University of Glasgow
School of Mathematics and Statistics

Investigating the Fuel Consumption Dataset

By David Quinlan

Contents

1 Introduction

The fuel dataset contains details of fuel consumption in the USA for the year 1974. This dataset contains the fuel consumption records for each of the 48 contiguous states of the USA, with 5 variables recorded for each of the 48 observations. The variables contained within this dataset are explained in more detail below:

- **Income:** Mean income of each state, given in thousands of dollars.
- **Roads:** Total length of roads, given in thousands of miles.
- **Drivers:** Percentage of the adult population of a state with a drivers licence.
- **Fuel:** Fuel consumption recorded as gallons/person.
- **TaxRate:** Recorded as the number of cents of tax paid per gallon.

It is of interest to examine the relationship between the tax rate and fuel consumption. The main goals of this investigation are:

- To use formal and informal methods to build a regression model for fuel consumption and estimate the relationship between fuel consumption and the tax rate.
- To check if the model assumptions are valid for the fitted regression model.
- Finally, to fully interpret and evaluate the results of the most appropriate model.

The details of this investigation are stated in the sections below.

2 Initial Exploratory Analysis

Before fitting a linear model between the response variable Fuel and the explanatory variables TaxRate, Income, Drivers and Roads, it is important to fully examine and explore the dataset.

First, boxplots were created to graphically examine the key summary statistics of each of the continuous variables. Boxplots can also be used to identify possible outliers present in each of the variables. The boxplots of the continuous variables from the fuel dataset are shown below in Figure 1.

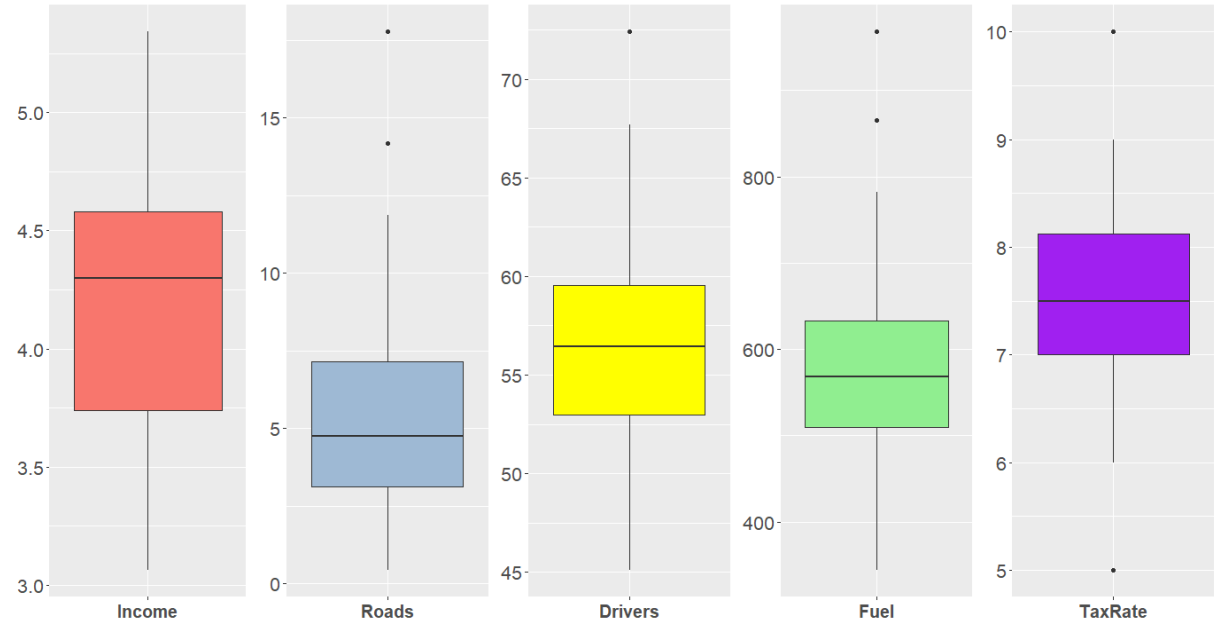

Figure 1: Boxplots of the contineous variables contained within the fuel dataset

From the boxplots it is clear that there are some possible outliers present in the Roads, Drivers, Fuel and TaxRate[1] variables. However, these possible outliers will not be removed for the moment as they may prove to be important when fitting a regression model later in this investigation.

Similarly, histograms were created to examine the distribution of the continuous variables contained within the dataset. These histograms are shown below in Figure 2.

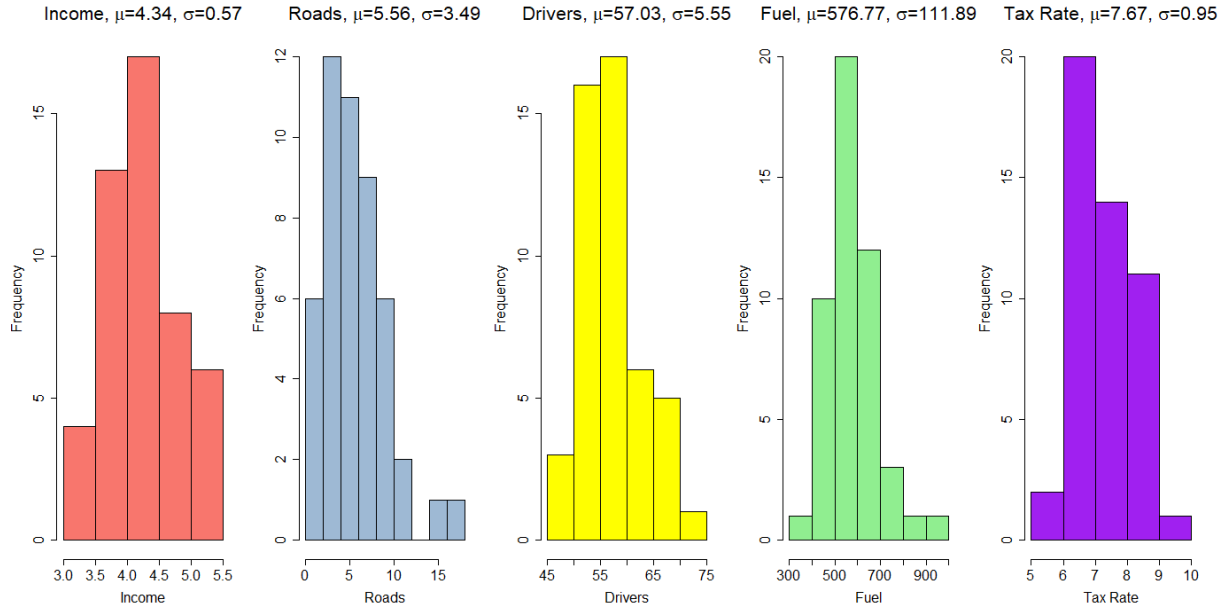

Figure 2: Histograms of the contineous variables contained within the fuel dataset

From the above histograms, the Income, Drivers and TaxRate variables appear to be approximately normally distributed. The Roads variable appears to slightly skewed to the right. Finally, it is important to note how the variable Fuel is approximately normally distributed, if this had not been the case it may have been necessary to

[1]The TaxRate variable was considered to be a categorical variable, however, due to some states possessing a unique tax rate, when fitting a linear regression model, these states were identified as having large leverage and were considered as being possible outliers in the model. I wished to avoid this issue and therefore, decided to consider the TaxRate variable as a continuous variable.

transformation this response variable to ensure that the constant variance assumption holds when examining the linear models diagnostics plots later in this investigation. The means and standard deviations of each variable are presented above their corresponding histogram in the plot shown above.

It is important to examine the relationship between the response variable Fuel and the explanatory variables. To examine these relationships scatterplots were created for each pairwise combination of variable contained within the dataset. These scatterplots are presented below in Figure 3.

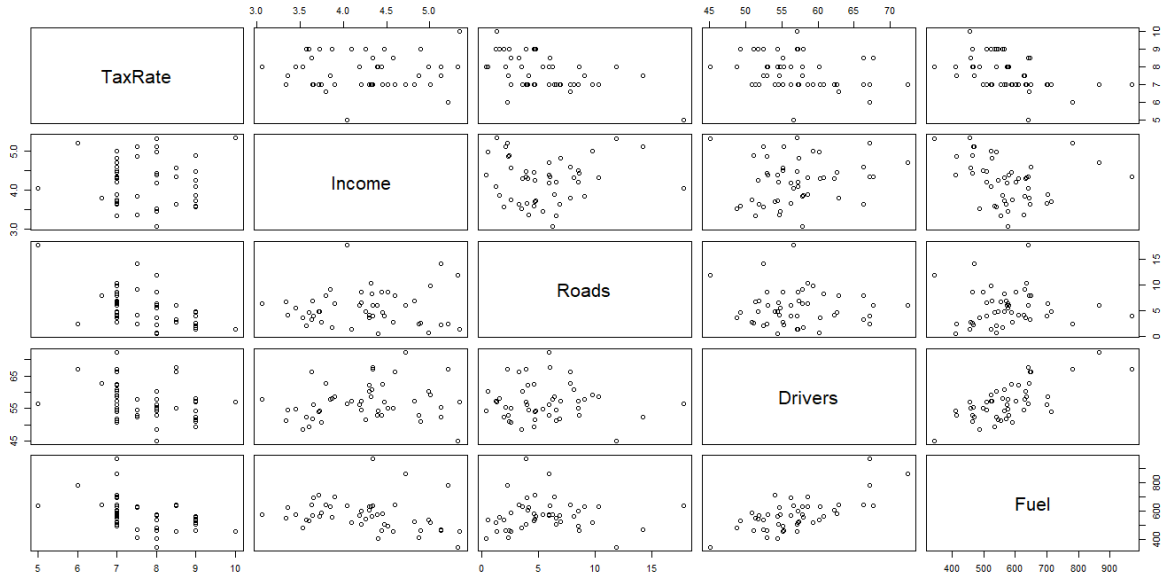

Figure 3: Scatterplots of each pairwise combination of variable contained within the fuel dataset.

Examining the scatterplots, there does not appear to be many significant relationships between the variables. The only clear relationship that appears to exist is a strong positive linear relationship between the Drivers and Fuel variables. This relationship had a correlation coefficient of approximately 0.70.

3 Methods

3.1 Multiple Linear Regression

A multiple linear regression model is a modelling approach whereby a linear relationship is fitted between the continuous response variable and a number of explanatory variables.
The linear regression model is given by the following formula:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \quad \text{For } i = 1, ..., n. \tag{1}$$

Where β_0 is the intercept term and $\beta_1, ..., \beta_p$ are the coefficients of the explanatory variables.

The linear regression assumptions are given as follows:

- There is a linear relationship between the response variable and the explanatory variables.
- The values $x_{i1}, ..., x_{ip}$ are fixed values.
- The explanatory variables are independent (i.e. no presence of multicollinearity).
- Error terms are independent and $N(0, \sigma^2)$ distributed with zero mean and variance σ^2.

4 Results and Discussion

In this section a linear regression model was used to examine the relationship between the response variable Fuel and the explanatory variables Roads, Income, TaxRate and Drivers. An adjusted R^2 value of 0.64 was obtained for the full model. This indicates how this model explained approximately 64% of the variation present in the response variable Fuel. Examining the summary statistics, the Roads variable obtained a p-value of 0.486 indicating that it was statistically insignificant at a 5% level of significance. Therefore, it was removed from the model and the model was re-fitted.

After re-fitting the model, it was found that all of the explanatory variables were statistically significant at a 5% level of significance. This reduced model obtained an adjusted R^2 value of 0.65, this indicates how the Roads variable added little to no additional information to the full model and would suggest that removing the Roads variable was the right decision. Such an R^2 value indicates how the variables Income, Drivers and TaxRate explain approximately 65% of the variance present in the response variable Fuel.

The next step was to examine the model diagnostics to establish if this model violated the linear models assumptions. The reduced models diagnostic plots were created and are shown below in Figure 4.

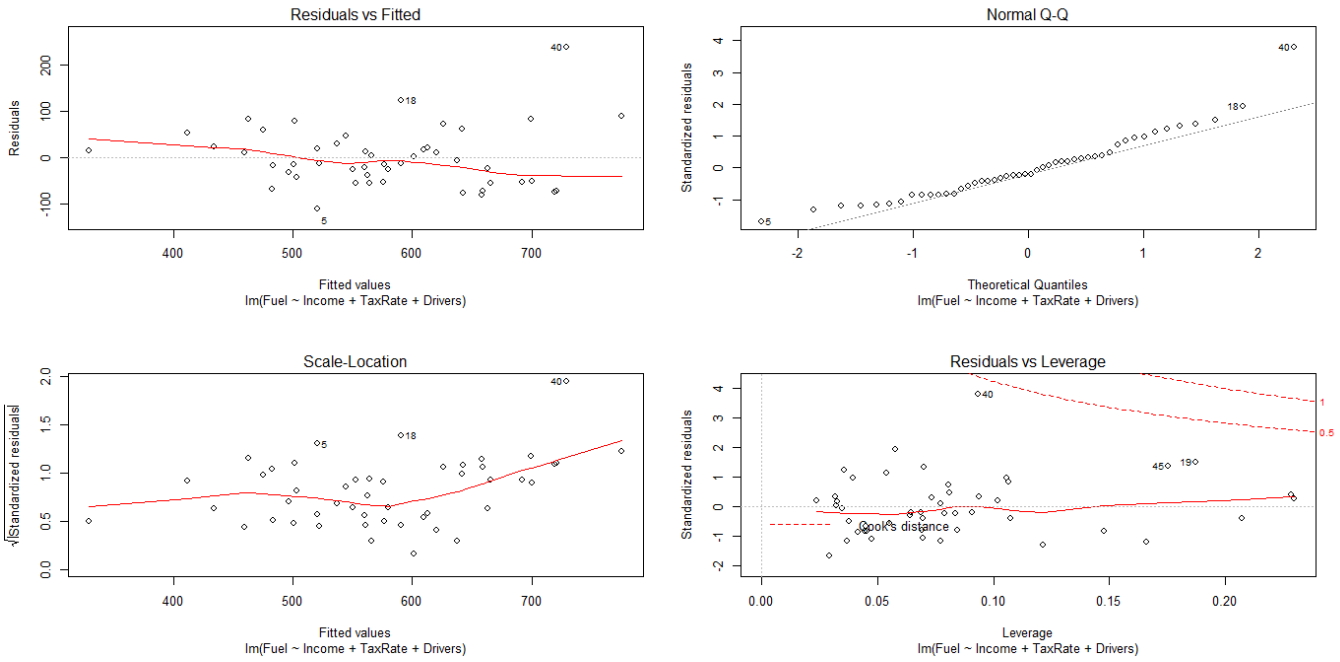

Figure 4: Diagnostic plots for the reduced model $Fuel \sim Income + Drivers + TaxRate$.

Examining the residuals vs fitted values plot above, it is clear that the residuals are scattered evenly above and below 0, with the line of best fit approximately following the line $y = 0$. There does not appear to be any structure present in the residuals vs fitted values plot which indicates how the linear model has accounted for all of the structure within the dataset. Examining the QQ-plot the residuals appear for the most part to be approximately normally distributed, however, some of the residuals appear to deviate away from the 45 degree line when examining the QQ-plot in more detail, this would be a slight cause for concern. It might be worth investigating the presence of outliers within the model. Examining the scale-location plot it is possible to identify if the constant variance assumption holds. In this case it appears that the constant variance assumption holds when examining the spread of the residuals across the range of fitted values. Looking at the residuals vs leverage plot, observation 40 appears to have a large standardized residual value with a leverage value of approximately 0.1. It might be worth further investigating the influence this possible outlier is having on the model and its assumptions.

The Cook's distances of each observation were identified and examined in more detail. A Cook's distance plot was created for the above model, with a cut-off value given by the formula $4/n$, where $n = 48$ in this case. The Cook's distance plot is shown below in Figure 5.

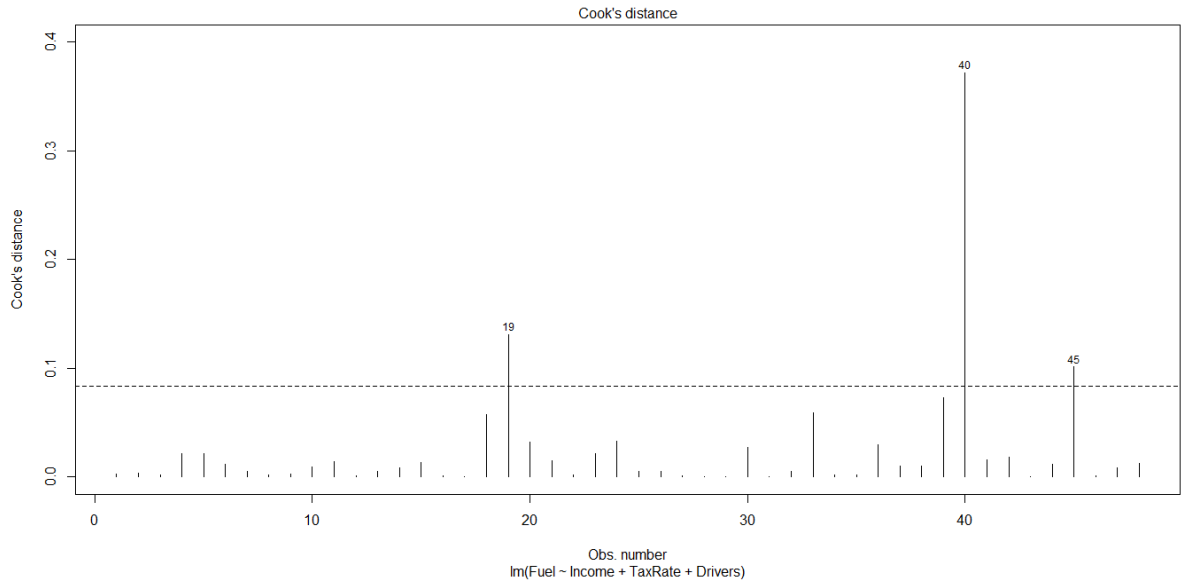

Figure 5: Plot of the Cook's distances including a reference cut-off line at $4/n$.

From the Cook's distance plot shown above one can see how observation 40 has a very large Cook's distance value, with a medium Cook's distance value obtained for observation 19, observation 45 also appears to just cross the cut-off line. In the interest of removing the least number of observations as possible, it was decided to remove observations 40 and 19 from the model and to then re-fit the model. Unfortunately, the dataset does not contain details of which states these observations represent thus it is not possible to identify any potential reasons why these states may have been outliers. Examining the characteristics of both observations one can see that these states both have very high levels of fuel consumption per person in addition to having very high percentages of the adult population with a drivers licence. Examining the Fuel boxplot again one can see how the these values are the two outliers present.

The new model obtained an adjusted R^2 value of 0.6578, indicating how removing these values has slightly improved the model fit compared to the previous model which contained the two outlier values. However, a more noticeable difference can be seen between the two models when examining the diagnostics plots of the new model shown below in Figure 6.

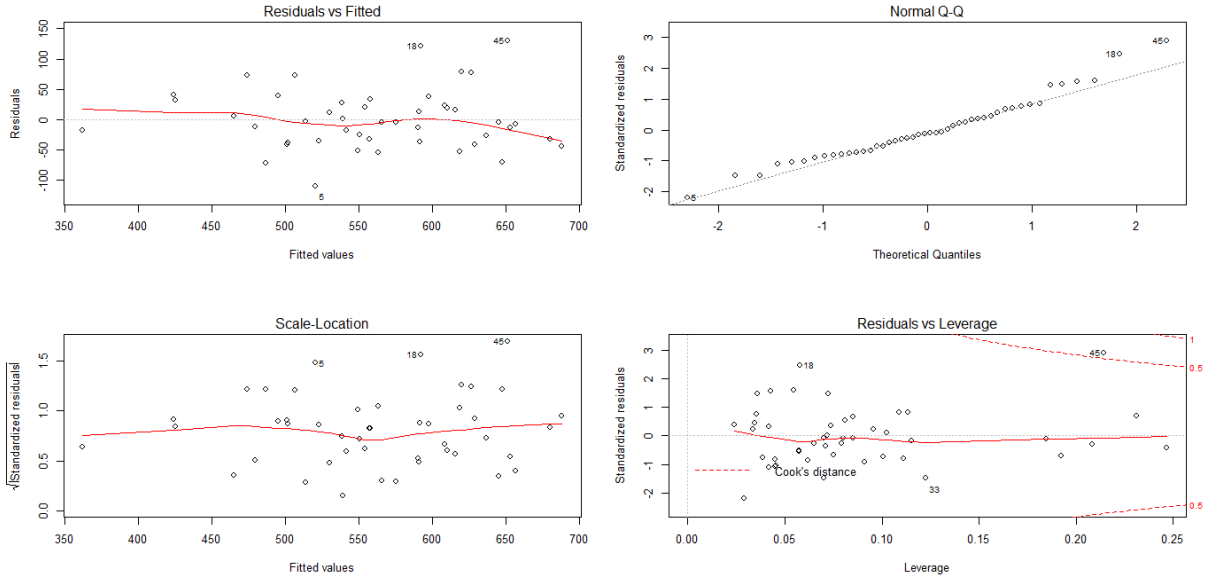

Figure 6: Diagnostic plots of the model $Fuel \sim Drivers + Income + TaxRate$ after observations 40 and 19 were removed.

Examining the diagnostic plots in more detail, one can see when observing the residuals vs fitted values plot that the residuals are evenly scattered above and below 0 with the line of best fit indicating how the residuals have a 0 mean. The residuals appear to be normally distributed when examining the QQ-plot with most of the residuals holding the 45 degree line, some of the wider residuals deviate slightly from the 45 degree line, however, given the small number of data points it can often be difficult to show that the residuals are perfectly normally distributed. From the scale-location plot the constant variance assumption appears to have been satisfied with little evidence to suggest a change from constant variance across the range of fitted values. It is clear from these diagnostic plots that the linear model assumptions have been satisfied.

4.1 Model Interpretation

In this section the final models parameters and coefficients will be fully interpreted and explained. Table 1 below is a table of the final models parameter values along with each coefficients associated standard error and p-value.

Table 1: Table of the coefficients, standard errors and associated p-values of the final model $Fuel \sim Income + Drivers + TaxRate$.

	Dependent variable:
	Fuel
Intercept	501.224^{***}
	(13.186)
TaxRate	-28.613^{***}
	(8.194)
Drivers	10.163^{***}
	(1.607)
Income	-69.264^{***}
	(128.916)
Observations	46
R^2	0.681
Adjusted R^2	0.658
Residual Std. Error	51.012 (df = 42)
F Statistic	29.834^{***} (df = 3; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

From the above table one can see how each of the variables are statistically significant at a 5% level of significance.

The parameter interpretations of the final model are given below:

- **Intercept:** The intercept of this model is given as the value of the model whereby all other variables are 0, this is also defined as the null model. The intercept value of this model is given as 501.22, therefore, indicating that a state which has no fuel tax, a mean income of 0 and 0% of the adult population with a drivers license, that the average fuel consumption would be 501.22 gallons per person. However, it is improbable if not impossible to obtained such an outcome.
- **Income:** A one unit increase in Income causes the average fuel consumption of a state to decrease by 69.264 gallons per person, provided all other variables remain constant. This would suggest that as the mean income of a state increases, the fuel consumption of that state decreases.
- **Drivers:** Provided all other variables remain constant a one unit increase in the Driver variable causes the average fuel consumption of a state to increase by 10.163 gallons per person. This suggests that as the percentage number of adult drivers increases in a state that the fuel consumption increases.

- **TaxRate:** A one unit increase in the number of cents of tax paid per gallon causes a decrease in the mean fuel consumption by 28.613 gallons per person, provided all other variables remain constant. This suggests that as the rate of tax on a gallon of fuel increases the consumption of fuel decreases.

It is interesting to note how as the mean income of a state increases the fuel consumption of that same state decreases, I believe it would be interesting to further investigate why this is the case. It is also worth noting how the rate of tax effects the fuel consumption of a state with an increase in tax increasing the price of fuel and therefore, reducing the fuel consumption of a state.

5 Conclusion

To conclude, a linear model was fitted to the fuel dataset between the response variable Fuel and the remaining explanatory variables Income, Roads, Drivers and TaxRate. The Roads variable was identified as being insignificant and was removed from the model. Two observations were also identified as outliers and therefore, were removed from the re-fitted model. After these observations were removed, a more accurate model was obtained which satisfied the linear regression assumptions.

Examining the models parameter interpretations it is clear that as the tax rate increases on a gallon of fuel, fuel consumption decreases. Similar conclusions can be made for the Drivers variable with an increase in the percentage number of drivers increasing the fuel consumption of that state, this would make sense as the more people driving on the roads, the more fuel being consumed. Finally, it is interesting to note how as the mean income of a states population increases, the fuel consumption decreases. It would be interesting to investigate why this is the case. Would this effect be due to people investing in more fuel efficient cars or would an increases in the mean income of a state be associated with states with larger city hubs and central business districts where the salaries are higher. Thus, due to city life there is less need to use your own car, with greater public transport options available in these cities, therefore reducing the fuel consumption of that state. To conclude, the Income, Drivers and TaxRate variables are the best predictors of the fuel consumption of a state.