Aaron Daniel
Chia Huang
Licai Huang
Medhavi Sikaria

Signal Processing: Forecasting and Modeling

Abstract

Forecasting future events and statistics is problematic because the data set is a stochastic, rather than deterministic one. However, being able to accurately evaluate, predict, and estimate the outcomes of situations and events may potentially yield a large profit as well as allowing one to more accurately prepare for the future. By utilizing best fit models, we are able to test different models against our data set to determine which one most accurately represents the given information. By finding a model to best fit the data, one is able to determine the statistical chance of a given outcome and, often times, accurately forecast events.

Introduction

Businesses must be able to predict the changing market and constantly adjust accordingly to make most profit or adapt to future changes. For instance, the government uses forecasting algorithms to predict the future unemployment rate, airlines may need to predict the average number of passengers in the coming year, and stock brokers may need to predict the stock futures. These predictions are made based on historical sales data, market trend, economic statistics, and external factors. Time series modeling, a method to analyze data history, is used during the process of forecasting. Time series is a collection of data recorded over the period of time. Using the time series, a chosen model is fit to the data set using model estimation methods. Together, the validity and confidence of the selected model is studied to find the best fit model which will predict the future of the data.

Problem

Models act as a fundamental form of an existing structure which presents knowledge of that system in usable form. Many models solves the past data sets by forecasting future data sets. Given unique data sets, we need to find the best appropriate fit model for our given data sets. Our data sets, earnings of particular companies, need to be applied to multiple types of models to select the best fit one. Each model provides its best forecast of the data sets. The AR model, it is able to provide and forecast a linear segment of the data. The MA model forecasts randomness of the data sets, providing a multitude of solutions of the times series to forecast trends. The ARMA model provides and forecasts linear data, while maintaining elements of randomness to give better results than AR and MA model.

Since every model has different order, we need to determine the correct order for AR(p), MA(q), and ARMA(p,q). The AR model property is determined by the autocorrelation function, which gives repeating patterns or trends. The MA model property, given by the partial autocorrelation function, gives a sharp cutoff. For the ARMA model, both of the autocorrelation function and the partial autocorrelation function exponentially decrease depending on the order. In addition, each model poses a problem regarding seasonality. Seasonality of each time series model is defined as a repetitive and predictable movement around the trend lines. This makes forecasting difficult. Moreover, seasonality adjustment does not account for abnormal earning conditions or month-to-month changes in earnings. It is crucial that we notice seasonality factors and make calculation adjustments based on the past data, otherwise, futuristic data may not be accurate.

The best model is often either the AR or MA model, depending on the data sets. However, it is possible for the AR and MA term to nullify each other if used in conjunction. This means that we need a different method to determine the order for the ARMA model. We call this method the Information Criterion. The two best known forms are the Akaike information criterion (AIC) and Bayesian information criterion (BIC). The AIC and BIC determine the order of an ARMA model. With the ARMA model working, it will be able to forecast the future data sets however, before implementing our findings, we need to know the confidence interval of our forecast.

Idea

Our objective is to develop an innovative model to forecast future data sets given the previous data, represented as a time series. In time series analysis, autoregressive models (AR), moving-average models (MA), or autoregressive moving average (ARMA) models are often used to fit the time series data. The purpose of this analysis is to assist with more accurately predicting the future trend. In practice, we will determine which one of the three models that is a better fit for the data. AR, MA, and ARMA show different properties. In the AR model, we observe that the current state depends strongly and linearly the previous data. In MA model, the data sets do not strongly depend on the one or more previous states, but rather rely on randomness. In the ARMA model, we simply combine the features of the MA and AR models. We use the above properties to determine which model we will use to fit and predict the future of our data. In some cases, some time series show obvious cyclical and periodical behavior, otherwise known as seasonality. We plot the original data sets to observe the periodical nature. If we observe any seasonality, we use multiplicative seasonal model, to better predict future data.

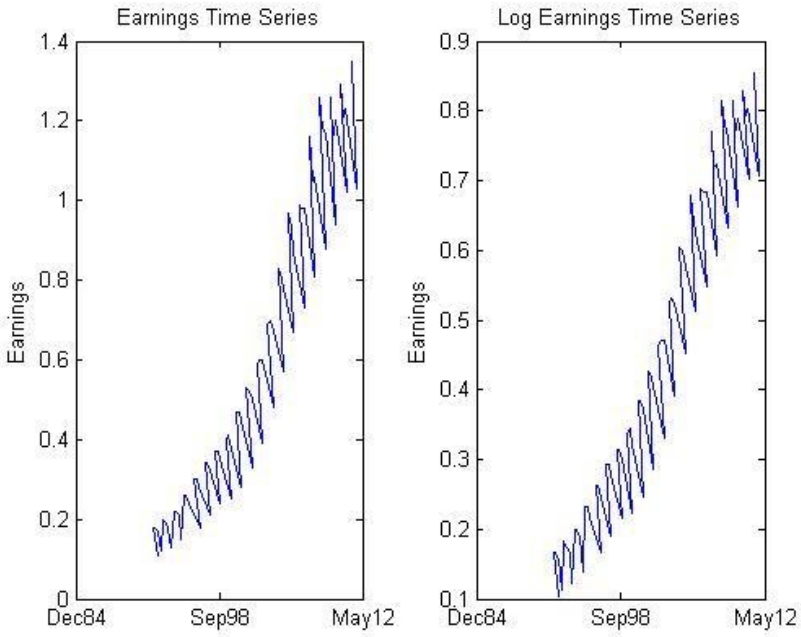
After determining the best fit model, we determine the appropriate order of the chosen model. For this purpose, we have two methods to minimize the order. For the first method, we plot the autocorrelation function (ACF) for AR model or the partial

autocorrelation function (PACF) for MA model to find the appropriate order. For the second method, we use several information criterion such as the Akaike information criterion (AIC) and Bayesian information Criterion (BIC). These two information criterion help us to minimize the order of the model and, at the same time, better fit the set of the data. The next step is to forecast next two years' stock price of Fedex and Johnson & Johnson based on the optimal model and predict it with 95% confidence intervals.

We then test our model to see whether it is optimal or not. In this section, we use two methods for testing. We simulate a time series with same number of data points as the given data sets, and then compare the simulated one with original one. We plot these data points side by side to see that how similar the two graphs are. In our data set, we have quarterly stock prices of Fedex from 1991 to 2006 and Johnson & Johnson from 1992 to 2011. We forecast the quarterly stock price for Fedex from 2007 to 2008 and for Johnson & Johnson from 2012 to 2013 and compare our predicted stock price with the actual stock price in the given period. As a result of this comparison, we can determine whether our fitted model is workable or not.

Details

The process of forecasting events begins with Time Series, or "collection of data taken over a sequence of times." (Thiesson, Bo; Maxwell, David; Chickering, Heckerman, David; Meek, Christopher) The stationarity of the series, in combination with the mean, or "expectation," and lag-k autocovariance function is determined and later used to calculate the covariance and finally return and log return of the series. After plotting the return and log return of our dataset it appears as follows:

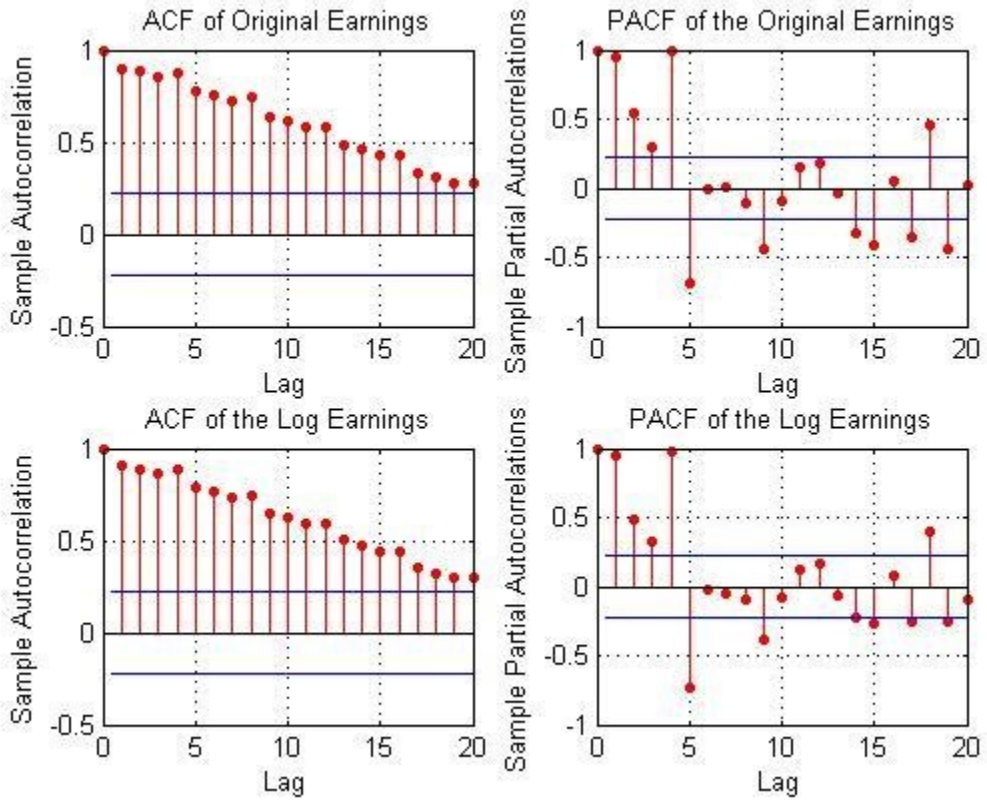


While a Time Series documents all data over a time period, it does not account for the importance of said data. For this we implement lag-k autocorrelation, where the relationship between the current and previous k states is highly correlated. The random variables are added to the model by adding white noise with a zero mean and finite variance. We can predict the current time (r_t) given the previous time (r_{t-1}) using the Autoregressive model with lag-1, otherwise known as the AR(1) model for short. The AR(1) model is represented by:

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t\,, \quad Cov(r_{t-1}, a_t) = 0$$

We fit the model parameters , φ_0 and φ_1, to fit the model to the statistical data.

As models get more advanced and the lag increases to numbers higher than one, it becomes necessary to change the AR(1) model to an AR(p) model where p = k (Ganapathy, Sriram; Hermansky, Hynek). The Partial Auto Correlation Function, or PACF, of a time series is a function of the Auto Correlation Function. It is utilized to determine the the “p” or “order” of the autoregressive model. The plot of the Auto Correlation and Partial Auto Correlation Functions for our given dataset appears thusly:

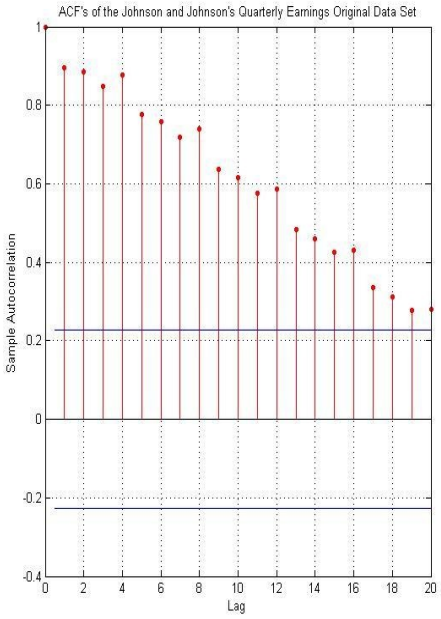


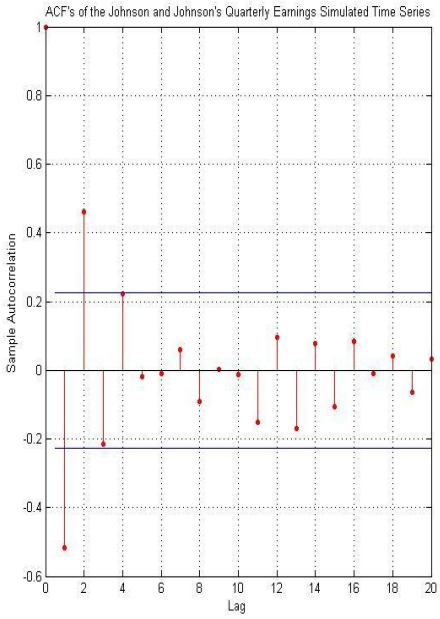
There are several methods by which to determine the order or an AR model such as the Akaike Information criterion or AIC. The AIC is defined as follows:

$$-\frac{2}{N}\log(\text{likelihood}) + \frac{2}{N} \times (\text{number of parameters})$$

where N is the estimate of the sample size of the dataset.

For some of the datasets encountered, the AR(p) model did not suffice to accurately represent the data as it would have infinitely many parameters. For datasets such as stock data a common model implemented is the Moving Average (MA) model. Such a model is beneficial because it is always weakly stationary since there exists a finite number of linear combinations of a sequence of white noise. We utilize the Auto Correlation Function or (ACF) to determine the order, represented by “q,” of the MA Model. After implementing the ACF of our original data set and simulated time series, we obtained the following:





For some applications, the MA and AR models require such a high number of parameters to represent the structure of the data, that they become impractical to use. In these situations, we utilize the Autoregressive Moving Average Model, or ARMA, which incorporates properties of both the AR and MA models, thereby implementing the significant properties of both model classes while keeping the number of parameters minimized. Since the ARMA model utilizes both AR and MA models, we denote the order of the ARMA model with both p and q, where p is the order of the AR portion, and q, the order of the MA portion (Broersen, P. M. T.). While the ARMA model is a combination of both MA and AR models, the fusion of the models actually invalidates both ACF and PACF methods of determining order. For the ARMA model, we must

implement the previously discussed, AIC. More specifically, the pair (p,q) which minimizes the AIC of a dataset often provides a sufficient model for the data.

By utilizing the aforementioned models, we can accurately forecast a time series. We begin to analyze a time series at time t=h, or the forecast origin. In order to predict the future of a time series, we attempt to interpret t = h+ℓ, where ℓ > 1 and where ℓ is called the forecast horizon (Gao, Wei). We utilize the minimum squared error loss function to accomplish this prediction:

$$E[(r_{h+\ell} - \hat{r}_h(\ell))^2 | F_h] \leq \min_g E[(r_{h+\ell} - g)^2 | F_h]$$

By setting t = h +1 and by implementing an AR(p) model, we obtain:

$$r_{h+1} = \phi_0 + \phi_1 r_h + \cdots + \phi_p r_{h+1-p} + a_{h+1}$$

with an associated forecast error of:

$$e_h(1) = r_{h+1} - \hat{r}_h(1) = a_{h+1}$$

Assuming we would like to predict further into the future, we set t = h+2 and implement the AR(p) model yielding:

$$r_{h+2} = \phi_0 + \phi_1 r_{h+1} + \cdots + \phi_p r_{h+2-p} + a_{h+2}$$

To look further into the future we find the general solution by setting t = h+ℓ and implement the AR(p) model yielding:

$$r_{h+\ell} = \phi_0 + \phi_1 r_{h+\ell-1} + \cdots + \phi_p r_{h+\ell-p} + a_{h+\ell}$$

For other types of datasets, we require an MA model. Given the above situations, we obtain each situations respective model as follows:

$$\hat{r}_h(1) = E(r_{h+1}|F_h) = c_0 - \theta_1 a_h \quad e_h(1) = r_{h+1} - \hat{r}_h(1) = a_{h+1}$$

$$\hat{r}_h(2) = E(r_{h+2}|F_h) = c_0 \quad \hat{r}_h(2) = E(r_{h+2}|F_h) = c_0$$

$$r_{h+\ell} = c_0 + a_{h+\ell} - \theta_1 a_{h+\ell-1} - \theta_2 a_{h+\ell-2}$$

Finally, for datasets with many parameters, we implement the previously discussed ARMA model. By setting t = h+1 we obtain:

$$r_{h+\ell} = c_0 + a_{h+\ell} - \theta_1 a_{h+\ell-1} - \theta_2 a_{h+\ell-2}$$

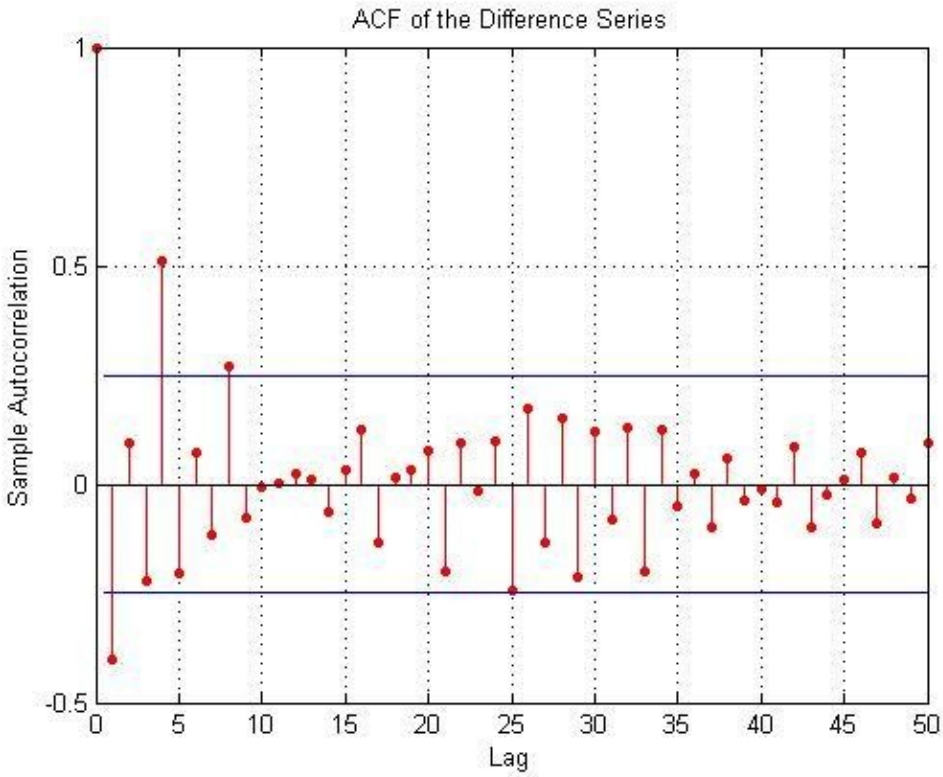
and given that we want the general solution for t = h + ℓ we find:

$$\hat{r}_h(\ell) = E(r_{h+\ell}|F_h) = \phi_0 + \sum_{i=1}^{p} \phi_i \hat{r}_h(\ell - i) - \sum_{i=1}^{q} \theta_i a_h(\ell - i)$$

Some time series, such as those attempting to represent financial data, often exhibit periodic or cyclical behavior which prevents one from forecasting. When attempting to forecast a time series, if seasonality exists within the data set, we must account for it. To do this we follow the procedure of Box, Jenkins, and Reinsel (1994). First, we calculate the differenced series:

$$(1 - B)(1 - B^{12})y_t$$

The plotted ACF of the differenced series, derived from our given data set is:



Where y_t is the original log-transformed data. We then find the regular and seasonal difference:

$$(1-B)y_t = y_t - y_{t-1}$$
$$(1-B^{12})y_t = y_t - y_{t-12}$$

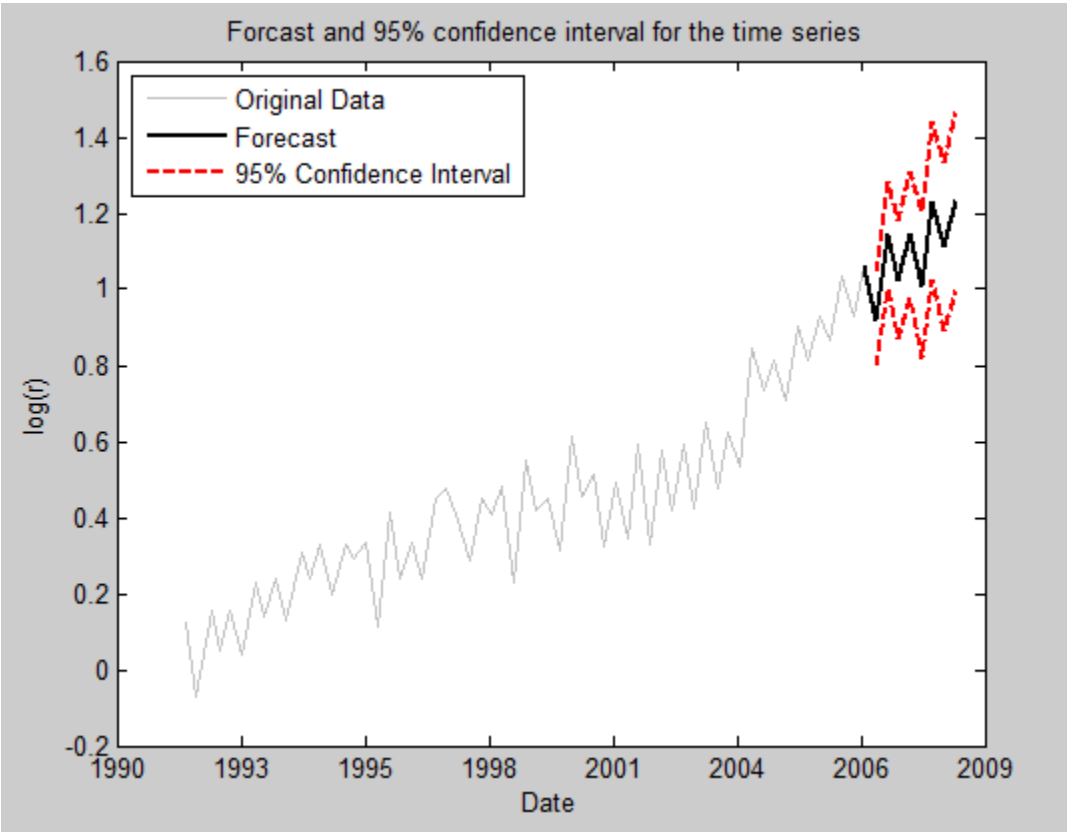
By combining these two equations we obtain the differenced series:

$$(1-B)(1-B^{12})y_t = (1-B)(y_t - y_{t-12}) = y_t - y_{t-1} - y_{t-12} + y_{t-13}$$

Finally, we are able to model the seasonal time series with the Multiplicative Seasonal Model and forecast our data with the following equation:

$$(1-B^s)(1-B)r_t = (1-\theta B)(1-\Theta B^s)a_t$$

Our forecasted results of the data are as follows:



Related Work

ARMA Time Series Modeling with Graphics Models By Bo Thiesson, David Maxwell Chickering, David Heckerman, and Christopher Meek

The paper explains about a special information criterion in order to determine the order p and q for the ARMA model. The ARMA model is a combination of AR model and MA model, however, the partial autocorrelation function and autocorrelation function could not be used to determine the order p and q for the ARMA model. The two information criterion which can determine the order p and q, the Akaike and Bayesian information criterion, were implemented in our research to determine our order for p and q.

Applications of Signal Analysis Using Autoregressive Models of Amplitude Modulation By Sriram Ganapathy and Hynek Hermansky

The research in this paper provided us with the approach to determine the order of an autoregressive model for the given time series. Partial autocorrelation function is needed to determine the unknown appropriate order p of an autoregressive model.

The Quality of Models for ARMA Processes By P. M. T. Broersen

The work explains methods measuring error and assess the quality of different models of a given ARMA process. It explains the different way to forecast and predict the time domain. We applied methods discussed by Broersen when attempting to determine our confidence interval.

Economic Time Series Forecasting: Box-Jenkins Methodology and Signal Processing Approach By Milan Marček

In his paper, Marcek discusses time series analysis and modeling with regards to the approach in signal processing. The fitted model will forecast the data sets into a plot which can be utilized to predict stock prices in the future. While this paper was not directly implemented in our research, it was useful as an early example of the uses of time series and modeling.

Forecasting of Nonlinear Signal Processing By Wei Gao

Discussing forecasting of a nonlinear signal processing, Gao's paper was very useful for reference when attempting to forecast nonlinear data. Unfortunately, our datasets were very different than Gao's so the research was not directly applicable, however, we were able to decipher most of the work and translate it to be compatible with our data.

Conclusion

When we first plot the chart of the log of the quarterly earnings and original quarterly earnings, both show obvious periodical behaviors. We observe that the periodicity is 4 for both of the data sets. Since these are quarterly earnings, it is reasonable to assume a periodicity of 4 from the beginning. When following the procedure of Box, Jenkins, and Reinsel (1994) to obtain the difference time series, confirming that the periodicity is 4 since with difference time series figure appears stationary. We assume that the ARMA model can capture the difference time series, so we use AIC or BIC to determine the order; the Quarterly earnings for Fedex is ARMA(3,3) and the earnings for Johnson and Johnson is ARMA(1,1). We simulate the time series with the same number of data points as the original difference data set. Compare the simulated one with the original one, a observe the similar trends between the two charts. We then apply the data set to the Multiplicative Seasonal Model to account for the seasonality. Comparing the actual data with our predicted data, we determine that our algorithm is feasible and correctly forecasts the future data within a 95% confidence interval and thus conclude that our model and algorithm is feasible for predicting a quarterly earnings time series.

After predicting the outcomes, we take the original data for the years that we forecast and plot it. After comparing the original data and the predicted data, we see how accurately our prediction matches the original data.

In the future, we will try to apply our model and algorithm to different types of data such as national unemployment rate or the national GDP while attempting to boost our confidence interval as high as possible.