Predictive Modeling of the Gasoline Consumption Dataset using Automated Variable Selection and Principal Component Analysis

Prepared by Amee Amin

Assignment 6

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. In this analysis, we compared predictive models for vehicle gasoline consumption using automated variable selection and principal component analysis-informed multiple regression.

The data comes from 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.

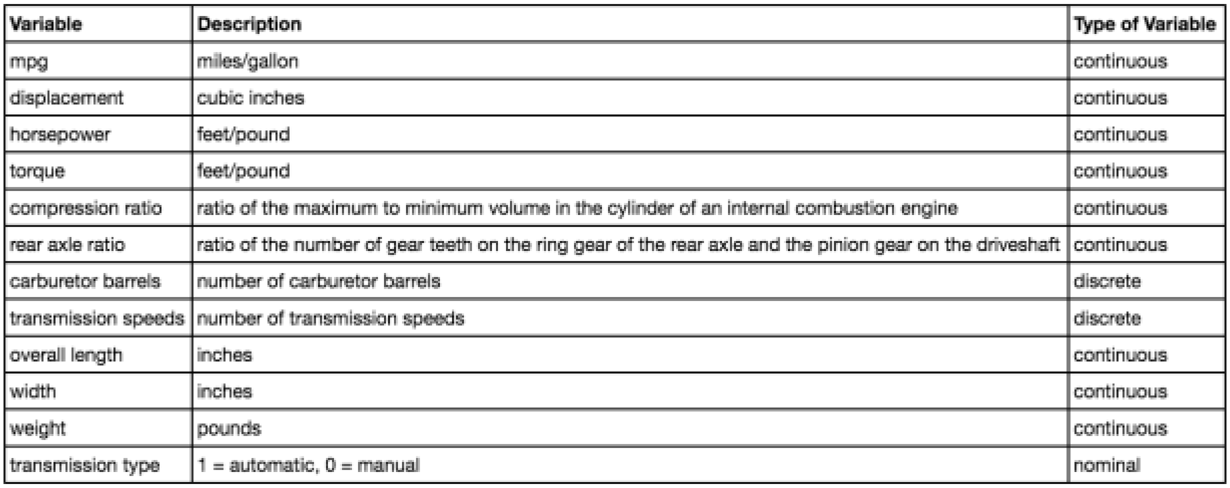
Using a log transformation of the response variable, we found that a PCA-informed model with a variance threshold of 95% had the highest r-squared value and eliminated multicollinearity. The PCA-informed model with a variance threshold of 70% had a lower r-squared value than the subset model using forward selection, which had the second highest r-squared value. The full model had no significant predictors and displayed characteristics of multicollinearity. We recommend that a PCA-informed model be further refined and used by management to predict vehicle gasoline consumption.

Section 1: Sample Data

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 its 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.

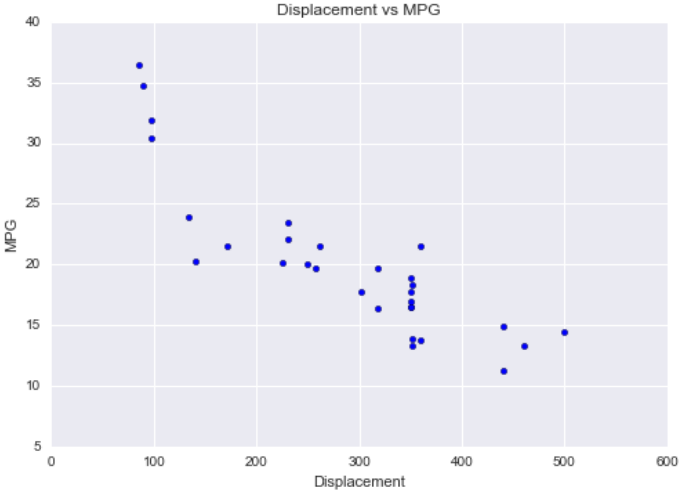
Section 2: Exploratory Data Analysis

Section 2.1: Scatterplots of Predictor vs Response Variable

We created scatterplots of each of the predictor variables with our response variable mpg to explore the linearity of their relationships.

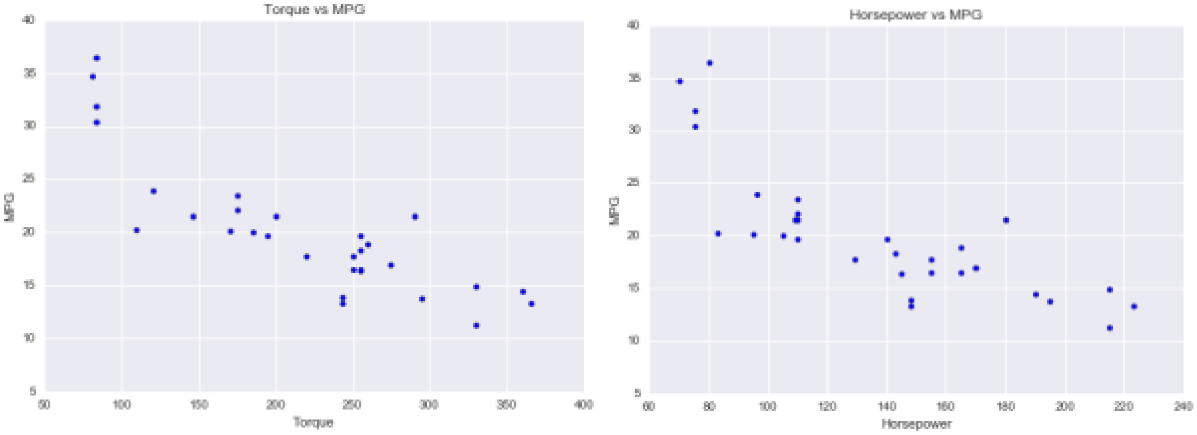
In the scatterplot below, we observed that mpg decreases as engine displacement increases, which is the pattern we expected. Most vehicles in our dataset had an engine displacement between 200 and 400 cubic inches. There are 4 vehicles which have a much higher mpg and an engine displacement below 100. 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 2.1.1: Scatterplot of Displacement vs MPG



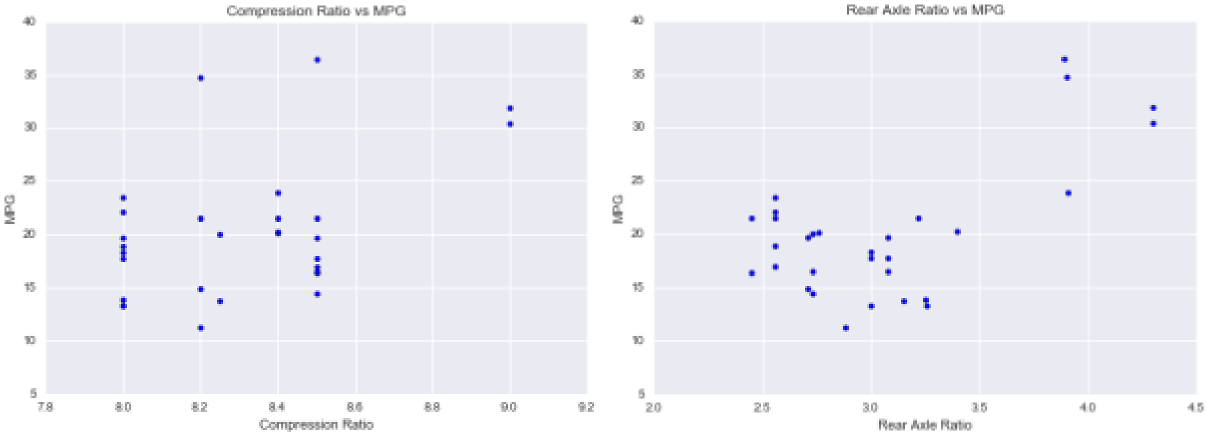
As a vehicle's horsepower and torque increase (below), we also observed that mpg tends to decrease. Looking at a side-by-side comparison of torque and horsepower to mpg, the variables have nearly identical scatterplots and we can expect to see some collinearity in our model. Given that torque is in the 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 2.1.2: 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 other engine components influence overall fuel economy more than compression ratio.

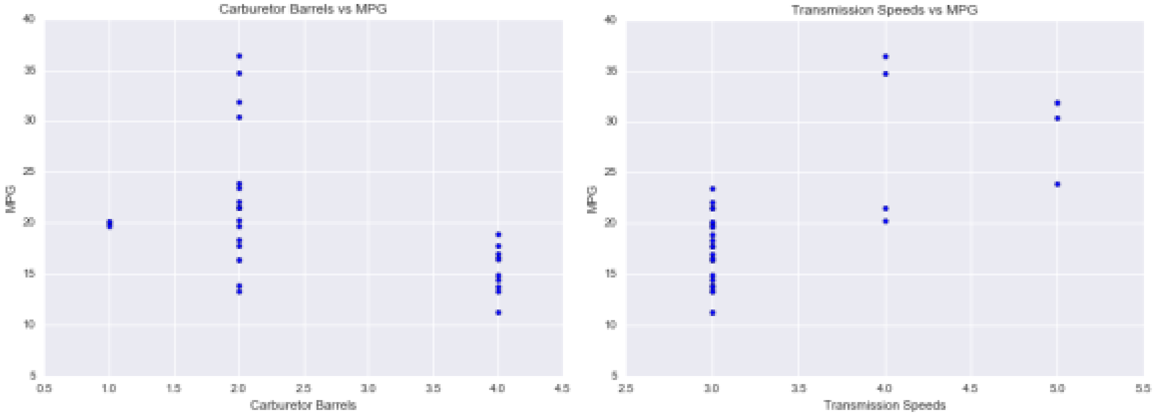
Fig 2.1.3: Scatterplots of Compression Ratio (left) and Rear Axle Ratio (right) vs MPG



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.

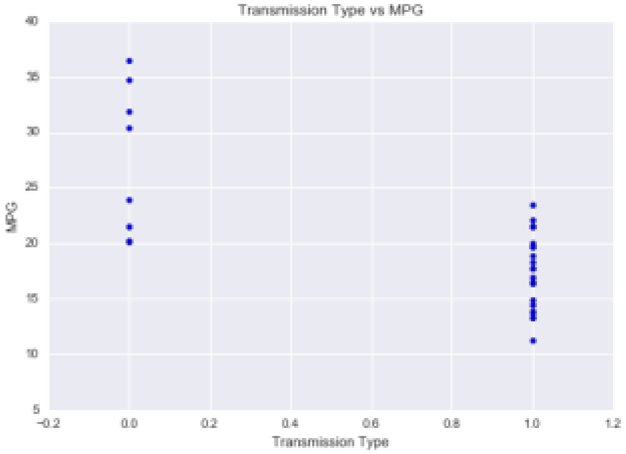
Scatterplots of carburetor barrels and transmission speeds with mpg below show the patterns we expected: mpg tends to decrease with an increasing number of carburetor barrels, and mpg tends to increase with an increasing number of transmission speeds. There seems to be a variation of 10-15 mpg for vehicles 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 2.1.4: Scatterplots of Carburetor Barrels (left) and Transmission Speeds (right) vs MPG



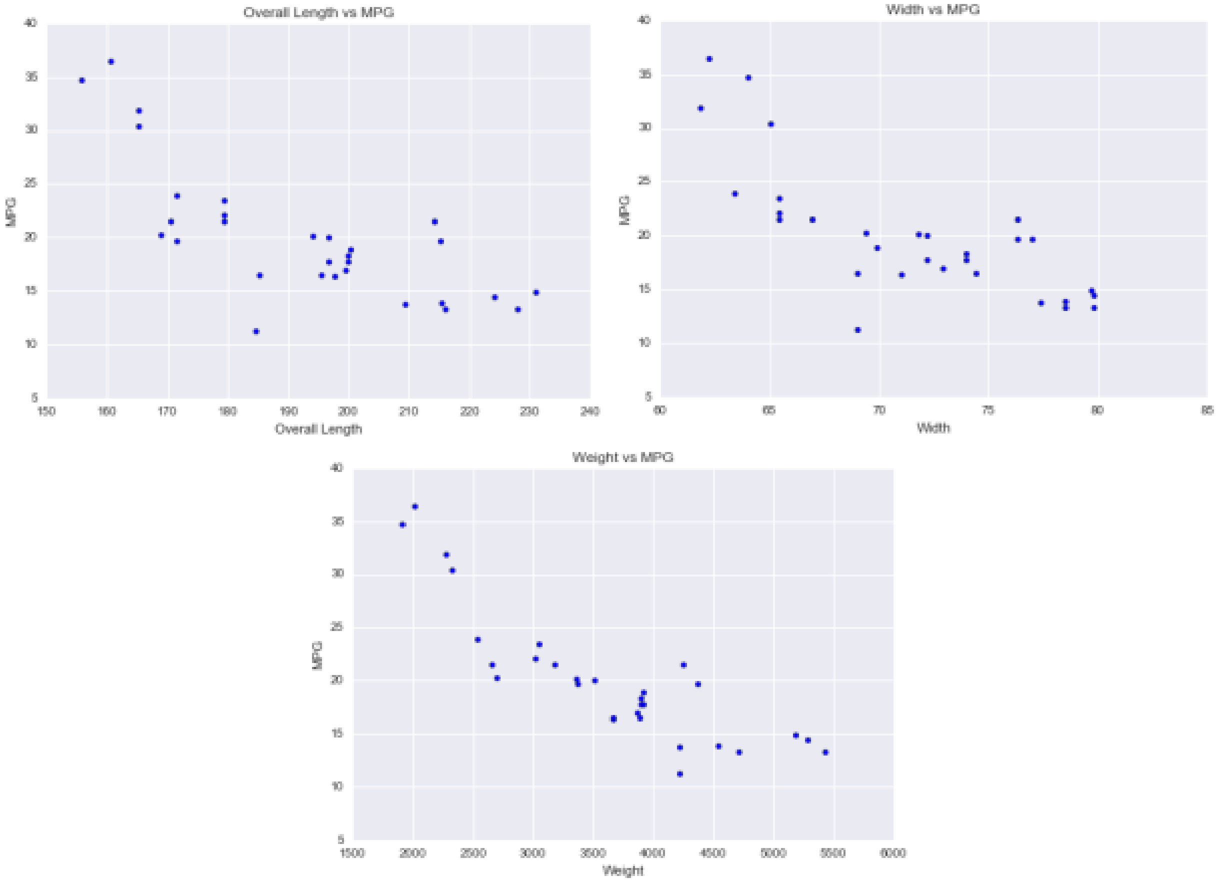
The scatterplot of transmission type versus mpg (below) shows that vehicles with automatic transmission clearly have lower mpg than vehicles with manual transmission. Vehicles with manual transmission have an mpg between 20 and 35, and vehicles with automatic transmission have an mpg between 10 and 25.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

Fig 2.1.5: 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.

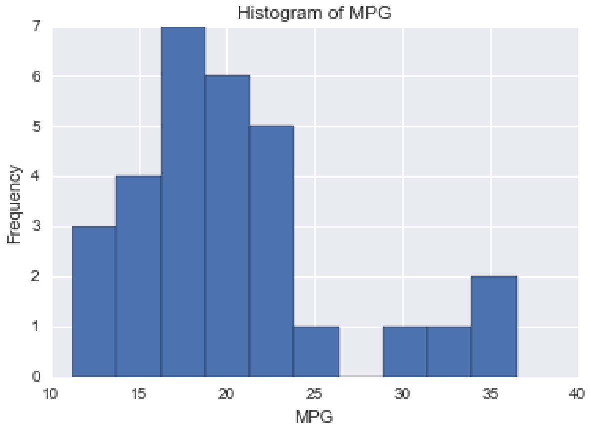
Fig 2.1.6: Scatterplots of Overall Length (Left), Width (Right), and Weight (below) vs MPG



Section 2.2: Transformation of the Response Variable

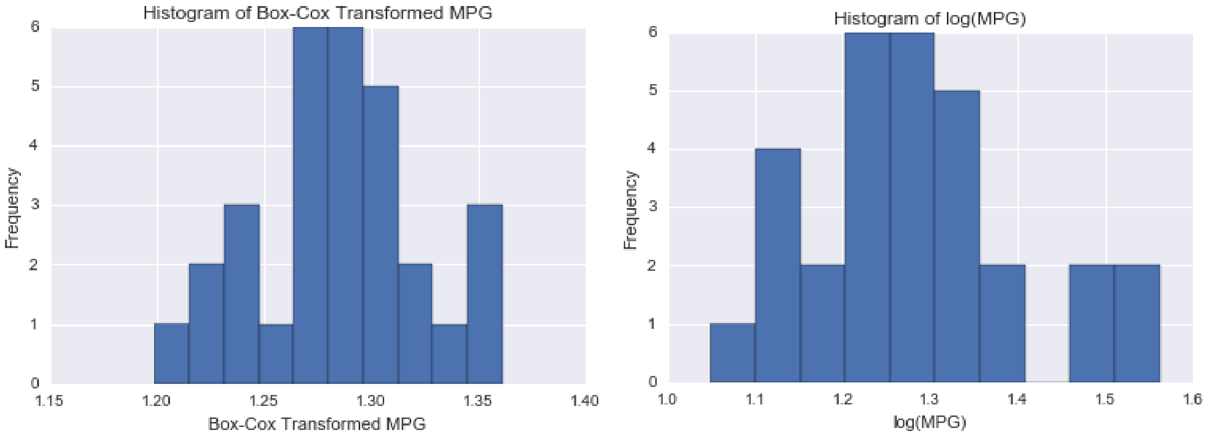
We observed the distribution of the response variable in order to determine whether our dataset validates the assumptions of linear regression. In order for the linear regression method to be accurate, a response variable should have at least a conditional normal distribution. In the histogram of mpg 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 2.2.1: 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 2.2.2: Histogram Comparison of Box-Cox (left) and Log (right) Transformations of MPG



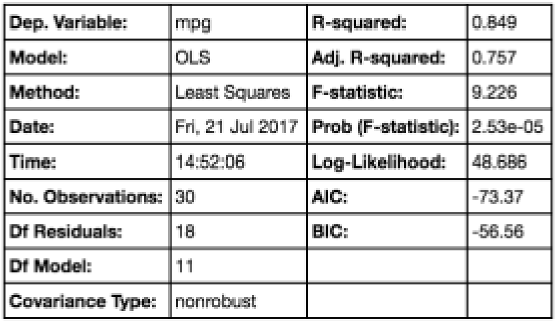
We performed three linear regressions using weight as the independent variable to determine which response transformation best predicted mpg. The correlation coefficient for each model was the following: weight versus mpg: 0.727; weight versus Box-Cox transformed mpg: 0.733; weight versus Log-transformed mpg: 0.750. Because the log transformation produced the largest correlation coefficient, we decided to use a log transformation of mpg for future models.

Section 3: Multiple Linear Regression Executive Summary

Section 3.1: 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.

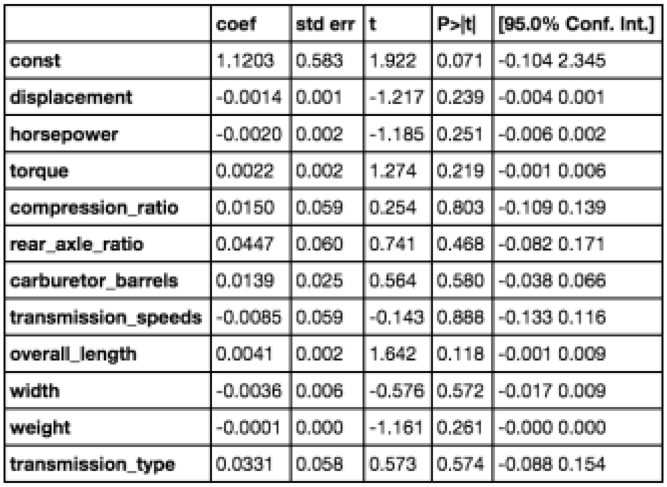
Table 3.1.1: OLS Regression Results for Multiple Regression – Full Model



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.

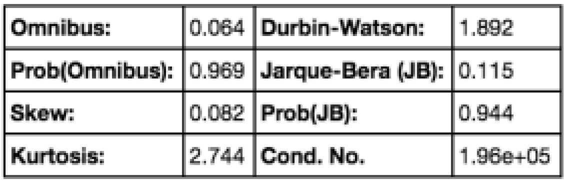
Table 3.1.2: OLS Regression Results for Multiple Regression – Full Model



	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 3.1.3: 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 3.1.1: QQ Norm Plot of Residuals, Multiple Regression – Full Model

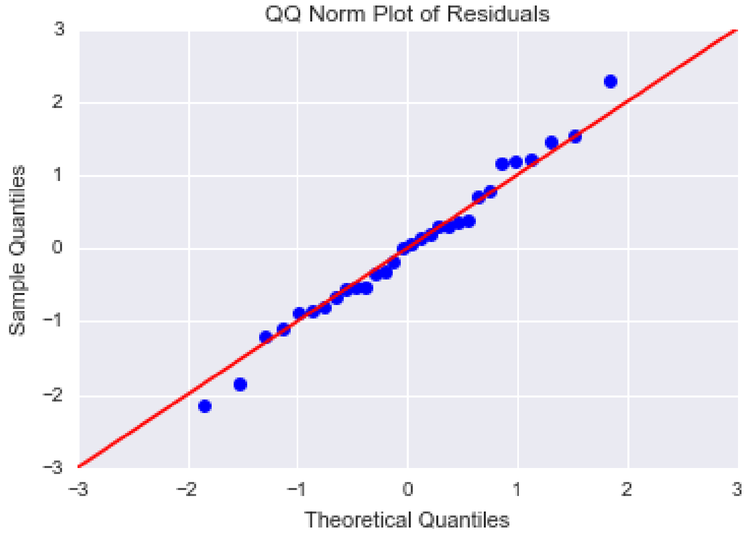
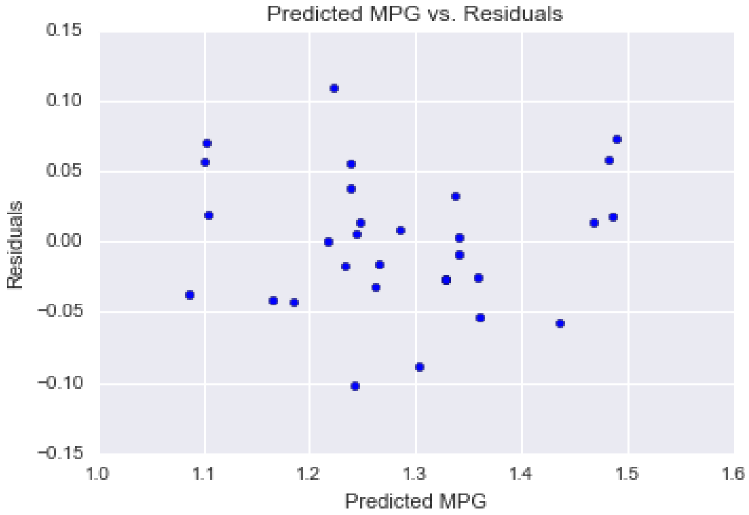


Fig 3.1.2: Predicted MPG vs Residuals, Multiple Regression – Full Model

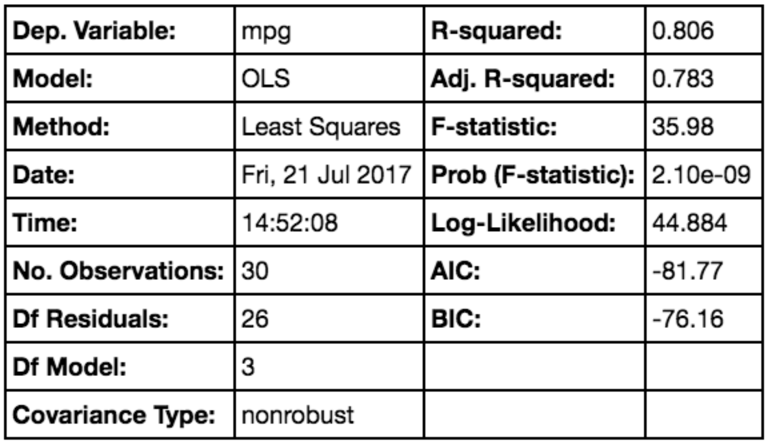


In general, the model is difficult to interpret, as it has a high correlation coefficient and residuals that follow a normal distribution, but no significant feature coefficients.

Section 3.2: 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.

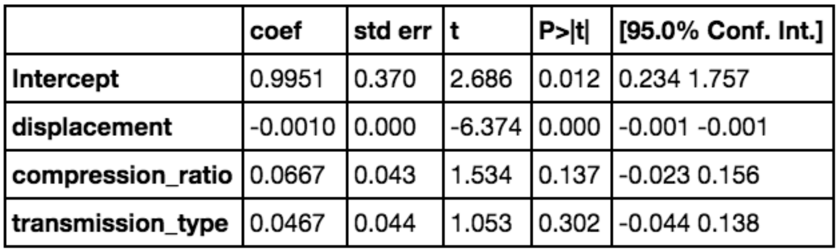
Table 3.2.1: OLS Regression Results for Multiple Regression – Subset Model



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.

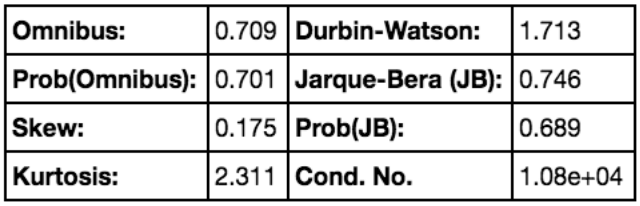
Table 3.2.2: OLS Regression Results for Multiple Regression – Subset Model



	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 3.2.3: 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 3.2.1: QQ Norm Plot of Residuals, Multiple Regression – Subset Model

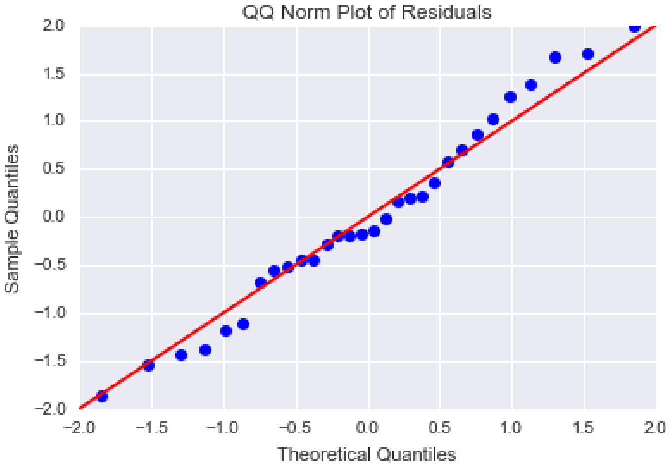
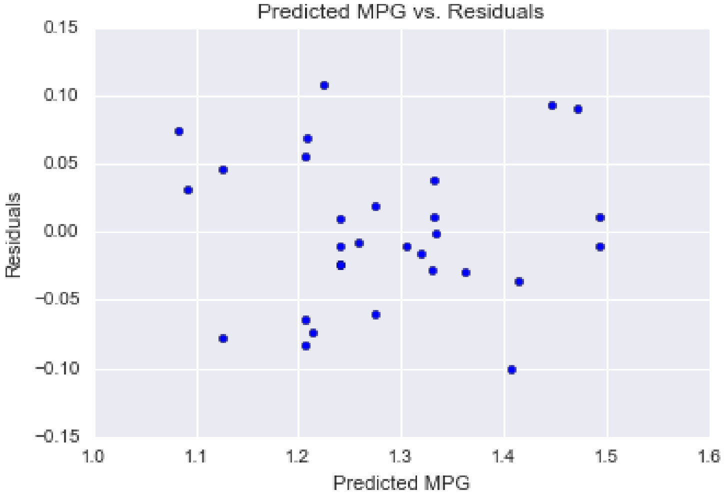


Fig 3.2.2: Predicted MPG vs Residuals, Multiple Regression – Subset Model



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: Principal Component Analysis

We performed a principal component analysis (PCA) on the explanatory variables to de-noise our data and remove variables that create collinearity. The principle components are the dimensions along which our data is most differentiated, as expressed by one, two, or more variables.

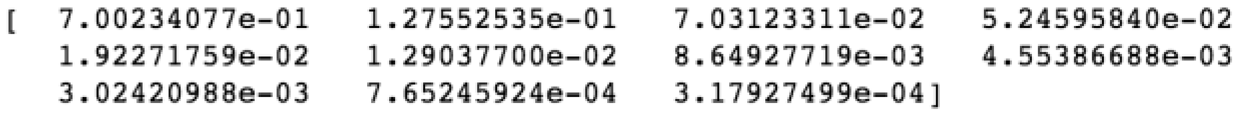
Before applying PCA, we preprocessed the data through mean normalization. We computed the mean μ of each feature, and then replaced each feature with X- μ. The result is that each feature has a mean μ of zero and unit variance. Preprocessing is important because different features have different scales. If we are comparing variation within features, each feature has to have a comparable range of values.

PCA reduces the number of variables by extracting the correlations between them. In terms of linear algebra, each of our 11 explanatory variables has a correlation with the other variables. PCA begins with a covariance matrix that undergoes an orthogonal transformation to produce a matrix of eigenvector and eigenvalue pairs. An eigenvector is a direction and an eigenvalue is a number representing the amount of variance there is in that direction. The number of eigenvectors and values is equal to the number of dimensions in a dataset, therefore our dataset has 11 eigenvectors and values.

The matrix of eigenvectors and values is then rearranged to reflect the order of decreasing eigenvalue, or a decreasing amount of variance. The direction in which there is the most variance, or the highest eigenvalue, becomes the principal component. The principal components are linear combinations of the original variables, weighted by their contribution to explaining variance in a specific orthogonal dimension. In layman's terms, the principal components reflect an underlying dimension that summarizes or accounts for the original variables. When computing PCA, a threshold is set to determine how much variance should be explained by the principal component(s). Additional components can be added, depending on whether the first component does or does not meet the threshold.

We applied PCA to our explanatory variables using the sklearn package. The total explained variance ratio for each component is listed below.

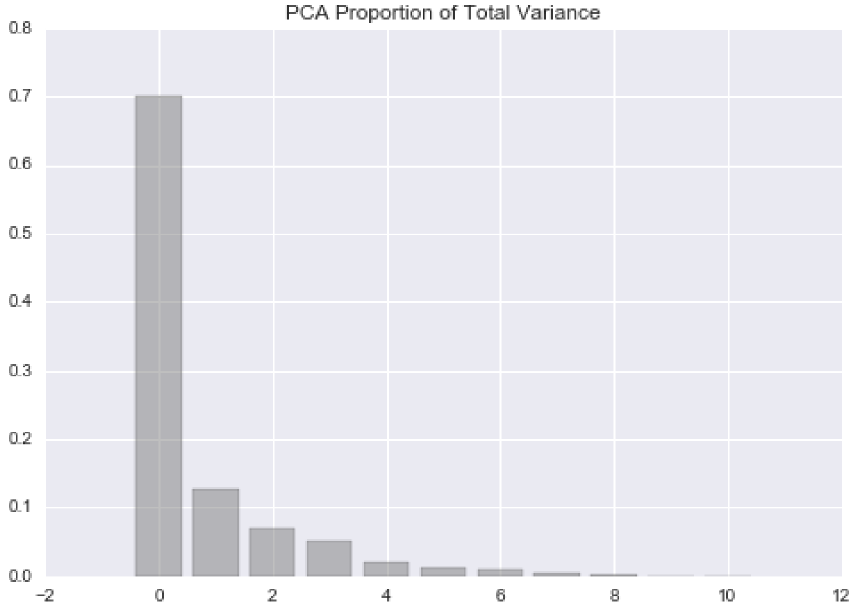
Table 4.1: PCA Output Explained Variance Ratio



```
[  7.00234077e-01   1.27552535e-01   7.03123311e-02   5.24595840e-02
   1.92271759e-02   1.29037700e-02   8.64927719e-03   4.55386688e-03
   3.02420988e-03   7.65245924e-04   3.17927499e-04]
```

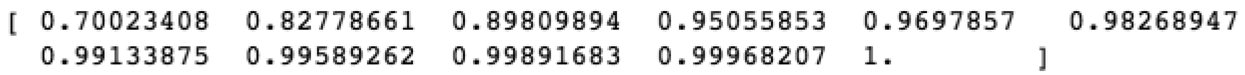
The explained variance ratio represents the proportion of variance explained by each of the selected components. The output above is ordered to reflect each component's explained variance ratio in decreasing order. The first component by far explains the most variance in our data.

Figure 4.1: Scree Plot of PCA



The scree plot above further shows how there is a huge jump in explained variance ratio from the first component to the second component. The difference between the first component and second component is greater than that of any further sequential components.

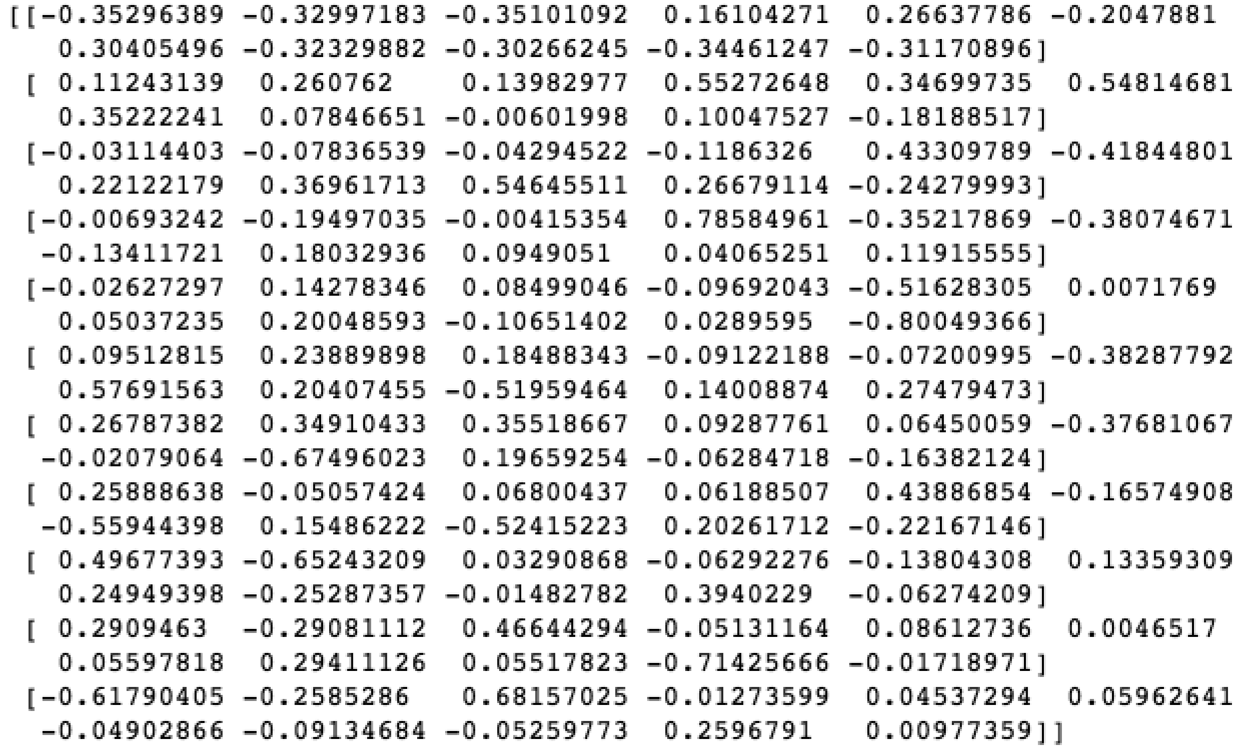
Table 4.2: PCA Output Cumulative Explained Variance Ratio



```
[ 0.70023408  0.82778661  0.89809894  0.95055853  0.9697857   0.98268947
  0.99133875  0.99589262  0.99891683  0.99968207  1.        ]
```

The cumulative explained variance ratio, shown in the table above, reveals the cumulative proportion of variance explained by each of the selected components. Generally, researchers look to determine the principal components that explain 0.95 of the variance. The first component reflects 0.700 of the variance in our data. The first and second components reflect 0.828 of the variance in our data. The first, second, and third components reflect 0.898 of the variance in our data. Finally, the first, second, third, and fourth components reflect 0.951 of the variance in our data.

Table 4.3: PCA Output Factor Loadings



```
[[-0.35296389 -0.32997183 -0.35101092  0.16104271  0.26637786 -0.2047881
   0.30405496 -0.32329882 -0.30266245 -0.34461247 -0.31170896]
 [ 0.11243139  0.260762    0.13982977  0.55272648  0.34699735  0.54814681
   0.35222241  0.07846651 -0.00601998  0.10047527 -0.18188517]
 [-0.03114403 -0.07836539 -0.04294522 -0.1186326   0.43309789 -0.41844801
   0.22122179  0.36961713  0.54645511  0.26679114 -0.24279993]
 [-0.00693242 -0.19497035 -0.00415354  0.78584961 -0.35217869 -0.38074671
  -0.13411721  0.18032936  0.0949051   0.04065251  0.11915555]
 [-0.02627297  0.14278346  0.08499046 -0.09692043 -0.51628305  0.0071769
   0.05037235  0.20048593 -0.10651402  0.0289595  -0.80049366]
 [ 0.09512815  0.23889898  0.18488343 -0.09122188 -0.07200995 -0.38287792
   0.57691563  0.20407455 -0.51959464  0.14008874  0.27479473]
 [ 0.26787382  0.34910433  0.35518667  0.09287761  0.06450059 -0.37681067
  -0.02079064 -0.67496023  0.19659254 -0.06284718 -0.16382124]
 [ 0.25888638 -0.05057424  0.06800437  0.06188507  0.43886854 -0.16574908
  -0.55944398  0.15486222 -0.52415223  0.20261712 -0.22167146]
 [ 0.49677393 -0.65243209  0.03290868 -0.06292276 -0.13804308  0.13359309
   0.24949398 -0.25287357 -0.01482782  0.3940229  -0.06274209]
 [ 0.2909463  -0.29081112  0.46644294 -0.05131164  0.08612736  0.0046517
   0.05597818  0.29411126  0.05517823 -0.71425666 -0.01718971]
 [-0.61790405 -0.2585286   0.68157025 -0.01273599  0.04537294  0.05962641
  -0.04902866 -0.09134684 -0.05259773  0.2596791   0.00977359]]
```

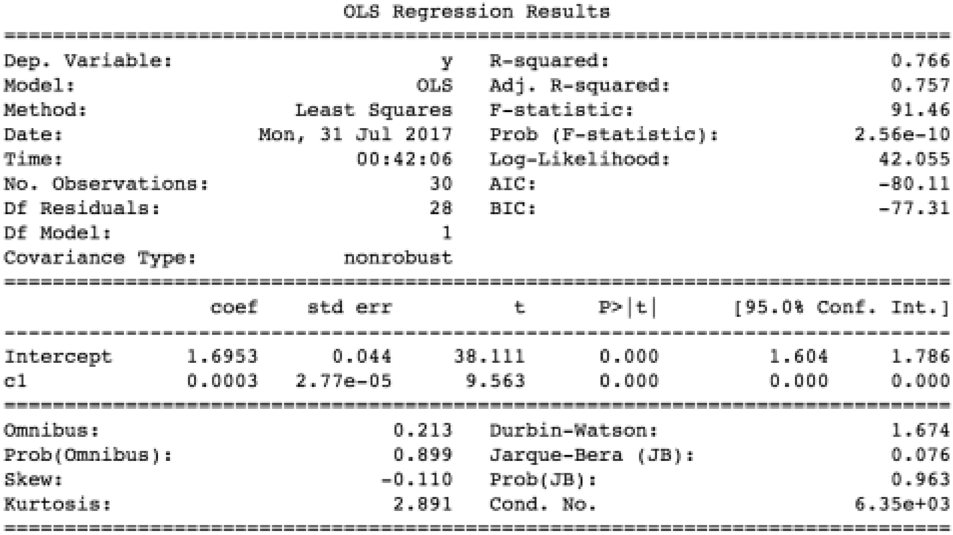
Each principal component consists of a set of factor loadings, listed in the table above. A factor loading represents the correlation between the original variable and the factor (the new underlying dimension). In the next section, we input the factor loadings for principal components that account for 70% and 95% of the variance into multiple regression models to predict mpg.

Section 5: Principal Component Regression

We performed two linear regressions on the principal components determined in Section 4.

First, we regressed a linear model on the first PCA component that captured 70% of variance in our data. The OLS output is shown below.

Table 5.1: OLS Regression Results for PCA-informed Linear Regression, 70% threshold

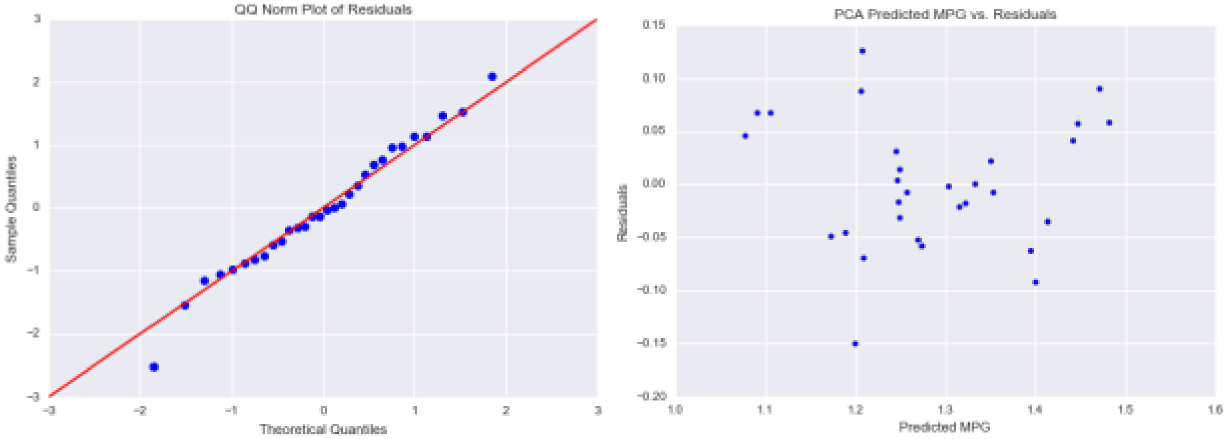


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.766
Model:                            OLS   Adj. R-squared:                  0.757
Method:                 Least Squares   F-statistic:                     91.46
Date:                Mon, 31 Jul 2017   Prob (F-statistic):           2.56e-10
Time:                        00:42:06   Log-Likelihood:                 42.055
No. Observations:                  30   AIC:                            -80.11
Df Residuals:                      28   BIC:                            -77.31
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      1.6953      0.044     38.111      0.000         1.604     1.786
c1             0.0003   2.77e-05      9.563      0.000         0.000     0.000
==============================================================================
Omnibus:                        0.213   Durbin-Watson:                   1.674
Prob(Omnibus):                  0.899   Jarque-Bera (JB):                0.076
Skew:                          -0.110   Prob(JB):                        0.963
Kurtosis:                       2.891   Cond. No.                     6.35e+03
==============================================================================
```

The correlation coefficient is 0.766, which indicates a strong positive linear relationship between the first principal component and mpg. The F-statistic is significant, so we can reject the null hypothesis that the regression coefficient is zero. The resulting distribution has a slightly negative skew and positive kurtosis.

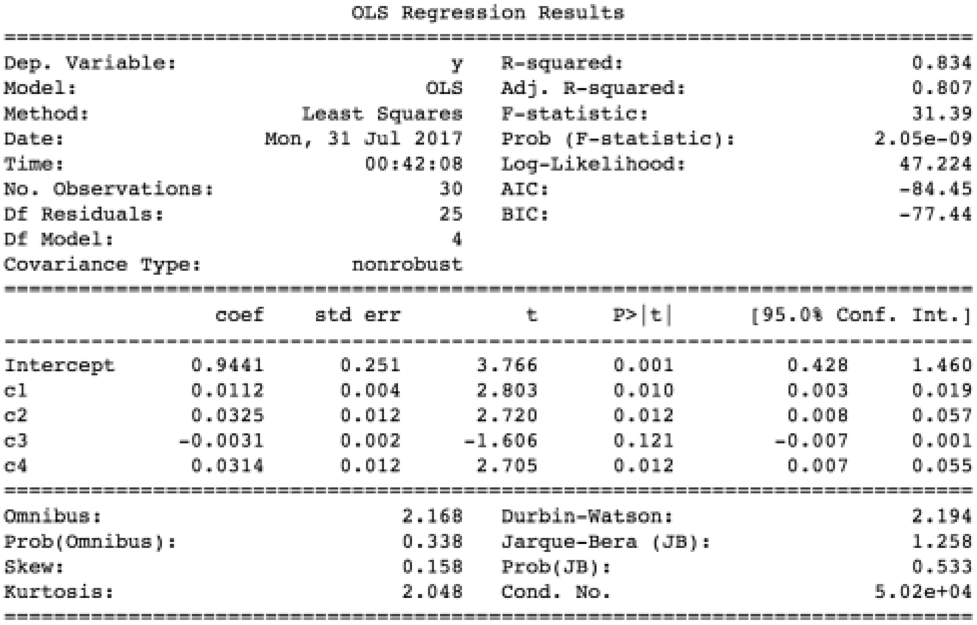
The QQ-Norm plot of the residuals (below, left), shows that the residuals closely follow a normal distribution with a slight skew at the tails. The scatterplot of fitted values versus residuals (below, right) does not display any obvious patterns, and the residuals appear to be approximately symmetrically distributed above and below 0. Both of these characteristics validate the assumption of residual normality for linear regression modeling.

Figure 5.1: QQ Norm Plot (left) and Fitted Value-Residual Scatter Plot (right)



Second, we regressed a linear model on the first, second, third, and fourth PCA components that captured 95% of variance in our data. The OLS output is shown below.

Table 5.2: OLS Regression Results for PCA-informed Linear Regression, 95% threshold

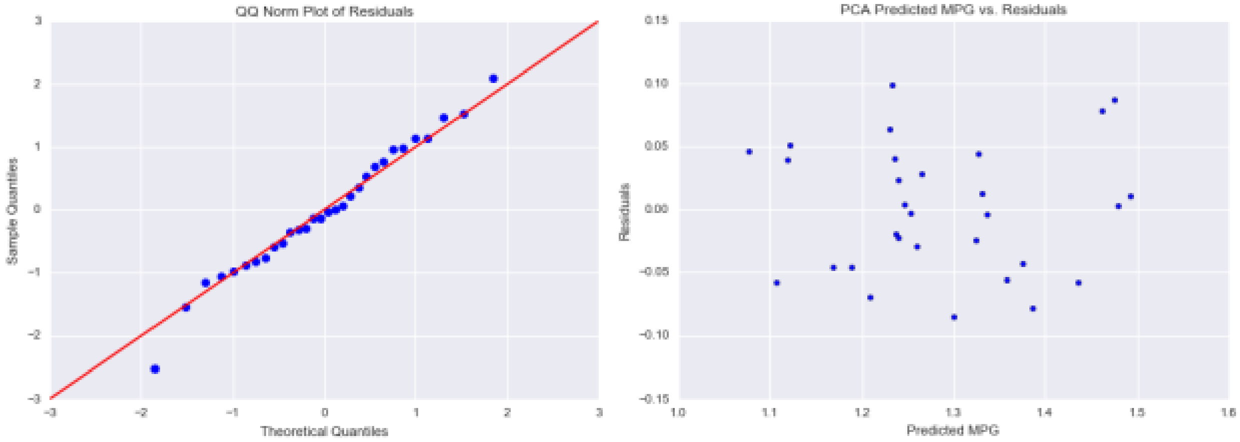


```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.834
Model:                            OLS   Adj. R-squared:                  0.807
Method:                 Least Squares   F-statistic:                     31.39
Date:                Mon, 31 Jul 2017   Prob (F-statistic):           2.05e-09
Time:                        00:42:08   Log-Likelihood:                 47.224
No. Observations:                  30   AIC:                            -84.45
Df Residuals:                      25   BIC:                            -77.44
Df Model:                           4                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      0.9441      0.251      3.766      0.001         0.428     1.460
c1             0.0112      0.004      2.803      0.010         0.003     0.019
c2             0.0325      0.012      2.720      0.012         0.008     0.057
c3            -0.0031      0.002     -1.606      0.121        -0.007     0.001
c4             0.0314      0.012      2.705      0.012         0.007     0.055
==============================================================================
Omnibus:                        2.168   Durbin-Watson:                   2.194
Prob(Omnibus):                  0.338   Jarque-Bera (JB):                1.258
Skew:                           0.158   Prob(JB):                        0.533
Kurtosis:                       2.048   Cond. No.                     5.02e+04
==============================================================================
```

The adjusted correlation coefficient is 0.807, which indicates a strong positive linear relationship between the first through fourth principal components and mpg. This correlation coefficient is higher than that of the first component linear regression, which is expected since we added variables to the model. The F-statistic is significant, so we can reject the null hypothesis that all of the regression coefficients are zero. The regression coefficients for the first, second, and fourth principal components are significant at a significance level of 0.05. However, the regression coefficient for the third principal component is not significant. The resulting distribution has a slightly positive skew and positive kurtosis.

The QQ-Norm plot of the residuals (below, left), shows that the residuals closely follow a normal distribution with a slight skew at the tails, similar to the first principal component linear regression. The scatterplot of fitted values versus residuals (below, right) also does not display any obvious patterns, and the residuals appear to be approximately symmetrically distributed above and below 0. The residuals appear to be more centrally clustered than those of the first principal component linear regression.

Figure 5.2: QQ Norm Plot (left) and Fitted Value-Residual Scatter Plot (right)



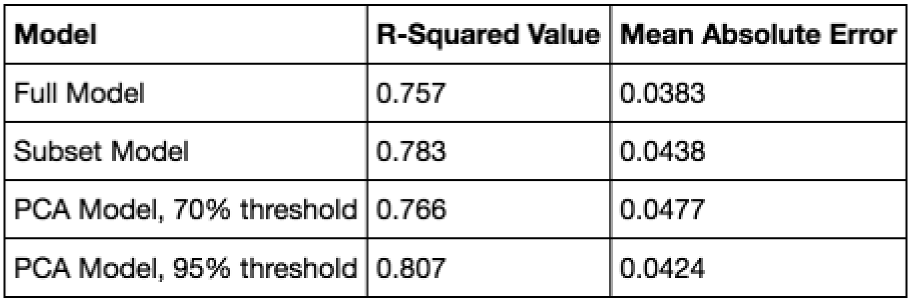
In the next section, we compare and contrast the four linear regression models produced in this analysis.

Section 6: Model Comparison and Recommendation

We built four different models to predict mpg in this analysis.

1. A multiple linear regression model including all features
2. A subset multiple linear regression model, using features determined by Forward Selection (an automated variable selection method)
3. A PCA-informed simple linear regression model, variance threshold of 70%
4. A PCA-informed multiple linear regression model, variance threshold of 95%

Table 6.1: Model Comparison of R-Square Values and MAEs



Model	R-Squared Value	Mean Absolute Error
Full Model	0.757	0.0383
Subset Model	0.783	0.0438
PCA Model, 70% threshold	0.766	0.0477
PCA Model, 95% threshold	0.807	0.0424

In order to compare these models, we listed their r-squared and mean absolute error values in the table above. The multiple regression full model had the lowest r-squared value, and the PCA model with a variance threshold of 95% had the highest r-squared value of 0.807, which is strongly positive. The PCA Model with a threshold of 95% only included four features while the multiple regression full model included 11 features; the PCA Model simplified the output model, although the regression coefficients in the PCA Model are more difficult to interpret since they reflect underlying dimensions that summarize the original variables. We may infer that PCA did reduce multicollinearity in our model. The subset model actually had a higher r-squared value than the PCA model with a 70% threshold. This suggests that the forward selection method produced a more accurate model than the PCA model with a 70% threshold, and a higher threshold should be tested in the future, such as 85%.

Interestingly, the multiple regression full model had the lowest mean absolute error. This may be due to the fact that this model is the most complex and over-fitting the data. The PCA Model with a threshold of 95% had the second lowest mean absolute error.

Overall, we would recommend the PCA Model with a threshold of 95% to be used to predict vehicle mpg, or gasoline consumption. This is because it produced the second-lowest mean absolute error in addition to the highest r-squared value. However, we noted that the PCA-informed regression is far more complicated to interpret and communicate than a subset model, because each principal component represents a set of factor loadings, not one explanatory variable. The subset model, in comparison is easier to explain because the ultimate model still reflects three distinct explanatory variables. Thus, if management seeks to prioritize model accuracy and avoid multicollinearity, the PCA model with a threshold of 95% variance should be used. If management seeks to prioritize a model that can be easily communicated and interpreted, then the subset model should be used.

Conclusion

In conclusion, we recommend a PCA-informed model with a variance threshold of 95% to predict vehicle gasoline consumption. We observed 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. PCA further improved our model by eliminating multicollinearity, although it made the outcome model more difficult to interpret.

We recommend that future analyses test different automated variable selection methods, such as backward elimination and stepwise regression, in order to further explore eliminating multicollinearity. Future analysis could also include additional engine features that are not commonly collinear. Furthermore, researchers could further refine the model with a cross-validation method. 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.