Time Series Analysis: Wisconsin

Amir Ostrowsky

Jesus Quintero

Evan Quash

Mohammad Khan

Cesar Perez-Salas

December 05, 2018

Contents

ABSTRACT

The goal for this project is to forecast a season (about 12 months) of the monthly employment figures of Wisconsin from 1961 to 1974. In our analysis of the time series data, we used various techniques in order to generate a stationary data. From lays within to logarithmic transformations and differencing to remove the seasonality and trend within our original data. As a result, it allows us to identify potential models by looking at ACF and PACF plots. We decided that the best SARIMA model is based on whether the model has significant coefficients and on the significant lags falling below the lower confidence interval and lowest AIC according to the stationary data. Where we obtained from transforming and differencing the original data.

(1) INTRODUCTION

The employment figures measure the extent to which available labor resources (people available to work) are being used. Throughout our Wisconsin employment data, we realize there's different types of employment: Seasonal employment which is caused by seasonal patterns in economic activity. Frictional employment shows temporary unemployment during the period when people are searching for a job. Structural unemployment when a required skill is often brought by technological changes that make many workers obsolete. While plotting our Wisconsin data we noticed our graph has an upward trend and seasonality. Under those circumstances, we chose to use logarithmic transformation over the box-cox transformation because $\lambda = 0$ lays within the 95% confidence interval for possible values of λ. In addition, the variance of the log transformation was significantly lower than the variance of the box-cox transformation resulting in the logarithmic transformation that allows us to reduce the variation of the data. We differenced at lag 12 to remove seasonality, then differencing again at lag 1 to remove the trend. After performing these procedures, our variance decreased drastically from our original variance. Then, we chose 3 possible SARIMA models based on the ACF and PACF graphs. Comparing the significant coefficients, significant lags under the lower edge of the confidence interval and AIC we decided the final model to be $SARIMA(0,1,0)x(1,1,0)_{12}$. Throughout forecasting, we were able to plot a potential trajectory with 95% confidence for 14 months that approximate close to the true value in the original dataset from September 1974 to October 1975.

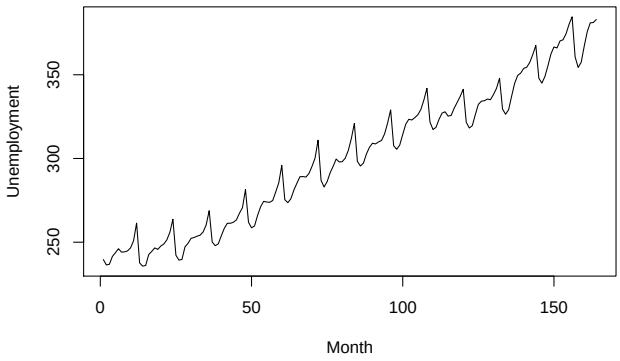
(2) EXPLORATORY DATA ANALYSIS

(2.1) Data Exploration

Our original dataset consists of monthly employment data from January 1961 to October 1975 in which we decided to removed the last 14 data points in order to compare forecasting accuracy. On *Figure 1* below you

can see the fixed plot of the unemployment in Wisconsin with a sample size of 164 months in total.

Figure 1: Wiscounsin employment data 1961−1974



As shown in our graph, there is an evident seasonality component because the dataset was collected monthly. In addition, there was an upward trend present throughout the sample dataset. Before applying any type of transformation, the variance of our original dataset is 2186.5. Since our dataset has a large variance as well as having an upward trend and seasonality. We can conclude that our time series needs to be transformed and differenced in order to make it stationary.

Figure 2: ACF and PACF of Wiscoun in Employment Data

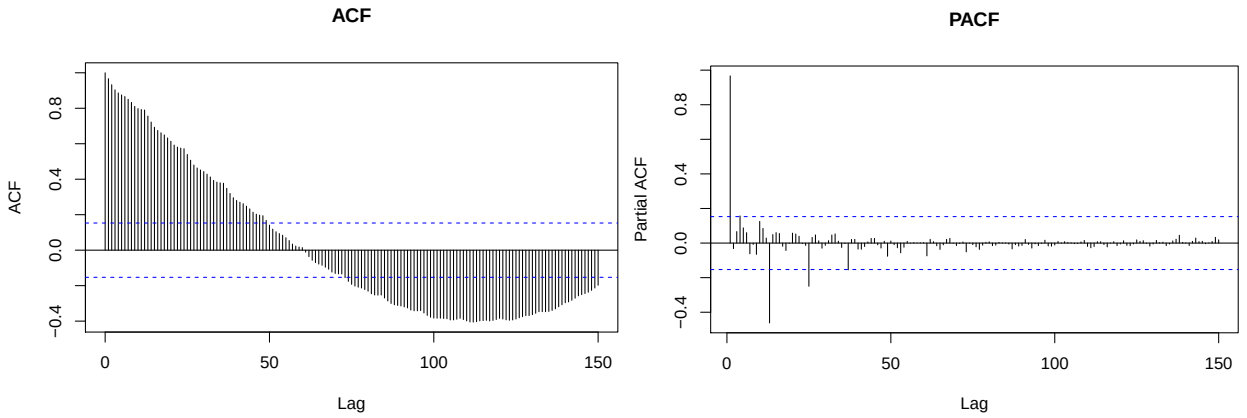


Figure 2 shows a strong seasonality component of period 12 and an increasing trend in the data. The variability of the data changes with time. A visual inspection of the autocorrelation function plot indicates that the employment series is nonstationary since the ACF decays very slowly. Suggesting some transformation to the data to render it stationary before choosing and fitting an ARMA time series model.

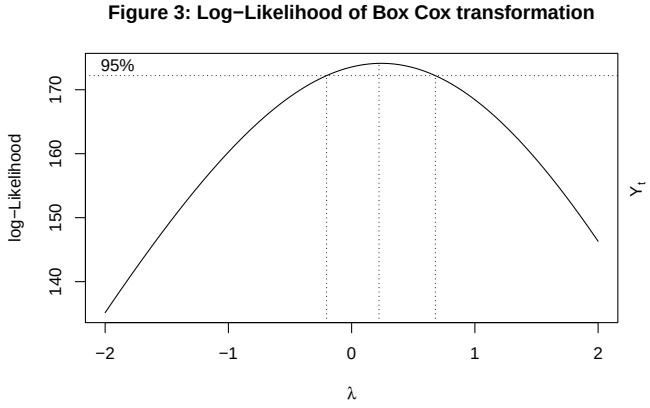
(2.2) Box-Cox Transformation

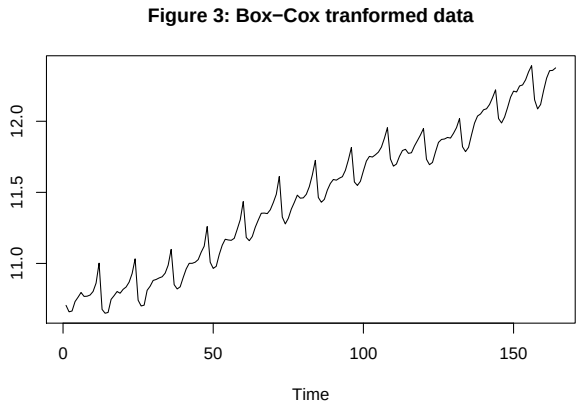
In order to make the data stationary and reduce the variance, we try a Box-Cox transformation. In Figure 3: Log Likelihood Box-Transformation, λ falls between 0 which makes the Box-Cox variable indeterminate

form 0/0. That being the case rewriting the Box-Cox formula gives us

$$X'_\lambda = \frac{(1 + \lambda log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + ...) - 1}{\lambda} \to \log(x)$$

corresponding to a logarithmic transformation. We can see the seasonal fluctuations and trend in the log-transformed time series showing constancy over time. In addition, the $log(X_t)$ transformation has smaller variance than Box-Cox(X_t) transformation.

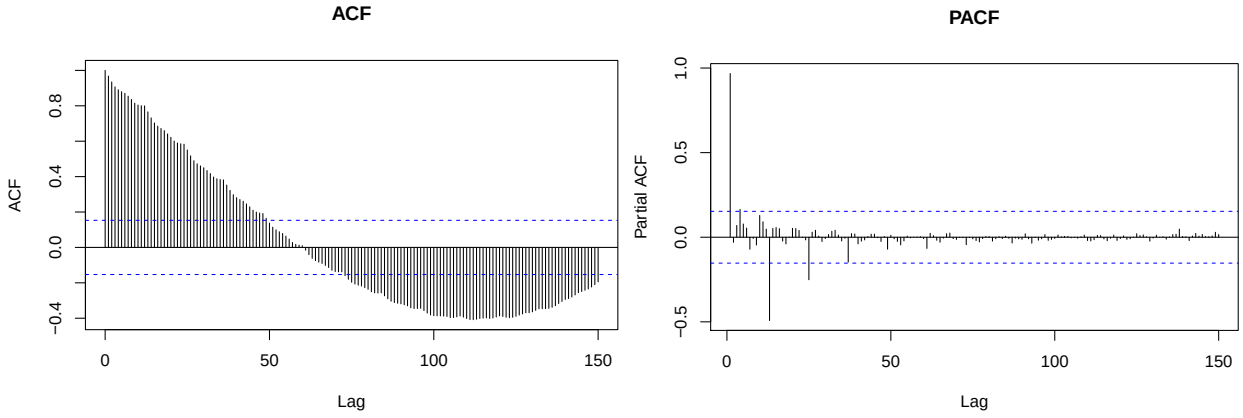




(2.2.2) Log Transformation

In order to make the data stationary and stabilize the variance, we used log transformation. Since it has the lowest variance compared to the Box-Cox transformation. The variance for the box-cox transformation is 0.247729 and was reduced to 0.0197 after using log transformation. Therefore, we picked log transformation as our best transformation.

Figure 4: ACF and PACF of Log transformed Data



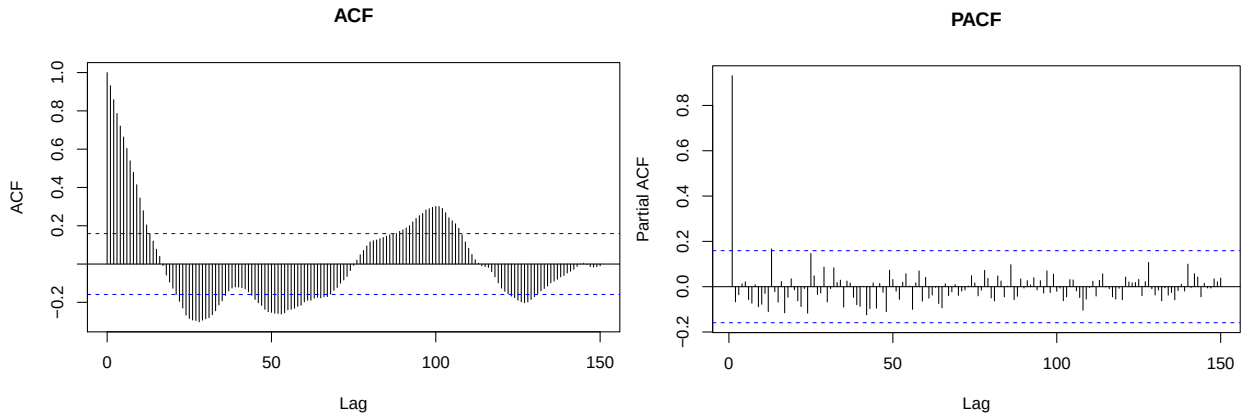
The slow decay of the autocorrelations in figure 4 is similar to the original dataset indicating that our data is still non-stationary. Thus, we further remove the seasonality and trend in the next step.

(2.3) Differencing

(2.3.1) De-Seasonalizing

We use differencing to deseasonalize and detrend the data. Since our data has a seasonal component of period 12, we used this to remove the seasonality of period Ű · 12. After taking the differenced, the variance decreased even more to 0.0002516346. Next, we remove the trend based on the deseasonalized data we obtained.

Figure 5: ACF and PACF of Deseasonalized Transformed Data

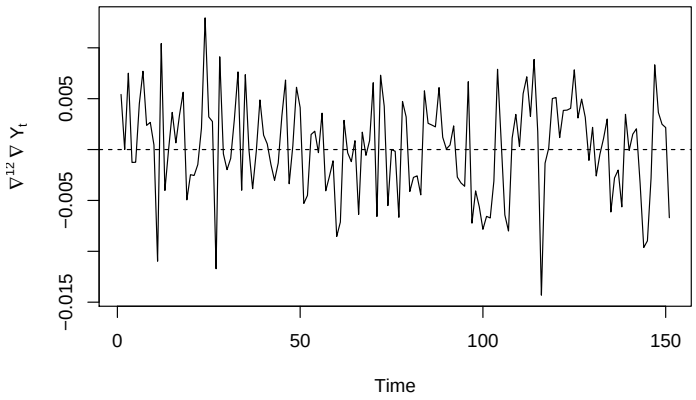


Since the ACF still has a cyclical pattern, we need to remove trending by differencing by lag 1.

(2.3.2) De-Trending

We use the differencing transformation to remove the trend from our data. The difference sequence is inverted using the prior values from the original sequence, as the primer for each transformation applying lag of 1. After de-trending, the variance is now 0.000002317563 which falls to a stationary time series. In order to check whether our time series is stationary. We used the Augmented Dickey-Fuller Test, to create a hypothesis test. Null Hypothesis is that X_t is non-stationary and the alternative hypothesis H_a is stationary. The test resulted in a significant p-value of 0.01951, so we reject the null hypothesis and conclude that our time series is stationary.

Figure 6: De−trended/seasonalized Time Series



We can further see Figure 6 above having no seasonality nor trend. The ACF and PACF of the log-transformed, deseasonalized, detrended data is shown in Model Building 3.1.

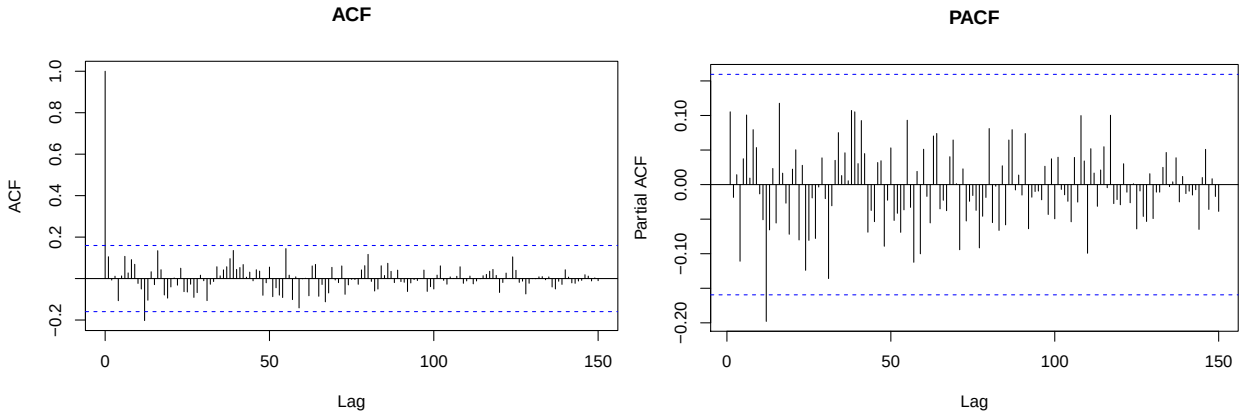
(3) MODEL BUILDING

After removing the trend and seasonality to produce a stationary series, we can fit the data into a SARIMA model. A SARIMA model is illustrated by $SARIMA(p,d,q)X(P,D,Q)_s$ where p = non-seasonal AR order, d = non-season differencing, q = non-seasonal MA order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and S = the length of the season. As the data was collected monthly from 1961 to 1975 we analyzed *Figure 7: ACF and PACF of log transformed, seasonalized detrended data* The autocorrelation function of seasonalized detrended data providing $D = 1$ and $S = 12$ since we difference at lag 12 then differenced again at lag 1 to remove trend giving $d = D = 1$. The trend elements can be chosen through careful analysis of ACF and PACF plots by looking at the correlations of recent time steps to p, q and P, Q based off ACF and PACF of the data below in *Figure 7.*.

(3.1) Analyzing ACF and PACF

We find the components of P and Q and p and q. First, we look at the first season lags (1-11) to find our p and q. None are significantly outside the confidence interval which would suggest $p = 0$ and $q = 0$ by looking at ACF. To determine the order of seasonal components P and Q we observe more notable lags at 12, 24 and 36 with a significant lag in the ACF graph at lag 12. When we differenced at lag 12, it resulted in an SMA component of 1 so P=1. If we compare lag 12 on PACF suggesting SAR component of 1, so $Q = 1$. In order to verify our assumption, we build three models based on $P = 1$ or 0 and $Q = 1$ or 0 to see which significant coefficients and low AIC.

Figure 7: ACF and PACF of log transformed, seasonalized detrended data



(3.2) Model Selection

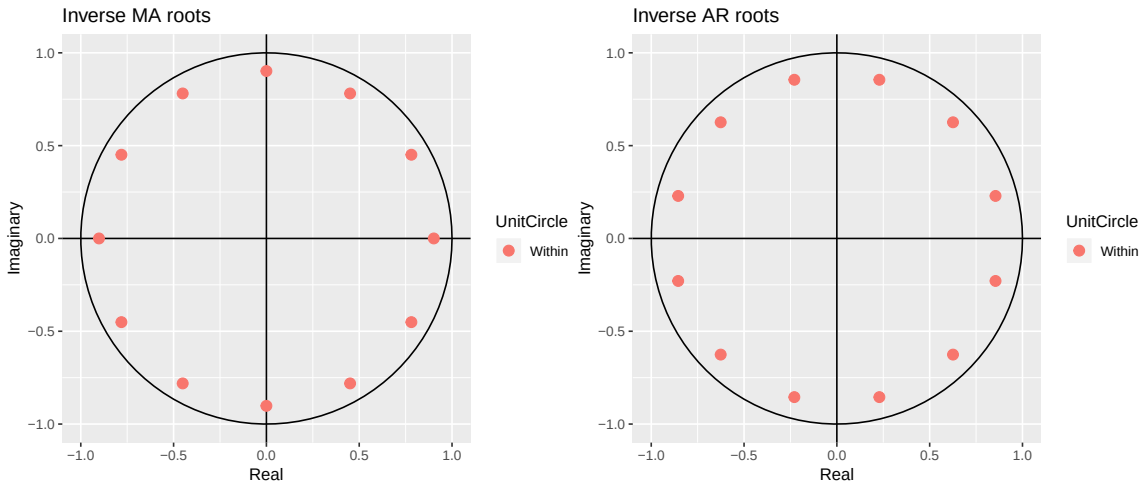
Based on our model building criteria, we select the best model for our forecast. We have three models $Model1 = SARIMA(0,1,0)X(0,1,1)_{12}$, $Model2 = SARIMA(0,1,0)X(1,1,0)_{12}$ and $Model3 = SARIMA(0,1,0)X(1,1,1)_{12}$. Of these possible models we identify which model with the lowest AICs have significant coefficients: coefficients whose absolute value is greater than their standard error values.

	Model 1	**Model 2**	Model 3
Confidence Interval	-0.2896	**-0.2308**	0.2023
SE	.6268	**.0939**	.2436
AIC	-.1186.48	**-1184.65**	-1185.02

The only model with significant coefficients is Model 2 based on the table above with $d = D = 1$, $P = 0$, $p = 0$, $Q = 1$, $q = 0$ and the second lowest AIC is $P = 1$ and $Q = 1$.

(3.3) Model Fitting

In the final analysis, we decided to pick Model 2 as our best SARIMA(0,1,0)X(1,1,0)$_{12}$. We examined the roots of the polynomials to check for causality and invertibility. After plotting the roots in red, it was apparent that the inverse roots lie within the unit circle for both models. We take the absolute values for all the coefficients that are less than 1 which means they are outside of the unit circle. Thus, we conclude model 2 is invertible and casual.



$$X_t + \phi_1 B X_t = Z_t + \theta_1 B Z_t$$

$$\phi_1 = 0.0939 \qquad \theta_1 = -0.2308$$

$$X_t(1 + (0.0939)B = Z_t(1 + (-0.2308)B)$$

$$1 + \phi_1 B = 0 \qquad 1 + \theta_1 B = 0$$

$$B = |-\frac{1}{\phi_1}| > 1 \qquad B = |-\frac{1}{\theta_1}| > 1$$

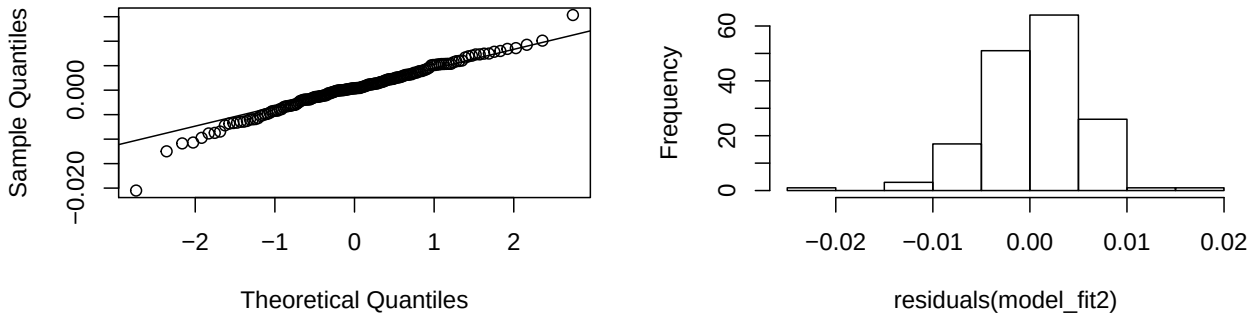
$$|\phi_1, \theta_1| < 1$$

(3.4) Diagnostics

(3.4.1) Normality

Another way of checking whether the condition of normality holds in our model is by plotting our residuals on a Q-Q (Quantile-Quantile) plot, as well as checking their frequencies using a histogram. Using the qqnorm() function we see our QQ plot is approximately linear which supports our assumption of normality. Moreover, it is observed that our histogram doesn't present outliers or skewness in the data, which would present itself in the form of tails in the graph.

Figure 7: QQ Plot and Histogram

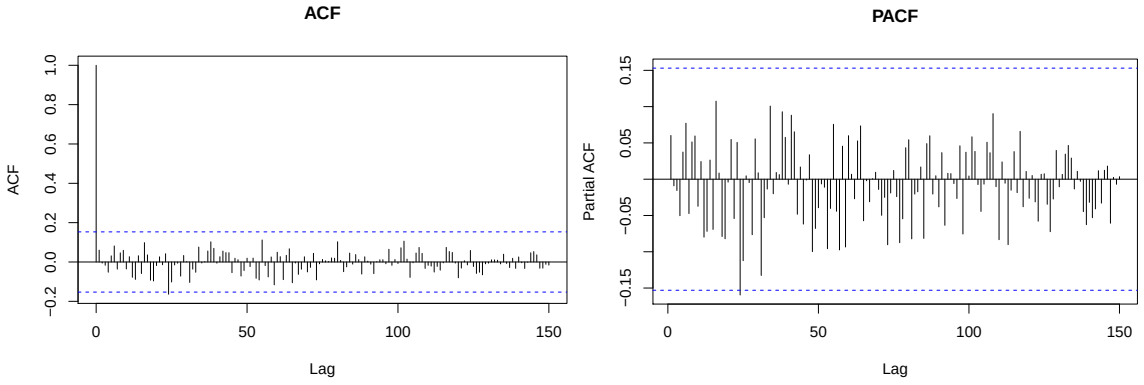


(3.4.2) Independence

To examine whether or not our model satisfies the condition of normality, we performed both the Box-Pierce and Ljung-Box test. As shown above, the p-values were calculated to be 0.6192 and 0.6224. Therefore, we fail to reject the null hypothesis and say with confidence that our model is normal since both p-values $>=.05$.

	Ljung-Box	Box-Pierce	Shapiro-Wilk
Test Statistic	$X^2 = 0.60667$	$X^2 = 0.59571$	W = 0.97692
df	1	1	-
p-value	**0.436**	**0.4402**	0.00766

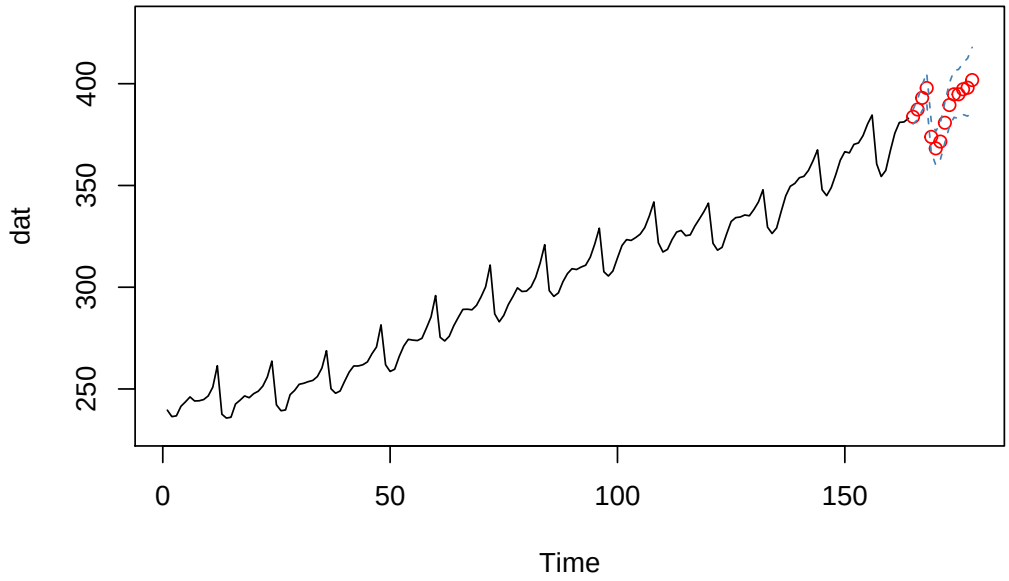
(3.4.3) Constant Variance



(4) FORECASTING

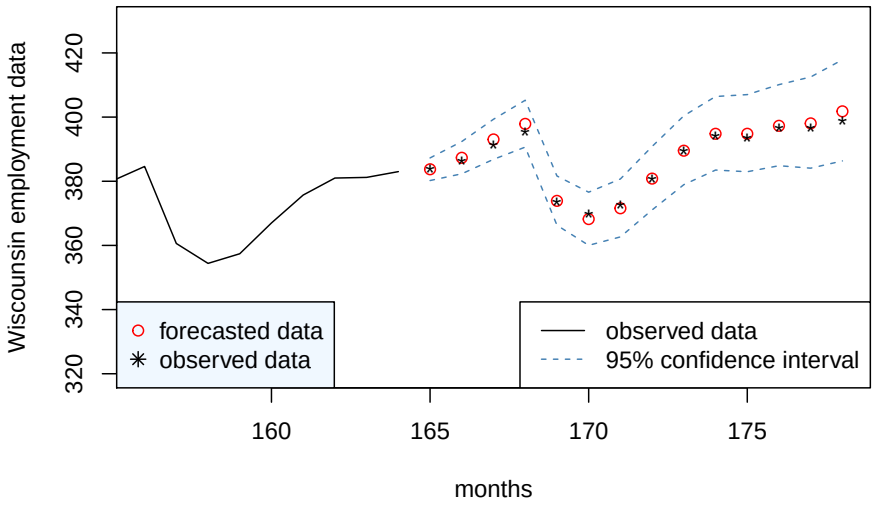
Using model 2, we back transformed the logarithmically transformation data by taking the exponential transformation of the logarithmically transformed data. We can observe the 14 forecasted values in the form of red circles in *figure 8* from September 1974 to October 1975, along with the 95% confidence intervals for these forecasted values. We cut the last 14 months as our testing data for comparison with our forecasted values.

Figure 8: Wiscounsin employment data forecasted



This final model captures the overall trend and seasonality of the data. Figure 9 shows a zoomed-in view of the forecasted values resembling the obvious patterns of our original data. The observed and forecasted values are very close throughout the season which proves the success of our final model.

Figure 9: Wiscounsin employment data forecasted



(5) Conclusion

Given the monthly data from January 1961 to August 1974, we were able to predict with a 95% confidence the employment figures until October 1975. We began our analysis of the data by removing all seasonality and trend by differencing at lag 12 and at lag 1, respectively. Then we attempted a Box-Cox transformation of the data but later found the logarithmic transformation was more appropriate with lower variance. We continued our analysis of the data by analyzing the the ACF and PACF graphs and were able to narrow our focus on 3 SARIMA models. With similar AICs we found that SARIMA(0,1,0)X(1,1,0)$_{12}$ had significant coefficients that captures best the overall trend and seasonality of the data. The coefficients for our model were calculated using mathematically by figuring out the roots for the MA and AR component of the model, and confirmed with the arima() function in R. Lastly, we were able to predict the final 14 months of our data with 95% confidence.

Our final model

$$Model2 = SARIMA(0,1,0)X(1,1,0)_{12}$$

We will like to thank Professor Bapat for working with us outside of the classroom, email and profoundly inspiring everybody to apply our Time Series knowledge into practice.

(6) REFERENCES

[1] Labour market, Source: Hipel and McLeod (1994), in file: wisconsi/trade, Description: Wisconsin employment time series, trade, Jan. 1961 – OCt. 1975 (https://datamarket.com/data/set/22l8/wisconsin-employment-time-series-trade-jan-1961-oct-1975#!ds=22l8&display=line)October 2018

[2] TheBalance, source: Types of https://www.thebalance.com/types-of-unemployment-3305522,Aug 14,2018. Used November 2018

[3] Hyndsight, Plotting the Characteristics roots for ARIMA models(https://robjhyndman.com/hyndsight/arma-roots/).Dec2 018

(7) APPENDEX

```
setwd('~/Desktop/project1')
dat0 = read.table('wisconsin-employment-time-series.csv', header = F, sep=',')
head(dat0)
tail(dat0)
dat0 <- ts(dat0[,2])
var(dat0)
dat <- dat0[-c(165:178)] #remove 14 data points from original data to compare forecasting accuracy
ts.plot(dat, xlab="Month", ylab= "Unemployment", main="Figure 1: Wiscounsin employment data 1961-1974")
acf(dat, lag.max=150, main="ACF")
pacf(dat, lag.max=150, main="PACF")



library(MASS)
t = 1:length(dat)
fit = lm(dat ~ t)
bcTransform = boxcox(dat ~ t, plotit = TRUE)
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
dat.bc = (1/lambda)*(dat^lambda-1)
title("Figure 3: Log-Likelihood of Box Cox transformation")

ts.plot(dat.bc,main = "Figure 3: Box-Cox tranformed data", ylab =
          expression(Y[t]))
#var(dat) This is the variance of the normal plot
```

```r
#var(dat.bc) variance after the Box

dat.log = log(dat)
var(dat)
var(dat.bc)
var(dat.log)

acf(dat.log, lag.max=150, main="ACF")
pacf(dat.log, lag.max=150, main="PACF")

# seasonalize: difference at lag 12
y12 = diff(dat.log, 12)
var(y12)

acf(y12, lag.max=150, main="ACF")
pacf(y12, lag.max=150, main="PACF")



# detrend: difference at lag 1
y1 = diff(y12, 1)

ts.plot(y1, main = "De-trended Time Series", ylab = expression(nabla~Y[t]))
abline(h = 0, lty = 2)

ts.plot(y1,main = "Figure 6: De-trended/seasonalized Time Series",ylab =
          expression(nabla^{12}~nabla~Y[t]))
abline(h = 0,lty = 2)

library(tseries)
adf.test(y1, k =12)
acf(y1, lag.max=150, main="ACF")
pacf(y1, lag.max=150, main="PACF")

# with d = D = 1, P = 1, p = q = 0, Q = 1 we have the lowest AIC selecting SARIMA(0,1,0)X(0,1,1) model.

model_fit1 = arima(dat.log,order = c(0,1,0),seasonal = list(order = c(0,1,1),period = 12), method = "ML
```

```r
model_fit1

# with d = D = 1, P = 0, p = 0, Q = 1, q = 0 we have the 1st lowest AIC selecting SARIMA(0,1,0)x(0,1,1)

model_fit2 = arima(dat.log, order = c(0,1,0),seasonal = list(order = c(1,1,0),period = 12), method = "M
model_fit2

# with d = D = 1, P = 1, p = 0, Q = 1, q = 0 we have the 2nd lowest AIC selecting SARIMA(0,1,0)x(1,1,1)

model_fit3 = arima(dat.log, order = c(0,1,0),seasonal = list(order = c(1,1,1),period = 12), method = "M
model_fit3



library(dse)
library(forecast)
library(ggplot2)

fit <- Arima(dat.log,order=c(0,1,0), seasonal=list(order = c(0,1,1),period=12), xreg = 1:length(dat.log
# SARIMA(0,1,1)x(0,1,1) model roots
autoplot(fit)
autoplot(model_fit2)

Box.test(residuals(model_fit2), type = "Ljung")
Box.test(residuals(model_fit2), type ="Box-Pierce")
shapiro.test(residuals(model_fit2))
op = par(mfrow = c(1,2))
qqnorm(residuals(model_fit2), main = "")
qqline(residuals(model_fit2))
hist(residuals(model_fit2),main="")
title("Figure 7: QQ Plot and Histogram",line = -1, outer=TRUE)

Box.test(residuals(model_fit2), type = "Ljung")
Box.test(residuals(model_fit2), type ="Box-Pierce")
shapiro.test(residuals(model_fit2))

op = par(mfrow = c(1,2))
```

```r
acf(residuals(model_fit2), lag.max=150, main="ACF")
pacf(residuals(model_fit2), lag.max=150, main="PACF")

# full view
len <- length(model_fit2)
dat_last14 <- dat0[165:178] # last 14 entries from time series data
mypred <- predict(model_fit2, n.ahead=len, newxreg = (length(dat)+1):(length(dat)+14))
#model_ts <- ts(model_fit2)
#back transforming the log transformed data by taking the exponential function of the data
pred.orig <- exp(mypred$pred)
pred.se.upper <- exp(mypred$pred+1.96*mypred$se)
pred.se.lower <- exp(mypred$pred-1.96*mypred$se)
ts.plot(dat, main = "Figure 8: Wiscounsin employment data forecasted",
        xlim=c(1,length(dat)+14), ylim=c(230,430))
points((length(dat)+1):(length(dat)+14),pred.orig, col="red")
lines(pred.se.upper,lty=2,col="steelblue")
lines(pred.se.lower,lty=2,col="steelblue")
#points((length(dat)+1):(length(dat)+14),dat_last14, pch="*", col="black")

# zoomed in view
ts.plot(dat, main = "Figure 9: Wiscounsin employment data forecasted",
        xlim=c(length(dat)-8,length(dat)+14), ylim=c(320,430), xlab="months",
        ylab="Wiscounsin employment data")
points((length(dat)+1):(length(dat)+14),pred.orig, col="red")
lines(pred.se.upper,lty=2,col="steelblue")
lines(pred.se.lower,lty=2,col="steelblue")
points((length(dat)+1):(length(dat)+14),dat_last14, pch="*", col="black")
legend("bottomright",c("observed data", "95% confidence interval"),
       bg="white", lty = c(1,2), lwd = c(1,1), col=c("black", "steelblue"))
legend("bottomleft",c("forecasted data", "observed data"),
       bg="aliceblue", pch=c(1,8), col=c("red", "black"))
```