Mark Chisholm
Time Series Student Project
Spring 2012

Year-End US Unemployment Rate

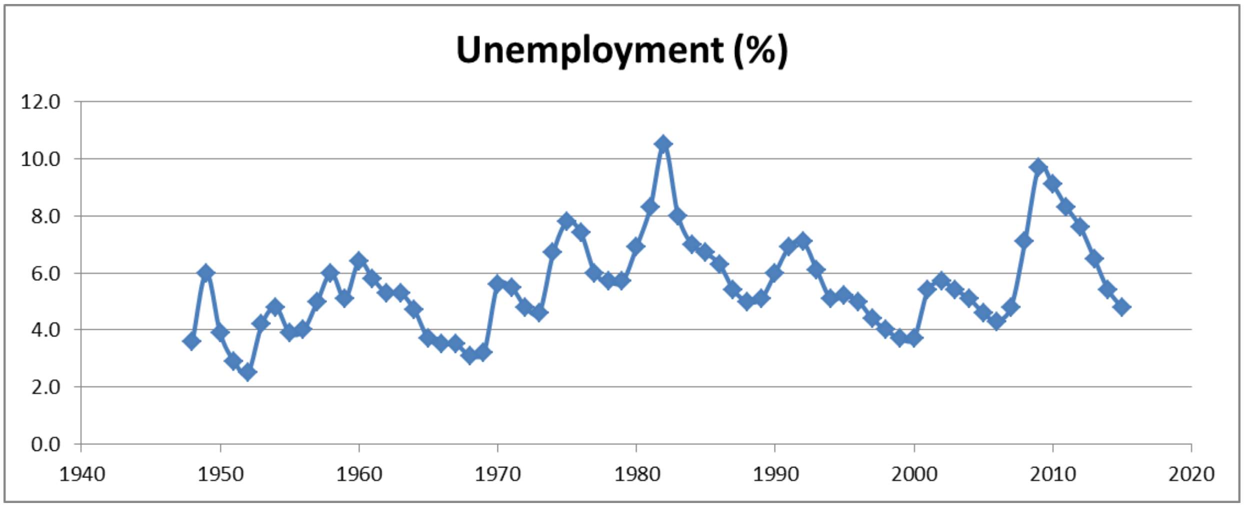
Background & Objective

This project aims to propose a time-series model from the ARIMA family that can describe the year-end unemployment rate in the United States. Year-end unemployment can be influenced by many factors, such as changes in technology, demographics and overall economic activity. While these factors are important to consider, the scope of this analysis is solely to select the simplest ARIMA model that reasonably describes the US year-end unemployment rate.

Data

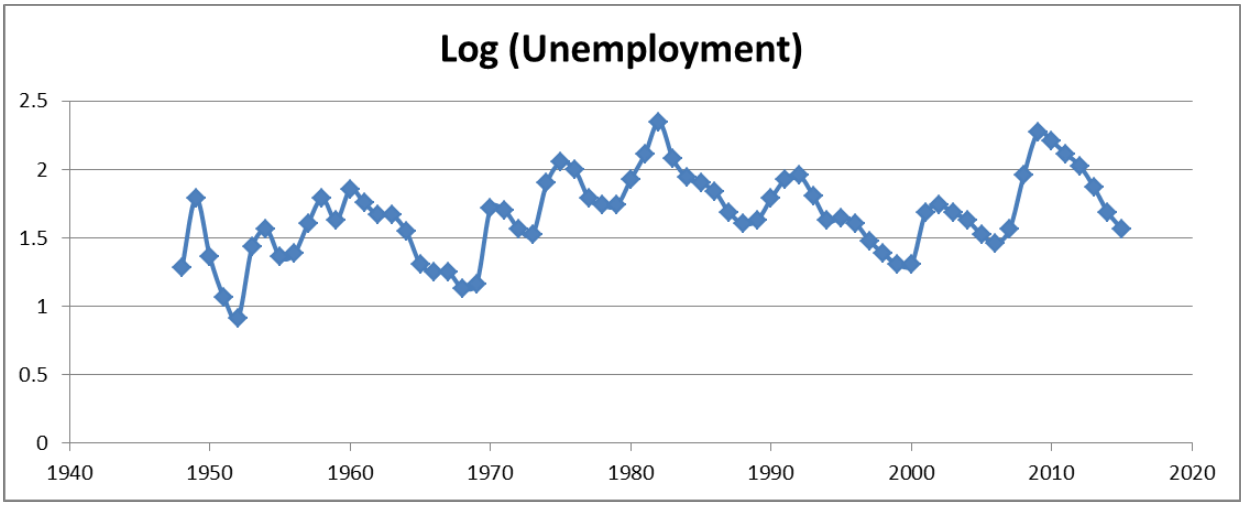
Data for this project was gathered from the Current Population Survey published by US Bureau of Labor Statistics (Series ID LNU04000000). It is not seasonally adjusted, so as to examine the information in as "raw" a state as possible. Data was collected from Year-end 1948, through Year-end 2015, so this series has 68 observations. Figure 1 below shows the unemployment rate over time. Note that the unemployment rate was recorded in whole numbers by the BLS, so a 5% unemployment rate was recorded as 5.0 in the data, not 0.05.

Figure 1 – Unemployment Rate (%)



Natural logarithms of the data series were taken as the unemployment rate is expressed as a percentage of the working population. Note that without taking a log transform, low rates of unemployment will be less dispersed than higher rates, and higher rates of unemployment might seem more extreme relative to lower rates. Figure 2 below shows the unemployment rate after a natural log transform.

Figure 2 – Natural Log Transform of the Unemployment Rate

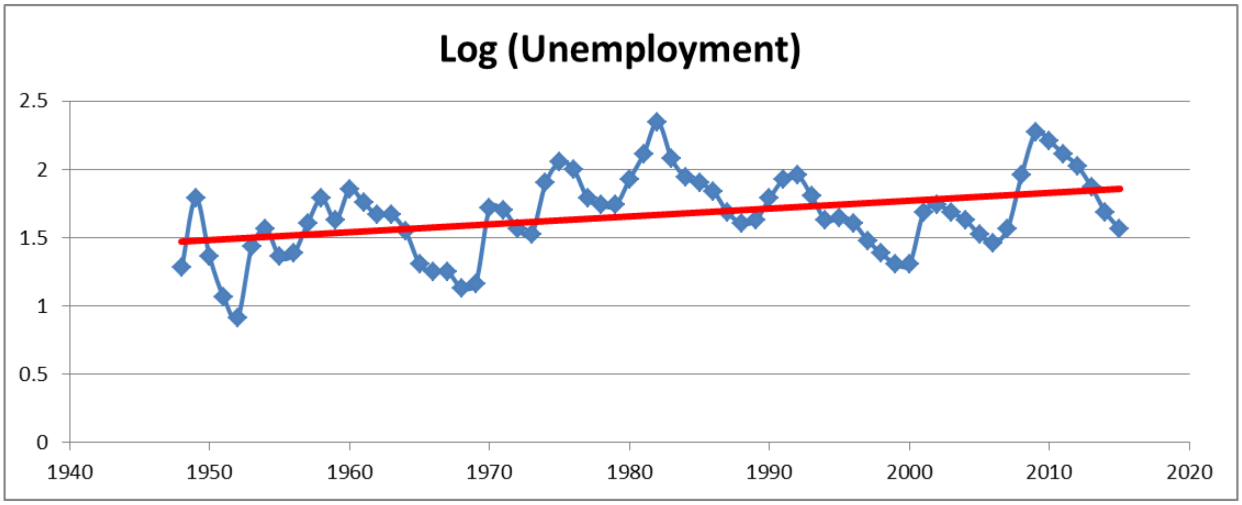


Stationarity

It might be reasonable to expect that over the long-run, after observing many iterations of the economic cycle, the level of employment would be stationary over time around some mean level of employment.

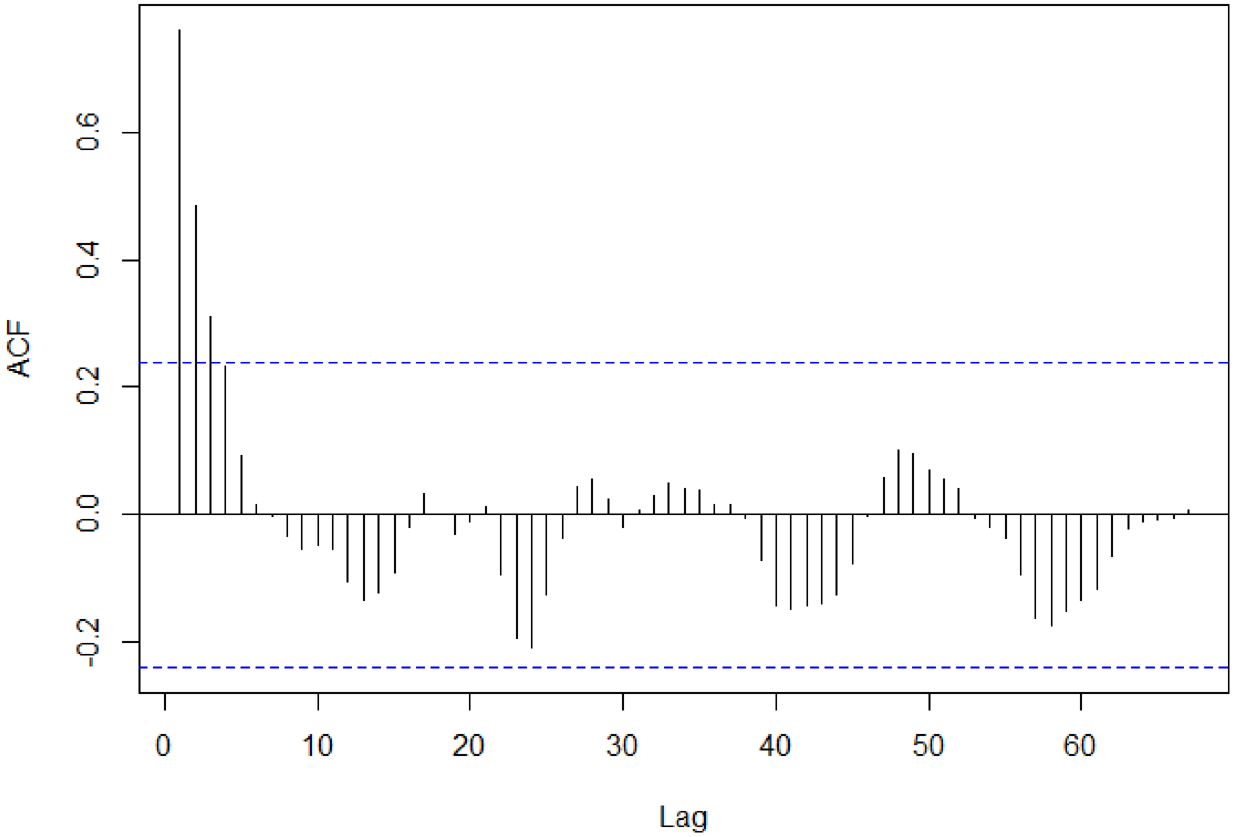
A visual inspection of Figure 3 suggests that the series might not be stationary, which shows the time series with a least squares trend-line superimposed.

Figure 3 – Log Transform of Unemployment With Least Squares Trend Line



While the trend-line in Figure 3 appears to have a positive slope, another way to assess stationarity is to examine a correlogram of the series, shown in Figure 4 below.

Figure 4 – Correlogram of Log-transformed Unemployment Rate



The autocorrelation decays to zero near lags 6 and 7. It then alternates between periods of relatively minor positive and negative autocorrelation among subsequent lags. Given that the autocorrelation is strong for recent lags and is comparatively less strong at older lags, it could be argued that the series is indeed stationary.

The behavior in the ACF in Figure 4 is more consistent with an autoregressive process than a moving average process given the decay. There does not appear to be any abrupt cut-off in autocorrelation after a particular lag, which would be consistent with a MA process. Instead there is a steady decay in the ACF.

The first difference is shown in Figure 5, and Figure 6 provides a correlogram associated with the first difference. Figure 5 appears to show that the first difference is a stationary process, as there is no obvious trend in the series. The ACF in Figure 6 shows a relatively strong negative correlation with lags 2, 3, and relatively strong positive correlation with lag 21. It is not clear if these are spurious given the small sample size. The relatively minor autocorrelations for older lags suggest that the first differences are stationary.

Figure 5 – First Difference of Log Transformed Unemployment Rate

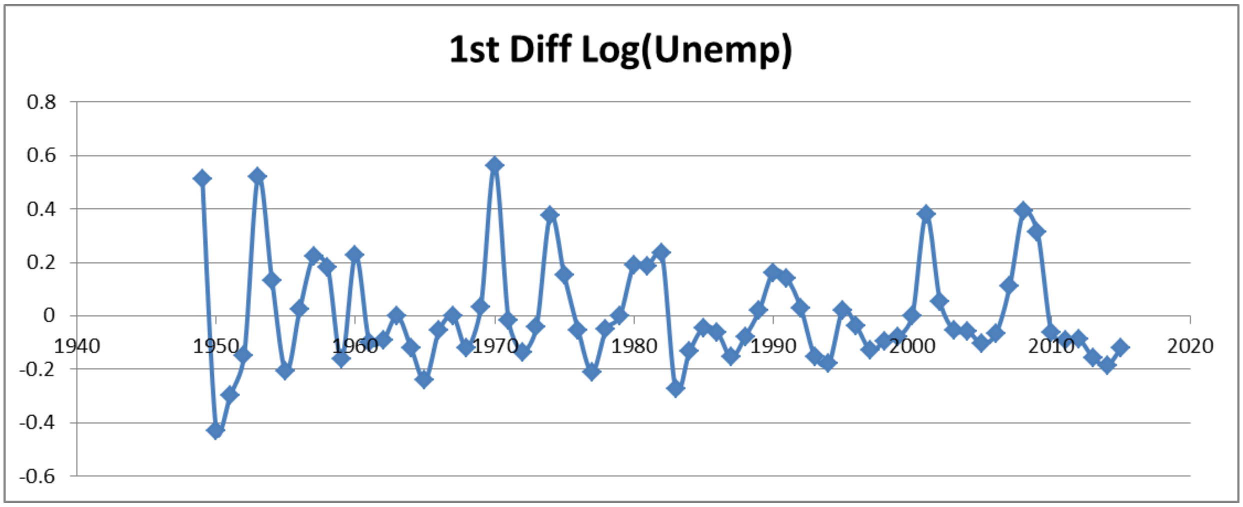
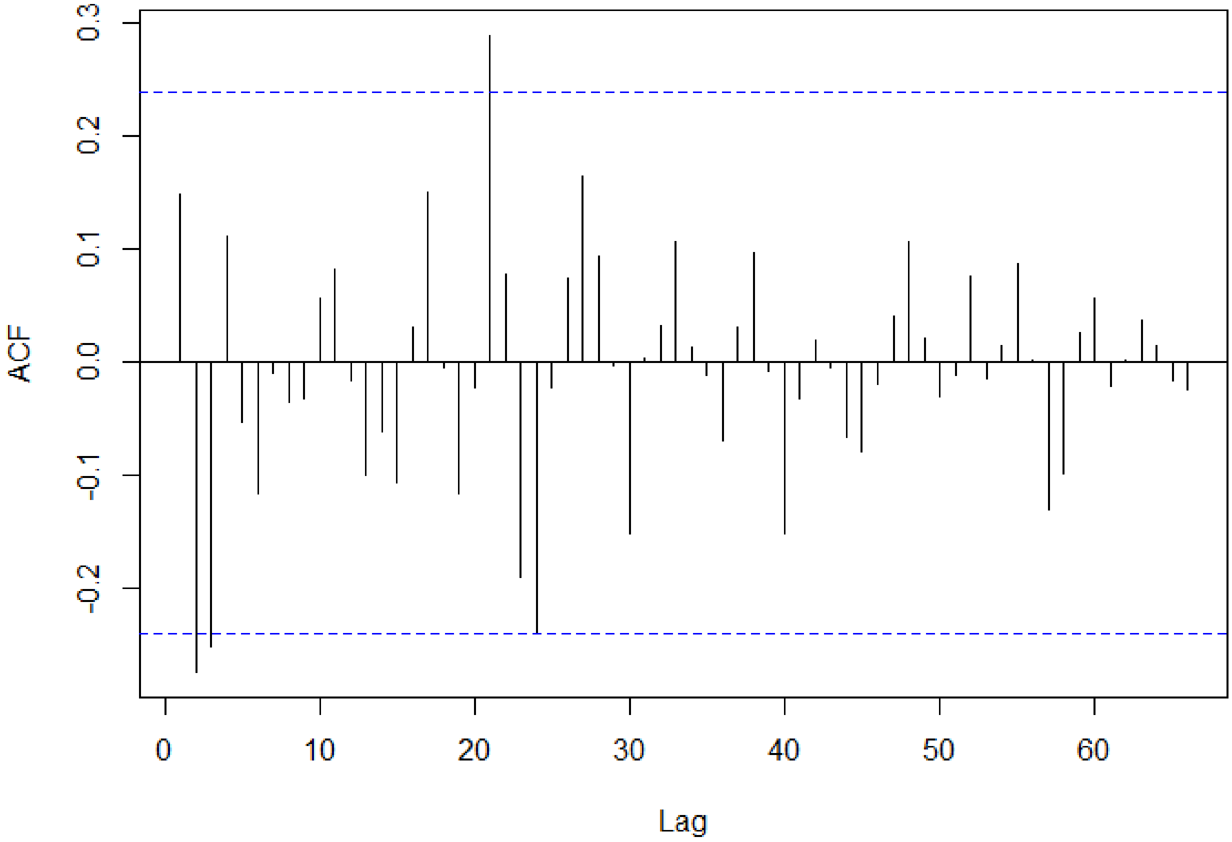


Figure 6 – Correlogram of First Difference of Log Transformed Unemployment Rate



Economic time series can exhibit the behavior of a random walk, with stock prices being a well-known example of this. Therefore it would not be unreasonable to initially propose that a random walk generated the observed series in Figure 2. A random walk has the form:

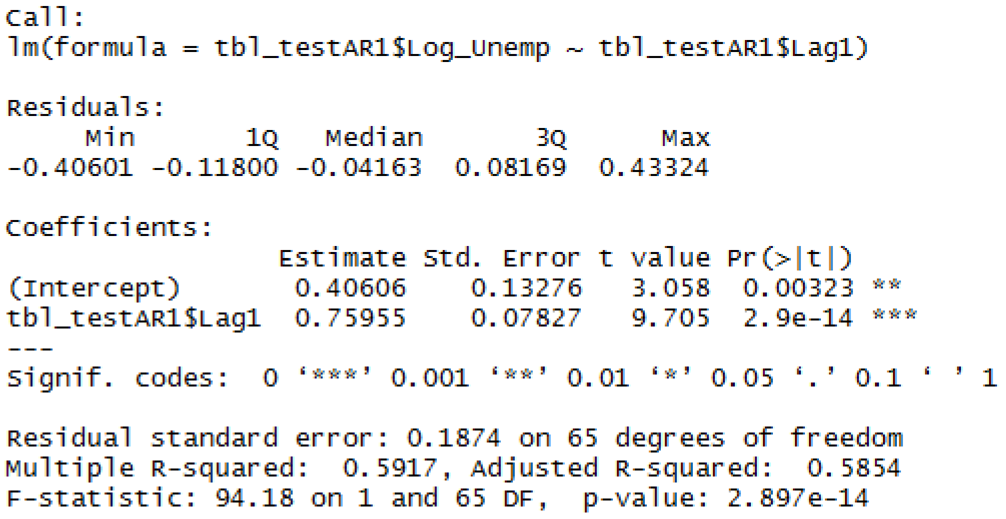
$$Y_t = Y_{t-1} + e_t$$

Linear regression was performed to determine if the log-transformed unemployment rate behaves like a random walk. Figure 7 shows the model results from fitting to an equation of the form:

$$Y_t = \mu + \varphi Y_{t-1} + e_t$$

This is also the form of an autoregressive process of order 1. If the series was a random walk, the tbl_testAR1$Lag1 estimate (which corresponds to the φ parameter above) would be near 1.0. This parameter has an estimate of 0.75955 and a standard error of 0.07827. Adding twice the standard error results in an upper bound of 0.91609, so it is reasonable to conclude that this series does not behave like a random walk. Random walks are not stationary.

Figure 7 – Fitting an AR(1) process to the Log-transformed Unemployment Rate



```
Call:
lm(formula = tbl_testAR1$Log_Unemp ~ tbl_testAR1$Lag1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.40601 -0.11800 -0.04163  0.08169  0.43324 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       0.40606    0.13276   3.058  0.00323 ** 
tbl_testAR1$Lag1  0.75955    0.07827   9.705  2.9e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1874 on 65 degrees of freedom
Multiple R-squared:  0.5917, Adjusted R-squared:  0.5854 
F-statistic: 94.18 on 1 and 65 DF,  p-value: 2.897e-14
```

It seems reasonable to conclude that the log-transformed unemployment rates are stationary on the basis that:

- The autocorrelation function does not show any significant autocorrelation beyond recent lags.
- The process is not entirely consistent with a random walk, which would make it nonstationary.
- The expectation that unemployment rates do not continually rise or fall in the long-run.

Model Specification

The purpose of this analysis is to determine the simplest time-series model that reasonably explains the variation in this process, and not to select a model that is any simpler. No differencing will be performed as it is not unreasonable to expect the log-transformed unemployment rates to be stationary.

As discussed in the previous section, the ACF of this series appears to be consistent with an autoregressive process. Cryer and Chan state that the partial autocorrelation function shows the correlation between the series at time (T) and time (T-k), after removing the effect of intervening variables at times (T-1), (T-2), …, (T-k+1). The PACF can help determine the order of an autoregressive process, and for the log-transformed unemployment rate is shown in Figure 8 below. It suggests that an AR(1) process is appropriate.

Figure 8 – PACF of Log-Transformed Unemployment Rates

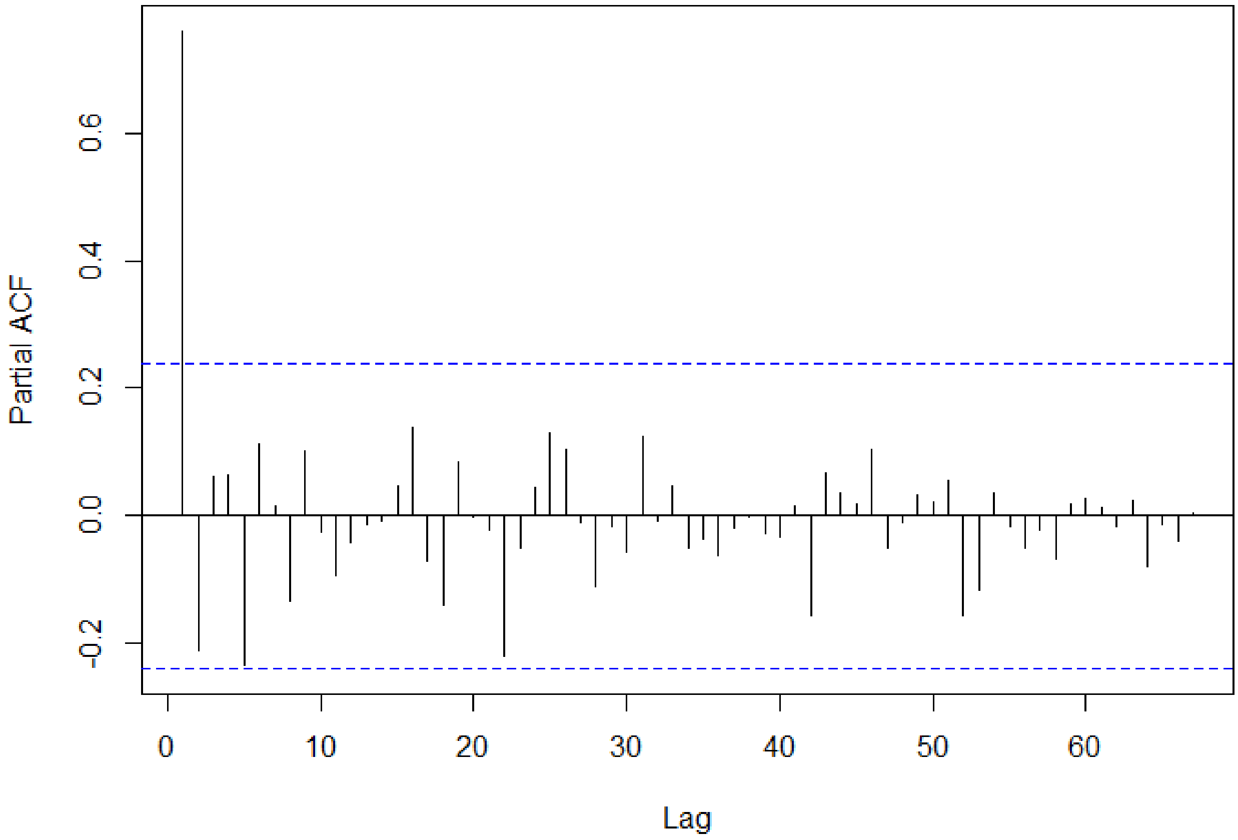


Figure 7 already shows the fit of an AR(1) model to the log-transformed unemployment rate. The fit is given by this formula:

$$Y_t = 0.40606 + 0.75955\, Y_{t-1} + e_t$$

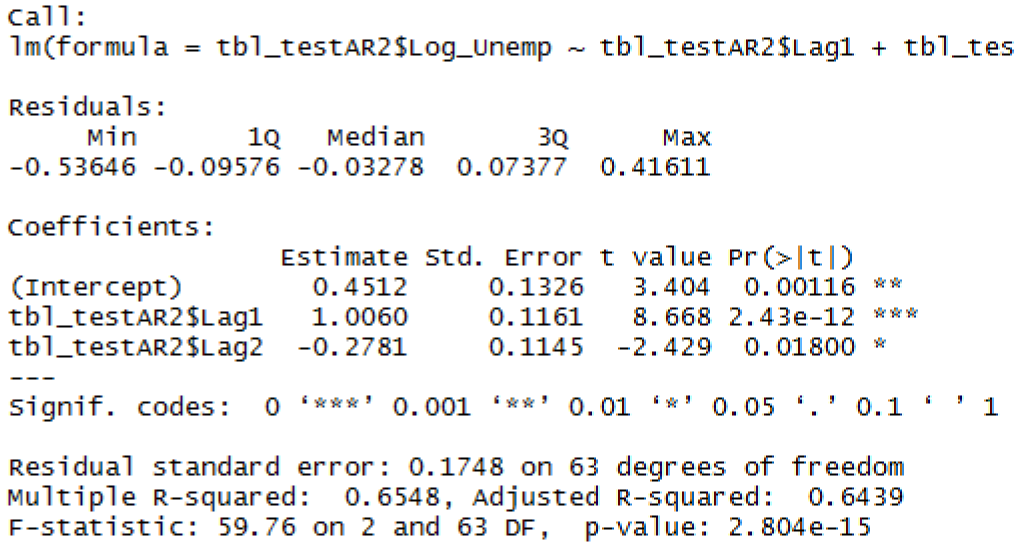
Note that the intercept is 2 standard errors greater than zero, and so it is kept in the model.

For comparison, an AR(2) model may also be considered. The results of such a model are in Figure 9, and are specified by this equation.

$$Y_t = 0.4512 + 1.0060\, Y_{t-1} - 0.2781\, Y_{t-2} + e_t$$

All of the parameter estimates in the AR(2) fit are 2 standard errors away from zero, and as such these parameters are included in the AR(2) model. However, it is worth keeping in mind that the standard errors of the AR(2) parameters may be understated due to correlation among the lags. This correlation does not affect the estimates, only their estimated standard error.

Figure 9 – Fitting an AR(2) Model to Log-Transformed Unemployment Rates



```
Call:
lm(formula = tbl_testAR2$Log_Unemp ~ tbl_testAR2$Lag1 + tbl_testAR2$Lag2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.53646 -0.09576 -0.03278  0.07377  0.41611 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        0.4512     0.1326   3.404  0.00116 ** 
tbl_testAR2$Lag1   1.0060     0.1161   8.668 2.43e-12 ***
tbl_testAR2$Lag2  -0.2781     0.1145  -2.429  0.01800 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1748 on 63 degrees of freedom
Multiple R-squared:  0.6548, Adjusted R-squared:  0.6439 
F-statistic: 59.76 on 2 and 63 DF,  p-value: 2.804e-15
```

Model Assessment

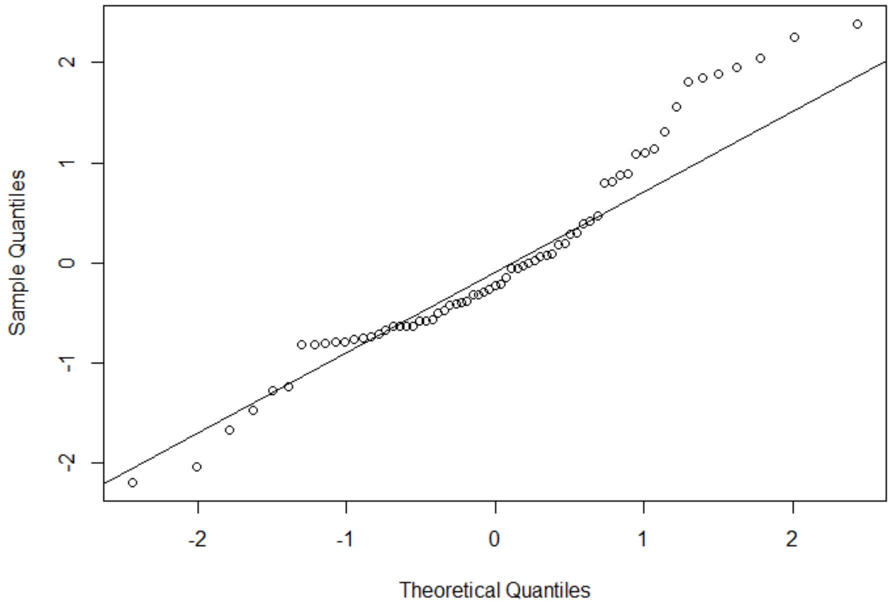
When interpreting the model diagnostics, it is important to consider the small sample size (68 observations in the data set). Less observations makes it difficult to interpret whether a diagnostic is influenced by random fluctuations or is indicative of a poor model choice.

Figures 10-13 show the QQ-Plot, histogram of residuals, plot of residuals, and autocorrelation function of the residuals for the AR(1) fit.

Figures 14-17 show these plots for the AR(2) fit.

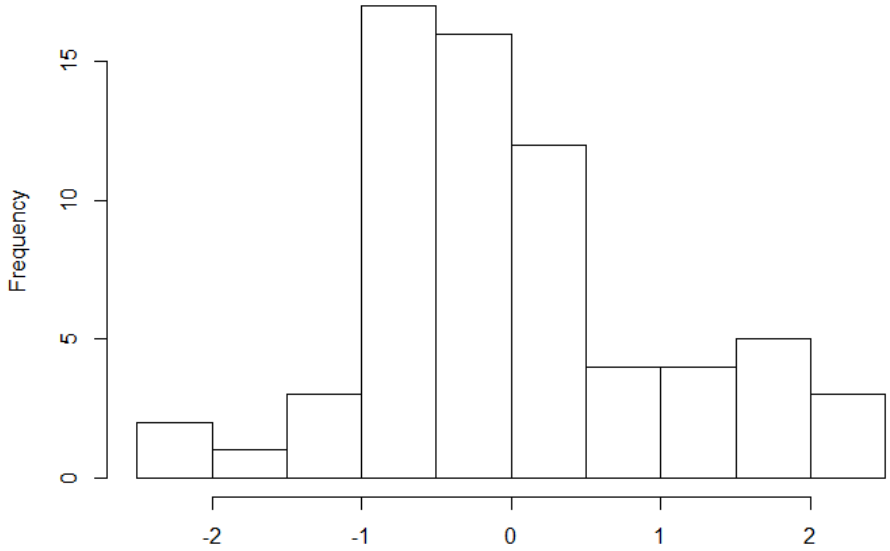
Interpretation is provided below each plot.

Figure 10 – QQ-Plot of AR(1) Fit



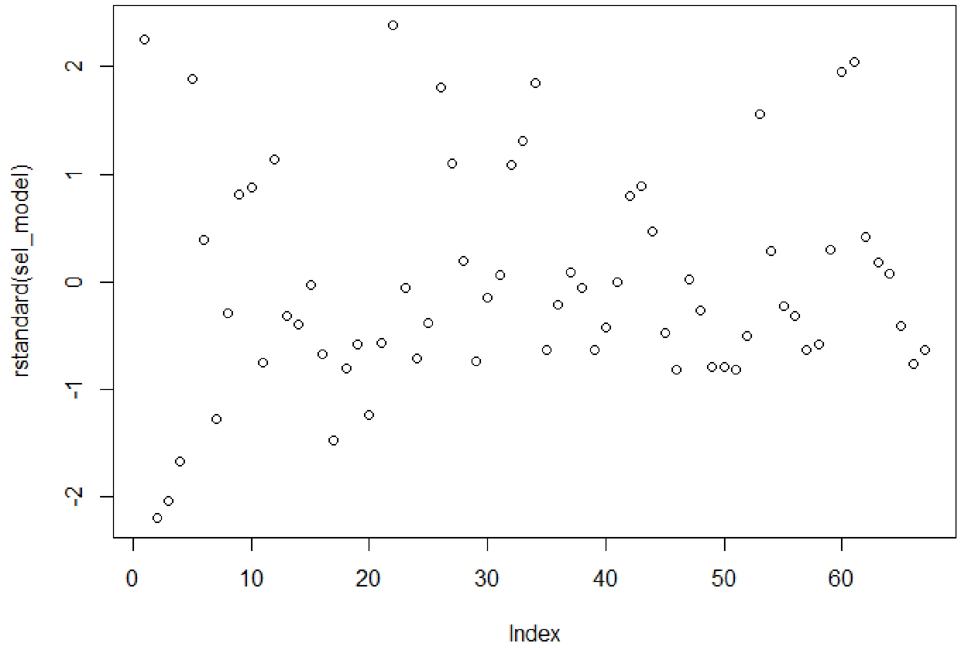
The QQ-plot does not display severe non-normality, although the some of the residuals might give the distribution a possible right-tail. The plot is not ideal, although it appears that the majority of points are near the reference line.

Figure 11 – Histogram of Residuals from AR(1) Fit



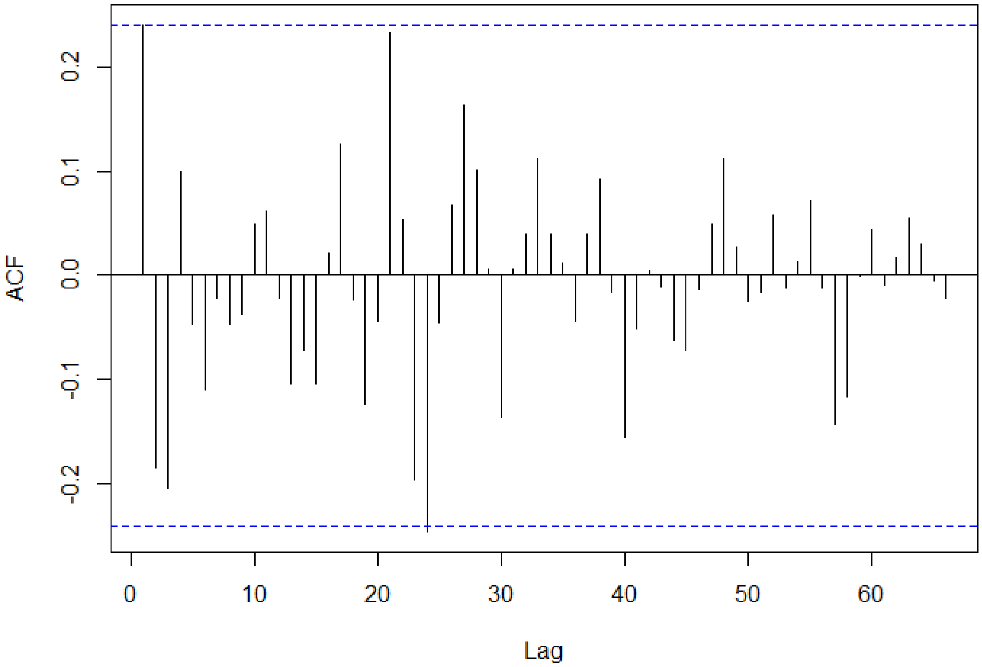
The histogram is composed of few data points – it is difficult to state whether the residuals of the AR(1) fit are significantly non-normal.

Figure 12 – Residuals from AR(1) Fit



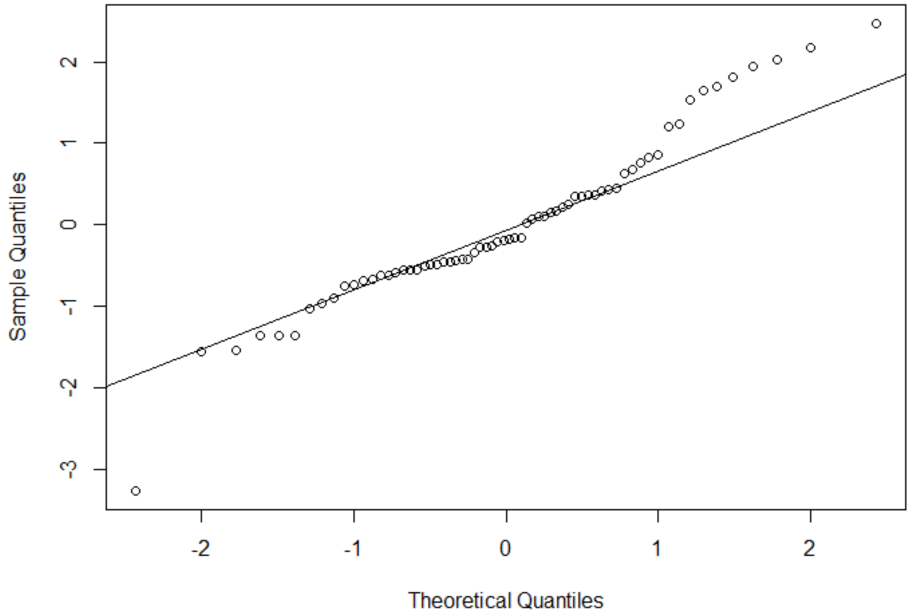
The residual plot indicates that they are randomly arranged. Some of the residuals on the lower end of the horizontal-axis appear to be more extreme than the residuals on the higher end of the horizontal-axis. There are too few data points to indicate if this is evidence of heteroscedasticity.

Figure 13 – ACF of AR(1) Residuals



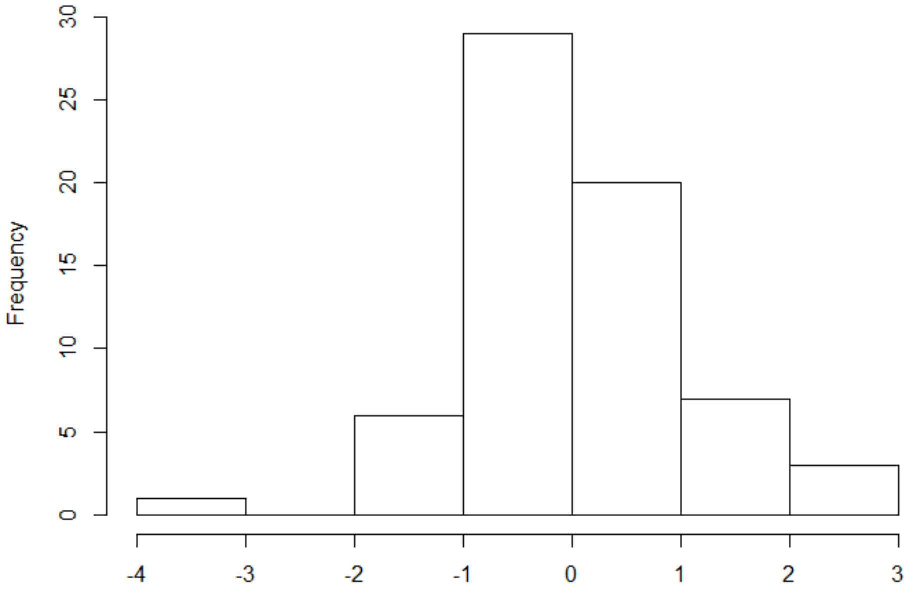
The ACF of the residuals do not show any strong autocorrelation at previous lags, although lags 1 and 24 do appear more prominent.

Figure 14 – QQ-Plot of AR(2) Fit



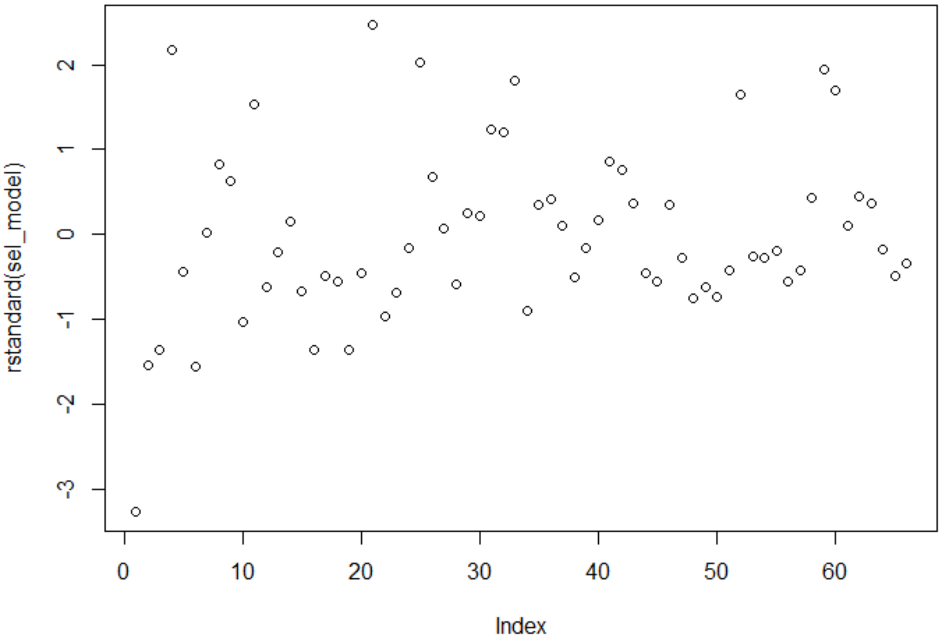
The QQ-Plot suggest that a normal distribution is reasonable, despite the distribution of the residuals possibly being right-tailed. The majority of points appear to follow the reference line.

Figure 15 – Histogram of Residuals from AR(2) Fit



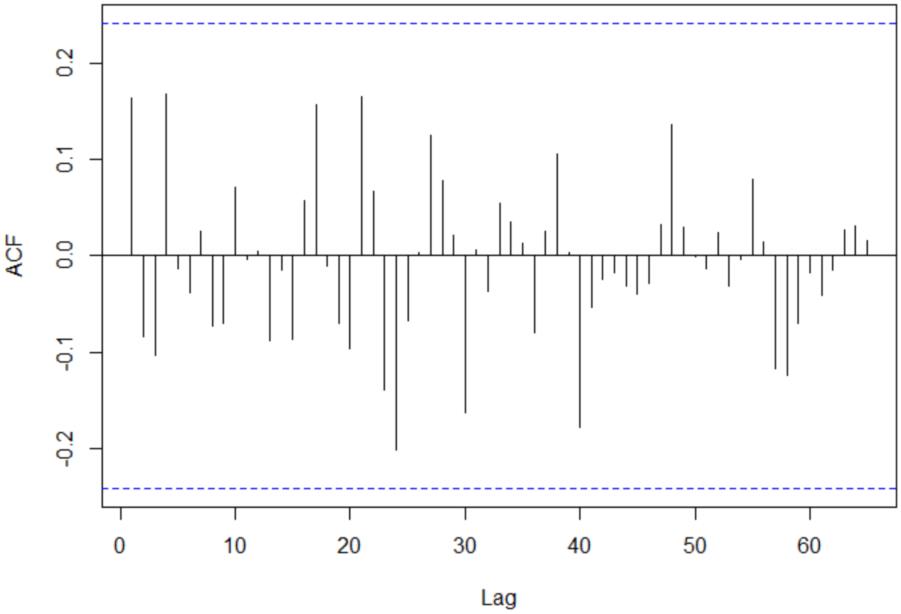
The histogram seems reasonably bell-shaped given the small sample size. One of the points (also seen on the QQ-Plot in Figure 14) appears to be an extreme negative outlier.

Figure 16 – Residuals from AR(2) Fit



The residuals appear to be somewhat random, although it is possible there is heteroscedasticity is the AR(2) residuals.

Figure 17 – ACF of AR(2) Residuals



The ACF of the AR(2) residuals shows them to be relatively uncorrelated.

Plots of the actual and predicted unemployment rates (not log-transformed) are shown in Figures 18 and 19 for the AR(1) and AR(2) fits, respectively. The fitted values are the one-step ahead forecasts.

Figure 18 – AR(1): Actual Unemployment (%) vs Predicted Unemployment (%)

Actual unemployment is the blue line, predicted is the red line

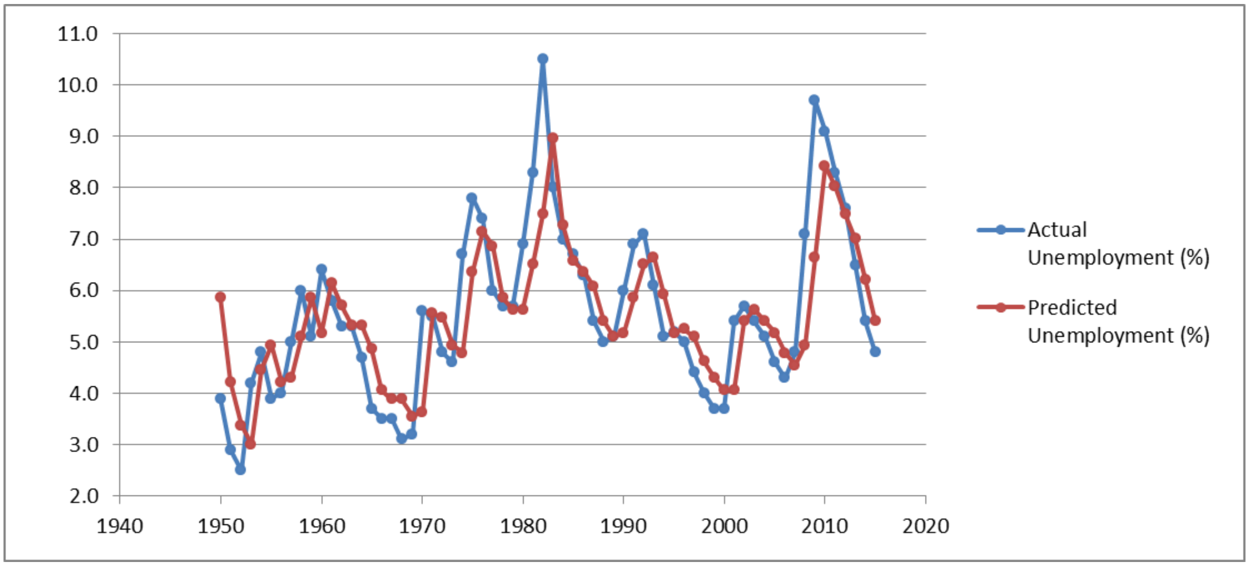
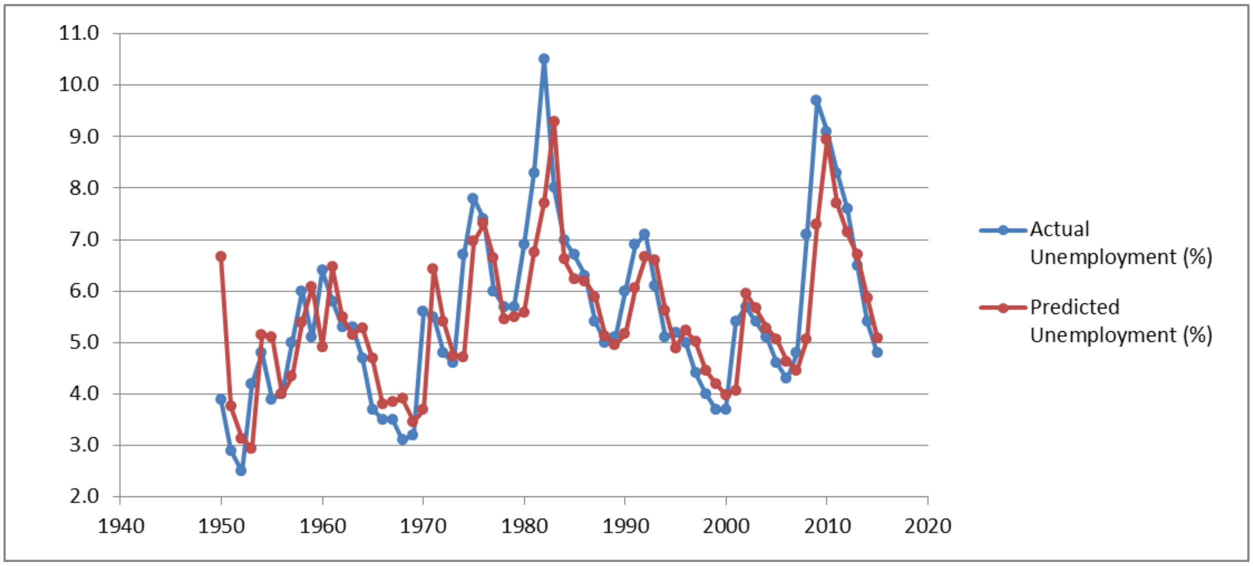


Figure 19 – AR(2): Actual Unemployment (%) vs Predicted Unemployment (%)

Actual unemployment is the blue line, predicted is the red line



To facilitate comparison, the axes in Figures 18 and 19 are on the same scale and show the same years. At first glance, these models seem to perform similarly.

Another check of the models is to simulate the AR(1) and AR(2) processes that were fitted to describe the log-transformed unemployment rates, and then see if the sample ACF and PACF of the simulated series seem consistent with the observed ACF and PACF. These simulated data sets had 68 observations, which is the number of observations in the unemployment data.

The AR(1) model was simulated by seeding an initial value at the first time period. This was taken from the first log-transformed observation in the unemployment data. This seed was needed because the AR process requires a previous observation in order to estimate the next. After seeding the first value, the fitted model was used to derive the subsequent data points. The error terms at each point are Normal with a mean of 0 and standard deviation of 0.1874 (taken from the residual standard error in Figure 7).

A similar procedure was followed to simulate the AR(2) process, except the first two values were seeded from the first two observations as this AR process requires the two most recent observations to predict the next. Also, the standard deviation of the error terms is 0.1748 (taken from the residual standard error in Figure 9).

Figures 20 and 21 show the sample ACF and PACF of the simulated AR(1) process. Figures 22 and 23 show the sample ACF and PACF of the simulated AR(2) process.

Figure 20 – ACF of Simulated AR(1)

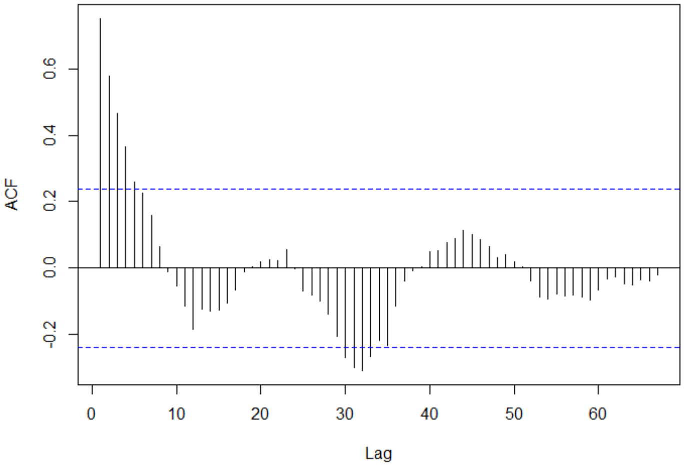


Figure 21 – PACF of Simulated AR(1)

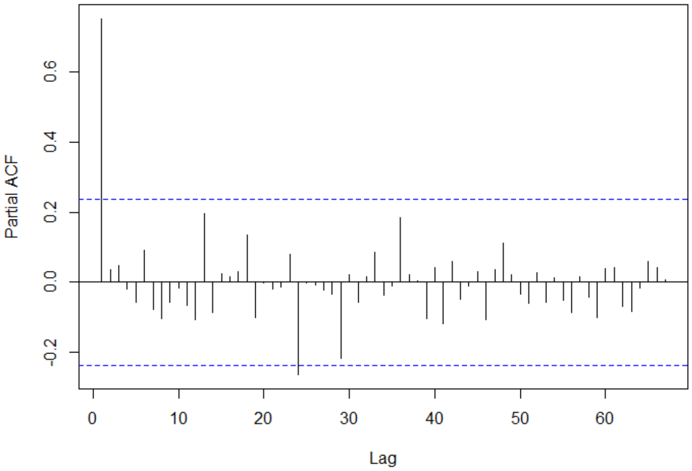


Figure 22 – ACF of Simulated AR(2)

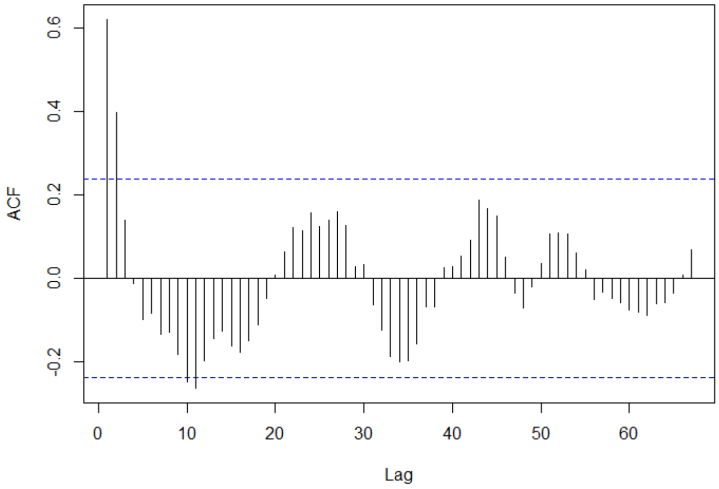
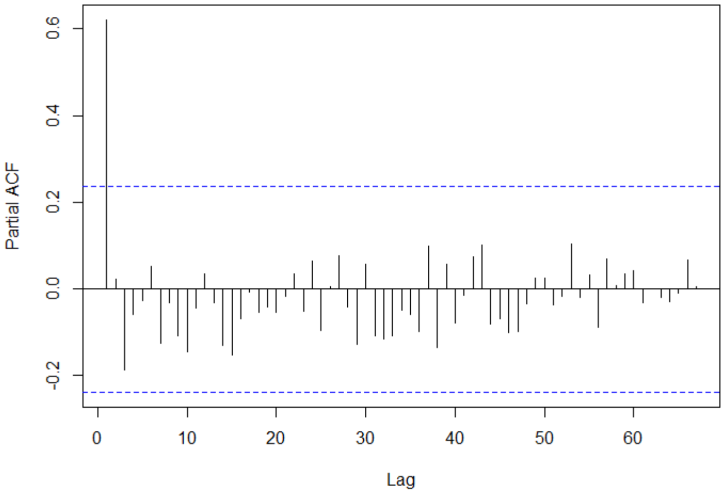


Figure 23 – PACF of Simulated AR(2)



Interpretation of the ACFs and PACFs is difficult because of the small number of data points in the sample. For example, while Figures 20 and 21 represent the ACF and PACF of one simulation, another simulation produced an ACF and PACF that looked very similar to AR(2)-generated plots in Figures 22 and 23. Also, different iterations of the AR(2) simulation produced ACF and PACF plots which looked like the AR(1)-generated ACF and PACF plots in Figures 20 and 21.

Note that the AR(2) PACF in Figure 23 would apparently suggest an AR(1) process as there is no prominent partial autocorrelation at lag 2. A subsequent simulation showed a negative partial autocorrelation at lag 2, and another showed a positive partial autocorrelation at lag 2.

These simulations demonstrate that an AR(1) or AR(2) with 68 observations could have generated the sample ACF and PACF plots in Figures 4 & 8.

Proposed Model and Commentary

I propose an AR(1) model of year-end unemployment rates (*U*) at time *t* as follows:

$$E[\log(U_t)] = 0.40606 + 0.75955 \log(U_{t-1})$$

The adjusted R-squared for the AR(2) model (shown in Figure 9 as 0.6439) is higher than the adjusted R-squared for the AR(1) model (shown in Figure 7 as 0.5854). This suggests an AR(2) model is a better fit, even after the addition of a parameter.

However, the purpose is to produce the simplest model that explains the variation in the observed data. It is plausible that the fitted AR(1) and AR(2) models could have generated the 68 observed unemployment rates, and corresponding ACF and PACF plots. The AR(1) is the simpler model that explains the observations.

Some possible improvements to this model include:

- Examine unemployment using monthly data instead of year-end data. The selected AR(1) model uses the December unemployment rate in the most recent year to predict the December unemployment rate in the upcoming year. If an event (or events) were to occur between these two dates that would result in a substantially different prediction of unemployment, then a model based on monthly data could reflect this sooner.
- Incorporate other variables in the model that could be related to unemployment, such as demographic, political or economic data.

On a final note, one interesting observation about the selected AR(1) model is a possible Bayesian interpretation. In the selected model, the expected log-transformed unemployment rate (*Y*) at time *t* is given by:

$$E(Y_t) = 0.40606 + 0.75955\, Y_{t-1}$$

If the latest observation at *t*-1 has credibility *z*, then this could be written as:

$$E(Y_t) = (1 - z)\tau + z\, Y_{t-1}$$

Under this interpretation, the latest observation is assigned a credibility-weight of 76%, and the prior expectation is assigned 24% credibility. Then the prior expectation of log-transformed unemployment rates is τ = 1.7083. Exponentiating this results in a value of approximately 5.5, suggesting that our prior expectation of long-run unemployment is 5.5% (as stated in the Data section, the BLS records unemployment rates as whole numbers). This matches the simple average of all the unemployment rates in the series, which was also 5.5%. It would be interesting to investigate whether all stationary AR(1) processes with φ greater than 0 but less than 1 can be interpreted in terms of a credibility parameter *z* and a prior mean parameter τ.