Assignment 6: Final – ACA Report – Auto-Correlation Model: ACA

Predict 411

Section 56

Winter Quarter

****

School of Continuing Studies

Northwestern University

****

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

****

Program Analyst

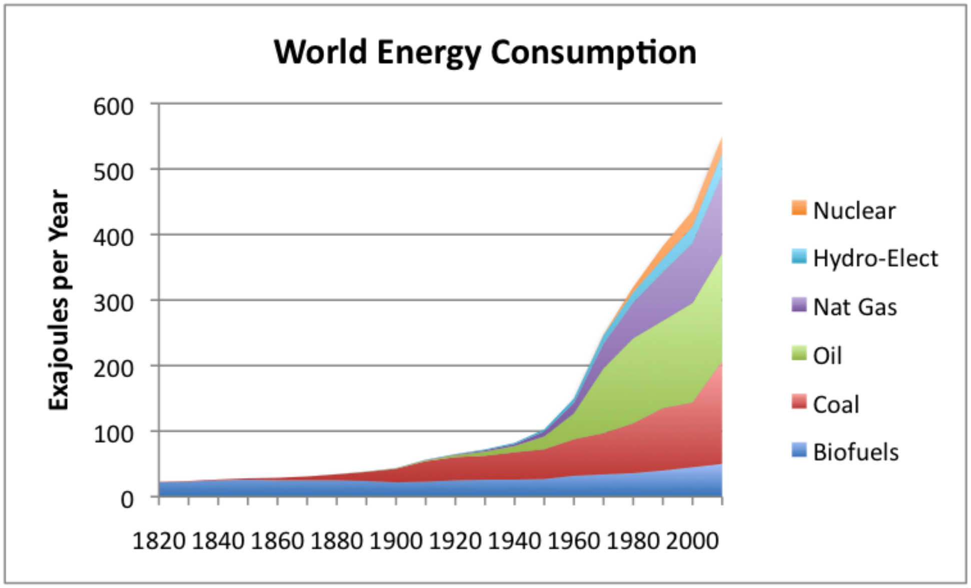
Wooddale Church

6630 Shady Oak Road

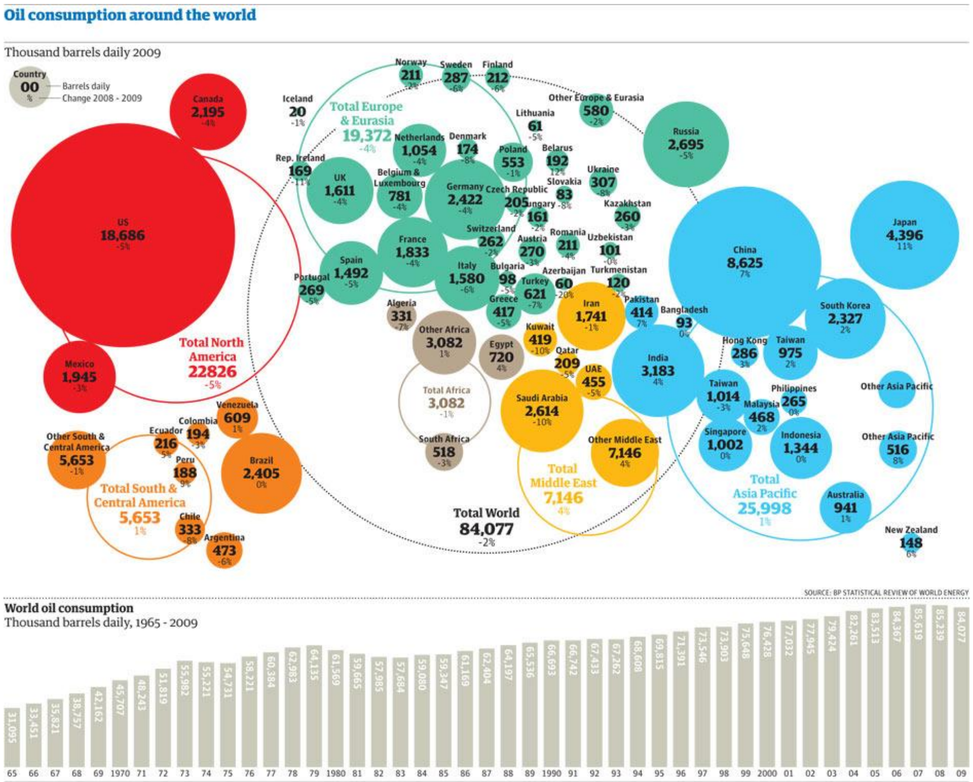
Eden Prairie, MN 55344

Executive Summary

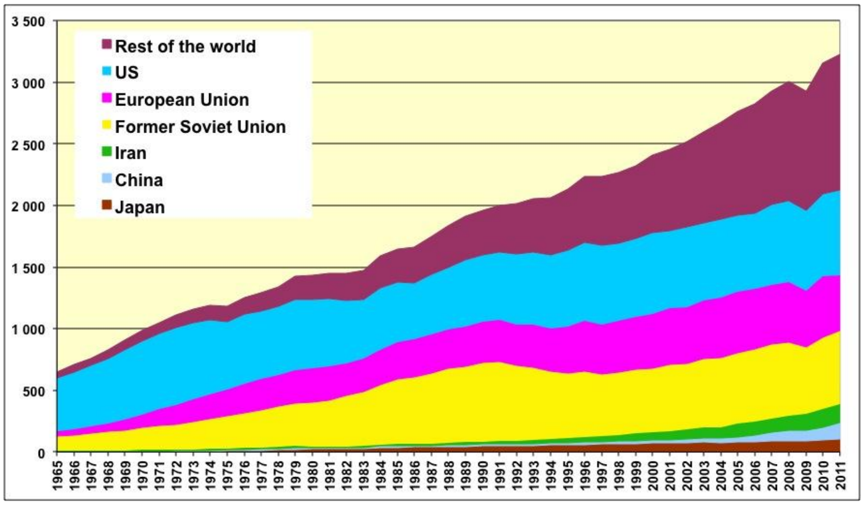
The world's appetite for energy and gasoline has grown considerably over the past 70 years. In the United States (US), a growing population and increased disposable income have a positive correlation with gasoline consumption. While conducting this initial Exploratory Data Analysis (EDA) on gasoline consumption, two models were used. The first model was the full model and had 10 independent variables, and the second model was reduced and consisted of 5 independent variables. Natural log transformations were used on all the variables, such that the distributions were more normal. It was found that the reduced model suffered from heteroscedasticity, and more specifically autocorrelation due to the time series nature of the data, thus leading to correlated residuals. This initial finding was further validated statistically through the Durbin-Watson, Lagrange Multiplier, Box and Pierce, and Ljung's tests. Generalized Least Squares (GLS) was initially used to fit the heteroscedastic data, but given the fact that the variance-covariance matrix was unknown Feasible Generalized Least Squares (FGLS) was used to estimate the matrix. After the second iteration or second-order autocorrelation lag, the heteroscedasticity was no longer an issue and this model proved to be the best fit with sound statistical output. Heteroscedasticity is a serious issue within regression analysis, but FGLS proves to assuage the issues.



Introduction In the last 70 years worldwide energy consumption has grown at an exponential rate. From the graphic to the left, it can be seen that Natural Gas, Oil, and Coal have grown the most. The estimates are based on Vaclav Smil estimation and can be further explored in the bibliography. As undeveloped countries modernize, their appetite for energy, especially fossil fuels, increases (ourfiniteworld.com). Specifically, Asian countries now dominate the oil consumption fuels as seen by the light blue bubbles (Ritholtz). The bottom bar graph demonstrates a steady, growing consumption of oil for the past 75 years. Notice how oil consumption for the past 75 years is not nearly as



quadratic as for the last 175 years. This assignment focuses on US consumption of gasoline. For the latter part of the 20th century, gasoline was considered "American" energy since the US dominated the consumption of gas. The graphic to the left shows that in 1965 the US consumed more than 60% of the gasoline on the planet (Jean-Marc Jancovici). Gradually, the world has caught up, and I attribute this to the world wide use of gas consuming vehicles for transportation.



The objective of this report is to conduct an EDA and explore heteroscedasticity in regard to auto-correlation by utilizing the gasoline dataset, of which ample background information has already been supplied above. OLS has many advantages when analyzing data, but it also has assumptions that are hard to validate such that the output is credible. An assumption in OLS is that the residuals, error terms, are not correlated. When the errors are correlated, it is known as autocorrelation. Time series data tends to suffer from autocorrelation given that the data is successive, meaning it follows a chronological occurrence (Chatterjee & Hadi). Another occurrence of autocorrelation arises when a relevant variable is omitted from the equation. Yet, when the variable is added to the equation, the autocorrelation disappears.

Autocorrelation is similar to heteroscedasticity and creates the following issues:

1. The regression coefficients do not have minimum variance as they should when using OLS.

2. The standard errors give a false impression of precision.
3. As a result of the standard errors, confidence intervals and tests for significance are erroneous.

As one can see, autocorrelation distorts the data and inhibits an accurate output from typical modeling techniques. In order to guard against autocorrelation one must understand specific tests used for detecting it, as well as modeling techniques to assuage the affects. As a result of this study, economists will better understand autocorrelation and how to deal with this occurrence in data sets. Time series analysis is synonymous with economics, and thus needs to be quarantined and resolved in order to produce results that are statistically sound. In my brief time as a data scientist in training, OLS has been the training/entry level model for which training wheels are to a bike. Given that OLS has many assumptions which limit its usefulness, I look forward to exploring techniques and diagnostics, such as FGLS, that are more robust in nature and pragmatic for analysis.

Analysis

In order to meet the objective of exploring the relationship between the dependent variable, gasoline consumption, and independent variables, an exploratory data analysis must be conducted. This EDA will start with a basic OLS model. From that model, analysis will be made and if the OLS assumptions do not hold I will move on to FGLS. As a data scientist in training, I am inculcating a paradigm of which to study data. While this paradigm is redundant report to report, it is training me to have the correct mindset. I will be utilizing the EDA paradigm and structure put forth by Bruce Ratner found in his book *Statistical and Machine-Learning Data Mining*:

> Problem/Objective: Conduct an EDA and explore heteroscedasticity in regard to auto-correlation by utilizing the gasoline dataset.

Data: The data has been aggregated and has been supplied from management. There are no missing values.

Analysis: I will describe the data via simple descriptive statistics at first. After the initial analysis, I will start with fitting the data with an OLS model. I am familiar with this model, and will better understand the data through this process. If/when issues arise from the state of the data as a result of not meeting the OLS assumptions I will select a different model in order to adhere to sound statistical principle. The residuals will be analyzed in conjunction with the time series plot. Different diagnostic tests will be utilized to analyze autocorrelation if it is present in the model.

Model: When issues arise from fitting the data with OLS, management has instructed the use of a FGLS model. The advantage of using this model is results that are statistically valid and adhere to the principles of data modeling.

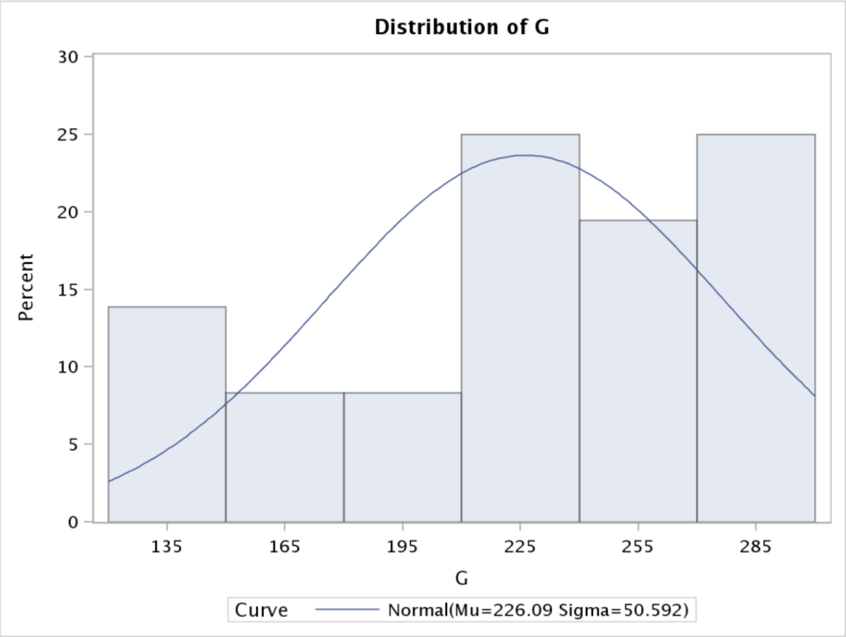
Results/Interpretation: Once the model has been created, an assessment of the model adequacy will be conducted to discern how well the model fits that data and the statistical backing of the model.

A properly executed EDA for management must reflect that the data was the driving force behind constructing the model. The steps outlined above are ordered such that the data drives building the best model, and the analyst's personal bias is mitigated.

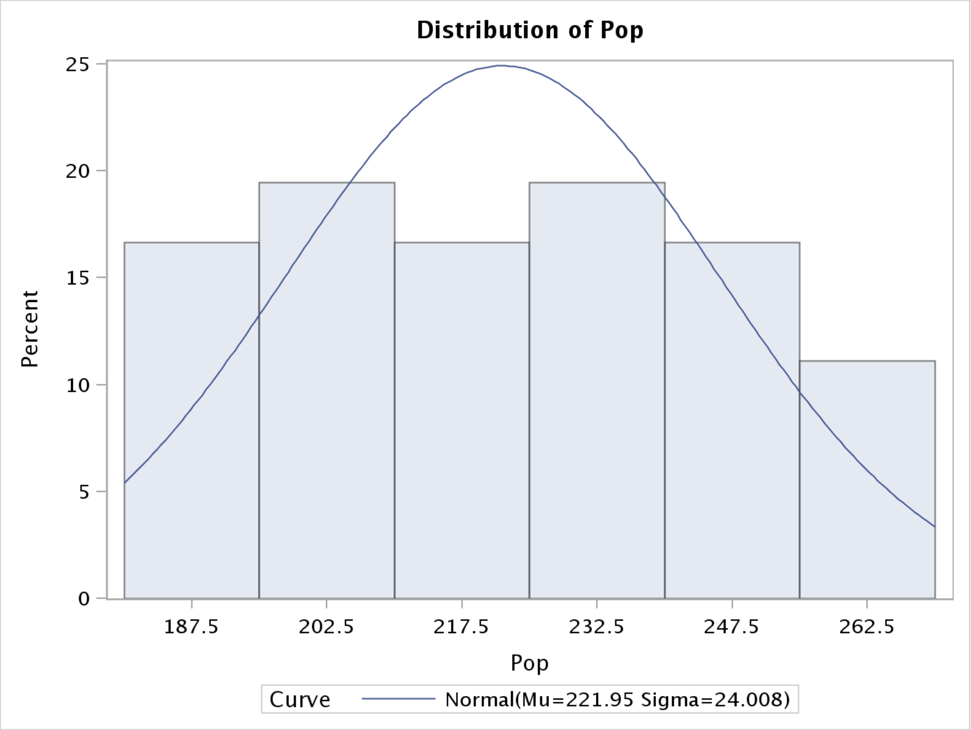
Data

Following the outline above, exploring the data is the next step for the EDA. The variable G is the total U.S. gasoline consumption, computed as total expenditure divided by the price index, the independent variables are the following: Pg is price index for gasoline, y is per capita disposable income, Pnc is the price index for new cars, Puc is the price index for used cars, Ppt is the price index for public transportation, Pd is the aggregate price index for consumer durables, Pn is the aggregate price index for consumer nondurables, Ps is the aggregate price index for consumer services and Pop is the U.S. total population in millions. In this data set, each year represents a row and there are 36 years represented starting in 1960 and ending in 1995. Given that there are 11 total variables and 36 unique rows, the sum of observations equals 396. Utilizing log transformations alter the data to a fairly standard shape (Ajmani). Specifically, this technique is used for positively skewed data and the result of the log transformation moves the majority of the data such that it follows a normal distribution.

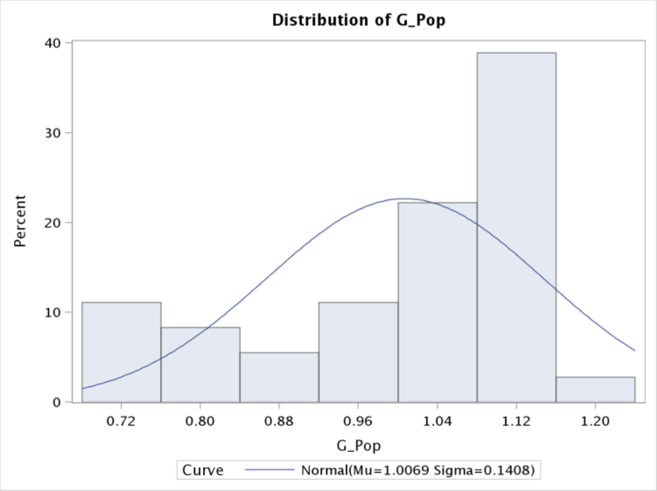
G is the total U.S. gasoline consumption, computed as total expenditure divided by the price index. Appendix 1 shows the SAS output generated from Proc Univariate in regard to the descriptive statistics for all the variables. In order to understand this complex variable, one must first take apart the variables associated with the finished variable. The average gasoline consumption over the 36 years is 226, and a median of 235, which demonstrates mild skewness. The standard deviation was 50, variance was 2559, and there was a slight negative skewness of -.57. The range is 168, and it should be noted the minimum consumption occurred in 1960 and the greatest consumption occurred in 1995. In general, as time increased so did gasoline consumption. The distribution of gasoline consumption can be visually seen from the graphic above. Over 70% of the consumption occurred at a rate greater than 225, which explains the negative skewness.



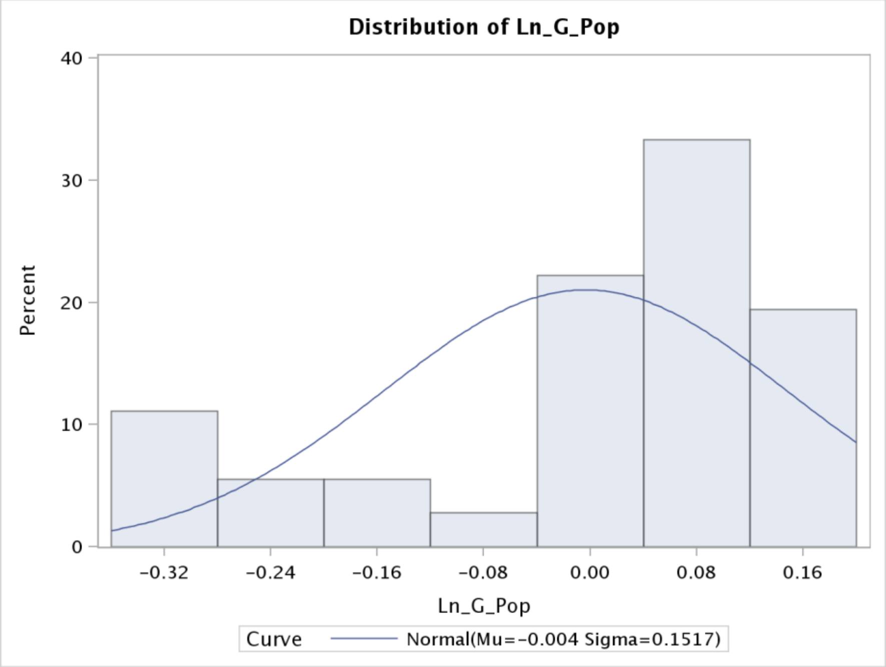
Pop is the U.S. total population in millions. The average population over the 36 years was 221, and a median of 221, which demonstrates little to no skewness. The standard deviation was 24; variance was 576, and a slight negative skewness of .018. The range is 82, and it should be noted the minimum population occurred in 1960 and the greatest population occurred in 1995. In general, as time increased so did the population. The distribution of gasoline consumption can be visually seen from the graphic. Over 70% of the population was between 202 and 212. Understanding these past two variables is pivotal to understanding the response variable. As stated above, G is total gas consumption divided by population. Now that both individual variables have been analyzed, the analysis of them combined will be easier to understand.

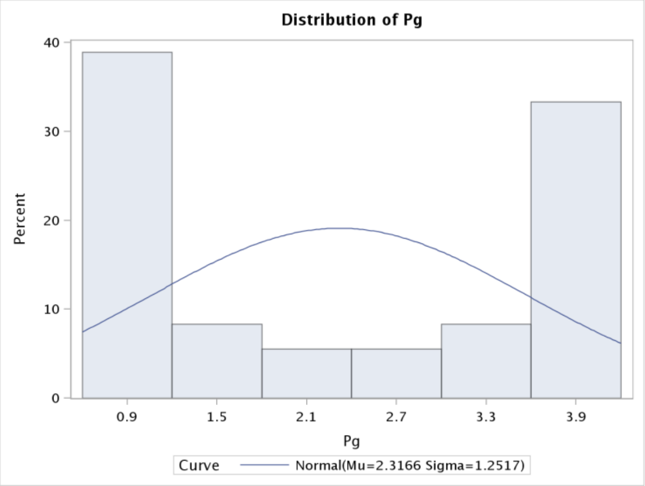


The response variable for this EDA is computed as Gasoline Consumption divided by Pop. In addition, management has recommended using the natural log transformation with this variable. Initially, I will analyze this variable without the log transformation and then I will include the log transformation. The response variables average over the 36 years was 1.006, and a median of 1.053, which demonstrates little skewness. The standard deviation (SD) was .14; variance was .02, and a slight negative skewness

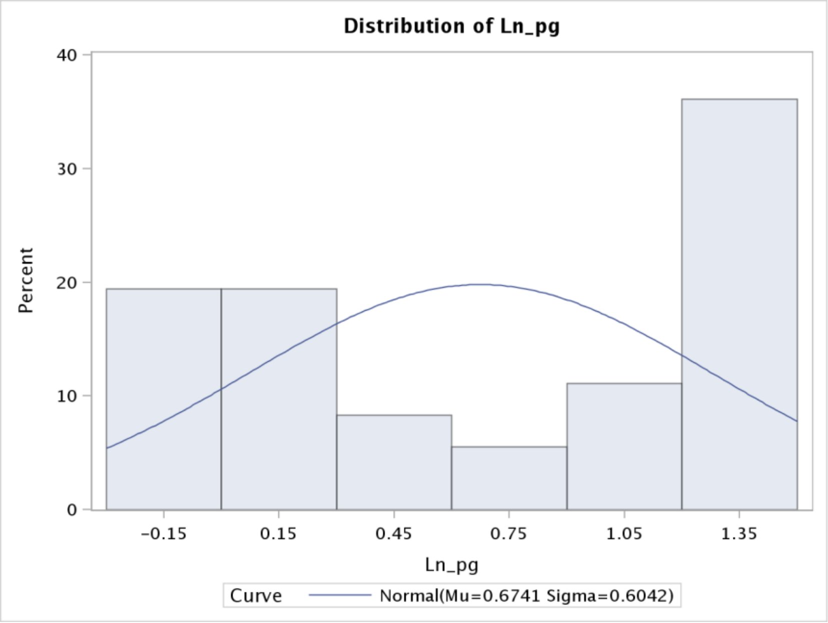


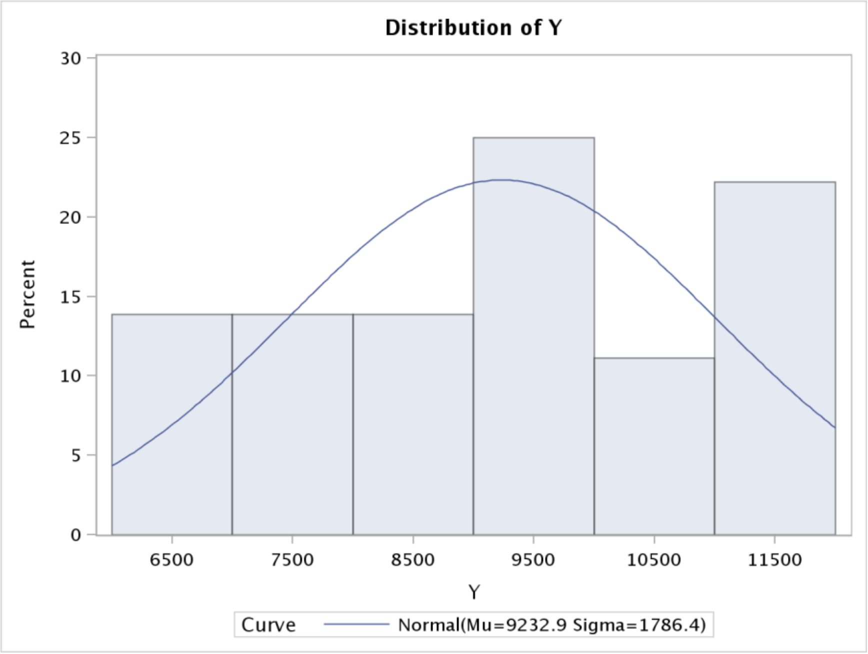
of .963. The range is .45, and the min and max occurrences are not as clear cut as previous variables. The distribution shows that about 40% of observations are around 1.12, which is why a log transformation is necessary. After the log transformation the response variables has the following descriptive statistics: mean - .003, median- .051, SD-.15, variance-.023, and a greater skewness that is -1.11. The range is .49. The numbers do not necessarily show the complete side of this transformation. After the log transformation, notice how the data demonstrates a more normal distribution, which is desired when conducting regression analysis. From this point forward as variables need log transformations, the only information I will show about the variables prior to the transformation is the distribution. After the transformation of log variables, I will analyze in more detail the descriptive statistics.





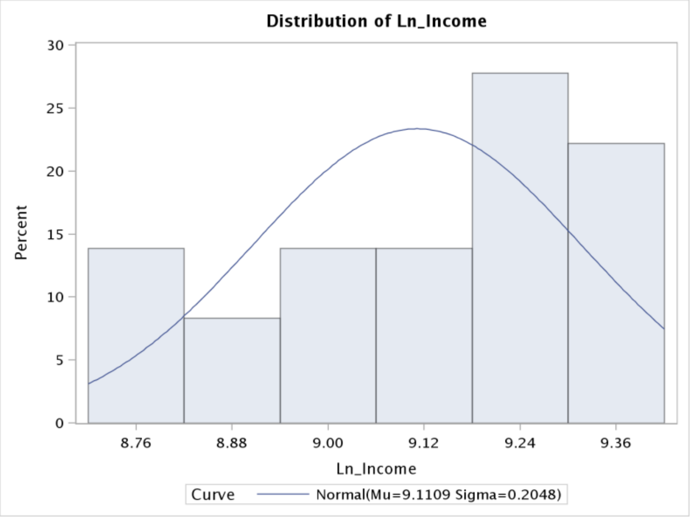
Pg is the price index for gasoline. Before the log transformation, the distribution has high bookend percent distributions. In my opinion, this distribution is almost the inverse of what a normal distribution should look like. After the log transformation of Pg, the mean is .674 with a range of 1.50 and a SD of .604. Given the range of values, there is a rather larger SD. This is not a very well distributed variable, and is rather ambiguous. I would not be surprised if it is troublesome further into the analysis. The median is .653, which is rather close to the mean and demonstrates a better overall distribution. From the histogram, the log transformation is less volatile, and the range is not as great. The overall distribution is not pretty, but this variable will be better to work with after the log transformation.



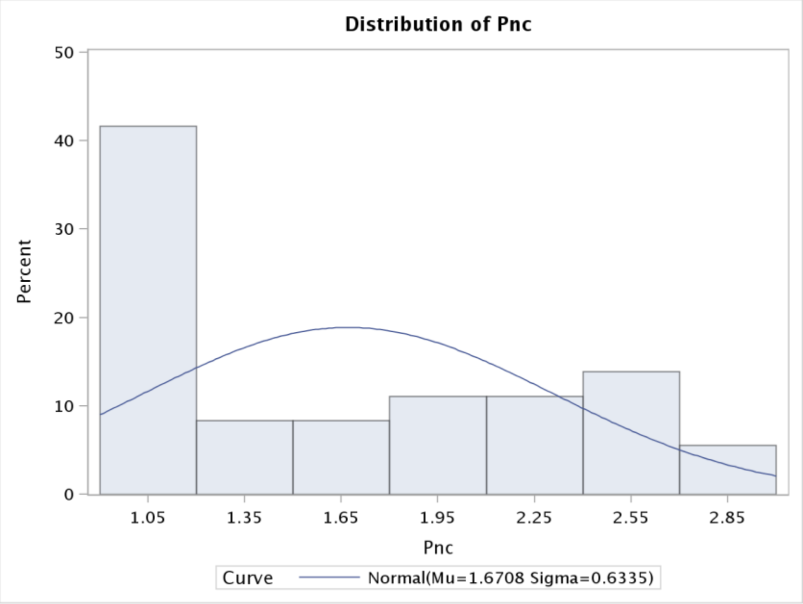


Y is per capita of disposable income. The pre-log distribution does not look all that bad. The only preliminary issue I see is the slight positive skew and the rather large range.

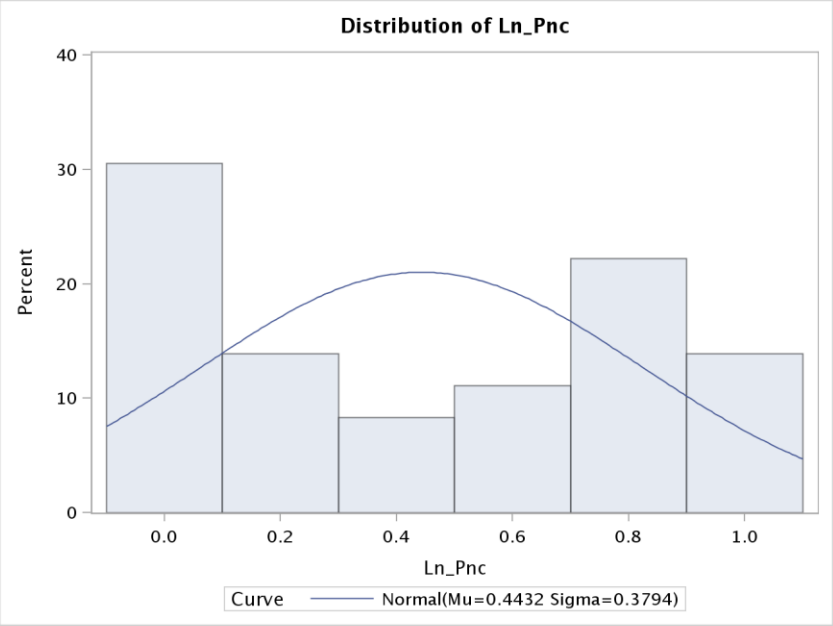
After the log transformation of Income, which was Y, the mean is 9.110 with a range of .681 and a SD of .204. Given the range of values prior to the transformation, this variable now has a much smaller range and is better suited for regression analysis. The standard deviation is also much smaller, and there is only a slight negative skewness of -.565. After the log transformation, the variable does not have the well distributed curve, but it will be easier to work with. The median is 9.110, which is the same as the mean and demonstrates a better overall distribution.



Pnc is the price index for new cars. Before the log transformation, the distribution has a high positive skew. One can see from the histogram that nearly 40% of the values are 1.05. In my opinion, this variable is troublesome and in need of a better distribution.

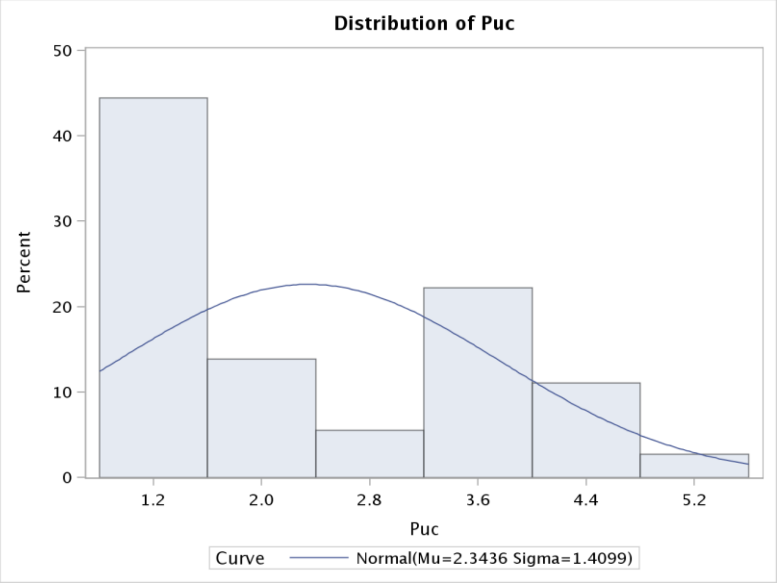


After the log transformation of Pnc, the mean is .443 with a range of 1.044 and a SD of .379. Given the range of values, the SD is just okay. I would prefer to see a tighter distribution.

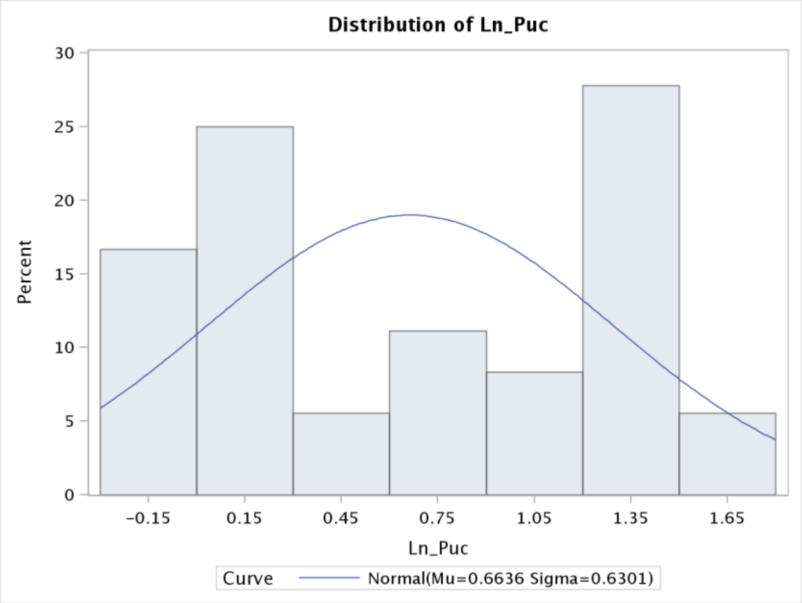


This is not a very well distributed variable, and is rather ambiguous. I would not be surprised if it is troublesome further into the analysis. The median is .393, which is rather close to the mean and the skewness is less severe in the overall distribution. From the histogram, the log transformation is less volatile, and the range is not as great. The overall distribution is not pretty, but this variable is better to work with after the log transformation.

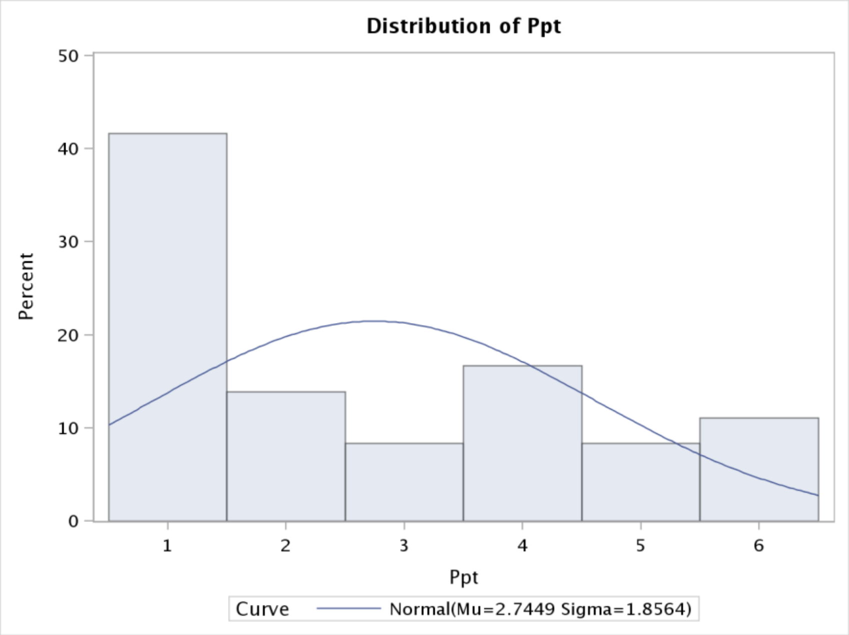
Puc is the price index for used cars. Before the log transformation, the distribution has a high positive skew. One can see from the histogram that nearly 45% of the values are around 1.2. In my opinion, it would be hard to use this variable in regression analysis because the distribution is skewed severely.



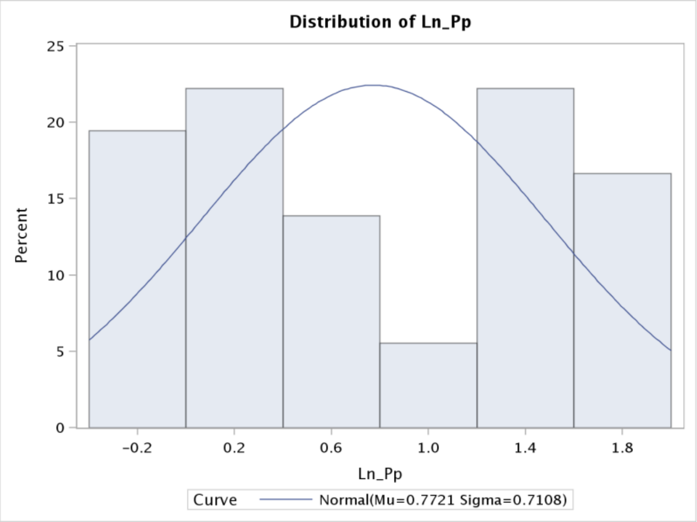
After the log transformation of Pnc, the mean is .663 with a median of .613, which demonstrates a tighter distribution. The SD is .630 and the range is 1.832. Given the range of values, the SD is just okay. 66% of values fall between .03 and 1.29, which is a rather large spread given the range. I would prefer to see a tighter distribution. From the histogram, the log transformation is less volatile, and the range is not as great. The overall distribution is not pretty, but this variable is better to work with after the log transformation.



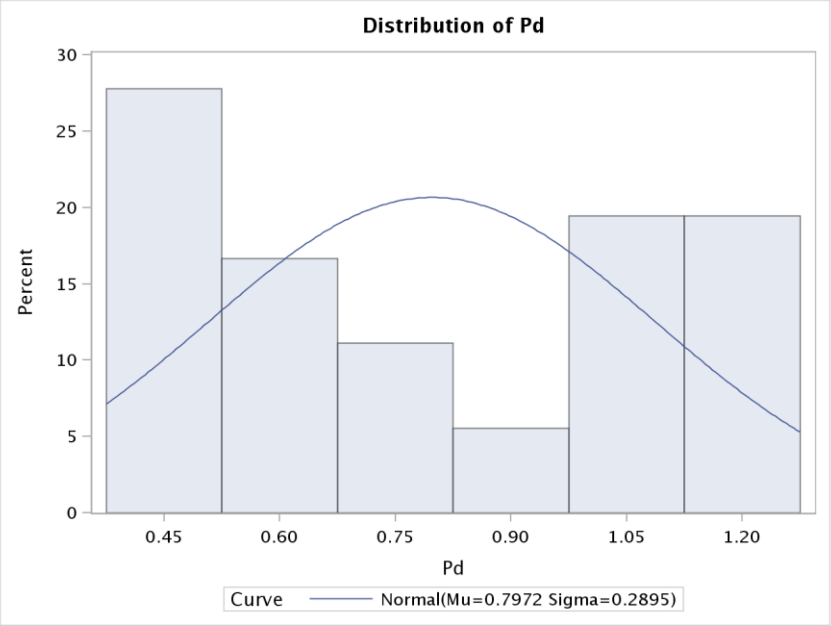
Ppt is the price index for public transportation. This variable has a very similar distribution to that of the Price index for new cars. Before the log transformation, the distribution has a high positive skew. One can see from the histogram that nearly 40% of the values are 1. In my opinion, this variable is in need of a more normal distribution. After the log transformation of



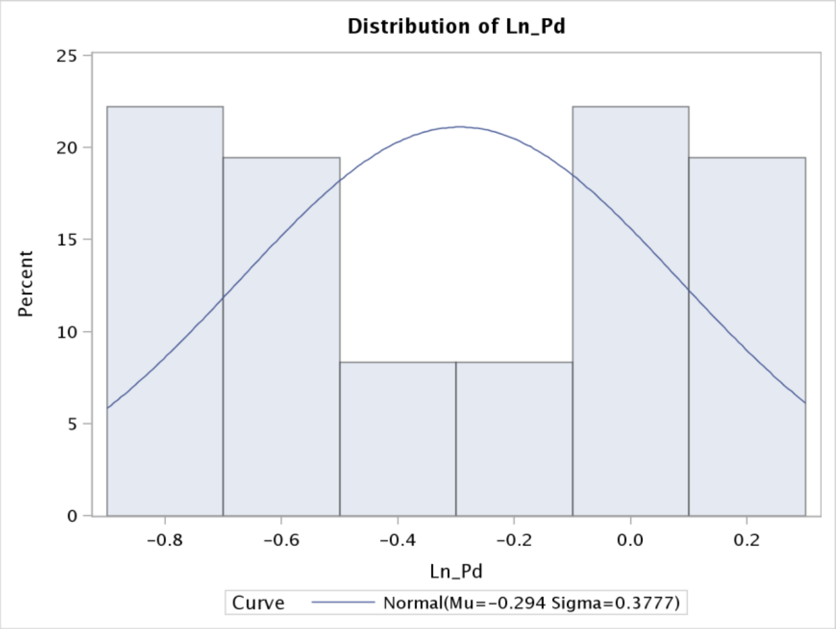
Ppt, the mean is .772 with a median of .615, which demonstrates an okay distribution. The SD is .710 and the range is 2.068. Given the range of values, the SD is just okay. I would prefer to see a tighter distribution, and one that does not look like the inverse of a normal distribution. From the histogram, the log transformation is less polemic, and the range is not as great. The overall distribution is not pretty, but this variable is better to work with after the log transformation. From my perspective, it looks like 80% of the values fall on the bookend sides of this distribution.



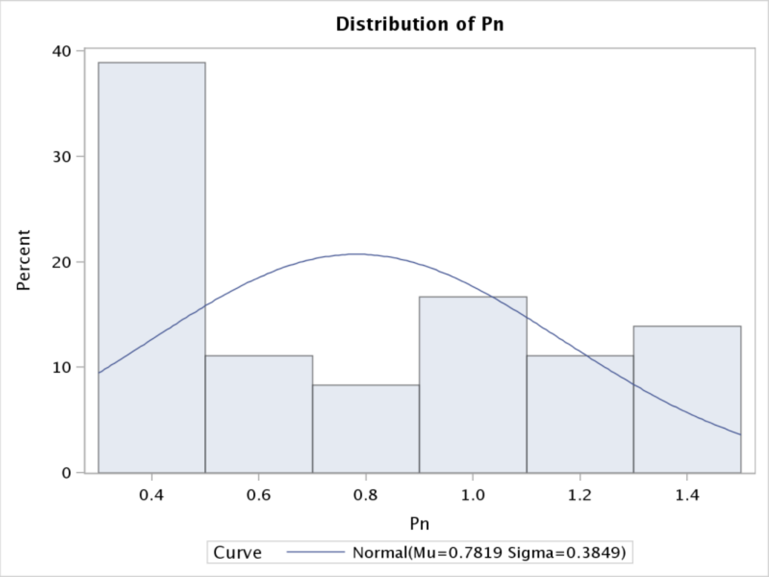
Pd is the aggregate price index for consumer durables. This variables distribution has little to no normal bell curvature. Before the log transformation, the distribution has a mild positive skew, but 40% of the values are on the tail end of the distribution. One can see from the histogram that it would be difficult to include this variable in regression analysis based on its awkward distribution.



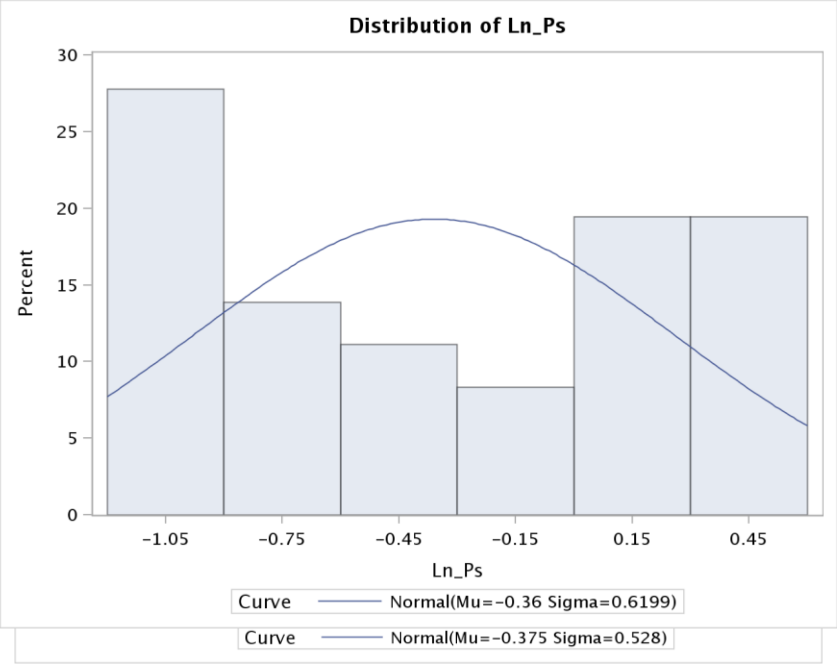
After the log transformation of Pd, the mean is -.294 with a median of -.290, which



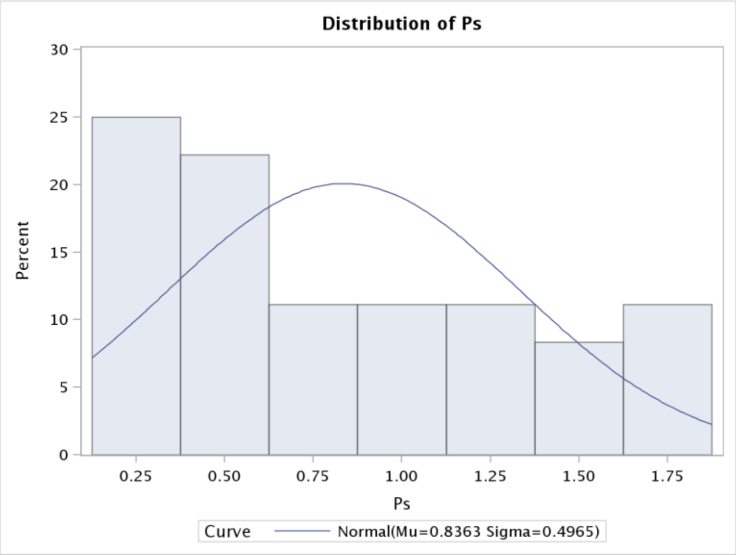
demonstrates a tighter distribution than previous variables. The SD is .377 and the range is 1.026. Given the range of values, the SD is better than the other variables. From the histogram, the log transformation is less volatile, and the range is not as great. The overall distribution is not pretty, but this variable is better to work with after the log transformation. With this data set, it seems that many of the variables have a distribution that is high on both sides of the curve but not in the middle of the distribution.



Pn is the aggregate price index for consumer nondurables. This variable has a very similar distribution to the variable price index for new cars. Before the log transformation, the distribution has a high positive skew. One can see from the histogram that nearly 40% of the values are .4. In my opinion, this variable is in need of a transformation that will yield a more normal distribution. After the log transformation of price index for consumer nondurables, the mean is -.375 with a range of 1.455 and a SD of .527.



Given the range of values prior to the transformation, this variable now has a much smaller range and is better suited for regression analysis. The standard deviation is also much smaller, and there is only a slight negative skewness of -.075. After the log transformation, the variable does not have the well distributed curve, but it will be easier to work with. The median is -.364, which is close to the mean and demonstrates a better overall distribution. Ps is the aggregate price index for consumer services. It can be seen from the initial histogram that it has a positive skew that flattens out as the index increases. A log transformation is needed based on the fact that 45% of the observations are .5 or less. The issue, as seen in other variables within this dataset, is that in order to conduct a regression analysis the data needs to follow a somewhat normal distribution. After the log transformation of aggregate price index for consumer services, the mean is -.360 with a range of 1.777 and a SD of .619. Given the range of values

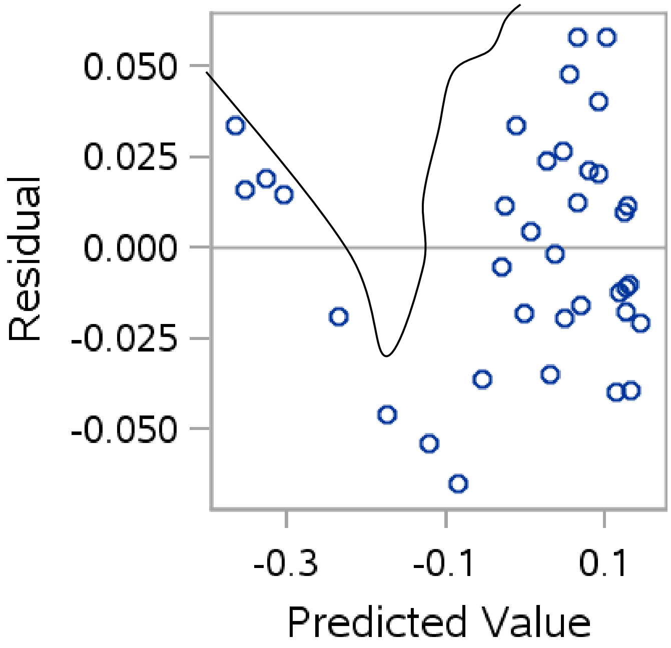


prior to the transformation, this variable now has a much smaller range and is better suited for regression analysis. The standard deviation is also much smaller, and there is only a slight positive skewness of .083. After the log transformation, the variable does not have the well distributed curve, but it will be easier to work with. The median is -.396, which is close to the mean and demonstrates a better overall distribution.

After doing an initial analysis of the variables distributions in this data set, all but one variable were in desperate need of a data transformations such that their distributions would be more normal. The only variable in my opinion that did not require a log transformation due to distribution was Y, per capita of disposable income.

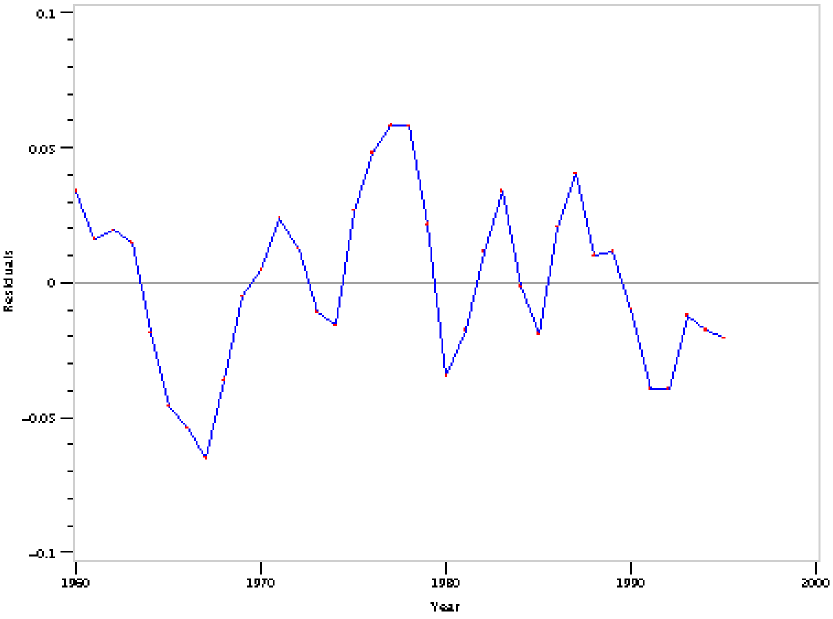
Results:

Management has instructed to use two models when fitting the data. The first model is the full model with all 9 variables, and the reduced model is comprised of just the first five variables. In the first part of the results, I am looking for heteroscedasticity given that I am working with time series data. Starting with the residuals and their association with time is the first step. Run an OLS model shows a solid R with statistically significant variables. The residuals plotted against the predicted values shows a relatively random distribution. Notice how the points rather make a V. This

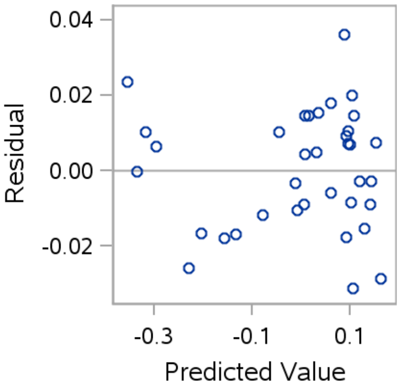


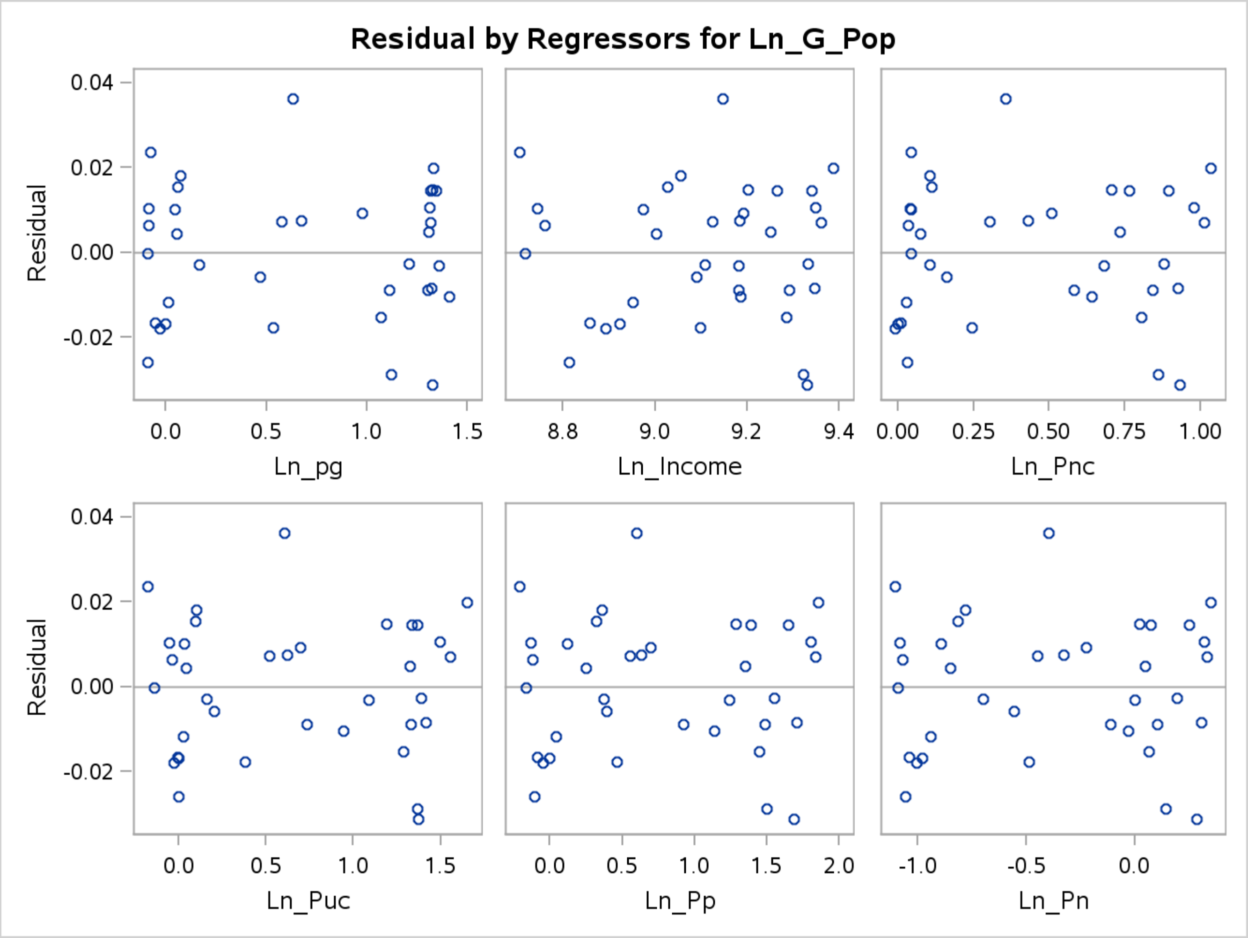
signals a rather non-spherical distribution and should illicit further investigation. The next step is to analyze the residuals against time. If the data suffers from heteroscedasticity, than there will not be a random disturbance.

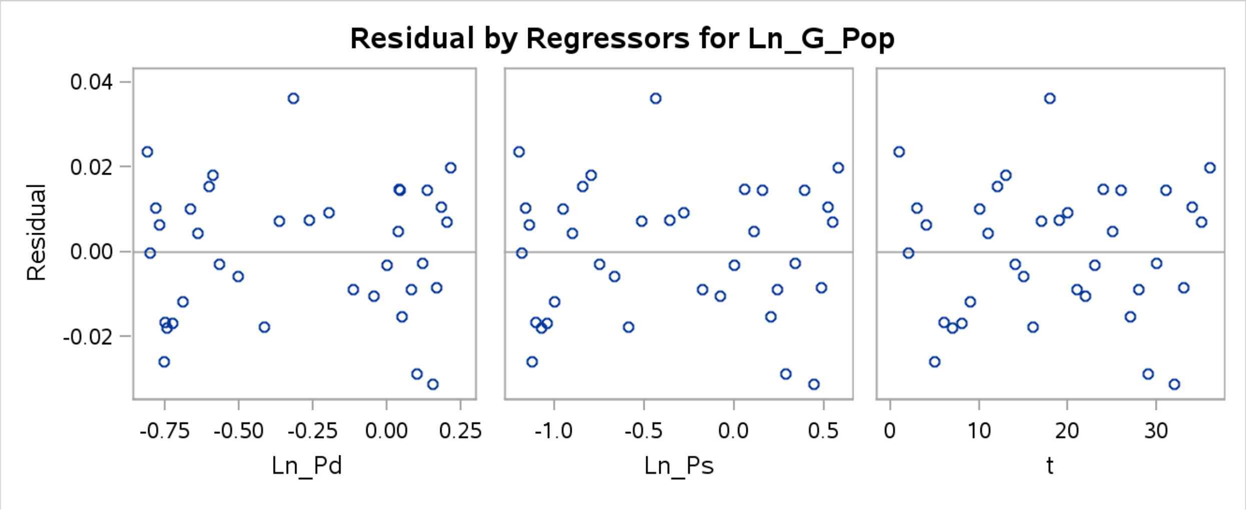
The graph to the right is the plot of the residuals against time. It is clear the residuals are highly correlated with time. When dealing with time-series data, the residuals will sometimes not be spherical and thus OLS will not generate the most precise linear relationship. Specifically, the observations in time-series data do not occur randomly, but occur at specific intervals based on time. This scatter plot clearly shows this effect taking place. As a result of the initial heteroscedastic findings in the scatter plot, additional test will have to be conducted to validate the findings.

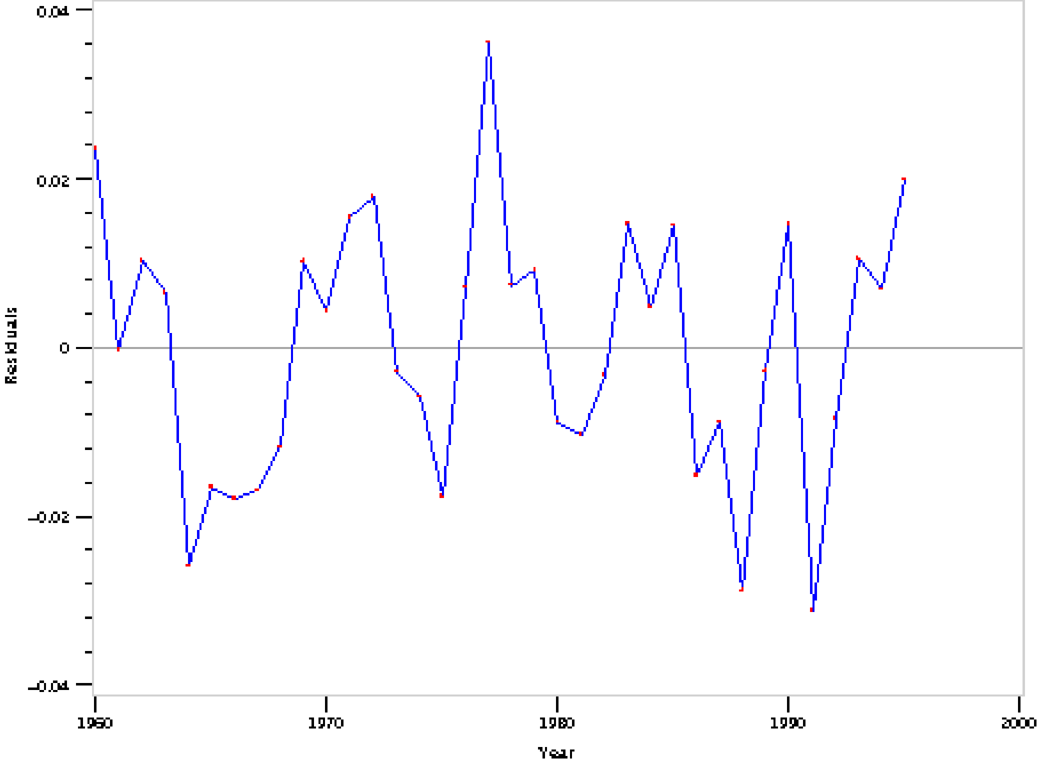


Utilizing the full model is helpful for gaining perspective on autocorrelation. After running an OLS model in SAS, see appendix 2, one can see this model is statistically sound and has a strong correlation coefficient. I want to look at the residuals and assess to see if there is autocorrelation.




Residual by Regressors for Ln_G_Pop
Residual
0.04
0.02
0.00
-0.02
0.0
0.5
1.0
1.5
Ln_pg
8.8
9.0
9.2
9.4
Ln_Income
0.00
0.25
0.50
0.75
1.00
Ln_Pnc
Ln_Puc
2.0
Ln_Pp
-1.0
-0.5
Ln_Pn





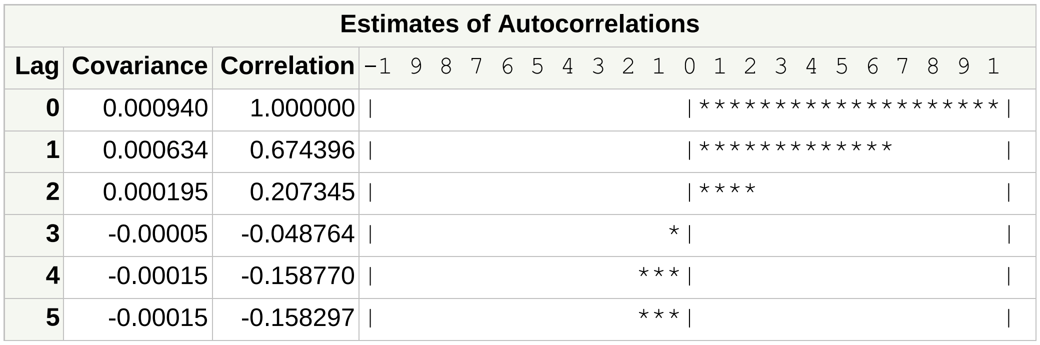
The residuals for the full model look fine. Individually, I cannot see a clear indication of heteroscedasticity. In the last scatter plot, I can see a definable pattern but I don't think it is definitive enough to claim heteroscedasticity.

The Durbin-Watson test is a hypothesis based assessment that tests the residuals for un-correlation (null) or correlation (1). This test is based on the serial correlation assumption that follows the spherical residuals nomenclature, mainly stationary and normally distributed errors with mean of zero (Stern, NYU). Given the scatter plot above, I would expect to reject the null hypothesis after running the Durbin-Watson test. The test is done by taking the least squares estimates coefficients and residuals, and mathematically assessing for difference. For the reduced model, the Durbin-Watson test has a value of .6047 and a statistically significant p-value. This test is a great initial test, but the shortcomings are: assumption that the form of the model is known, the test can be uncertain, and the hypothesis test does not prove that autocorrelation is the only reason the model suffers from heteroscedasticity. The Lagrange multiplier test has been used in past EDAs, and allows the option to test autocorrelation beyond the first-order autocorrelation. Appendix 2 shows the output from this test, and there is autocorrelation that is statistically significant up to the fourth-order.

Box and Pierce have created a hypothesis test with a chi-square distribution that verifies whether time series data is random or ordered. Ljung has modified the procedure to include degrees of freedom (Oxford). This test is robust in that it takes into account lags. For each lag, a correlation value is calculated, squared, added with the other lags, and multiplied by the number of observations. The Ljung test simply divides the correlation values, per lag, by the number of observations, yielding a higher overall value. From the table below, the calculation would be:

36* (.674^2 + .207^2 + .048^2 + .158^2 + .158^2) =19.816 –B&P Statistic

36 * 38 ((.674^2/35) + (.207^2/34) + (.048^2/33) +(.158^2/32) +(.158^2/31) = 21.755- Ljung

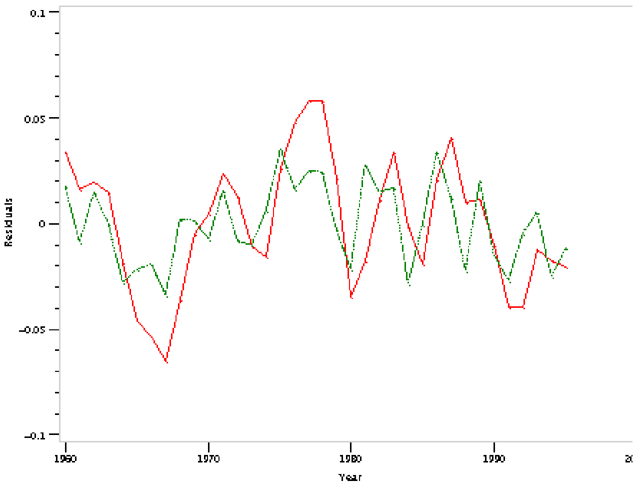


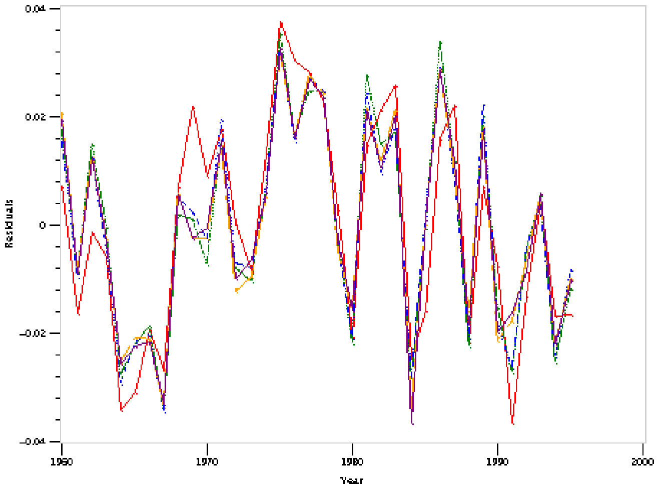
<table>
<tr><th colspan="4">Estimates of Autocorrelations</th></tr>
<tr><th>Lag</th><th>Covariance</th><th>Correlation</th><th>-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1</th></tr>
<tr><td>0</td><td>0.000940</td><td>1.000000</td><td>| |********************|</td></tr>
<tr><td>1</td><td>0.000634</td><td>0.674396</td><td>| |************* |</td></tr>
<tr><td>2</td><td>0.000195</td><td>0.207345</td><td>| |**** |</td></tr>
<tr><td>3</td><td>-0.00005</td><td>-0.048764</td><td>| *| |</td></tr>
<tr><td>4</td><td>-0.00015</td><td>-0.158770</td><td>| ***| |</td></tr>
<tr><td>5</td><td>-0.00015</td><td>-0.158297</td><td>| ***| |</td></tr>
</table>

Both test statistics have validated the assumption of heteroscedasticity, thus OLS should not be used to fit the model.

Generalized Least Squares (GLS) is often used when the presence of heteroscedasticity and autocorrelation is present in a dataset. The assumptions for GLS are more relaxed than OLS, and include an assumption that the errors are nonspherical. There is a rather large shortcoming with using GLS in that one must know the variance-covariance matrix of the disturbances, which are never actually known (Eastern Michigan). Feasible Generalized Least Squares (FGLS) takes the sample data and estimates a variance-covariance matrix. Given that the data is heteroscedastic and specifically this set is working within time series, it will be assumed the errors are correlated and thus non-spherical. FGLS essentially transforms the residuals, similar to Weighted Least Squares, such that they are spherical and satisfy the assumptions of linear regression. FGLS transformed the model with the purpose of obtaining more efficient estimators and standard errors than OLS.

The caveat with using FGLS is the variance-covariance matrix requires an estimate of each element/observation. There are not enough degrees of freedom to take estimates of all the different variances and covariance (Eastern Michigan). Working around this issue is feasible based on the ability to pin-point the nature of the autocorrelation. From using the Autoreg procedure in SAS, it was found that the second autocorrelation model, AR 2, proved to be the best fit model. Essentially, the second iteration of FGLS produced a model that did not suffer from heteroscedasticity from a Durbin-Watson statistical standpoint. Appendix 2 highlights the model specifications, and it can be seen that the model has a strong correlation coefficient, and a few of the coefficients are statistically sound.





After fitting the model with FGLS, to the second autocorrelation or iteration, the residuals are responsible for the green line. As one can see in the first scatter plot, the errors are less polemic and center around 0. The second scatter plot shows the model being fit beyond the autocorrelation of 2 all the way to 5 autocorrelations. It can be seen there is little improvement, and stopping at the second iteration was appropriate.

Future Work

Further recommendations on how this study can be improved upon are the following:

- This EDA utilized many different assessment tools for heteroscedasticity and it was rather overwhelming grasping the concepts. Perhaps focusing on just two instead of four would help comprehension.
- The response variable was hard to conceive, and utilizing a less complex variable would help grasp the implications of the overall EDA.

- Other countries probably have similar records for their economies, and it would be insightful to compare data and models.

Through this initial EDA, coupled with the future work recommendations, economists would gather pertinent information in regard to gasoline consumption and key consumption indicators.

References

Works Cited

Armstrong. "Autocorrelated Errors." *University of Pennsylvania*. Wharton School, n.d. Web. 15 Feb. 2013. <http://armstrong.wharton.upenn.edu/dictionary/definitions/box-pierce%20test.html>.

Chatterjee, Samprit, and Ali S. Hadi. *Regression Analysis by Example*. 5th ed. New York: Wiley, 2000. Print.

Churvich. "The Durbin-Watson Test." *Stern NYU*. NYU, n.d. Web. 15 Feb. 2013. <http://pages.stern.nyu.edu/~churvich/Forecasting/Handouts/DWTest.pdf>.

"Economics 515." *Economics*. Eastern Michigan University , n.d. Web. 15 Feb. 2013. < http://people.emich.edu/j...515_out_general_model.doc>.

Jancovici, Jean-Marc. "Using gas? But what for?." *Manicore*. N.p., n.d. Web. 12 Feb. 2013. <http://www.manicore.com/anglais/documentation_a/oil/gas_use.html>.

Oxford Disctionary. Oxford: Oxford Press, 2012. Print.

Ritholtz, Barry . "Oil Consumption Around the World | The Big Picture." *The Big Picture*. N.p., n.d. Web. 12 Feb. 2013. <http://www.ritholtz.com/blog/2010/06/oil-consumption-around-the-world/>.

Tverberg, Gail. "World Energy Consumption Since 1820 in Charts | Our Finite World." *Our Finite World | Providing a wide view of what may be ahead*. N.p., n.d. Web. 12 Feb. 2013. <http://ourfiniteworld.com/2012/03/12/world-energy-consumption-since-1820-in-charts/>.