STA457 Final Project
Report

Using Time Series Analysis of US Unemployment Rate 1948-2016 to Forecast Unemployment Rate

Submitted by

Marcos Jaen Cortes

Under the guidance of
Tharshanna Nadarajah

Winter Semester 2022

Abstract

Unemployment is an economic concept that affects all agents of the economy directly or indirectly, and consequently, it is of great importance to be able to forecast it in order to plan ahead. In this paper it is shown that an ARIMA model can be used to forecast short-term unemployment rate values for the US to great precision and accuracy. To do this, the author of this report uses historical monthly unemployment rate figures from the US encompassing more than 67 years. Seasonal and non-seasonal differencing are used to make the data stationary, and various ARIMA models are proposed to model unemployment. An in-depth analysis is then executed to identify the best-fit model that satisfies assumptions and best models the data to later use this model to forecast 10 months of unemployment rate figures and do an spectral analysis. Implications of results and weaknesses of methods and results is then discussed with an emphasis on the assumption of Ceteris paribus, or the fact that this model assumes the lack of existence of exogenous economic shocks, which is a very limiting and powerful assumption.

Keywords: unemployment, unemployment rate, ARIMA, forecasting unemployment.

Contents

List of Figures

0.1 Introduction

0.1.1 Background

Unemployment is an economic phenomena of such importance that a lot of literature has being devoted to it. The term unemployment refers to the situation when job-seeker, for more than 3 months, actively searches to provide its labour to the market, however is unable to find work. Unemployment rate is an economic indicator that measures the number of unemployed people as a percentage of the labour force population (those age between 16-65). It is normal and healthy to have some minor level of unemployment in an economy, this level is usually denoted as the natural unemployment rate. Studies have shown that this rate is unique to every economy and is dependent of various factors, but most economies have a natural unemployment rate of 4 to 5 percent.

0.1.2 Motivation

Thus, due to the importance that this indicator has in our lives and my interest in economics I have decided to forecast the US unemployment rate using historic data and compare my results to the actual figures. The data I will use comes from the R library astsa and is named UnempRate. It has the monthly U.S. unemployment rate in percent unemployed from Jan, 1948 to Nov, 2016. The purpose of this study is to evaluate to what extent it is possible to predict the unemployment rate of a nation using its historical unemployment figures and advance statistical methods.

0.1.3 Literature

Research regarding unemployment rate points out that it is hard to forecast unemployment without majors assumptions as unemployment can be heavily influenced by various exogenous variables such as but not limited to foreign direct investment, economic downturns or economic expansions, technology, etc. Papers such as, "Annual Estimates of Unemployment in the United States, 1900-1954" by Stanley Lebergott support this conclusion and add that study of this indicator should focus more on the effects it has on the economy and its agents rather than in forecasting the indicator itself due to the difficulty and inaccuracy of the latter.

0.1.4 Report Statement

Studies such as this one are of great importance as forecasting unemployment is of great utility as it enables better governmental budget and policy planning for example. It also permits citizens to better allocate their disposable income and plan ahead. In short, unemployment is an economic phenomena that affects anyone and everyone that is part of a market structure; thus been able to forecast it is of great benefit to society as a whole.

0.2 Statistical Methods

0.2.1 Exploratory Analysis

Our dataset has 827 observations, each corresponds to the unemployment rate of the US on a specific month between Jan, 1948 and Nov, 2016 inclusive. Plotting this time series we obtain the following graph.

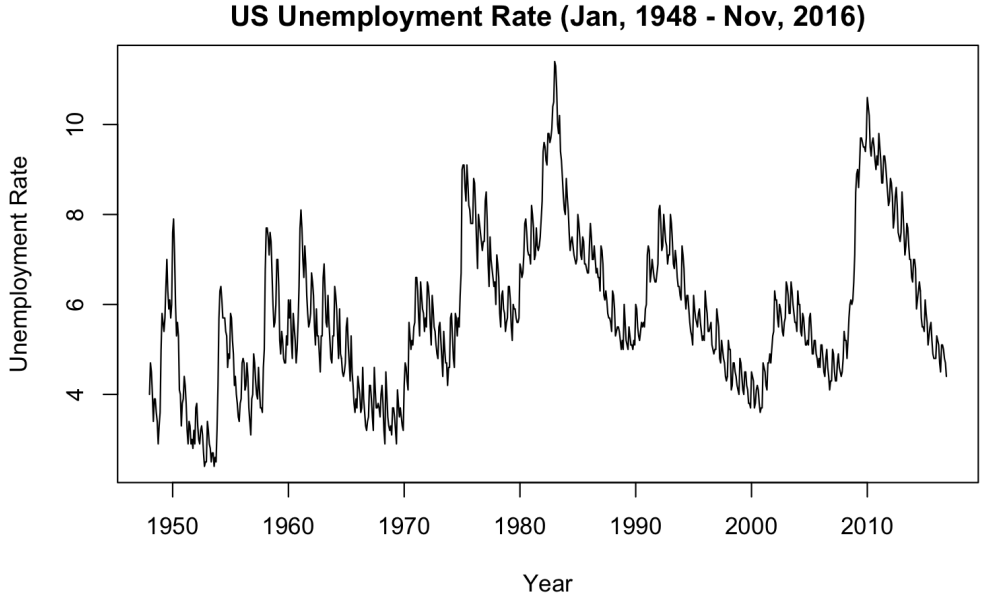

Figure 1: Unaltered Plot of Data

Clearly, the mean is not constant and the process is not currently stationary. Further inspection shows the sample ACF decays to zero extremely slowly as h increases, meaning that differencing is needed to convert data to

a stationary process. Thus, we proceed to take the first difference of the data and obtain the following:

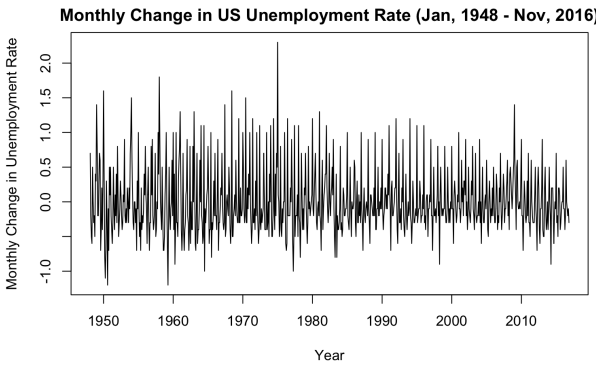

Figure 2: Plot of First Difference

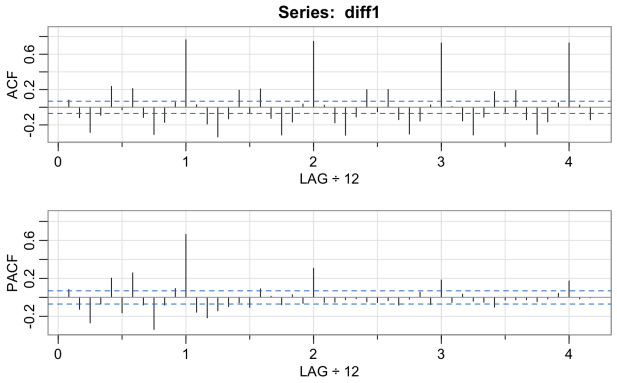

Figure 3: ACF and PACF of First Difference

From figure 2.2, we can see the mean is now roughly constant around zero. The variance also is no longer drifting. Figure 2.3 suggests a seasonal trend every 12 months as we peaks every lag 12 on the ACF. We are able to confirm this hypothesis by getting a month plot of our first differenced data, which tells us we must take a seasonal difference. As we are working with monthly data, which entails having 12 periods in a season, the seasonal difference of Y at period t would be $Y_t - Y_{t-12}$. Thus the first difference of the seasonal difference is equivalent to $(Y_t - Y_{t-1})$ - $(Y_{t-12} - Y_{t-13})$. For our purposes, this refers to the amount by which the change in unemployment rate from the previous month to the contemporary month is different from the change observed one year earlier.

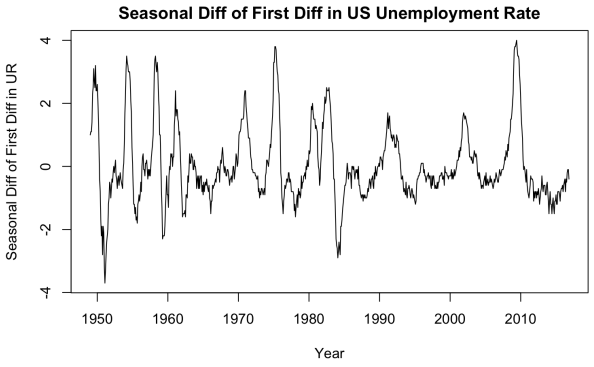

Figure 4: Plot of Seasonally Diff of First Diff

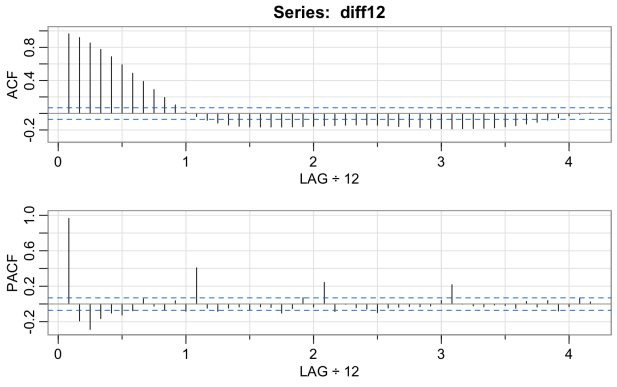

Figure 5: ACF and PACF of Figure 2.4

Figure 2.4 looks like a random walk with roughly constant mean around

zero and not much drifting in variance. Thus, we may assume our data is now stationary. Now we can focus on building our model.

Figure 2.5, suggest the seasonal ACF plot goes to zero after lag 1 and the PACF tails off, indicating P = 0 and Q = 1. The nonseasonal PACF cuts off after lag 3, thus we propose an $ARIMA(3,1,0)\text{x}ARIMA(0,1,1)_{12}$ model. Another alternative model we propose for the sake of variety and comparison is $ARIMA(3,1,3)\text{x}ARIMA(0,1,1)_{12}$, the difference between the former and the latter model is that the latter has a non-seasonal MA(3) while the former has a non-season MA(0).

0.3 Results

0.3.1 Analysis of Proposed Models

We will start by analysing our first proposed model, $ARIMA(3,1,0)\text{x}ARIMA(0,1,1)_{12}$. For the sake of simplicity, we will refer to this model as Model 1.

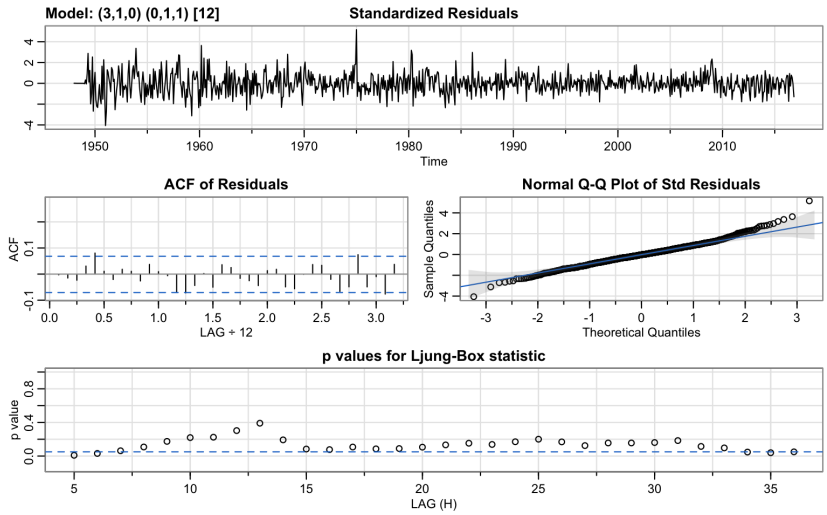

Figure 6: Tests of Model 1

Inspection of the standard residuals show no obvious patterns, however there are at least 2 outliers exceeding 3 standard deviations from the mean. The ACF Residuals plot doesn't show any significant spikes. Hence, we conclude that there are largely no apparent departures from the model randomness assumption. The Normal Q-Q Plot of Residuals also support that there are no apparent departures from the normality assumption aside from

some outliers in the tails, which show some departure from normality. However, most of the p-values for Ljung-Box statistics are at or above the significant level, so we accept the null hypothesis that the residuals are independent. Below, I present other important information regarding Model 1.

Model 1 Fit			
Component	Coefficient Estimate	Standard Error	P-Value
AR(1)	0.1148	0.0351	0.0011
AR(2)	0.2023	0.0345	0.000
AR(3)	0.0900	0.0350	0.0103
Seasonal MA(1)	-0.7674	0.0256	0.000

From the values in the table we can derive that all coefficients are non-zero as all our AR terms are have a p-value less than the significance level of 0.05 and our MA terms has a p-value less than the significance level. I should also highlight that this model had an AIC of -18.08 and a variance of residuals of 0.05582. So our model is thus,
$(1-B)(x_t - .1148_{(.0351)}x_{t-1} - .2023_{(.0345)}x_{t-2} - .0900_{(.0350)}x_{t-3}) = (1-B)(1-B^{12})(w_t - .7674_{(.0256)}w_{t-1})$.

In non-technical language, our model tells us that we can forecast a future unemployment rate value by knowing the past four unemployment rates and multiplying them by some constant depending on the date of the observation. Moreover, it also has some differencing meaning that we will have to take the difference of the values we obtain, and white noise terms, one of which is also multiplied by a constant.

Now we will analyze the second proposed model, $ARIMA(3,1,3)\text{x}ARIMA(0,1,1)_{12}$. Let's call this model, Model 2. Model 2 satisfies the standard residuals test and it only has 1 outliers exceeding 3 standard deviations from the mean. It also satisifies the ACF Residuals plot test and the Normall Q-Q Plot test, hence, we satisfy the randomness and normality assumption. Model 2 also performs better than Model 1 for the Ljung-Box statistic test.Below, I present other important information regarding Model 2.

Model 2 Fit			
Component	Coefficient Estimate	Standard Error	P-Value
AR(1)	-0.4783	0.0448	0.0000
AR(2)	0.0674	0.0576	0.2425
AR(3)	0.8191	0.0446	0.0000
MA(1)	0.6412	0.0562	0.0000
MA(2)	0.1585	0.0726	0.0293
MA(3)	-0.6440	0.0559	0.0000
Seasonal MA(1)	-0.7586	0.0256	0.000

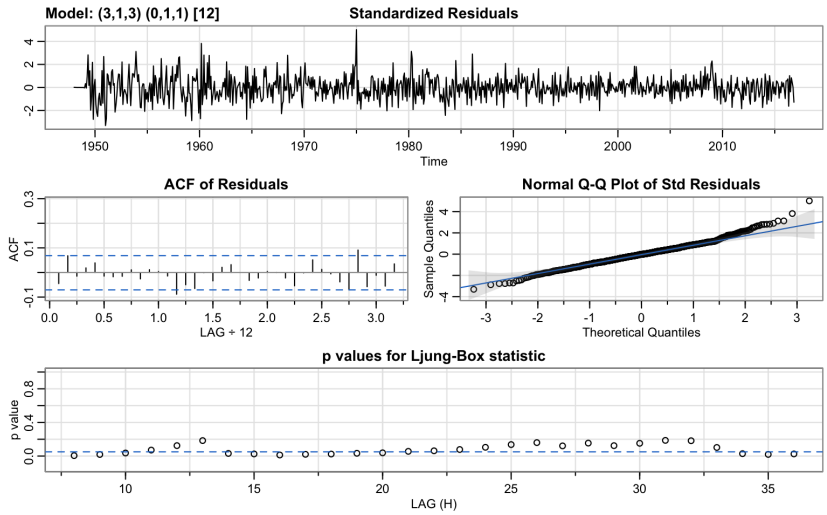

Figure 7: Tests of Model 2

From the values in the table we fail to reject the hypothesis that the AR(2) coefficient is non-zero as their p-values is greater than the significance level of 0.05. The AIC of this model is -20.97 and a variance of residuals of 0.05507.

Using AIC criteria, Model 2 is better fit and some assumptions are better satisfied, however, some components of the model have coefficients with p-values that fail our hypothesis test. Thus, we decide in favour of Model 1. Likewise, we will not discuss or analyze Model 2 anymore.

0.3.2 Forecasting

Now that we have a final model, we can use this model to forecast future unemployment numbers in the US and compare them with actual unemployment figures. Graphically we obtain the following from our model:

The table showed below showcases our predicted values, using a 95 percent confidence interval.

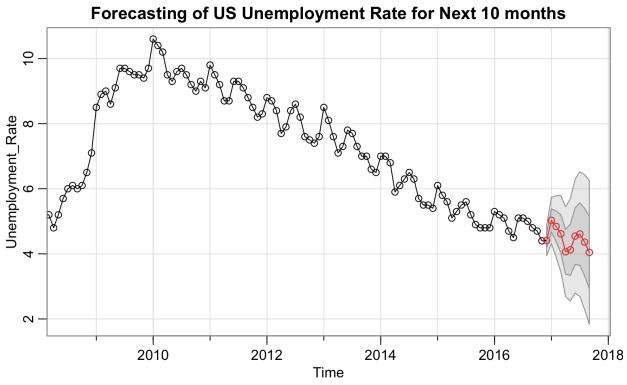

Figure 8: Forecast of US Unemployment Rate

Predicted Values		
Prediction	Lower Bound	Upper Bound
4.412259	3.949198	4.875320
5.027862	4.334392	5.721333
4.842440	3.914901	5.769979
4.616658	3.466651	5.766666
4.063778	2.711130	5.416425
4.125746	2.584995	5.666497
4.543923	5.666497	6.258816
4.609209	6.258816	6.486094
4.362433	2.334006	6.390860
4.043646	1.872779	6.214514

We can see that after the second month, the prediction of next months are within the confidence interval of previous months and more importantly, the standard deviation of the confidence interval increases really fast decreasing the precision of our predictions. Now I will compare the predicted values of my model versus the actual values.

Predicted Values vs Real Values	
Prediction	Real Value
4.412259	4.5
5.027862	5.1
4.842440	4.9
4.616658	4.6
4.063778	4.1
4.125746	4.1
4.543923	4.5
4.609209	4.6
4.362433	4.5
4.043646	4.1

My predicted values are extremely accurate to the extent that only 1 out of 10 is has a difference of more than 0.1 with the real or actual value, which is my second last predicted value. This shows my model was a success as it was able to accomplish the goal it was built for.

0.3.3 Spectral Analysis

Unemployment closely follows the business cycle, hence, it is of importance to execute an spectral analysis and find the three dominant frequencies on our data.

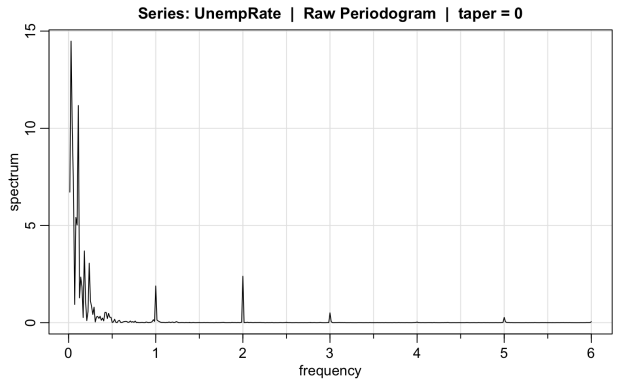

Figure 9: Spectral Estimation of Unemployment Data

From figure 3.4, and with the help of some statistics we can obtain the information below, with a confidence interval of 95 percent.

Spectral Analysis Data				
Dominant Frequency	Period	Spectrum	Lower Bound	Upper Bound
0.0278	36.0000	14.4829	3.926100	572.0440
0.1111	9.0000	11.1720	3.028600	441.2704
0.0417	24.0000	9.9625	2.700685	393.4977

From our spectral analysis data we can see that we cannot establish the significance of the first peak since the periodogram ordinate is 14.4829, which lies in the confidence intervals of the second and third peak. Similarly, we cannot establish the significance of the second peak since the periodogram ordinate is 11.1720, which lies in the confidence interval of the first and third peak and we cannot establish the significance of the third peak since the periodogram ordinate is 9.9625, which lies in the confidence interval of the second peak and fist peak.

0.4 Discussion

ARIMA models are an efficient way of forecasting short-term economic indicators such as unemployment rate, however, we must first acknowledge the weaknesses that this approach has, such as the assumption of Ceteris paribus, meaning that we assume that all the other things that may affect unemployment are held constant to some degree, which is unreasonable as things such as a war or a pandemic may heavily influence unemployment.

Regarding the model, the weaknesses it has are that it is really hard to explain it in a non-technical way, the model itself has significant outliers in its error terms, it lacks precision for long-term forecasting, it assumes Ceteris paribus as explained earlier,and it only works for the US economy, it is not universal. However, as shown by the results we obtained from the model, we can say that even with this limitations, the model was a success as it very accurately and precisely predicted future values.

I believe that more research should be done in the area of exogenous shocks and how they can be predicted or modeled using economic indicators, as that is one of the biggest weaknesses of the majority of economic models, and any breakthrough in this field of knowledge will greatly positively impact the field of economics and econometrics.