2009

Factors related to the unemployment rate: A statistical analysis

Reanna Collins
University of Northern Iowa

FACTORS RELATED TO THE UNITED STATES' UNEMPLOYMENT RATE: A STATISTICAL ANALYSIS

Introduction

Current United States economic conditions have everyone searching for ways to solve the downward spiral that has instigated a crisis in the minds of many people. Rapidly increasing unemployment rates are key reasons leading to this feeling of emergency. A better understanding of unemployment trends over the past few decades and factors correlated to the rate could serve as future indicators of fluctuations in the unemployment rate. A statistical analysis was conducted to determine if any relationship between the unemployment rate and nine selected variables existed in order to identify trends in the unemployment rate and to produce a model that could be used to predict the unemployment rate for a given year. This analysis was done to examine previously suggested variables associated with the unemployment rate and also to find if any new variables might be associated with the rate.

Background

Unemployment Rate Details. How the unemployment rate is calculated is essential in first being able to understand the data that will be presented. The unemployment rate is defined as the proportion of the nation's non-institutionalized population sixteen years and older that is out of work, actively looking for a job, and available for work (Frumkin, 1998). Institutionalized people confined to facilities such as nursing homes or jails are not included in the rate. U.S. armed forces members are also excluded so that the measure represents only the civilian labor force (Frumkin, 1998).

The unemployment rate is computed mathematically as the number of unemployed persons divided by the civilian labor force and multiplied by 100%. A person is considered employed for a given month if the person did at least one hour of work during the reference week, the seven

day week that includes the twelfth of that month; this could be as a paid employee or self-employed. An alternative to the one hour is if the person worked at least fifteen hours at non-paid jobs in a family business. It also includes people temporarily absent due to vacations, illness, or other reasons. People with more than one job are only counted once. Citizens of foreign countries who are temporarily living and working in the United States are also counted as long as they are not working or living on the premises of an embassy. Unemployed people are those who were available to work and actively looked for work during the four-week period prior to the reference week. Merely looking at job advertisements or attending job-training programs or courses do not count as actively searching for work, but is deemed passively searching. Actively searching for work includes activities such as contacting an employer for or participating in an interview, contacting an employment agency, sending out resumes, answering or placing a job advertisement, or checking a union or professional register. People not pursuing a job at all, such as students or stay at home mothers, or are passively looking for jobs are not considered part of the work force and are not included in the unemployment figure at all (Frumkin, 1998).

The U.S. Bureau of Labor Statistics figures the unemployment rate monthly based on a survey of a sample of approximately 50,000 households called the Current Population Study. According to Frumkin (1998), about 47,000 households respond each month. The households are utilized in a rotation. One-fourth of the households chosen one time are not chosen the next time, and each household is interviewed for four months and then dropped from the rotation during the next eight months before being re-included one last time in the survey for another four month period. Frumkin (1998, p. 209) also states that due to sampling error, a plus or minus of 0.1 percentage point in monthly movement of the unemployment rate is not statistically

significant 66% of the time. A change of 0.2 is considered significant, as well as a consistent change of 0.1 upward or downward in consecutive months. If raised to 95% of the time, an unemployment rate would have a range of plus or minus 0.2. This means that monthly movement would have to change by 0.3 in either direction to be statistically significant. This helps when trying to analyze whether or not to consider a monthly change in the unemployment rate measured in this manner as being noteworthy.

The survey method used by the Bureau of Labor Statistics is also said to produce an upward bias in unemployment statistics (Hoel, Clarkson, & Miller, 1983). Other countries require people to physically go to an agency to be recorded as unemployed, making their rate much lower. For example, a 2.5% unemployment rate in the United Kingdom is approximately equivalent to a 5% rate in the United States. Also, amendments to the food stamp and Aid to Families with Dependent Children programs in the 1970s facilitated a larger number of people being considered unemployed. These amendments required recipients of these welfare programs to register for work. If the number who were forced to register for work to receive benefits but cannot or do not wish to work is large, then the unemployment rate will increase even though the number of persons actually seeking work may remain relatively constant. It is hard to determine the extent to which these registration requirements constitute an over-counting in unemployment due to those who have no intention to work or are unemployable (Hoel, Clarkson, & Miller, 1983).

The duration of unemployment is another aspect to consider when investigating the unemployment rate. Duration of unemployment is described as the length of time workers are unemployed. When the average duration of unemployment rises due to workers remaining unemployed longer, the unemployment rate can rise even when the number of workers becoming

unemployed per week remains constant. If the average duration increases by a week or two, the unemployment rate may reflect this by a 0.5% increase (Hoel et al., 1983, p. 193). Some policy advisors believe an unemployment rate including only those who have been unemployed for a longer duration, around six months, is more beneficial when considering courses of action in policies. This idea is due to the fact that most unemployed workers find work within fourteen weeks during times of recession (Hoel et al., 1983, p. 194). This shows that short term spikes in unemployment might not be as important, at least in terms of policy making, as focusing on if a factor like overall duration of unemployment is causing the increased rate.

Factors previously found to be related to the Unemployment Rate. A few factors considered to affect the unemployment rate are discussed by Frumkin (1998, p. 211-212). He lists long term birthrate cycles as having an important impact on the labor force, but they lag in the effect due to the time that must pass for the children to reach working age. Immigration and death rates are less important according to him. Frumkin (1998) also states that the unemployment rates for men were 0.5 to 1.5 percentage points lower than those for women from 1950 to the 1980s, when that differential disappeared and later reversed itself in the 1990s. Teenagers' unemployment rates were two to three times larger than adults during the entire period from 1950-1996. Paul Flaim (as cited in Frumkin, 1998, p. 214) conducted a study and found that the unemployment rate in 1979 would have been 4.4% instead of 5.8%, a difference of 1.4%, if the labor force composition would have remained the same between 1959 and 1979. Flaim (Frumkin, 1998) found that most of the rise in unemployment due to demographic factors over the period resulted from the increased number of teenagers in the population. He also found that out of the 0.6 percentage point reduction in the unemployment rate in the 1980s, 0.5 percentage point, or 80% of the reduction, was due to the decrease in the proportion of teenagers

in the labor force. This displays a relationship between demographic factors and the unemployment rate.

The unemployment rate is also assumed to have a connection with wage growth. More specifically, unemployment has been found to have a negative correlation with wage growth. The economic expansion ending in the early 2000s was one of the longest and delivered the lowest unemployment rates in previous decades, yet nominal wage growth remained relatively contained. Some found the failed acceleration of wages to suggest a breakdown in the historical relationship between unemployment and wage growth. According to Aaronson and Sullivan (2000), the relationship has always been a loose one and if the unemployment rates are broken down amongst states, then the relationship still holds. However, Aaronson and Sullivan (2000) do recognize that a shift in the relationship between unemployment and wage growth would not be terribly surprising. This response is due to the many changes in the labor market in recent years. The general drop in the level of job security, the aging of the work force, higher levels of education, the growth of temporary services employment, the use of fax machines and the internet in job search, and even the increase in the prison population could each be changing the relationship between unemployment and wage growth (Aaronson & Sullivan, 2000). Another thing that could be affecting the wage growth correlation is the increase in other compensatory packages other than wages paid to employees. Aaronson and Sullivan (2000) assert that the fraction of employee compensation paid in the form of wage and salary accruals fell from 92.4% in 1959 to 83.4% in 1980, then to a minimum of 81.0 % in 1994. Since 1994, the fraction of compensation paid in the form of wages and salaries has increased again to 83.9% in 1999. All of these changes in wages and employment could be changing the current connection between wage growth and unemployment.

Gross domestic product (GDP) has also been found to share a relationship with the unemployment rate. Zuckerman (2004) claims that unemployment and gross domestic product are linked, but the connection has been evolving. He reports that usually a 1% decrease in the GDP results in a reduction of 1.3% to 1.5% in private employment. This trend was broken during the recession in the early 2000s when a 0.5% decline in GDP was followed by an almost 3% shrinkage of jobs. Zuckerman (2004) also stated that most of the jobs lost in the factory industry are gone for good, and jobs in the 2003 recovery also shifted from higher to lower paying positions. The results from this recession/recovery are possibly impacting the current economic situation and unemployment spike.

Escalated unemployment rates and criminal activity could share a relationship as well. Moyer (2009) alleges there is a loose relation between unemployment and crime, if there is one at all. During the Great Depression, the crime rate plummeted, which contradicts the initial notion that crime spikes during economic hardships. Bruce Weinberg (as cited in Moyer, 2009, para. 3) states, "People sitting in their houses don't make great targets for crime." While during the "Roaring Twenties," when the economy was booming and people were spending lots of money, the crime rate was high. Two other spikes in the crime rate later in the century were during the early 1970s when a boom was occurring and during the early 1990s when there was a recession taking place. The link tying these two swells in the crime rate has less to do with the economy, but more to do with the escalation of drug use during these periods, heroin in the 1970s and crack cocaine in the 1990s. However, Weinberg and some colleagues conducted an analysis on young males with no more than a high school education, because they are the demographic group that commits most crimes. This examination exposed a direct connection between unemployment rates and the incidence of property crimes, which include felonies such

as burglary, auto theft and robbery. They also discovered that hard times lead to more domestic abuse.

On the other hand, murder, a horrendous criminal activity, has never been well linked to unemployment rates or other standard economic indicators. Rosenfeld and Fornango (as cited in Moyer, 2009) say this is because of a human emotion factor. They reveal a connection between murder rates and the Consumer Sentiment Index, a survey of how people view their current financial situation and how hopeful they are about the future; the lower the index score, the higher the crime rate. A recent study has also shown that economic stimulus can act as a salve, because in the 1930s communities that spent more on public works programs had lower crime rates than the other communities. This helps back the recent governmental economic stimulus package, which included several public works programs, especially since most Americans' hopes were diminishing by the end of the Bush administration.

Another study involving criminal rates and economic downturns was presented by Karen Heimer and Janet Lauritsen (as cited in Galluzzo, 2009). Their analysis concentrated on victimization of Latinos and African Americans versus non Latinos and African Americans during and after periods of economic decline. An association relating economic downturns and victimization rates among minorities over the period of 1973-2005 was established. This coincides with Weinberg's research, where breaking down the society into different demographics can allow there to be trends where the crime rate rises with downturns in the economy.

Immigration is another demographically linked variable that has been found to have some association to the unemployment rate. Employing United States census data covering 1940 to 1980, a study was done to examine the impact of immigration on unemployment and earnings

among racial minorities in the United States (Kposowa, 1995). This study found that increases in immigration in some periods of U.S. history had significant negative effects on employment levels among racial minorities in the United States. Specifically, results of the regression analysis showed that in 1970, a standard deviation change in immigration increased unemployment among minorities by nearly 14%, while in 1980 unemployment increased by nearly 10% given a standard deviation increase in immigration. These results, contrary to Frumkin's (1998), justify some relationship with the unemployment rate and immigration.

Henry Blodget (2008) discussed the comparison of today's recession with the Great Depression in a recent article. He noticed a problem with the argument that today's problems are not the same as during the Great Depression, because the unemployment rate spiked to around 25%, while currently the rate is only around 10.2%. Part of the big difference is due to a change in the method used to calculate the unemployment rate, and if the rate was computed the same way now as it was during the Great Depression, then the unemployment rate today would be much higher. In 1933, the definition of the labor force included individuals in the military, prisoners, and also individuals as young as fourteen (Lieberman, 2008). Even more significant was the difference in the description of the unemployed, which at the time was not limited to those actively seeking employment. Many reported that the changes in society also brought changes to the ruler with which unemployment is measured. More women are encouraged to enter the workforce now compared to during the 1930s, when several who were in the workforce were not even counted. All of these disparities make it complicated to compare the two time periods' unemployment rates alone without taking into consideration other factors. The inconsistency in the type of data available between then and now also causes a dilemma when trying to include data from that time period into a current analysis.

Hypotheses

Before any formal statistical analysis was performed, the reviewed literature and previous knowledge suggested that each of the chosen factors may correlate with the unemployment rate in the following manner:

Independent Variable	Relationship Type
CPI	Inverse
Immigrants Admitted	Direct or None
Burglary	Direct
Minimum Wage Adjusted for Inflation	Inverse
Public Debt	Direct
NASDAQ Low	Inverse
Trade Balance	Direct
Production Workers' Average Hourly Salary	Inverse
Average Gas Price	N/A

The CPI may be inversely related to the unemployment rate because inflation and the unemployment rate are usually inversely related in the short run, and inflation influences the CPI directly. Frumkin (1998) found immigration to be an unimportant factor in the unemployment rate, while a study done by Kposowa (1995) suggests there is a direct relationship between immigration and unemployment among minorities. Based on these two findings, there may be no relationship overall or a possible direct relationship between immigration and the unemployment rate. A loose direct relationship may be found between the unemployment rate and the amount of burglaries in a year, because burglary is a main criminal act associated with unemployment due to the hard times encountered by the unemployed. However, there was contradictory evidence found by Moyer (2009). The minimum wage and hourly wage of production workers may have an inverse relationship with the unemployment rate because when unemployment is low, the economy is generally doing well, which usually leads to higher wages. Although, Aaronson and Sullivan (2000) said the relationship is a loose one based on the changes in compensation packages. Public debt might be directly correlated with the

unemployment rate because the unemployment rate is usually a good economic condition indicator, and if the economy is suffering, then the government might increase the amount it is borrowing. Also, if the economy is doing poorly, unemployment tends to rise. The NASDAQ low for the year could be found inversely related with the unemployment rate, because one would expect the stock market to be down when unemployment is high because the economy is weak. The United States' trade balance might be directly related to the unemployment rate due to a trade surplus being positive (higher) and a trade deficit being negative (lower). A larger trade deficit is probably associated with higher unemployment because in that case, the United States is importing more than it is producing and exporting. The average gas price of all gas types could have either relationship with unemployment if it has any relationship at all. In order to see whether or not these hypotheses hold any validity, an analysis was performed involving the unemployment rate and these variables.

Methodology

Only data from 1975-2007 was included in this analysis because of the previously stated reason of data inconsistencies over long time spans. Nine covariates were chosen to determine if any have a statistically significant relationship with the unemployment rate. These covariates include the United States consumer price index (CPI), the total immigrants admitted to the United States, the amount of burglaries, the minimum wage adjusted for inflation, the public debt, the NASDAQ low, the United States trade balance, the hourly earnings of production workers, and the average price of all gas types. Summary statistics of all the variables were first computed. Next, a linear regression model was found in Statistical Analysis System (SAS) using multiple model building techniques. The SAS code used for this analysis is listed in Appendix A. Model assumptions were then verified in order to make the model valid. Furthermore, the data

were searched for damaging points. After the linear regression aspect was complete, more in-depth correlation checks were performed in order to inspect for problems in the model due to time correlations; the data were observed each year and consecutive years might affect each others' unemployment rates. The results were then summarized and conclusions were produced.

Analysis

Summary Statistics. All of the raw data for the dependent and independent variables for this analysis are listed in Appendix B. After acquiring the data, summary statistics were computed, such as means, medians, standard deviations, sample sizes, minimum values, and maximum values. The summary statistics are provided in the table below:

	U.S. Unemployment Rate	U.S. Consumer Price Index	Total Immigrants Admitted to the U.S.	Burglary	U.S. Fed. Min. Wage Adjusted for inflation in 2008	Public Debt (in Billions)	NASDAQ Low	U.S. Trade Balance (in Millions)	Hourly Earnings of Production Workers	Avg. Price of All Gas Types (in Cents)
Min	4.00	53.80	385378	2050992	5.53	533.20	60.70	-753283.00	4.73	56.70
Max	9.70	207.30	1826595	3795200	8.80	9007.70	2340.68	12404.00	17.43	274.60
Mean	6.23	132.15	800259	2824952	6.85	3854.44	789.79	-194546.70	10.72	129.04
Median	6.00	136.20	703542	2984434	6.56	3665.30	376.87	-104065.00	10.52	120.50
St. Dev	1.4393	44.9653	326234	535299	0.9412	2601.73	741.86	226361.39	3.6050	50.8084
Sample Size	33	33	33	33	33	33	33	33	33	33

The dependent variable, the unemployment rate, varied from 4.00%, which occurred in 2000, to 9.70%, which occurred in 1982, and had a mean of 6.23%; this means that the average unemployment rate between 1975 and 2007 was 6.23% and had a range of 5.70%.

Linear Regression. To assess the trend in the unemployment rate and determine if there are any factors related to the trend, a linear regression analysis including model building techniques and assumption verifications were conducted. Linear regression models the relationship between two or more variables by fitting the line (plane) that best explains the dependent variable from the explanatory variables. Linear regression does this by finding the line (plane) that minimizes the sum of the squares of the vertical distances between the actual

points and the fitted line (plane). The procedures used to decide which set of covariates makes up the best fitting linear regression model include forward selection, backward elimination, stepwise selection, r-squared values or plot, MSE plot, and Mallow's C(p) criterion.

The forward selection procedure began by including the most statistically significant variable. Statistical significance is determined by the p-value, which is the probability of seeing a result as rare as the one observed in a collection of random data in which the variable had no effect. The smaller the p-value is, the more significant the result. During the forward selection process, the variable with the next smallest p-value was added to the model and the regression was performed again. This iteration continued until none of the variables' p-values fell beneath a predetermined cutoff value. A reasonable cutoff value of 0.15 was used in this regression, meaning that there is only a 15% chance of observing a result at least as rare. The first covariate inserted into the model was the NASDAQ low with a p-value of less than 0.0001. In the second step of the forward selection process, the average price of gas was added due to its p-value of 0.0810. Next, the total amount of immigrants admitted to the United States was included in the model with a p-value of 0.0538. No other variables met the 0.15 significance level for entry into the model. The resulting model from the forward selection procedure includes the NASDAQ low, the average price of all gas types, and the total amount of immigrants admitted into the United States.

Opposite of the forward selection process, the backward elimination procedure was conducted next. It started with a model including every covariate chosen for evaluation and eliminated the covariate with the highest p-value in each step until no variables were left with p-values greater than the cutoff value. Thus, only statistically significant variables were left in the regression model. The same 0.15 cutoff value was used for this process as was used in the

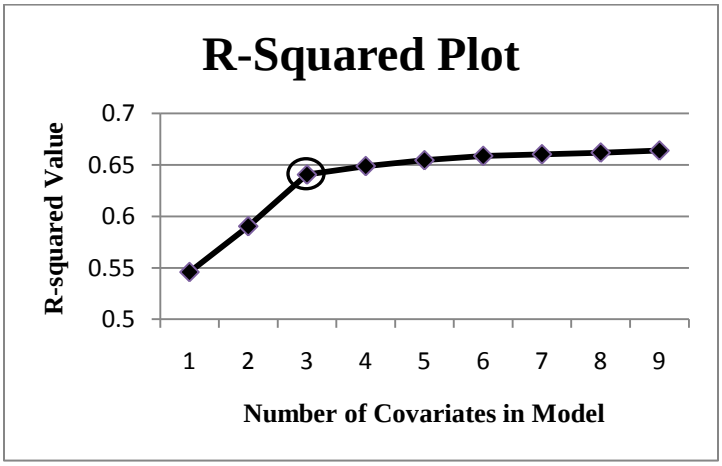
forward selection. The United States Federal minimum wage adjusted for inflation in the year 2008 was the first variable removed from the model because of its p-value equal to 0.7190. Next, the hourly earnings of production workers variable was removed from the model with a p-value of 0.7340. A new regression was conducted each time a variable was discarded, similar to when a variable was added in the forward selection process. This recalculation is the reason why the p-value of the second variable removed is higher than the p-value of the first variable removed; once the first variable was removed, the p-value of the second increased in the new regression model. Subsequently, public debt in millions was the next variable discarded with a p-value equal to 0.7388. The burglary count was the next variable eliminated from the model due to its p-value of 0.5742. Step five in the backward elimination procedure removed the consumer price index variable with a p-value of 0.4929. Finally, the last variable disposed of was the United States trade balance with its p-value equal to 0.4370, after which no variables possessed a p-value greater than 0.15. Identical results to the forward selection process were found by the backward elimination procedure; the model containing the NASDAQ low, the average price of all gas types, and the total amount of immigrants admitted into the United States comprised the end result.

Next, the stepwise selection method was performed. This technique is a variation on forward selection. This method begins the same as the forward selection process, by starting with the one variable model with the lowest p-value. The difference is that at each stage in the process, after a new variable is added, a test is made to check if some previously added variable can be deleted without significantly increasing the residual sum of squares (RSS), which is the sum of squares of the residuals and measures the discrepancy between the data and an estimation model. A cutoff value of 0.15 was used as a limit yet again. The best one variable model found

by the stepwise selection was the model containing the NASDAQ low due to its p-value of <0.0001. Next, the possible two variable models were considered, and the model containing the NASDAQ low and the average price of all gas types was determined to be the best. Three variable models were evaluated after that, and the model with the NASDAQ low, the average price of all gas types, and the number of immigrants admitted to the United States was deemed the best. Finally, four variable models were constructed, but none of them met the criteria of having all covariate p-values below 0.15, so the best three variable model is the model chosen by the stepwise selection procedure. This is the same model selected by both the forward selection and backward elimination methods.

Another technique, the r-squared method, is based on a value expressing the squared correlation of the predicted value from the proposed model to the actual value. This value was used to judge the goodness of fit of a model. The r-squared value, also referred to as the coefficient of determination, is a measure of association ranging from 0 to 1; the closer the value is to 1, the more correlated the predicted value is to the actual value, or the more relevant the regression model is. R-squared is defined as the ratio of the sum of squares explained by a regression model and the total sum of squares around the mean; the equation is $(1 - \frac{SSE}{SST})$. It is referred to as the proportion of variation explained by the model. This method involves evaluating the r-squared value for every model and comparing them to find the model with the highest r-squared value where if another variable was added to the model, then the r-squared value does not increase significantly. This idea follows parsimony, the concept of using the model with the smallest amount of variables that gives the most information, in other words, the model with the highest r-squared value and least amount of variables. The model containing all covariates will always have the highest r-squared value, but parsimony eliminates the covariates

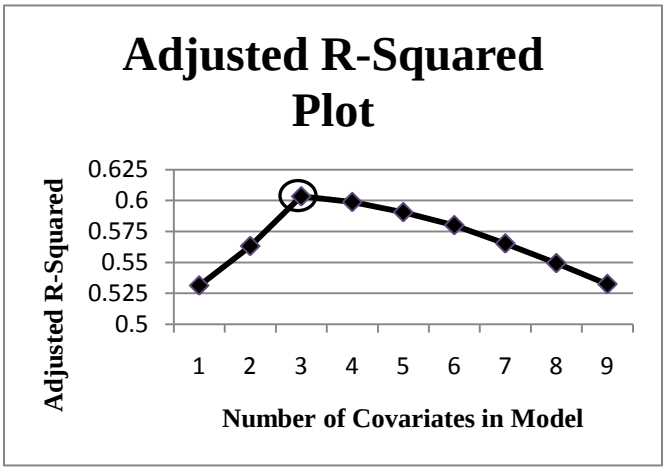
which do not increase correlation significantly. In order to gauge the r-squared values more easily, the r-squared values were separated into subsets based on the number of variables in the model, and only the model with the highest r-squared value within each subset was assessed further. A graph containing the best fitting model (model with the highest r-squared value) with one variable, best model with two variables, all the way up to the best model with nine variables, is displayed below.



As shown in the graph, the model with three variables is the elbow in the graph, which means that it has the highest r-squared value taking into account parsimony; from the best two variable model to the best three variable model, the r-squared value increases significantly, but from the best three variable model to the best four variable model, the r-squared value only increases slightly. The best three variable model turns out to be the same model established by the previous three methods. This model is the model with covariates including the NASDAQ low, the number of immigrants admitted to the United States, and the average price of all gas types.

An adjusted r-squared technique using the same principles as the r-squared technique described previously was also conducted. Adjusted r-squared is a modification of r-squared that adjusts for the number of explanatory terms in a model. The equation for the adjusted r-squared

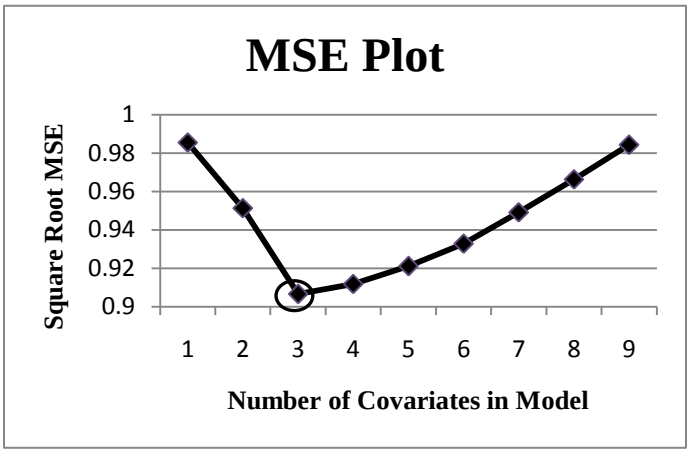
is $1 - \left(\frac{SSE}{SST} * \frac{n-1}{n-p-1}\right)$. Unlike r-squared, the adjusted r-squared increases only if the new term improves the model more than would be expected by chance. The adjusted r-squared value can be negative, and will always be less than or equal to the corresponding r-squared value. The graph below displays the adjusted r-squared values for the same subsets as was used before in the r-squared technique.



The model with three covariates is the best fitting model because it is the peak in the graph. This three covariate model is the same model found by all the previous methods; the model containing the NASDAQ low, the average price of all gas types, and the number of immigrants admitted to the United States.

A mean squared error (MSE) plot was also used to decide upon a model. MSE equals $\frac{SSE}{n-p}$, where SSE is the sums of squared error, n is the sample size, and p is the number of covariates plus 1, or the number of parameters in the model (covariate slopes plus the intercept). This approach is opposite of the r-squared method because the smaller the MSE, the better the model. A smaller MSE means there is less error. Sometimes the model with the absolute smallest MSE is not the best model to use due to parsimony, so a technique similar to the technique used in the r-squared method was performed but using the smallest values, not the

largest values. In practice, the square root MSE is usually used to keep the numbers more manageable. A plot of the smallest square root MSE for each number of covariates is provided below. Once again, the same model with three covariates is the elbow in the graph.



Mallows C (p) criterion was another method used to judge the best fitting model. This method defines a set of reasonable models based on the following criterion: the C (p) for a given model with p covariates is approximately equal to p for the model. C (p) equals $\frac{SSE_p}{MSE_{full}} - (n - 2p)$, where SSE_p is the sums of squared error of the model with p parameters, MSE_{full} is the mean squared error for the model with all possible covariates, n is the number of observations, and p is the number of parameters. The Mallows C (p) for the full model always equals p because the equation would be $\frac{SSE_{full}}{MSE_{full}} - (n - 2p)$, and $MSE_{full} = \frac{SSE_{full}}{n-p}$. Accordingly, the full model will always be included in the set defined by the criterion due to the equation equaling p exactly. Consistent with these guidelines, the model chosen by all other methods was also included in the set of reasonable models defined by the Mallows C (p) criterion.

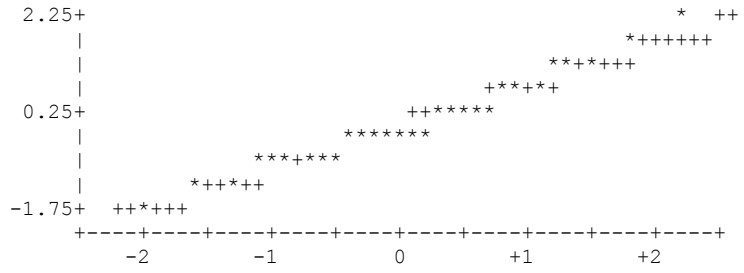
The same model was preferred by all model building techniques. This means that the model favored by all techniques was decided upon to fit the given data the best. The model

containing the NASDAQ low, the number of immigrants admitted to the United States, and the average price of all gas types as the independent variables, and the unemployment rate as the dependent variable was chosen to represent the best relationship between the covariates and the unemployment rate by this linear regression analysis.

In order to utilize the model found by this linear regression analysis, some model assumptions had to be checked. These regression model assumptions include: X-values are measured without error, each observation is equally reliable, residuals are normal with constant variance, appropriateness of a linear model, and independence of the residuals. The first two assumptions, X-values are measured without error and each observation is equally reliable, cannot be tested statistically, but depend on the sample collection methods. The data came from dependable sources, so the first two assumptions are assumed to hold true.

Normality of the residuals can be verified graphically and formally. Graphically, normality of the residuals is demonstrated by a normal probability plot of the residuals. A straight line in the graph suggests normality of the residuals. The plot of the residuals below illustrates a nearly straight line, so the assumption is valid.

Normal Probability Plot
Legend: * = 1 observations + = 2 observations

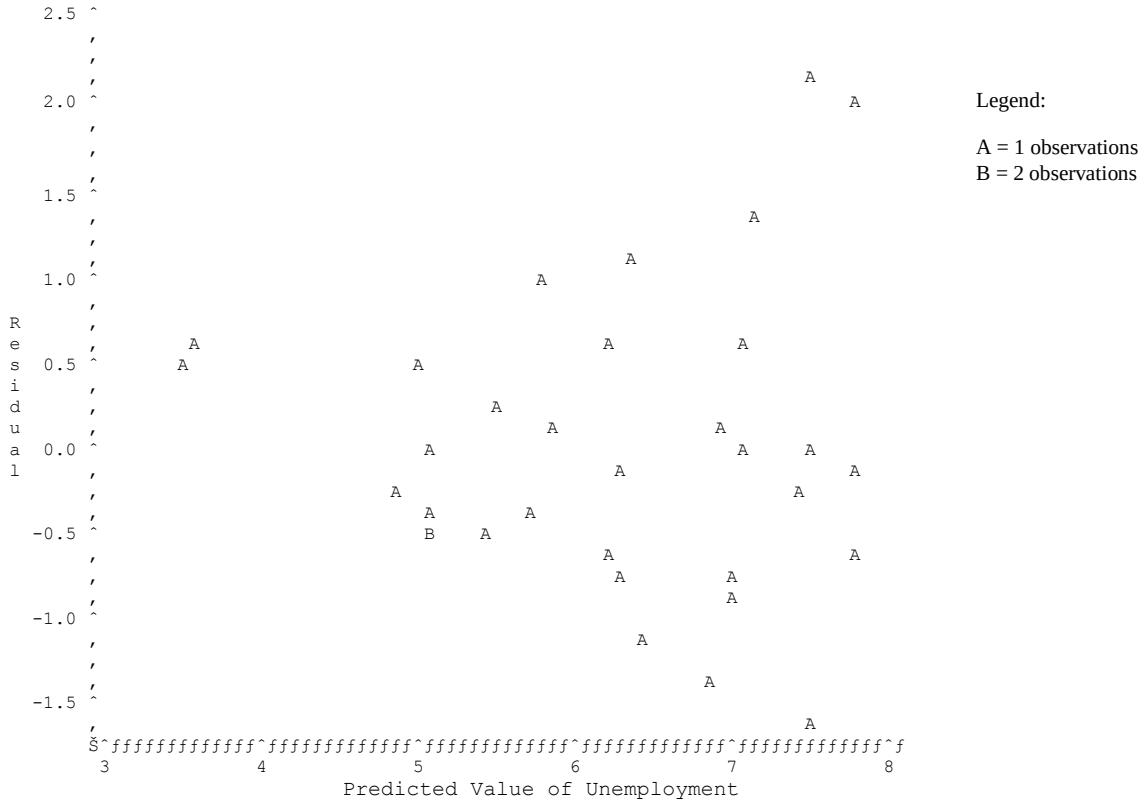


A formal correlation test involving the $R_{e,q}$ value can also be conducted to check for normality. $R_{e,q}$ is the correlation between the residual and its normal score, the corresponding standard normal quartile. If the $R_{e,q}$ is less than a given cutoff value from a table of critical values for the

coefficient of correlation, then reject the assumption of normal residuals. $R_{e,q}$ for the preferred model equals 0.98526, which is greater than the cutoff value of 0.966 for α=0.01 and n=33, so the assumption of normal residuals was once again confirmed acceptable.

Constant variance of the residuals was the next assumption to verify. This can be done two ways, graphically by a plot of the predicted values against the residuals or formally by Breusch-Pagan tests for each covariate. A predicted vs. residuals plot also shows whether a linear model is appropriate or not. The plot is examined for patterns; a fan-shape indicates there are non-equivalent variances, while non-linear patterns indicate problems with having a linear model. The predicted vs. residuals plot below appears to be slightly fan-shaped, so the variances of the residuals may not be constant. However, there does not appear to be any non-linear patterns in the graph, so a linear model is not found inappropriate. The plot is displayed below:

Plot of Predicted vs. Residuals



Breusch-Pagan tests were conducted to further evaluate the variances. In a Breusch-Pagan test, e_i^2 is regressed on X_i for each X. The test statistic equals $\frac{SSR_{Xi}/2}{(SSE/n)^2}$, where SSR_{Xi} is the sums of squared regression from X_i being regressed on X, SSE is the sums of squared error from the preferred model, and n is the number of observations. This test statistic follows a Chi-squared distribution with 1 degree of freedom; reject the assumption of constant variance if the test statistic is greater than the cutoff value determined by the Chi-squared distribution. The results from running all three Breusch-Pagan tests with $\alpha = 0.05$ are summarized in the table below.

Covariate	Test Statistic	Cutoff Value	Decision
NASDAQ Low	5.321573	3.84	Reject Constant Variance
Gas Price	2.289452	3.84	Accept Constant Variance
Immigrants Admitted	2.188324	3.84	Accept constant Variance

Using this alpha level, results in the assumption of constant variance being rejected for the NASDAQ low covariate. This helped confirm the results found with the predicted vs. residual plot. However, if the alpha level is lowered to 0.01, the cutoff value is increased to 6.63 and the following outcome is produced:

Covariate	Test Statistic	Cutoff Value	Decision
NASDAQ Low	5.321573	6.63	Accept Constant Variance
Gas Price	2.289452	6.63	Accept Constant Variance
Immigrants Admitted	2.188324	6.63	Accept Constant Variance

All of the test statistics are below the new cutoff value. These results show the importance in choosing an alpha level when validating assumptions. Using the results from the alpha level of 0.01, the variances are considered to be constant for this analysis.

The last assumption to validate is the independence of the residuals, which can be determined by a runs test. A runs test requires examining the residuals in time sequential order to determine if any patterns exist. This test is important for this particular model due to the

observations being evenly spaced over time, and due to the question surrounding whether one year's value impacts the next year's value. A lack of independence is depicted by long strings of positive or negative residuals or by too many alternating signs of residuals. The test statistic equation is $[\ 1 - \left(\frac{2*n_1*n_2}{n_1+n_2}\right) + \frac{1}{2}] / \sqrt{\frac{(2*n_1*n_2)(2*n_1*n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$, where r equals the number of runs of signs, n_1 equals the number of positive residuals, and n_2 equals the number of negative residuals. For the preferred model, r is 12 n_1 is 15, n_2 is 18 and the test statistic is -1.378429. The test statistic follows a standard normal probability distribution so the cutoff value for α = 0.05 is ±1.96. Independence of residuals is a valid assumption because the test statistic of -1.378429 is in between -1.96 and +1.96.

After checking all of the model assumptions, the data were analyzed for individual damaging points caused by influential observations, large residuals, and high leverages. Influential observations are determined by judging how much the model changes by removing that observation; the greater the change the more influence it has. There are three ways to judge influence, DFITS, Cook's Distance, and DF Betas. DFITS measures how much $\hat{Y}_i$ changes when the i^{th} observation is removed. The i^{th} observation is influential if its |DFITS| is greater than $2*\sqrt{\frac{p}{n}}$, which equals $2*\sqrt{\frac{4}{33}} \approx 0.696311$. According to the DFITS procedure, observations 8, 9 and 17 are influential. Cook's Distance is another way to judge influence, and it measures the influence of the i^{th} observation on all n values, thirty-three in this case. The i^{th} observation is influential if its Cook's Distance is greater than the median of the F distribution with degrees of freedom numerator equal to p (4 in this analysis) and degrees of freedom denominator equal to n – p (29 in this analysis). As a result, the cutoff value for the Cook's Distance is approximately 0.890, and no observations were deemed influential. DF Betas measure how much removing the

i^{th} observation changes the β parameters. The cutoff value for the DF Betas is $\pm\frac{2}{\sqrt{n}}$, which is approximately ±0.348155. Since DF Betas measure the influence on each β, there are influential values for each β individually, and these influential values are summarized in the table below.

Parameter	Influential Observation(s)
β_0	17, 25
β_1	9, 16, 17
β_2	9, 25
β_3	9, 17, 25

None of the observations are shared by all the β parameters, so no observations are deemed influential overall by the DF Betas approach.

Large residuals, which are the distance between the predicted value and the real value, were investigated. Large residuals can be determined using studentized deleted residuals, which follow a t-distribution with n – p – 1 (28 for this analysis) degrees of freedom. For this analysis, the cutoff value is approximately 1.701 with alpha level 0.1. Using this cutoff value, observation 5, 8, and 9 were found to have large residuals.

After large residuals, high leverages were examined. High leverages are points that are remote in the X-space. From the SAS output, Hat Diag H was used to decide if the value has a high leverage. The value is $2\left(\frac{n}{p}\right)$ (0.242424 for this analysis). Points with high leverage are 16, 17, 25, 26, 32, and 33 according to this method.

Damaging points are determined by meeting all three criteria: influential, large residual, high leverage. After looking at the results, no points are damaging because no points meet all three criteria. Even if there were observations deemed damaging, no observations should ever be removed unless there is a problem from data error.

One last diagnostic was to check for multicollinearity, which happens when covariates are highly correlated. There are two ways to test for multicollinearity. One way is called the variance inflation factor (VIF). If the VIF is greater than 10 for an observation there is a multicollinearity problem. None of the VIFs were larger than 10, so no problem was found in this check. Another way to check for multicollinearity is by a condition index. If the index is larger than 30, then there appears to be a multicollinearity problem. No problems were found in this check either, so no multicollinearity problems were found for this model.

Time Series: Are there problems with the linear regression model due to undetected time correlations? The linear regression model established by all of the model building techniques seems viable since all of the model assumptions can be verified, including constant variances of the residuals if the alpha level is set to 0.01. However, since the dependent variable is being measured repeatedly over time, there may be correlation between the measurements for which the regression analysis fails to account. This would appear as correlation, or lack of independence, of the errors during the regression analysis. A runs test was previously conducted to verify independence of the residuals or errors, but since it only takes into account the signs of the residuals, and not the magnitude, it may fail to detect some correlation between the residuals. To address this problem, autocorrelation and partial autocorrelation plots were created, also called ACF and PACF plots. These plots display the correlation between the residuals at different time lags.

The autocorrelation of a time series Y at lag k is the coefficient of correlation between Y_t and $Y_{t-k.}$ The ACF plot will always equal 1 at lag 0 because this is showing that a residual is completely correlated with itself. Autocorrelations should be near 0 for all other time-lag separations. If the residuals are correlated, or non-random, then one or more of the

autocorrelations will be significantly non-zero. A cutoff value, positive and negative, is displayed on the ACF plot, and all other lags should have a correlation in between the cutoff value marks to signify no correlation. For this analysis, lag 1 does not fall within the cutoff value, along with lags 3, 4, 5, 8, and 9. This plot is displayed in Appendix C.

The partial autocorrelation is the amount of correlation between two variables that is not explained by their mutual correlations with a specified set of other variables. It is the amount of correlation between a variable and a lag of itself that is not explained by correlations at all lower-order-lags. The PACF plot has cutoff values marked similar to the ACF plot and the correlation values should fall within the marks, the closer to 0 the better. In the PACF plot for this analysis, lags 1, 2, 4, and 6 have correlation values outside the cutoff value range. This plot is displayed in Appendix D.

Both the ACF and PACF plots suggest that a time series analysis is warranted, so Auto-Regressive Integrated Moving Average (ARIMA) modeling was performed. ARIMA models are noted as ARIMA (p,d,q), where the p is the auto-regressive term, d is the differencing term, and q is the moving average term. Differencing terms are only used for non-stationary processes, when the ACF plot dies down very slowly. The ACF plot for this analysis does not fall into this category, so the differencing term was left at 0. The ACF and PACF plots help identify the p and q components in the ARIMA model. The table below describes the general way to identify which ARIMA model to use.

Model	**ACF Plot**	**PACF Plot**
ARIMA (p,0,0)	Dies down	Cuts off after lag p
ARIMA (0,0,q)	Cuts off after lag q	Dies down
ARIMA (p,0,q)	Dies down	Dies down
ARIMA (p,d,q)	Does not die down	Does not die down

The ACF and PACF plots for this analysis do not fit perfectly into any of these categories, so several ARIMA models were fitted and compared based upon Akaike Information Criterion (AIC) and the resulting ACF and PACF plots from the ARIMA model. AIC is a penalized likelihood, and the lower the AIC is, the better the model is. The AIC for model k equals $n*\log(SSE_k) + 2P_k$, where n is the number of residuals, SSE_k is the sums of squared error for model k, and P_k is the number of parameters for model k. SSE_k represents the goodness of fit of the model, and P_k represents a penalty term for how many parameters are used in the model. As the number of parameters increases, the SSE_k will decrease, but the P_k will increase, which is why it is called a penalized likelihood. ACF and PACF plots are also created in order to make sure the problem of correlated residuals is resolved with the ARIMA model. The table below displays the results from the fitted ARIMA models.

Possible Model	**AIC**	**Comments about ACF & PACF plots**
ARIMA (1,0,0)	70.06815	Both ok
ARIMA (2,0,0)	61.29249	Both ok
ARIMA (3,0,0)	61.02022	Both ok
ARIMA (4,0,0)	54.8522	Both ok
ARIMA (5,0,0)	54.01416	Both ok
ARIMA(6,0,0)	49.28251	Both ok
ARIMA (7,0,0)	49.78799	Both ok
ARIMA (0,0,1)	65.70913	PACF plot lag 4 exceeds cutoff value
ARIMA (0,0,2)	67.69519	PACF plot lag 4 exceeds cutoff value
ARIMA (1,0,1)	65.66281	PACF plot lag 7 exceeds cutoff value
ARIMA (1,0,2)	67.51379	PACF plot lag 7 exceeds cutoff value
ARIMA (2,0,1)	53.84297	Both ok
ARIMA (2,0,2)	55.84295	Both ok
ARIMA (3,0,1)	55.29058	Both ok

The ARIMA model chosen is bolded in the table above, and detailed results of the model are in appendix E. ARIMA (6,0,0) was decided as the best model because it has the lowest AIC value and the ACF and PACF plots show no significant correlation at any of the lags. In order to get perfect results, the ARIMA model and the previously selected linear regression model should

be fitted at the same time in SAS. The signs of the slopes in the new model created by this process would be expected to remain the same as in the linear regression model found previously. The main reason for this analysis is to see what variables are correlated with the unemployment rate and the type of relationship they share, so this process was not deemed necessary because the relationships established by the linear regression model should not change.

Results

Linear Regression Equation. In general, the linear regression equation is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$; the subsequent table explains the notations.

Notation	Meaning
$\hat{Y}$	Estimate for the unemployment rate
$\hat{\beta}_0$	Intercept of the line
$\hat{\beta}_1$	Slope for the NASDAQ low
X_1	Value for the NASDAQ low
β_2	Slope for the number of immigrants
X_2	Value for the number of immigrants
$\hat{\beta}_3$	Slope for the average price of all gas
X_3	Value for the average price of all gas

The model preferred by the previously conducted model building techniques provides the estimates for the intercept and slope coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$. In linear regression, the size of the coefficient for each independent variable is equivalent to the size of the effect that variable has on the dependent variable, and the sign on the coefficient (positive or negative) reflects the direction of the effect. For the best fitting model, $\hat{\beta}_0$ is 6.97312, $\hat{\beta}_1$ is -0.00196, $\hat{\beta}_2$ is -0.00000113, and $\hat{\beta}_3$ is 0.01328. Following the guidelines stated before, the NASDAQ low has an inverse relationship with the unemployment rate due to the corresponding slope, $\hat{\beta}_1$, being a negative number. If the NASDAQ low is increased by 1, the unemployment rate would decrease by 0.00196 on average, which displays how the size of the slope matches the amount of effect the variable has on the dependent variable, the unemployment rate. The number of immigrants

admitted to the United States also has an inverse relationship with the unemployment rate due to its negative slope. If the number of immigrants admitted to the United States increased by 1, the unemployment rate would decrease by 0.00000113 on average. Alternatively, the average price of all gas types has a direct relationship with the unemployment rate because of its positive slope. If the average price of all gas types increased by 1, the unemployment rate would increase by 0.01328 on average. All of these individual effects are contingent on holding the other variables constant while examining it separately.

This linear equation can be used to predict the unemployment rate in a given year. Plugging in numbers for the average price of all gas types, the number of immigrants admitted to the United States, and the NASDAQ low will produce an estimate of the unemployment rate for that year. If a more accurate result is needed, the linear regression model and the ARIMA model can be fitted together in SAS in the future.

Relationships Established by this Analysis. The table below displays the relationships believed to be held between the unemployment rate and the chosen independent variables before this analysis and the relationship types found by the analysis.

Dependent Variable	Relationship Type Believed Before Analysis	Relationship Type Found After Analysis
CPI	Inverse	Insignificant
Immigrants Admitted	Direct or None	Inverse
Burglary	Direct	Insignificant
Minimum Wage Adjusted for Inflation	Inverse	Insignificant
Public Debt	Direct	Insignificant
NASDAQ Low	Inverse	Inverse
Trade Balance	Direct	Insignificant
Production Workers' Average Hourly	Inverse	Insignificant
Average Gas Price	Not Sure	Direct

Several variables were found insignificant by this analysis, some of which have been stated significant in previous research; these included the CPI, the amount of burglaries, hourly wages,

public debt, and the trade balance. The three variables that were deemed to be significantly correlated with the unemployment rate include the NASDAQ low, the amount of immigrants admitted to the United States and the average price of all gas types.

The NASDAQ low was found to have an inverse relationship to the unemployment rate by this analysis, which is the same as the relationship thought to have existed based on prior knowledge. An inverse relationship indicates that a downward movement in the NASDAQ low is associated with an increase in the unemployment rate, and an upward shift in the NASDAQ low is associated with a decline in the unemployment rate. Logically, this relationship makes sense because one would naturally think that if the low of the NASDAQ is extremely low one year, this indicates that the economy is doing poorly so the unemployment rate is probably high. On the other hand, if the NASDAQ low is relatively high, than the economy is most likely doing well so the unemployment rate is presumably low. Correlation, not causation is implied by this relationship; this means that a change in the NASDAQ low from year to year does not necessarily cause a change in the unemployment rate, and vice versa. The two variables tend to move opposite of each other; this could mean that one variable causes the change in the other, another variable all together causes both to change, or that the relationship is completely coincidental. Although a change in the NASDAQ low does not necessarily cause a change in the unemployment rate, an inverse relationship does appear to exist according to this analysis, so the NASDAQ low can be used as an indicator of what change might appear in the unemployment rate for that year.

Another variable found to be significant by this analysis is the number of immigrants admitted to the United States yearly. Based on the literature previously reviewed, this variable was thought to have either no relationship or a direct relationship with the unemployment rate.

Frumkin (1998) did not think any significant relationship existed between immigration and the unemployment rate, while a study in Kposowa (1995) found that increases in immigration had a direct relationship with unemployment among minorities throughout some periods of history in the United States. However, different results were observed in this statistical analysis. An inverse relationship was established between the number of immigrants admitted to the United States and the unemployment rate in this analysis. This is contradictory to most people's first notion about immigration. A lot of people think that as immigration increases, these immigrants take jobs from the currently employed and do more work for less money. Although, rationally if the immigrants come in legally and fill employment openings, they increase the number of employed and would thus lower the unemployment rate so an inverse relationship could make sense. Once again this does not mean that immigration causes a change in the unemployment rate, but rather that the two variables tend to move opposite of each other. This means that the number of immigrants admitted to the United States each year could be used as an indicator of that year's unemployment rate.

Furthermore, the average price of all gas types was observed to have a significant relationship with the unemployment rate based on this analysis. Before the analysis was conducted, no specific relationship was thought to exist between the two variables. The analysis revealed a direct relationship between the average price of all gas types and the unemployment rate. This means that high gas prices are correlated with high unemployment rates. High unemployment rates may be associated with high gas prices due to the fact that gas prices effect a lot of businesses; when gas prices are high business costs increase, which may lead to cuts in costs elsewhere such as labor reductions. Yet again, this correlation does not imply causation, which means high gas prices do not necessarily cause the unemployment rate to rise, and vice

versa. This direct relationship can serve as another indicator of the unemployment rate for a certain year.

Conclusion

The statistical analysis conducted found three variables that may serve as future indicators of the unemployment rate: the NASDAQ low, the number of immigrants admitted to the United States, and the average price of all gas types. The NASDAQ low and the number of immigrants admitted both have an indirect relationship with the unemployment rate, while the average price of all gas types shares a direct relationship with the unemployment rate. This means that a higher NASDAQ low and number of immigrants admitted are correlated with a lower unemployment rate, and that a higher average price of gas is correlated with a higher unemployment rate. Although these three variables may be used as future indicators of the unemployment rate, one must be careful of extrapolation. The relationships were found using data collected from 1975-2007 and can be used to approximate the unemployment rate during this period, but may not provide correct results when applied to data prior to 1975 and subsequent to 2007. Even with the risk of extrapolation, linear regression models are reasonable tools for predictions. Hopefully, the relationships discovered between the significant covariates and the unemployment rate through this statistical analysis can prove to be helpful in forecasting future changes in the unemployment rate. More research is left to be done in this field, such as combining the ARIMA model with the regression model to get a more accurate model, trying these processes with other variables suspected to have a relationship with the unemployment rate, and a more in depth study into why these relationships exist. In spite of the future work that can be done, the relationships found in this statistical analysis can serve as indications of an increasing unemployment rate and also as indications of an unemployment rate declining back to

its normal state; this could serve useful in the upcoming months as a hint to when the economy will begin recovering and also in the future to serve as warning signals of potential downturns.

References

Aaronson, D. & Sullivan, D. (2000). Unemployment and wage growth: Recent cross-state evidence. *Chicago Fed: Economic Perspectives, 24*. Retrieved from http://www.chicagofed.org/publications/economicperspectives/2000/2qep4.pdf

Average hours and earnings of U.S. production workers, 1969-2006. (2009). In *The world almanac and book of facts 2008*. (p.101). New York: Readers Digest Trade Publishing.

Blodget, H. (2008, December 5). Great depression unemployment didn't hit 25 percent overnight. Retrieved February 24, 2009, from Yahoo! Finance Web site: http://finance.yahoo.com/tech-ticker/article/140365/Great-Depression-Unemployment-Didnt-Hit-25-Percent-Overnight

Bureau of Labor Statistics. *Labor statistics from the current population survey*. Available from: http://www.bls.gov/cps/

Chan, J. S. K. & Choy S. T. B. (2008). Analysis of covariance structures in time series. *Journal of Data Science 6*, 573-589. Retrieved from http://proj1.sinica.edu.tw/~jds/JDS-432.pdf

Crime in the U.S. (2009). In *TIME almanac 2009*. (p.692). Chicago: Encyclopedia Britannica Inc.

Fernandez, G. C. J. (2006). *Model selection methods in proc mixed.* Retrieved from http://www.ag.unr.edu/gf/ashs/allmixedwuss2006.pdf

Frumkin, N. (1998). *Tracking America's economy*. New York: M.E. Sharpe.

Galluzzo, G. (2009, February 15). *Study finds recession associated with increases in minority victims of crime.* Retrieved February 24, 2009, from University of Iowa News Services Web site: http://www.eurekalert.org/pub_releases/2009-02/uoi-sfr020909.php

Gasoline retail prices, U.S. city average, 1974-2007. (2009). In *The world almanac and book of facts 2008*. (p.106). New York: Readers Digest Trade Publishing.

Hoel, A. A., Clarkson, K. W., & Miller, R. L. (1983). *Economics sourcebook of government statistics*. Lexington, MA: LexingtonBooks.

Kposowa, A. (1995). The Impact of immigration on unemployment and earnings among racial minorities in the United States. *Ethnic & racial studies, 18*(3), 605-628. Retrieved from http://www.popline.org/docs/1171/250511.html

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Table B-6. *Applied linear regression models*. New York: McGraw-Hill Companies, Inc.

Lieberman, M. (2008, October 28). *Jobless rate is a deceptive depression comparison*. Retrieved from http://www.foxbusiness.com/story/markets/economy/jobless-rate-deceptive-depression-comparison/

Moyer, M. (2009). Stick 'em up. *Scientific American, 300*(3), 15-16. Retrieved from Academic Search Elite database.

NASDAQ stock market, 1971-2007. (2009). In *The world almanac and book of facts 2008*. (p.60). New York: Readers Digest Trade Publishing.

Public debt of the U.S. (2009). In *the world almanac and book of facts 2008*. (p.55). New York: Readers Digest Trade Publishing.

Total immigrants admitted to the US, 1901-2007. (2009). In *TIME almanac 2009*. (p.649-650). Chicago: Encyclopedia Britannica Inc.

Trends in U.S. foreign trade, 1970-2006. (2009). In *The world almanac and book of facts 2008*. (p.68). New York: Readers Digest Trade Publishing.

U.S. consumer price index, 1913-2007. (2009). In *TIME almanac 2009*. (p.719-720). Chicago: Encyclopedia Britannica Inc.

U.S. federal minimum wage rates, 1952-2008. (2009). In *TIME almanac 2009*. (p.712). Chicago: Encyclopedia Britannica Inc.

U.S. unemployment rates. (2009). In *TIME almanac 2009*. (p.715). Chicago: Encyclopedia Britannica Inc.

Zuckerman, M. (2004, February 9). The case of the missing jobs. *U.S. News & World Report, 136*(5), 67-68. Retrieved from Middle Search Plus database.

APPENDIX A

SAS Code

```
Options ls=80;
Data UnemploymentRate;
Input Unemployment CPI Immigrants Burglary MinWage PubDebt NASDAQ Trade HrEarnings
Gas;
Cards;
8.5 53.80 385378 3625300 8.46 533.2 60.70 12404 4.73 56.7
7.7 56.90 499093 3108700 8.76 620.4 78.06 -6082 5.06 59.0
7.1 60.60 458755 3071500 8.22 698.8 93.66 -27246 5.44 62.2
6.1 65.20 589810 3128300 8.80 771.5 99.09 -29763 5.88 65.2
5.8 72.60 394244 3327700 8.65 826.5 117.84 -24565 6.34 88.2
7.1 82.40 524295 3795200 8.15 907.7 124.09 -19407 6.85 122.1
7.6 90.90 595014 3779700 7.98 997.9 170.80 -16172 7.44 135.3
9.7 96.50 533624 3447100 7.52 1142.0 158.92 -24156 7.87 128.1
9.6 99.60 550052 3129900 7.29 1377.2 229.88 -57767 8.20 122.5
7.5 103.9 541811 2984434 6.98 1572.3 223.91 -109072 8.49 119.8
7.2 107.6 568149 3073348 6.74 1823.1 245.82 -121880 8.74 119.6
7.0 109.6 600027 3241410 6.62 2125.3 322.14 -138538 8.93 93.1
6.2 113.6 599889 3236184 6.39 2350.3 288.49 -151684 9.14 95.7
5.5 118.3 641346 3218077 6.13 2602.3 329.00 -114566 9.44 96.3
5.3 124.0 1090172 3168170 5.85 2857.4 376.87 -93141 9.80 106.0
5.6 130.7 1535872 3073909 6.30 3233.3 322.93 -80864 10.20 121.7
6.8 136.2 1826595 3157150 6.76 3665.3 352.85 -31135 10.52 119.6
7.5 140.3 973445 2979884 6.56 4064.6 545.85 -39212 10.77 119.0
6.9 144.5 903916 2834808 6.37 4411.5 645.02 -70311 11.05 117.3
6.1 148.2 803993 2712774 6.21 4692.8 691.23 -98493 11.34 117.4
5.6 152.4 720177 2593784 6.04 5224.8 740.53 -96384 11.65 120.5
5.4 156.9 915560 2506400 6.56 5413.1 978.17 -104065 12.04 128.8
4.9 160.5 797847 2460526 6.95 5526.2 1194.39 -108273 12.51 129.1
4.5 163.0 653206 2332735 6.84 5656.3 1357.09 -166140 13.01 111.5
4.2 166.6 644787 2100739 6.70 5674.2 2193.13 -265090 13.49 122.1
4.0 172.2 841002 2050992 6.48 5807.5 2332.78 -379835 14.02 156.3
4.7 177.1 1058902 2116531 6.30 6228.2 1387.06 -365126 14.54 153.1
5.8 179.9 1059356 2151875 6.20 6783.2 1114.11 -423725 14.97 144.1
6.0 184.0 703542 2154834 6.06 6783.2 1271.47 -496915 15.37 163.8
5.5 188.9 957883 2144446 5.91 7397.1 1752.00 -607730 15.69 192.3
5.1 195.3 1122257 2154126 5.71 7932.7 1904.00 -711567 16.13 233.8
4.6 201.6 1266129 2183746 5.53 8507.0 2020.39 -753283 16.76 263.5
4.6 207.3 1052415 2179140 6.11 2340.7 2340.68 -700258 17.43 274.6
;
Proc reg data=UnemploymentRate;
Model Unemployment= CPI Immigrants Burglary MinWage PubDebt NASDAQ Trade
HrEarnings Gas / selection=forward slentry=0.15;
```

```
Proc reg data= UnemploymentRate;
Model Unemployment= CPI Immigrants Burglary MinWage PubDebt NASDAQ Trade
HrEarnings Gas / selection=backward slstay=0.15;

Proc reg data= UnemploymentRate;
Model Unemployment= CPI Immigrants Burglary MinWage PubDebt NASDAQ Trade
HrEarnings Gas / selection=stepwise slentry=0.15;

Proc reg data= UnemploymentRate;
Model Unemployment= CPI Immigrants Burglary MinWage PubDebt NASDAQ Trade
HrEarnings Gas / selection=rsquare rmse;

Proc reg data= UnemploymentRate;
Model Unemployment= CPI Immigrants Burglary MinWage PubDebt NASDAQ Trade
HrEarnings Gas / selection=cp;

Proc reg data= UnemploymentRate;
Model Unemployment= CPI Immigrants Burglary MinWage PubDebt NASDAQ Trade
HrEarnings Gas / selection=adjrsq;

Proc reg data= UnemploymentRate;
Model Unemployment= Immigrants NASDAQ Gas;
Output out=next r=resid p=yhat;

Data final;
Set next;
Sqrres = resid**2;

Proc rank normal=blom;
Var resid;
Ranks nscore;

Proc univariate plot data=final;
Var resid;

Proc plot data=final;
Plot resid*yhat;

Proc corr;
Var resid nscore;

Proc glm data=final;
Model sqrres=Immigrants;
Proc glm data=final;
Model sqrres=Nasdaq;
Proc glm data=final;
```

```
Model sqrres=Gas;

Proc reg data=UnemploymentRate;
Model Unemployment=Immigrants  Nasdaq Gas / r p influence Collin vif;

Run;
```

APPENDIX B

Raw Data

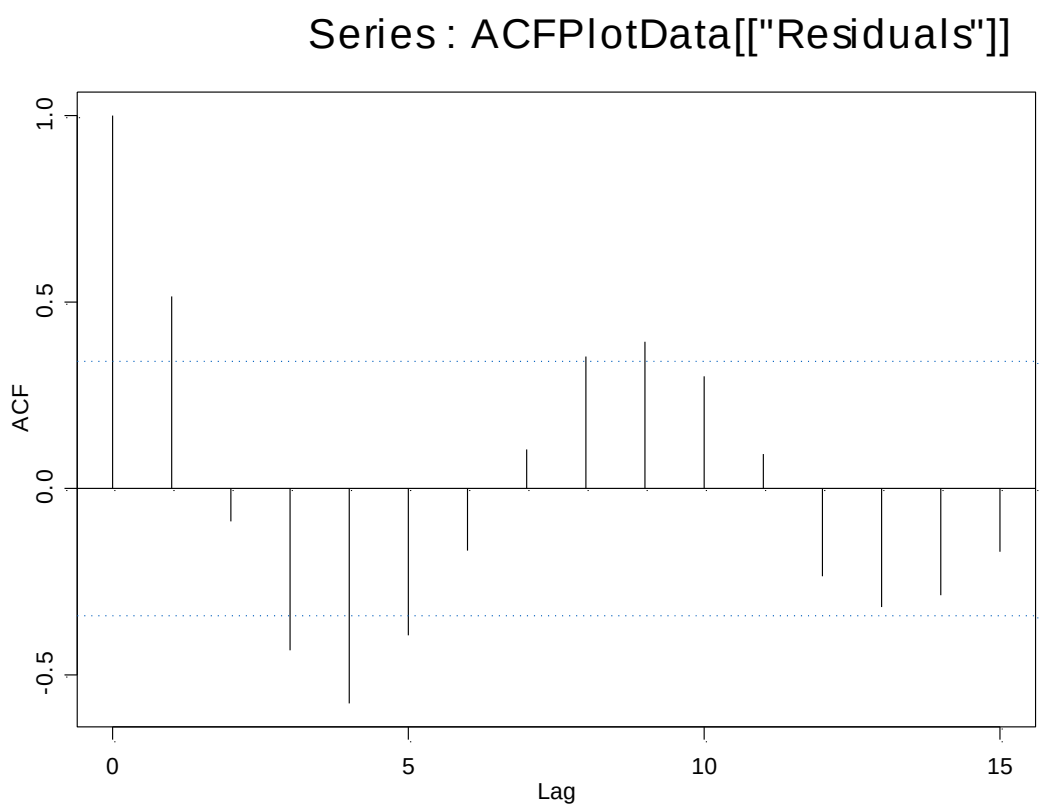
Year	*US Unemployment Rate*	*US Consumer Price Index*	*Total Immigrants Admitted to the US*	*Burglary*	*US Fed. Min. Wage Adjusted for inflation in yr 2008*	*Public Debt (in Billions)*	*NASDAQ Low*	*US Trade Balance (in millions)*	*Hourly Earnings of Production Workers*	*Avg Price of All Gas Types (in cents)*
1975	8.50	53.8	385,378	3,625,30	8.46	533.2	60.70	12,404	4.73	56.7
1976	7.70	56.9	499,093	3,108,70	8.76	620.4	78.06	-6,082	5.06	59.0
1977	7.10	60.6	458,755	3,071,50	8.22	698.8	93.66	-27,246	5.44	62.2
1978	6.10	65.2	589,810	3,128,30	8.80	771.5	99.09	-29,763	5.88	65.2
1979	5.80	72.6	394,244	3,327,70	8.65	826.5	117.84	-24,565	6.34	88.2
1980	7.10	82.4	524,295	3,795,20	8.15	907.7	124.09	-19,407	6.85	122.1
1981	7.60	90.9	595,014	3,779,70	7.98	997.9	170.80	-16,172	7.44	135.3
1982	9.70	96.5	533,624	3,447,10	7.52	1,142.0	158.92	-24,156	7.87	128.1
1983	9.60	99.6	550,052	3,129,90	7.29	1,377.2	229.88	-57,767	8.20	122.5
1984	7.50	103.9	541,811	2,984,43	6.98	1,572.3	223.91	-109,072	8.49	119.8
1985	7.20	107.6	568,149	3,073,34	6.74	1,823.1	245.82	-121,880	8.74	119.6
1986	7.00	109.6	600,027	3,241,41	6.62	2,125.3	322.14	-138,538	8.93	93.1
1987	6.20	113.6	599,889	3,236,18	6.39	2,350.3	288.49	-151,684	9.14	95.7
1988	5.50	118.3	641,346	3,218,07	6.13	2,602.3	329.00	-114,566	9.44	96.3
1989	5.30	124.0	1,090,172	3,168,17	5.85	2,857.4	376.87	-93,141	9.80	106.0
1990	5.60	130.7	1,535,872	3,073,90	6.30	3,233.3	322.93	-80,864	10.20	121.7
1991	6.80	136.2	1,826,595	3,157,15	6.76	3,665.3	352.85	-31,135	10.52	119.6
1992	7.50	140.3	973,445	2,979,88	6.56	4,064.6	545.85	-39,212	10.77	119.0
1993	6.90	144.5	903,916	2,834,80	6.37	4,411.5	645.02	-70,311	11.05	117.3
1994	6.10	148.2	803,993	2,712,77	6.21	4,692.8	691.23	-98,493	11.34	117.4
1995	5.60	152.4	720,177	2,593,78	6.04	5,224.8	740.53	-96,384	11.65	120.5
1996	5.40	156.9	915,560	2,506,40	6.56	5,413.1	978.17	-104,065	12.04	128.8
1997	4.90	160.5	797,847	2,460,52	6.95	5,526.2	1,194.39	-108,273	12.51	129.1
1998	4.50	163.0	653,206	2,332,73	6.84	5,656.3	1,357.09	-166,140	13.01	111.5
1999	4.20	166.6	644,787	2,100,73	6.70	5,674.2	2,193.13	-265,090	13.49	122.1
2000	4.00	172.2	841,002	2,050,99	6.48	5,807.5	2,332.78	-379,835	14.02	156.3
2001	4.70	177.1	1,058,902	2,116,53	6.30	6,228.2	1,387.06	-365,126	14.54	153.1
2002	5.80	179.9	1,059,356	2,151,87	6.20	6,783.2	1,114.11	-423,725	14.97	144.1
2003	6.00	184.0	703,542	2,154,83	6.06	6,783.2	1,271.47	-496,915	15.37	163.8
2004	5.50	188.9	957,883	2,144,44	5.91	7,379.1	1,752.00	-607,730	15.69	192.3
2005	5.10	195.3	1,122,257	2,154,12	5.71	7,932.7	1,904.00	-711,567	16.13	233.8
2006	4.60	201.6	1,266,129	2,183,74	5.53	8,507.0	2,020.39	-753,283	16.76	263.5
2007	4.60	207.3	1,052,415	2,179,14	6.11	9,007.7	2,340.68	-700,258	17.43	274.6

APPENDIX C

ACF Plot

*** Autocorrelations for series ACFPlotData[["Residuals"]] ***
Call: acf(x = ACFPlotData[["Residuals"]], type = "correlation", plot = F)
Autocorrelation matrix:

	lag	ACFPlotData
1	0	1.0000
2	1	0.5153
3	2	-0.0882
4	3	-0.4340
5	4	-0.5766
6	5	-0.3939
7	6	-0.1668
8	7	0.1044
9	8	0.3539
10	9	0.3935
11	10	0.3006
12	11	0.0924
13	12	-0.2356
14	13	-0.3183
15	14	-0.2862
16	15	-0.1694

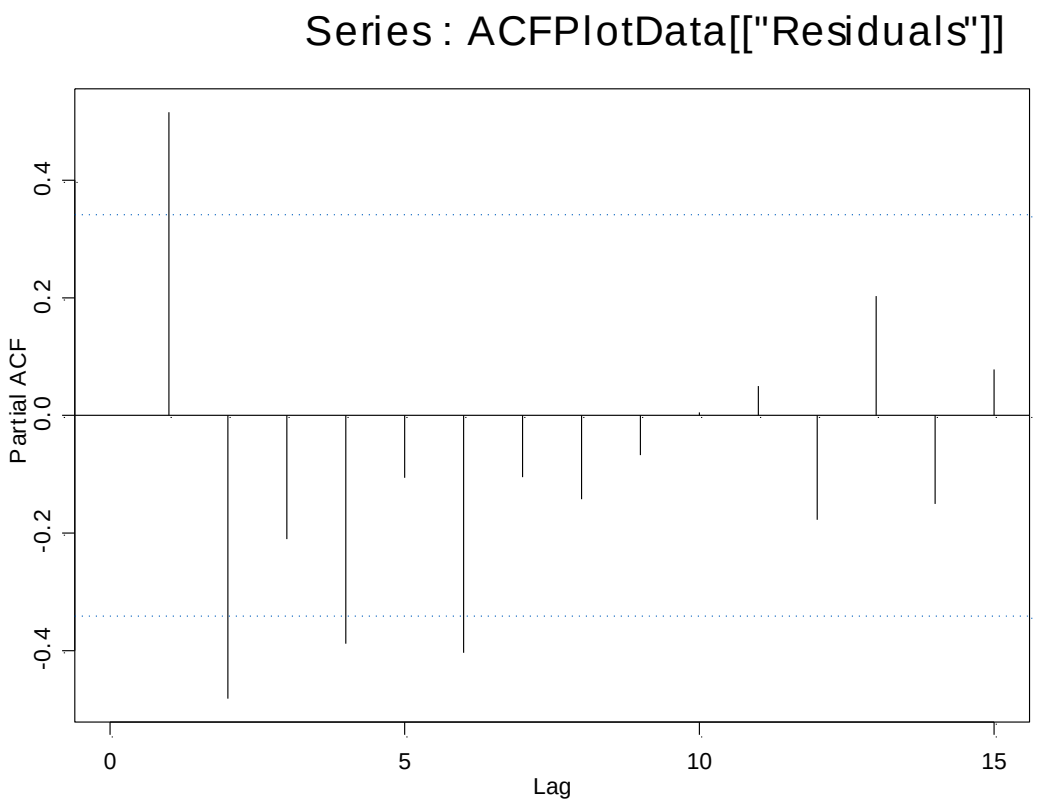


APPENDIX D

PACF Plot

*** Partial Correlations for series ACFPlotData[["Residuals"]] ***
Call: acf(x = ACFPlotData[["Residuals"]], type = "partial", plot = F)
Partial Correlation matrix:

	lag	ACFPlotData
1	1	0.5153
2	2	-0.4815
3	3	-0.2102
4	4	-0.3879
5	5	-0.1060
6	6	-0.4034
7	7	-0.1049
8	8	-0.1423
9	9	-0.0675
10	10	0.0053
11	11	0.0499
12	12	-0.1776
13	13	0.2027
14	14	-0.1505
15	15	0.0781



APPENDIX E

ARIMA Model Chosen

*** ARIMA Model Fitted to Series ACFPlotData[["Residuals"]] ***

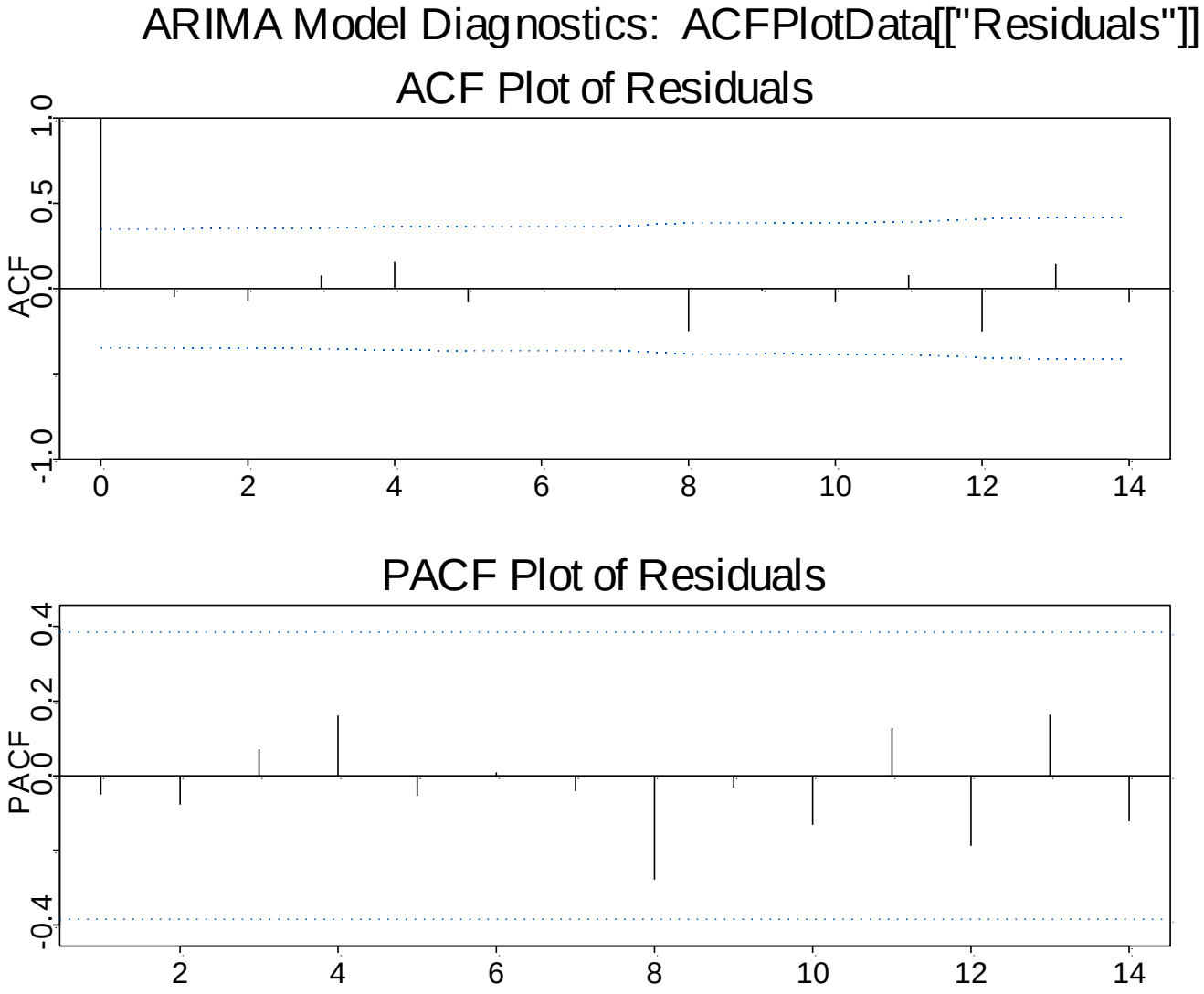
Call: arima.mle(x = `ACFPlotData[[\"Residuals\"]]`, model = model, xreg = xreg, max.iter = nlmin.max.iter, max.fcal = nlmin.max.fcal)
Method: Maximum Likelihood
Model : 6 0 0
Period: 1

Coefficients:
AR : 0.44863 -0.61846 0.06408 -0.52047 0.09813 -0.4211

Variance-Covariance Matrix:

	ar(1)	ar(2)	ar(3)	ar(4)	ar(5)	ar(6)
ar(1)	0.030469345	-0.015085553	0.014788582	-0.001373974	0.009630928	0.003362501
ar(2)	-0.015085553	0.037567217	-0.023470306	0.015620472	-0.007774325	0.009630928
ar(3)	0.014788582	-0.023470306	0.041700793	-0.023702884	0.015620472	-0.001373974
ar(4)	-0.001373974	0.015620472	-0.023702884	0.041700793	-0.023470306	0.014788582
ar(5)	0.009630928	-0.007774325	0.015620472	-0.023470306	0.037567217	-0.015085553
ar(6)	0.003362501	0.009630928	-0.001373974	0.014788582	-0.015085553	0.030469345

Optimizer has NOT converged
Due to: iteration limit
AIC: 49.28251


ARIMA Model Diagnostics: ACFPlotData[["Residuals"]]
ACF Plot of Residuals
ACF
1.0
0.5
0.0
-1.0
0
2
4
6
8
10
12
14
PACF Plot of Residuals
PACF
0.4
0.2
0.0
-0.4
2
4
6
8
10
12
14
ARIMA(6,0,0) Model with Mean 0