ANALYSIS OF WISCONSIN EMPLOYMENT JAN.1961 – OCT. 1975

ANDREW ZHANG
PSTAT 274
Thursday 10AM

Abstract

The objective of this project was to identify the best autoregressive model for the Wisconsin Employment Time Series. This involved utilized Box-Jenkins methodology, which includes detecting stationarity, detecting seasonality, differencing, identifying p,q for an ARMA model, and forecasting. After following these steps, we can conclude that an ARMA(0,2,1) is the best model for the data. Interestingly, there were 4 models that were possible fits as they all passed stationarity and seasonality tests, however, we used Occam's Razor to determine that the best model was the simplest one.

Introduction

Employment has always been a solid indicator of economic growth and development in various parts of the world. The following data set displays Wisconsin's employment from the years January 1961 to October 1975.

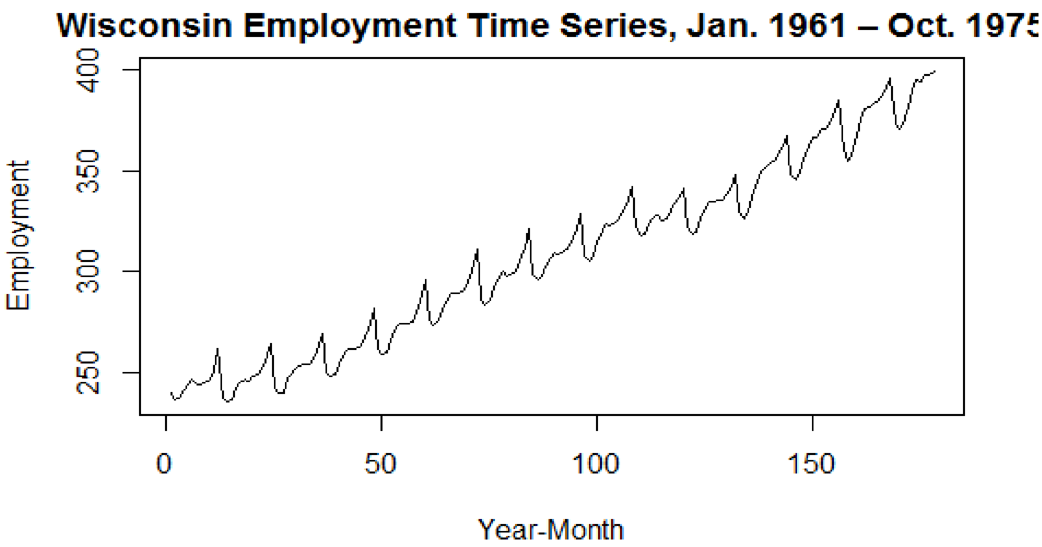

Figure 1

The purpose of working with this data set is to forecast the potential growth of the employment curve and predict what the best model may be. Given how the data seems seasonal and trended, we would predict that the data continues to rise and follow its past patterns. The plan is to forecast by first checking whether the data has a trend and/or is seasonal. Following that, we will test for differencing by checking the variances of the ACF and PACF while examining the plots to see if any significant spikes remain. Finally, we will use confidence intervals to determine the potential growth of the graph.

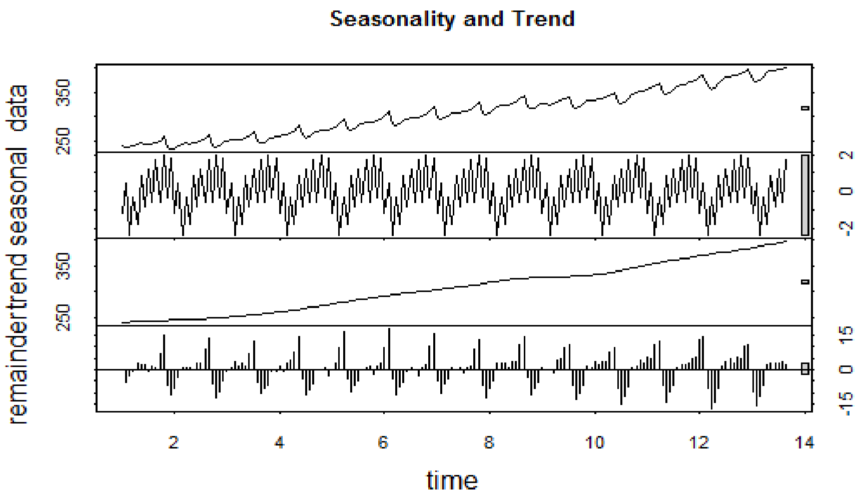

Figure 2

Figure 2 displays both trend and seasonality on a simplified graph. According to the decomposition, it seems like there is strong seasonality over the 15 years the data was collected. The trend line also indicates that there is a strong upward trend within the data. Due to this trend, we know that the original data is not stationary. Since we have confirmed both a seasonal component and a trend component, we can go ahead and apply a transformation.

Transformations

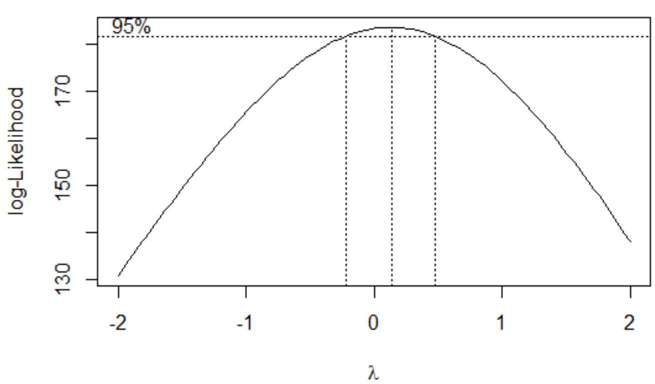

Figure 3

Figure 3 indicates the range for λ. In this case, λ = 0.1414141, but the interval indicates that λ=0 is a possibility, therefore, we cannot rule out a log-transformation.

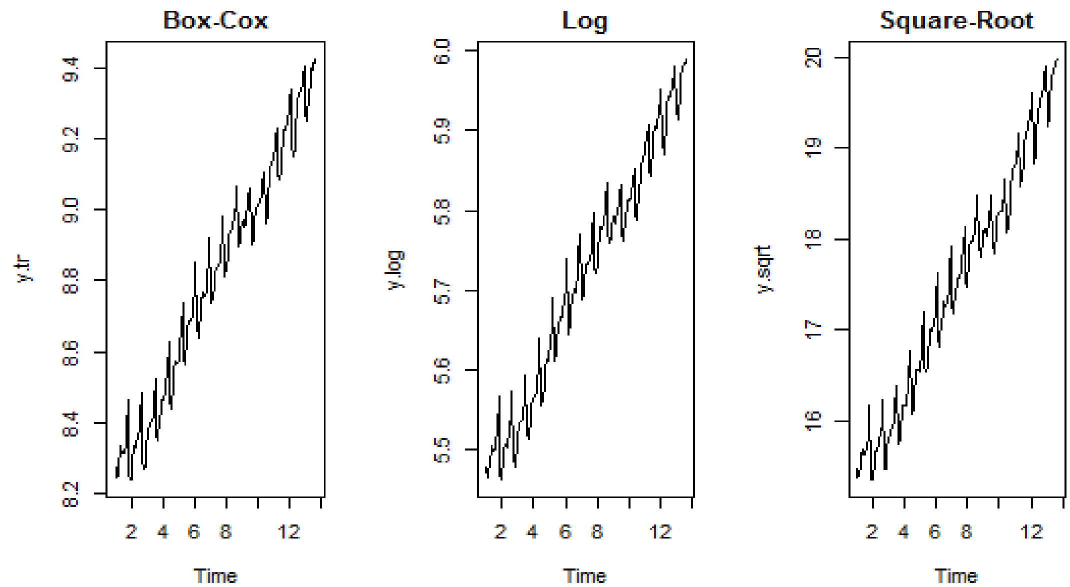

Figure 4

Figure 4 juxtaposes 3 different transformations done on the original data; Box-Cox Transformation, Log Transformation, Square-Root Transformation. As we can see, even after applying 3 different transformations, none of the models are stationary. Therefore, we can begin differencing to select an appropriate model.

Differencing/ACF and PACF

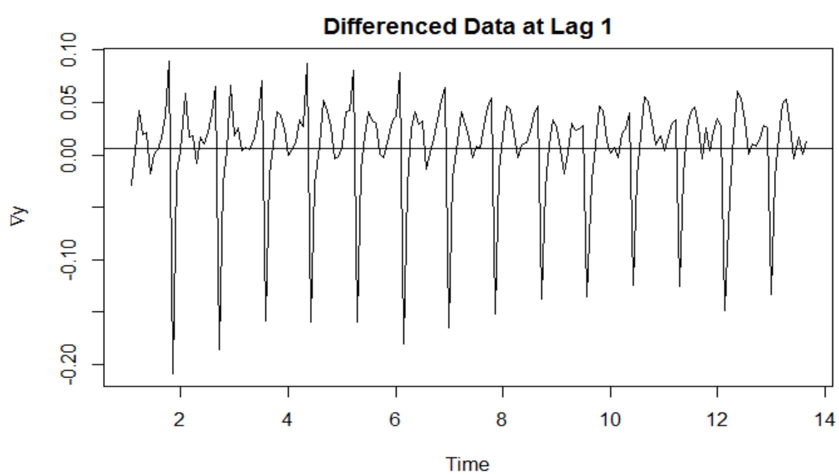

Figure 5

After applying the Box-Cox Transformation and still seeing that the model is not stationary, we apply a differencing at lag 1. Figure 5 depicts the result of the differencing. Although the model looks more stationary, there is still a significant seasonal component that seems to be affecting the variance, which is measured at 0.002850806.

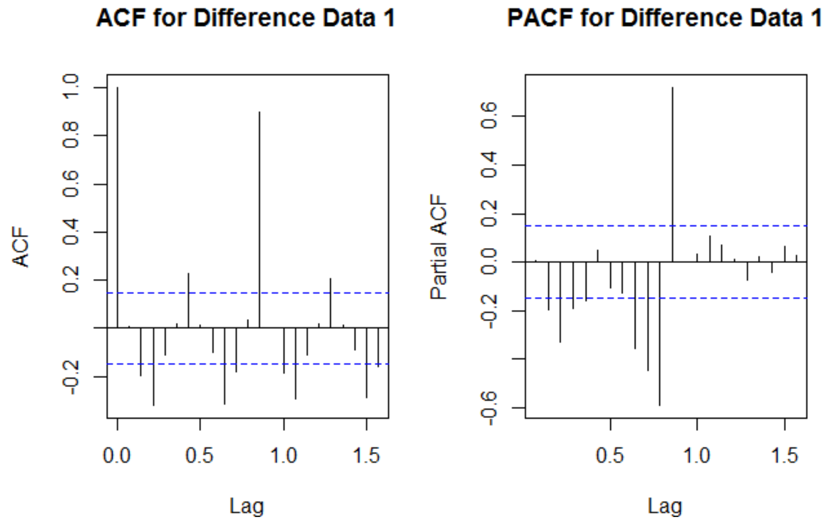

Figure 6

The ACF and PACF of the data differenced at lag 1 support the fact that the model is not completely stationary yet. According to Figure 6, we can continue differencing at lag 12 to see if not only the model becomes more stationary, but also if the variance goes down.

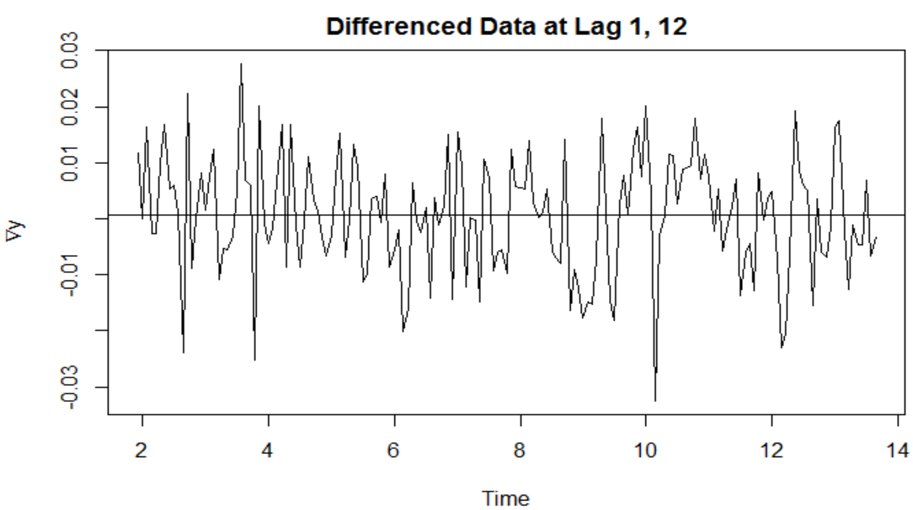

Figure 7

Figure 7 depicts the data after a second differencing at lag 12. The line across the model indicates the mean of the data, which if stationary should be at 0. It seems that the data is more stationary while keeping the variance smaller. In fact, the variance for this differenced model is 0.0001117976 which is less than the variance of the model differenced at lag 1.

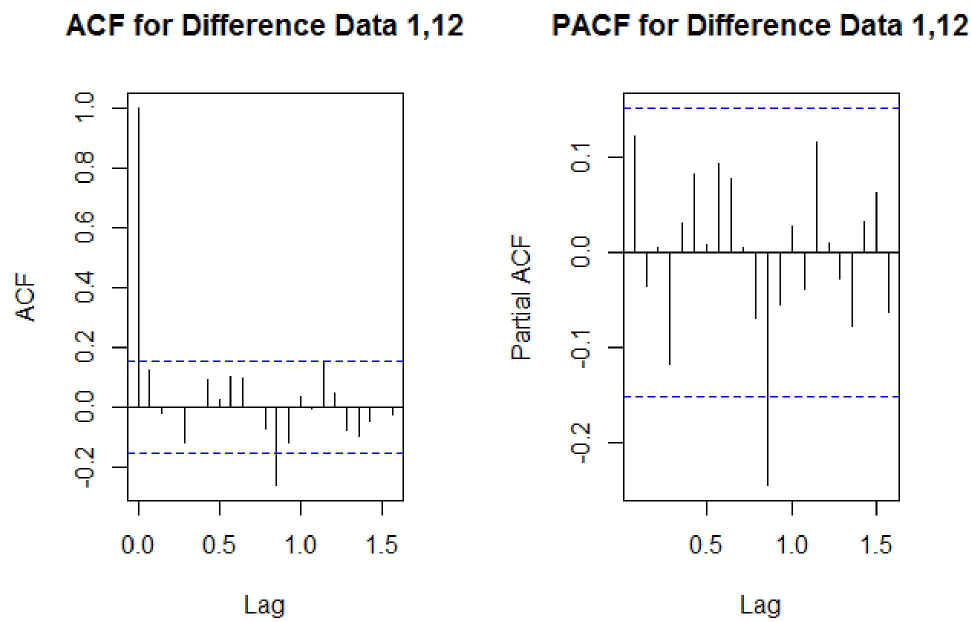

Figure 8

The ACF and PACF of the model differenced at lag 1 and lag 12 shows significant improvement compared to the ACF and PACF of the model only differenced at lag 1. The only concern here is that at lag 12, even after differencing, a spike remains in both the ACF and PACF plot. We can check this by differencing the model again at lag 12.

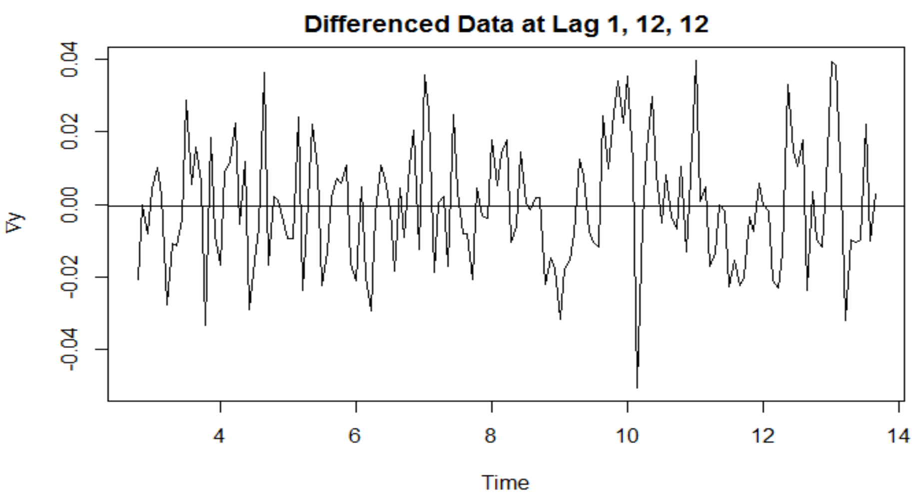

Figure 9

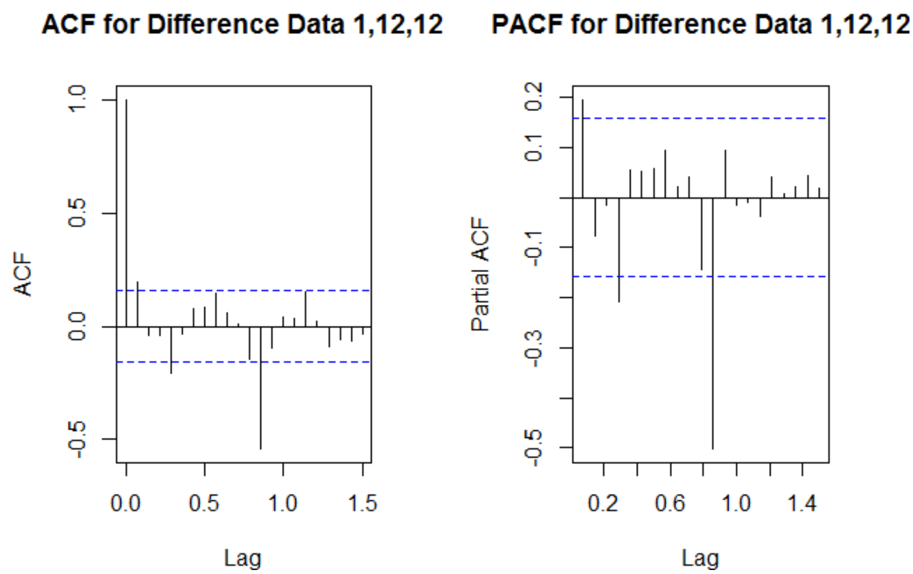

Figure 10

Figures 9 and 10 above depict the model and the ACF/PACF of the model differenced at lag 1, 12, and again at 12. As we can see, by comparing the model differenced at lag 1 and 12 to the model above, it seems that the variance increases slightly, while the ACF and PACF have spikes at lag 5 and 12. The variance for this model is 0.0002854621, which is greater than is 0.0001117976. Therefore, we can conclude that this is a case of overdifferencing and speculate that the model differenced at lags 1 and 12 is the most suitable.

Model Fitting

To determine which p,ql we should select, we can run a function that displays a range of AICcs for various ARMA functions. We should select a couple of ARMA models with the smallest AICc values in order to test whether they are appropriate for the data set.

p/q	0	1	2	3	4	5	6	7
0	-1030.034	-1030.692	-1028.759	-1026.770	-1026.515	-1024.368	-1024.444	-1022.707
1	-1030.528	-1029.012	-1027.000	-1025.021	-1024.364	-1022.179	-1023.064	-1020.826
2	-1028.663	-1027.014	-1024.808	-1022.726	-1024.088	-1021.922	-1020.827	-1018.562
3	-1026.570	-1024.962	-1022.781	-1032.388	-1021.945	-1019.653	-1023.611	-1016.263
4	-1026.819	-1024.739	-1023.805	-1021.954	-1032.844	-1026.867	-1024.604	-1023.570
5	-1024.842	-1023.686	-1021.670	-1019.371	-1030.583	-1025.659	-1019.857	-1021.201
6	-1023.768	-1021.548	-1019.947	-1023.660	-1025.446	-1023.490	-1022.776	-1019.602
7	-1021.571	-1019.321	-1017.690	-1019.877	-1023.738	-1021.675	-1019.854	-1018.683

This chart reveals multiple AICc values that indicate potential appropriate models. In this case, we can look at ARMA(0,1), ARMA(1,0), ARMA(3,3) and ARMA(4,4)

ARMA(0,1): $X_t = W_t + 0.1263W_{t-1}$

ARMA(1,0): $X_t - 0.1168X_{t-1} = W_t$

ARMA(3,3): $X_t - 0.559X_{t-1} + 0.1564X_{t-2} + 0.609X_{t-3} = W_t - 0.4593W_{t-1} + 0.1523W_{t-2} + 0.7193W_{t-3}$

ARMA(4,4): $X_t + 0.3047X_{t-1} + 0.0437X_{t-2} + 0.2872X_{t-3} + 0.8878X_{t-4} = W_t + 0.4211W_{t-1} + 0.1124W_{t-2} + 0.4211W_{t-3} + 0.9999W_{t-4}$

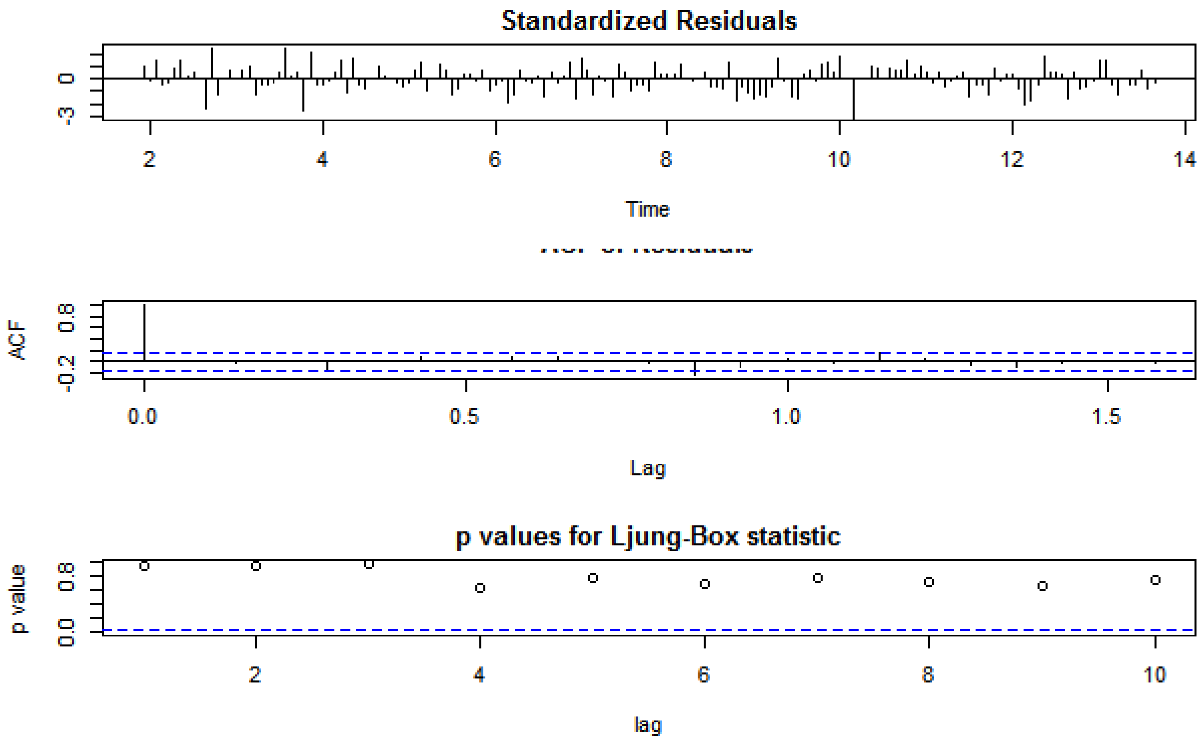

Figure 11

Figure 11 shows the summary statistics for the ARMA(0,0,1) model. Almost all ACF spikes are within the confidence interval, which is indicated by the dotted lines. The P-Values for the Ljung-Box test are all high, indicating that we do not need to reject this model. Although this figure only displays ARMA(0,0,1), the other 3 models look almost identical, with high P-Values. Therefore, we can confirm that all 4 models are potentially viable models for the dataset.

Model Diagnostics

In order to confirm which model is the best, we must run several tests to test for normality and independence. First, we can test normality by using the Shapiro Wilkes Test. All 4 models had P-values above 0.05, indicating normality. From there, we test for independence by using the L-Jung and Box-Pierce Tests. Again, all 4 models yielded P-values greater than 0.05, indicating independence. Knowing that all 4 models are suitable, we can use Occam's Razor to select ARMA(0,0,1) as the most appropriate model due to its simplicity.

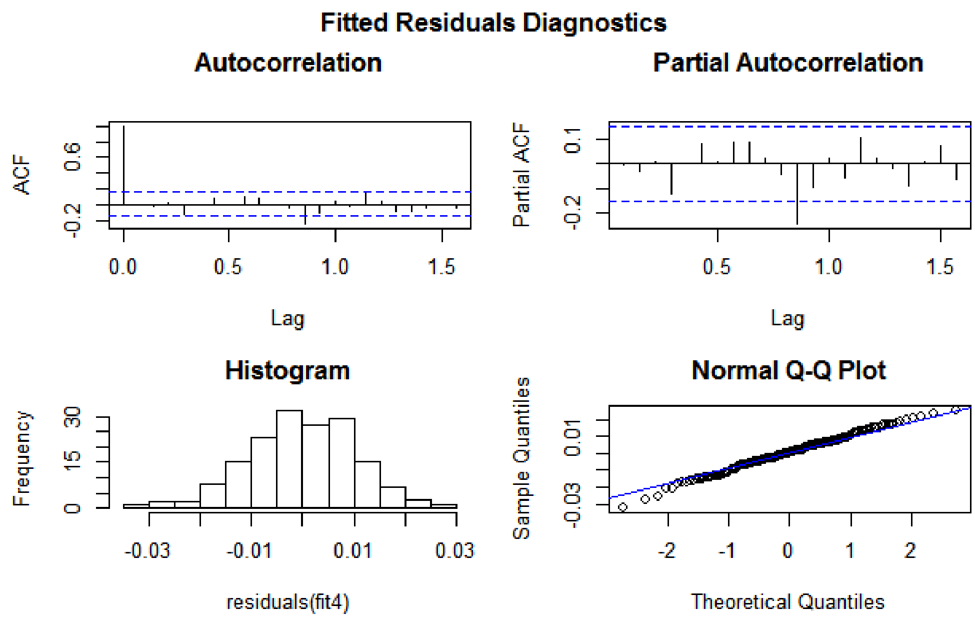

Figure 12

Figure 12 depicts some of the summary statistics for the residuals of ARMA(0,0,1). The ACF of the residuals are again, almost all within the confidence interval. Same goes for the PACF as almost all spikes are within the confidence interval. The histogram almost represents a normal bell curve while the QQ Plot follows a normal linear line.

Forecasting

95% Confidence Interval: (-0.03104026, 0.28355206)

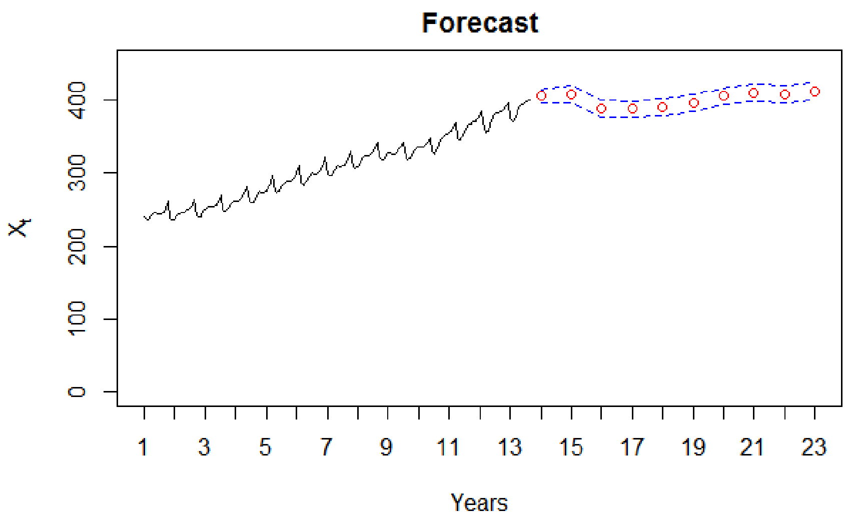

Figure 13

Figure 13 shows the predictions for the next 10 years of the Wisconsin Employment data. The red points indicate the potential future points, while the blue dotted lines indicate the range at which the red points can be placed. It seems that, based on these 10 points, the data no longer follows the seasonal and trend that the original data had followed for 15 years.

Spectral Analysis

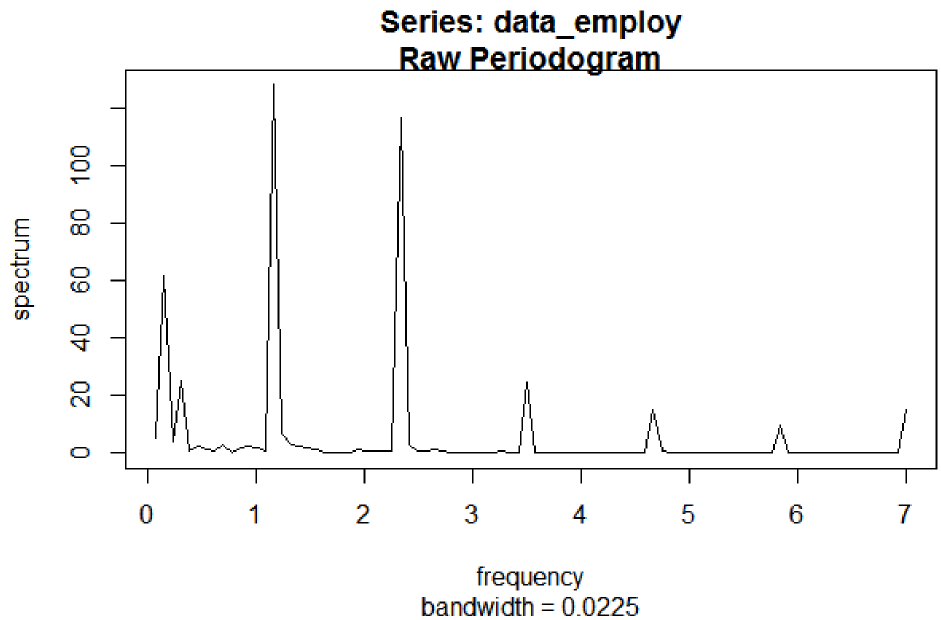

Figure 14

The figure above shows the results of the spectral analysis done on the data. As you can see, the periods are divided into partitions of frequency 1. Within each of the periods, we can clearly see that the spectrum spikes, potentially indicating weak seasonality. These spikes are most evident at periods 0,1,2, where the spikes are significantly stronger than the rest. This supports Figure 2 where we can clearly see the seasonal aspect depicted in the breakdown of the original dataset.

Conclusion

Following this project, it seems that the best model is an ARMA(0,2,1). Compared to the other models presented, it seemed to present the best ACF and PACF while maintaining the smallest variance. The fitted model also showed P-Values greater than 0.8 for Box-Pierce and Ljung-Box tests meaning that the model is sufficient. Finally, the forecasting for the next 10 points showed interesting results as they moved away from the upward seasonal trend that the data had followed for 15 years. We can speculate as to why this trend continues to persist within the data. In 1961, Wisconsin established a standard minimum wage for workers. According to the Wisconsin DWD, from 1961-1875, minimum wage increased from $0.85 to $2.00, while introducing wages for women in various fields of work. This increase in minimum wage indicates a significant economic growth that is reflected within the dataset.

References

Hipel and McLeod. "Wisconsin Employment Time Series, Trade, Jan. 1961-Oct.1975". 1994. Web. 14 Mar. 2017. https://datamarket.com/data/set/22l8/wisconsin-employment-time-series-trade-jan-1961-oct-1975#!ds=22l8&display=line

Wisconsin DWD. "Historical Resume of Minimum Wage Regulations in Wisconsin." N.P., N.D., Web. 14 Mar. 2017. https://dwd.wisconsin.gov/dwd/publications/erd/pdf/ls_39e_p.pdf

Appendix

```
employment = read.table("C:\\Users\\impor\\Desktop\\ANDREW\\COLLEGE\\STATISTI
CS\\PSTAT274\\Project\\wisconsin-employment-time-series.txt")
employment$V1 <- NULL
employ2 <- employment$V2
data_employ <- ts(employ2, freq = 14)
ts.plot(employment, xlab="Year-Month", ylab="Employment", main = "Wisconsin E
mployment Time Series, Jan. 1961 - Oct. 1975")

plot(stl(data_employ, s.window = "periodic"), main = "Seasonality and Trend")

#Original Histogram
hist(data_employ)
curve(dnorm(x, mean = mean(data_employ), sd = sqrt(var(data_employ))), add =
TRUE)

#Box Cox Transformation
require(MASS)

bcTransform <- boxcox(data_employ ~ as.numeric(1:178))

lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]

y.tr <- (1/lambda)*(data_employ^lambda -1)
y.log <- log(data_employ)
y.sqrt <- sqrt(data_employ)

op <- par(mfrow = c(1,3), mar = c(4,4.5,3,1))
ts.plot(y.tr, main="Box-Cox")
ts.plot(y.log, main = "Log")
ts.plot(y.sqrt, main="Square-Root")

#Differencing to find the best ACF/PACF Models
#Difference at lag1
y.tr.diff1 <- diff(y.tr,1)
ts.plot(y.tr.diff1, main="Differenced Data at Lag 1", ylab=expression(paste(n
abla,y)))
abline(h = mean(y.tr.diff1, col = "red"))

op <- par(mfrow=c(1,2), mar = c(4,4.5,3,1))
acf(y.tr.diff1, main="ACF for Difference Data 1")
pacf(y.tr.diff1, main="PACF for Difference Data 1")
var(y.tr.diff1)

#Difference at lag1 and lag12
y.tr.diff12 <- diff(y.tr.diff1, 12)
diff12 <- ts.plot(y.tr.diff12, main="Differenced Data at Lag 1, 12", ylab=exp
```

```r
ression(paste(nabla,y)))
abline(h = mean(y.tr.diff12, col = "red"))

op <- par(mfrow=c(1,2), mar = c(4,4.5,3,3))
acf(y.tr.diff12, main="ACF for Difference Data 1,12")
pacf(y.tr.diff12, main="PACF for Difference Data 1,12")
var(y.tr.diff12)

#Difference at lag1 and lag12 and lag12
y.tr.diff122 <- diff(y.tr.diff12, 12)
diff122 <- ts.plot(y.tr.diff122, main="Differenced Data at Lag 1, 12, 12", yl
ab=expression(paste(nabla,y)))
abline(h = mean(y.tr.diff122, col = "red"))

op <- par(mfrow=c(1,2), mar = c(4,4.5,3,3))
acf(y.tr.diff122, main="ACF for Difference Data 1,12,12")
pacf(y.tr.diff122, main="PACF for Difference Data 1,12,12")
var(y.tr.diff122)

#install.packages('qpcR')
library(qpcR)

aiccs <- matrix(NA, nr = 8, nc = 8)
dimnames(aiccs) = list(p=0:7, q=0:7)
for (p in 0:7){
  for (q in 0:7){
    aiccs[p+1, q+1] = AICc(arima(y.tr.diff12, order = c(p,0,q), method = "ML"
))
  }
}

aiccs

#Select ARMA(4,4), ARMA(3,3), ARMA(1,0), ARMA(0,1)

#Selecting appropriate model
library(forecast)

tsdiag(arima(y.tr.diff12, order = c(4,0,4)))
tsdiag(arima(y.tr.diff12, order = c(3,0,3)))
tsdiag(arima(y.tr.diff12, order = c(1,0,0)))
tsdiag(arima(y.tr.diff12, order = c(0,0,1)))

fit1 = arima(y.tr.diff12, order = c(4,0,4), method = "ML", xreg=1:length(y.tr
.diff12))
fit2 = arima(y.tr.diff12, order = c(3,0,3), method = "ML", xreg=1:length(y.tr
.diff12))
fit3 = arima(y.tr.diff12, order = c(1,0,0), method = "ML", xreg=1:length(y.tr
.diff12))
fit4 = arima(y.tr.diff12, order = c(0,0,1), method = "ML", xreg=1:length(y.tr
.diff12))
```

```r
#Normality
shapiro.test(residuals(fit1))
shapiro.test(residuals(fit2))
shapiro.test(residuals(fit3))
shapiro.test(residuals(fit4))

#Independence
Box.test(residuals(fit1), type = "Ljung")
Box.test(residuals(fit1), type = "Box-Pierce")
Box.test((residuals(fit1))^2, lag=13, type="Ljung")



Box.test(residuals(fit2), type = "Ljung")
Box.test(residuals(fit2), type = "Box-Pierce")
Box.test((residuals(fit2))^2, lag=13, type="Ljung")

Box.test(residuals(fit3), type = "Ljung")
Box.test(residuals(fit3), type = "Box-Pierce")
Box.test((residuals(fit3))^2, lag=13, type="Ljung")

Box.test(residuals(fit4), type = "Ljung")
Box.test(residuals(fit4), type = "Box-Pierce")
Box.test((residuals(fit4))^2, lag=13, type="Ljung")
#All tests have p-values greater than 0.05
#Since all fitted models pass, select the one with the smallest number of coe
fficients(fit4)

ts.plot(residuals(fit4), main="Plot of Residual of Fit4")

par(mfrow=c(1,4), mar = c(4,4.5,3,1), oma=c(0,0,2,0))
op <- par(mfrow=c(2,2))
acf(residuals(fit4), main="Autocorrelation")
pacf(residuals(fit4), main="Partial Autocorrelation")
hist(residuals(fit4), main="Histogram")
qqnorm(residuals(fit4))
qqline(residuals(fit4), col="blue")
title("Fitted Residuals Diagnostics", outer=TRUE)
par(op)

#Forecasting
fit.new <- arima(y.tr, order = c(0, 0, 1), seasonal = list(order = c(0, 2, 0)
, period = 12),  method = "ML")
pred <- predict(fit.new, n.ahead = 10)
pred.orig <- (pred$pred*lambda + 1)^(1/lambda)
pred.se <- ((lambda*pred$pred + 1)^((1-lambda)/lambda))*(pred$se)

op = par(mfrow = c(1,1))
plot(data_employ, xlim=c(1,23),ylim=c(0,450), xaxt='n', ylab = expression(X[t
]), xlab = 'Years', main='Forecast')
axis(1, at=seq(from=1, to=180, by=1))
```

```r
points(14:23, pred.orig, col="red")
lines(14:23, pred.orig+1.96*pred.se, lty=2, col="blue")
lines(14:23, pred.orig-1.96*pred.se, lty=2, col="blue")

#95% CI
ar1.se <- sqrt(fit4$var.coef[1])
c(fit4$coef[1] - 1.96*ar1.se, fit4$coef[1] + 1.96*ar1.se)

#Spectral Analysis
employ.per <- spec.pgram(employment, taper = 0, log = "no")
#Consider lag 1, 2, 3
```