did not have any significant feature coefficients, which makes the model very difficult to interpret. Although we decreased the number of predictors, the subset model still had a higher correlation coefficient and F-statistic. This indicates to us that there may be multicollinearity effects that are interfering in the full model

We would recommend that future analyses test different automated variable selection methods, such as backward elimination and stepwise regression, in order to further explore eliminating multicollinearity. Furthermore, researchers could further refine the model with a cross-validation method. Applying cross-validation could help compensate for bias that automated variable selection methods introduce. Last, the dataset for this analysis is small, with only 30 observations. We recommend gathering a larger dataset in order to build a more accurate model. Due to the small sample size, we should consider any significance found in this analysis with caution.

Conclusion

In conclusion, we observed that there is a strong linear relationship between displacement, compression ratio, transmission type and mpg. A log transformation of the response variable was applied to make its distribution more normal. The subset model using an automated variable selection method—forward selection—improved the correlation coefficient and the significance of the predictors, compared to the full model. Although many predictors showed linear relationships with mpg during our exploratory data analysis, none of the predictors in the full model were significant, suggesting that the predictors were collinear. Because the subset model had an improved adjusted correlation coefficient, higher F-statistic, and significant coefficient, we recommended the subset model for predicting mpg. Future studies should consider acquiring a larger dataset, applying other automated variable selection methods, and including additional engine features that are not commonly collinear.