The Impact of Gasoline Prices on U.S. Transportation

A Regression Analysis Study of the U.S. Retail Gasoline Price and the U.S. Annual Vehicle-Distance Travelled by Month (April 1993 – August 2014)

Project Report for:
STAT 301
Spring 2015
Weijie Shen
The Pennsylvania State University
May 1, 2015

By Dylan J. Humenik
May 1, 2015

Weijie Shen
Department of Statistics
The Pennsylvania State University

Dear Mr. Shen:

The report below, entitled *The Impact of Gasoline Prices on U.S. Transportation: A Regression Analysis Study of the U.S. Retail Gasoline Price and the U.S. Annual Vehicle-Distance Travelled by Month (April 1993 – August 2014)* represents the application of the introductory statistical concepts taught in STAT 301, the culmination of my learning in the class, and a bridge between the realm of statistics and the focus of my academic studies in energy, transportation, and economics. The report examines the retail gasoline prices in the United States for each month from April 1993 to August 2014 and the annual distance travelled by vehicles in the United States for each month during the same time frame for a possible correlation between the two sets of data.

The main statistical analysis method focused on in this report is regression analysis. Performing regression analysis on these two sets of data will allow for the examination of a possible correlation between the data sets. This analysis was completed using Minitab. The procedure that was used to accomplish this is outlined in detail below, as well as the results.

A number of other analyses were completed as well, also with the use of Minitab. Standardized residual and fitted value analysis were completed to further examine the correlation between the two sets of data. With the many variables that affect the price of gasoline and the number of miles driven, this report will uncover how strongly one is correlated with the other.

I gained many insights into the world of statistics during this semester of STAT 301, and I cherished this opportunity to further expand my knowledge as I advance toward my degree. Any questions about this report can be directed to me at dvh5219@psu.edu.

Sincerely,

Dylan J. Humenik
Abstract

This report analyzes the price of gasoline in the United States and the number of miles driven by vehicles in the United States from April 1993 through August 2014 and examines any correlation appearing between the two. The price of gasoline is critical in the transportation industry among private vehicle owners and commercial transportation services. The two sets of data analyzed in this report were pulled from U.S. government databases and then subjected to linear regression analysis. Further analysis included residual plot analysis. The analysis showed that a positive correlation exists between the price of gasoline and the number of miles driven in the U.S. from April 1993 to August 2014 but the correlation is not linear. The correlation varies with time and also with the different prices of gasoline (the correlation is stronger in periods of lower priced gasoline and weaker in periods of higher priced gasoline). It was also determined that the 2008 financial crisis had a big influence on the correlation. This suggests that while the price of gasoline and the number of miles driven do show a positive correlation, other economic factors are at play that can influence the correlation between the two.
# Table of Contents

**Introduction** .................................................................................................................. 4  
  The Price of Gasoline ........................................................................................................ 4  
  The Transportation Industry ............................................................................................. 4  

**Procedure** ....................................................................................................................... 6  
  Obtaining the Data ........................................................................................................... 6  
  Importing the Data into Minitab for Analysis ................................................................. 7  
  Performing a Linear Regression Analysis ....................................................................... 7  
  Obtaining Residual Plots for the Data ............................................................................ 7  
  Examining for a Correlation between the Gasoline Price and Number of Miles Driven... 8  

**Results** ................................................................................................................................ 10  

**Discussion** ....................................................................................................................... 15  
  Linear Regression ............................................................................................................ 11  
  Residual Plots .................................................................................................................. 14  

**Conclusions** .................................................................................................................... 18  

**Acknowledgements** ....................................................................................................... 20  

**References** ...................................................................................................................... 21  

**Appendix** ......................................................................................................................... 22  

# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Monthly U.S. Gasoline Price (April 1993 to August 2014)</td>
<td>10</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Monthly U.S. Annual Vehicle-Distance Travelled (April 1993 to August 2014)</td>
<td>11</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Combined Monthly U.S. Gasoline Price and Monthly U.S. Annual Vehicle-Distance Travelled (April 1993 to August 2014)</td>
<td>12</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Fitted Line Plot for Linear Regression</td>
<td>13</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Correlation Coefficient (r Value)</td>
<td>13</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Normal Probability Plot, Residual Histogram, Residual versus Fits Plot, and Residual versus Order Plot</td>
<td>14</td>
</tr>
</tbody>
</table>
Introduction

The Price of Gasoline

Gasoline is the clear aromatic liquid used to power most internal combustion engines since their development in the late 1800s. A large majority of these internal combustion engines are used in automobiles. Gasoline is refined from crude oil – the non-renewable fossil fuel resource occurring underground as a yellow to black liquid – and is the largest refinery “cut” among all other products derived from crude oil, including diesel fuel, jet fuel, heating oil, and heavy oils.

The price of gasoline is firmly linked with the price of crude oil. Therefore, many of the factors in the colossal worldwide oil and gas industry that influence the price of crude oil will also influence the price of gasoline in the U.S. The pricing of crude oil is generally determined from the daily spot price of West Texas Intermediate crude oil, as this is the New York Mercantile Exchange’s underlying commodity\(^1\). The price of a barrel of crude oil itself depends on the quality of the oil, along with many other factors that affect the global and regional supply and demand for crude oil. Some of these factors include the world’s global economic condition, the strength of the U.S. dollar, collusion among oil producers (including OPEC), geopolitical conflict in oil producing countries, and the strength of consumers’ demand of oil\(^2\).

The price of gasoline is one of the largest factors that determines the overall cost of ownership of a vehicle. The amount spent on gasoline also varies with the distance that a vehicle is driven. More miles driven means a higher expenditure on gasoline. This rate varies from vehicle to vehicle by the fuel economy.

The Transportation Industry

The transportation industry includes many means of transport, but the focus in this report will be on road transport by vehicles, specifically those vehicles that use gasoline. The transportation industry is a huge industry in the U.S. that employs many people and has an impact on many other industries. In 2007, it
was estimated that there are 254.4 million registered passenger vehicles in the U.S.\textsuperscript{3}. The transportation industry responds in many ways to environmental concerns and changing oil and gasoline prices. For example, in the 1970s and 1980s, fuel economy and emissions requirements were introduced that halted the muscle car era of the 1960s\textsuperscript{4}. In more recent years, higher efficiency engines and hybrid vehicles, the most famous of which is the Toyota Prius, were introduced due to similar concerns about gasoline prices and the environment. In 2008, rising gasoline prices and the financial crisis resulted in the mid-size sedan reclaiming its title from pick-up trucks as the best-selling vehicle type\textsuperscript{5}. The transportation industry is crucial to the American economy and is co-dependent with many other industries.
Procedure

The goal of this analysis is to effectively examine the comparison between the monthly average U.S. gasoline retail price and the monthly annual vehicle-miles driven in the U.S. Note that the “monthly annual vehicle-miles driven” refers to the annual vehicle-miles driven determined for each month – in other words, a value for March would refer to the number of miles driven from March to March of one year. This method of measurement allows for the annual value to be taken on a monthly basis. The term “vehicle-miles driven” refers to every mile that every vehicle in the United States has driven, and is an immense value in the trillions of miles, as anticipated. To complete this analysis, a number of methods were used, including:

- Obtaining the data
- Importing the data into Minitab (see Appendix) for analysis
- Performing a linear regression analysis
- Obtaining residual plots for the data
- Examining for a correlation between the gasoline price and number of miles driven

Obtaining the Data

The data for the analysis was to come from a trusted source. It was determined that government data would be the most trustworthy, accurate, and available source of data. The data for the monthly annual vehicle-miles driven was found on the U.S. government’s open data website, www.data.gov. Note that while this data was obtained on the open data website, the actual source of the data was published by the U.S. Federal Highway Administration. This set of data covered the time period of January 1990 to August 2014. The monthly U.S. gasoline price was obtained at the U.S. Energy Information Administration’s website. This set of data covered the time period of April 1993 to April 2015. Due to the slightly different timescales, the timescale analyzed in this report is the period covered by both sets of data: April 1993 to August 2014. Both of these sets of data were downloaded as an Excel file.
Importing the Data into Minitab for Analysis

Both sets of data were entered into Minitab for the time period specified above.

Performing a Linear Regression Analysis

The purpose of performing a linear analysis is to model the relationship between two variables. This can then be used to determine how much of a correlation, if any, exists between the two sets of data. A linear regression analysis was performed on the two sets of data with gasoline price on the x-axis and miles driven on the y-axis. The linear regression analysis includes an $R^2$ value (coefficient of determination) which is used to indicate how well a set of data fits into a statistical model, such as a line or curve. In this analysis, a line is used. A correlation coefficient was also obtained. This value, sometimes known as r, is the square root of the $R^2$ value and is also a great indicator of any correlation between the two sets of data. The r value varies between -1 and 1, where -1 is a totally negative correlation, 0 is no correlation, and 1 is a totally positive correlation.

Obtaining Residual Plots for the Data

Although the linear regression provided a suitable indication of whether or not a correlation existed between the price of gasoline and the number of miles driven, a few more tests were run on the data in the form of a series of residual plots. These plots will reveal more information about the data as described below. Four charts were obtained through Minitab to further analyze the existence of a correlation between the data.

The first of these plots is the normal probability plot. The normal probability plot was used to determine departures from normality in the data. Normality, or the normal distribution, is a prevalent continuous probability variable used to represent real-valued random variables whose distributions are not
known\(^8\). In a normal probability plot, the closer the data appears in a straight line, the closer it is to a normal distribution. Any deviation from a straight line implies a departure from normality.

The next plot in the residual plots is the histogram of the frequency of the standardized residuals of the data. This plot is also helpful in determining the correlation between the two sets of data, because it is an alternate means of representing the data in the normal probability plot. Higher frequencies around the center of the histogram (value of zero) indicate a more normal distribution.

Another plot in the residual plots is the residual versus fits plot. The purpose of the residual versus fits plot is to detect nonlinearity, unequal error variances, and outliers. A well behaved residual versus fits plot will have the following characteristics\(^9\):

- The data points distributed randomly around the 0 line, suggesting a linear relationship
- The data points form a horizontal band around the 0 line, suggesting that the variances of the error terms are equal
- No data points stand out, suggesting that no outliers exist.

The residual versus fits plot will help to determine if a correlation exists between the two sets of data by determining the whether or not the data is linear, contains equal error variances, or contains outliers.

The final plot of the residual plots is the residual versus order plot. This plot examines whether or not the data is serial dependent – in this case, time dependent. Any correlation between the error terms that are near each other in the sequence will show up as a correlation in this plot\(^10\). A residual versus order plot that exhibits any behavior other than randomly bouncing around the 0 line implies that the time variable has some sort of influence on the data. These four plots, which make up the residual plots, all help to further examine the data and analyze the possible correlation between the two sets of data.

**Examining for a Correlation between the Gasoline Price and Number of Miles Driven**

After obtaining the various plots and values as described above, the determination can be made as to whether or not a correlation exists between the price of gasoline in the U.S. and the number of miles driven.
driven in the U.S. This is a brazen indicator about the link between two giant industries in the United States; however, the correlation between these two sets of data does not imply any causation between one variable and the other. No claim may be made that asserts, for instance, that the price of gasoline causes U.S. citizens to drive less or more, or alternatively, the number of miles driven causes the price of gasoline to go up or down. The only claim that can be made is whether or not a correlation exists between the two. It is feasible likely that many other variables are at play in such a large U.S. economy containing the transportation and energy industries. The purpose of the other examinations (the residual plots) is to examine whether other variables have any influence on the correlation, including the variable of time.
Results

Below in Figure 1 is the data obtained for the monthly U.S. retail gasoline price for all grades and formulations for the period of April 1993 to August 2014. An $R^2$ value of 0.8419 was obtained for this graph:

![Monthly U.S. Gasoline Price](image)

**Figure 1** Monthly U.S. Gasoline Price (April 1993 to August 2014)
Below in Figure 2 is the data obtained for the monthly U.S. annual vehicle-distance travelled for the period of April 1993 to August 2014. An $R^2$ value of 0.8112 was obtained for this graph:

![Monthly U.S. Annual Vehicle Distance Travelled](image)

**Figure 2** Monthly U.S. Annual Vehicle-Distance Travelled (April 1993 to August 2014)
Below in Figure 3 is the combined data shown in the two graphs above, Figures 1 and 2, for the period of April 1993 to August 2014:

**Figure 3** Combined Monthly U.S. Gasoline Price and Monthly U.S. Annual Vehicle-Distance Travelled  
(April 1993 to August 2014)
The data above was imported into Minitab to obtain the following figures. Figure 4 below is the fitted line plot used for linear regression analysis in which gasoline price is on the x-axis and distance travelled is on the y-axis:

![Fitted Line Plot](image)

**Figure 4** Fitted Line Plot for Linear Regression

Below in Figure 5 is the output produced by Minitab that gives the r value, or the correlation coefficient:

<table>
<thead>
<tr>
<th>Correlation: U.S. Gasoline Price ($/gal), Vehicle-Distance Traveled/yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson correlation of U.S. Gasoline Price ($/gal) and Vehicle-Distance Traveled/yr = 0.768</td>
</tr>
<tr>
<td>P-Value = 0.000</td>
</tr>
</tbody>
</table>

**Figure 5** Correlation Coefficient (r Value)
Below in Figure 6 are the residual plots:

![Residual Plots for U.S. Gasoline Price ($/gal)](image_url)

**Figure 6** Normal Probability Plot, Residual Histogram, Residual versus Fits Plot, and Residual versus Order Plot
Discussion

The raw data obtained from the United States government is shown above in Figures 1, 2, and 3. The monthly U.S. gasoline price (Figure 1) shows a general upward trend with many up-and-down fluctuations. Perhaps the most noticeable fluctuation is the extreme drop in price that occurred in 2008. This was the year that the 2008 financial crisis began, and its effect can be noticed in this graph of gasoline prices. Similarly, the graph of the annual miles driven (Figure 2) shows an upward trend as well. There are no dramatic up-and-down spikes in the data as there were in the gasoline price. However, the miles driven graph is similar to the gasoline price graph in that it also experienced a reduction during the 2008 financial recession. Both of these upward trends may suggest some correlation, and this is better shown in Figure 3, where the two data sets are laid on top of one another. However, a more definitive analysis is needed to confirm whether or not a correlation exists between the two.

Linear Regression

The linear regression analysis that was performed using Minitab as shown in Figure 4 was done with U.S. gasoline price on the x-axis and annual vehicle-miles driven on the y-axis. The result is an upward trending plot with an $R^2$ value of 59%. However, the shape is rather peculiar in that it trends heavily upward for lower gasoline prices and then levels out at higher gasoline prices. This may suggest that the relationship is not linear; this will be explored below with the residual plots. Shown in Figure 5 is the Minitab output for the correlation coefficient (R value), which is given as 0.768. As stated above, a positive value for the R value indicates a positive correlation. This R value is fairly high as well (greater than 0.5) which indicates that the correlation is fairly strong. This means that it can be confirmed that there is a statistically significant correlation between the price of gasoline and the number of miles driven annually. In summary, it is observed that higher gasoline prices result in a greater number of miles driven annually; however, the shape of the graph may indicate that this relationship is stronger at some times and weaker at other times.
Residual Plots

To further analyze the observed correlation between the gasoline prices in the United States and the annual vehicle-miles driven in the United States, a number of other analyses were conducted with the two sets of data; these four further analyses make up the residual plots. Again, these residual plots were produced in Minitab, and they are shown above in Figure 6.

The first of the residual plots is the normal probability plot. In this plot, the data is not observed to not follow a straight line. This indicates that the data is not quite normally distributed – in other words, the data is not quite randomly distributed in a way such that it follows the normal distribution curve of a continuous probability variable. The fact that the data does not quite exactly follow this distribution suggests that there may be other factors that influence the distribution of the data – and therefore, other factors that influence one or both of the two datasets.

The second residual plot is the histogram plot. This plot shows the frequency of the standardized residuals of the data in a histogram format. It is essentially another way of representing the data in the normal probability plot as described above. If the data were normally distributed, the histogram would appear to peak around the 0 value while tapering off on either side of 0. The histogram plot of the two sets of data do not follow this pattern, and therefore it is evident again that the data does not follow a rough normal distribution. The frequency of the standard residual actually peaks around -0.8, and there are two other smaller peaks around 0.2 and 1.6. Again, this suggests that there may be other factors at play between these two datasets and confirms that the relationship is not strictly linear.

The third residual plot is the residual versus fits plot. This plot will determine the linearity, similarity between the variances, and the presence of outliers. As shown in the residual versus fits plot of the data, the plot is not distributed randomly around the 0 line and the data points do not form a horizontal band around the 0 line. Instead, the data in the plot follows a delineated downward-sloping line until after a fitted value of 2, and then abruptly rises up. This confirms the hypothesis that the data is not linear, and also confirms that the variances are not evenly distributed throughout the data. Again, there is a
compelling indication from this graph that the two sets of data are not strictly dependent on one another but instead influenced by other factors as well.

The last residual plot is the residual versus order plot. This plot puts the residuals in order of observation – in the case of the two datasets, this order is chronological. If the time variable has no influence on the datasets, then the plot would appear as randomly distributed around the 0 line; it is evident that this is not the case for the data. In fact, a decisive line appears in the first part of the graph, before beginning to jump up and down. This indicates that the data was time-dependent, especially before the recession occurred in 2008 (which is shown by the up-and-down spikes in the plot).
Conclusions

Two sets of data were analyzed for a possible correlation: the monthly price of gasoline in the U.S. and the monthly U.S. annual vehicle-miles driven. The time period analyzed was from April 1993 through August 2014. A number of statistical techniques were used to make the following conclusions, including regression analysis and residual plots. According to the analysis of the data above, the following conclusions can be drawn about the price of gasoline and the number of vehicle-miles driven in the time period of the study:

- A positive correlation exists between the price of gasoline and the number of vehicle-miles driven in the U.S.
- This correlation is not linear.
  - The strength of the correlation is stronger at lower gasoline prices and levels off at higher gasoline prices.
  - The strength of the correlation changes with time.
- The recession that began in 2008 had a large effect on the correlation between the two datasets.

In summary, it was observed that a higher gasoline price is usually accompanied by a greater number of miles driven in the U.S. from April 1993 to August 2014. This means that the two occurrences are not substitutable, i.e. higher gas prices do not result in a reduction of the number of miles driven as most people would expect (at least not for the gasoline prices seen from April 1993 to August 2014). Instead, it suggests that bigger political and socioeconomic factors account for the price of gasoline and the number of miles driven, such as the large number of factors that influence the strength of the U.S. economy. This is evident in the observed effect that the 2008 financial crisis had on the price of gasoline and the number of miles driven – both saw reductions during the crisis.

With such a large number of factors that could possibly influence the correlation between gasoline prices and the miles driven in the U.S. a number of analyses could be performed as future
follow-up work for this report. Different factors could be analyzed for possible correlations in much of the same manner as in this report. Economic factors such as the GDP, national interest rates, and stock market indices could be compared with either the price of gasoline or the number of miles driven in the U.S. The price of diesel fuel can also be analyzed to determine if the industrial side of the transportation industry experienced a similar correlation. Also, different timescales can be used. The correlation can be analyzed much further back in time than 1993 to see whether or not it was always positive. Also, shortly after the time period analyzed in this report ended, the price of gasoline in the U.S. dropped sharply in response to OPEC lowering its prices for crude oil\textsuperscript{11}. Whether or not this increased or decreased the number of miles driven can also be analyzed in future work. These future analyses can make clearer the complex relationship between gasoline prices and the number of miles driven in the U.S. as well as the relationship between other factors.
Acknowledgements

This report would not be possible without a number of individuals who gave me the resources and knowledge necessary to complete it. Weijie Shen instructed STAT 301, the course that provided me with the insights into the world of statistics that allowed me to conduct the reasonable and necessary methods that were used to analyze the data. My ENGL 202C class taught by Faith McDonald in the fall of 2014 instructed me on the composition of proper technical reports, and many of the elements used in this report are ones that I learned in that course. Any courses that have given me insight into the economy and the energy industry were useful as well, including EME 460, EME 500, EBF 301, EBF 200, and ECON 102. All of these courses were taken at Penn State University. I consulted with my classmate Cody Dumont during the creation of this report; his knowledge of Minitab saved me valuable time.
References

Appendix

Definitions of Selected Terms

2008 Financial Crisis – The worst financial crisis to hit the United States since the Great Depression. The crisis started from the bursting of the housing market bubble due to the vast amount of loans being given out to people who could not afford the homes being purchased with such loans. The crisis threatened many large institutions in the United States that were ultimately saved by controversial government bailouts. The 2008 financial crisis that occurred in the United States led to the 2008 – 2012 global recession. The crisis and resulting recession had extensive, dramatic, and far-reaching effects on many industries in the United States and the world. Its effects can be seen in both of the data sets analyzed in this report: the price of gasoline in the U.S. and the number of vehicle-miles driven in the U.S.

Linear Regression – A statistical approach modelling the relationship between two scalar variables. This approach allows one to see the correlation between the two variables. The data points in a linear regression plot can be fit to a line of best fit, upon which the slope of that line can be used to determine the correlation. An upward sloping line indicates a positive correlation, a negative sloping line indicates a negative correlation, and a line with no slope indicates no correlation. Such correlations observed in a linear regression analysis can be used to predict unknown future values of one variable if the corresponding values of the other variable are known or extrapolated.

Minitab – One of the most prominent and powerful statistics software packages available for analyzing statistics in an academic, research, or business setting. Minitab was developed in 1972 at the Pennsylvania State University.
Normal Distribution – Also known as the Gaussian distribution, the normal distribution is a continuous probability distribution that is the resulting distribution predicted by the central limit theorem – which states that the averages of random variables that are independently drawn from independent distributions are normally distributions. The normal distribution takes on the shape of a bell curve and has the equation:

\[ f(x, \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

Where:

- \( \mu \) is the mean of the distribution (also the median and mode)
- \( \sigma \) is the standard deviation (\( \sigma^2 \) is the variance)

r (also called the Correlation Coefficient) – a statistical value that measures the linear correlation between two variables or sets of data. Alternatively stated, it measures the linear dependence between the two sets of data. It lies between -1 and 1, where -1 is a totally negative correlation, 0 is no correlation, and 1 is a totally positive correlation. The full name for the coefficient is the Pearson Product-Moment Correlation Coefficient (PPMCC or Pearson’s r). The equation for \( r \) is:

\[ r = \frac{1}{n-1} \sum \left( \frac{X - \bar{X}}{S_X} \right) \left( \frac{Y - \bar{Y}}{S_Y} \right) \]

Where:

- \( n \) is the number of paired data points
- \( X \) is a data point from one data set
- \( \bar{X} \) is the average of the points from the data set containing the \( X \) values
- \( S_X \) is the sample standard deviation:

\[ S_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2} \]

- \( Y \) is a data point from the other data set
- $\bar{Y}$ is the average of the points from the data set containing the $Y$ values

- $S_Y$ is the sample standard deviation:

$$S_Y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

$R^2$ (also called the Coefficient of Determination) – a statistical value that indicates how well a set of data fits a statistical model – in the case of this report, a line. The value provides a measure of how well the observed data points are replicated by the line. The coefficient of determination is simply the square of the sample correlation coefficient $R$:

$$R^2 = (R)^2$$

Variance – a statistical measure of how far a set of values are spread out (a variance of zero means that the values are identical). The range of possible variances are greater than or equal to zero. Low variances mean the values are close and high variances mean that the values are spread out. The square root of the variance is the standard deviation, which is also a statistical value used to measure the variation in a set of data. The equation for variance is:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

Where:

- $\mu$ is the mean, or the expected value, calculated as:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} X_i$$