PSTAT 174 TIME SERIES PROJECT

Trisha Jo Dumayas

November 29, 2017

Table of Contents

Abstract

The housing market undergoes various increases and decrease in sales over time. Our data set provides monthly sales of new one-family houses sold in USA from 1973 to 1995. Using this data set, our goal is to forecast the amount of houses sold within the next 10 months and see how they compare to the true number of houses sold within those 10 months. Predicting these data points will give us a better understanding of the fluctuation of sales in the housing market. Analyzing the increases and decreases of sales is useful in deciding on an optimal time to buy or sell a house.

First I created a training set and did not include the last 10 data points to see if I could forecast values close to the true values. In order to forecast, our data must be stationary. Since the data set was not stationary, I applied a Box-Cox transformation. The Box-Cox transformed data displayed a trend, a seasonal component, and a changing variance so I used differencing. Examining the ACF and PACF after differencing the data, I identified three candidate models: SARIMA $(0, 0, 0)$ x $(2, 0, 1)_{12}$, SARIMA $(0, 0, 0)$ x $(4, 0, 1)_{12}$, SARIMA $(2, 0, 0)$ x $(2, 0, 1)_{12}$. For each model, I examined plots of the residuals, preformed diagnostic checks, compared AIC, and calculated causality and invertibility. From the tests, I concluded that Model 3 was most suitable model of the three to use in forecasting. Plotting forecasted future values using Model 3, it is clear that the true values are within the 95% confidence interval of the forecasted values. Thus, SARIMA $(2, 0, 0)$ x $(2, 0, 1)_{12}$ is an accurate model to forecast our data set.

Introduction

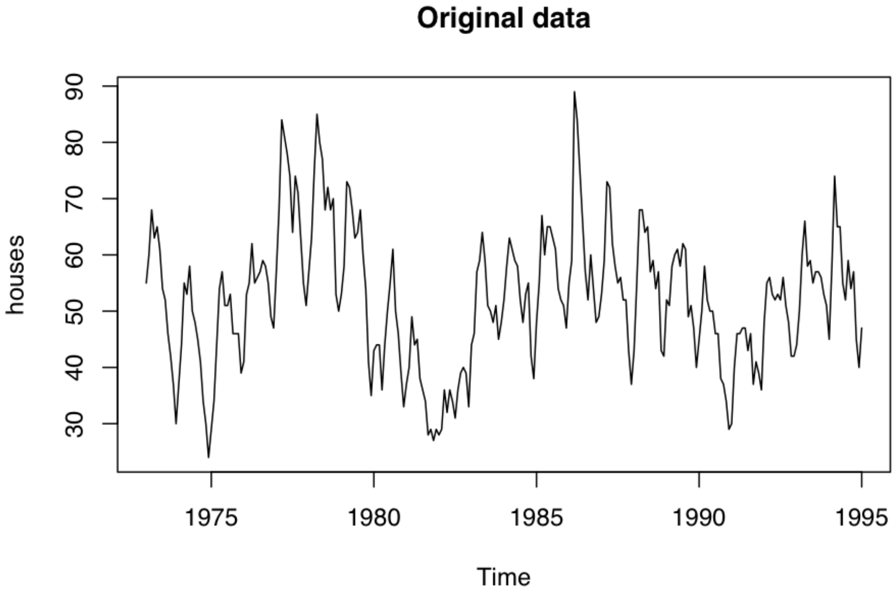
The housing market encounters fluctuations in sales due to shifts in the economy. This directly influences the increase and decrease of housing prices making it difficult to decide the best time to purchase or sell a home. This data set provides monthly sales of new one-family houses sold in the USA from 1973 to 1995. From the plot of the time series, we can identify certain times throughout the year when sales are at their lowest and highest. From the law of supply and demand we know that the more houses available, the more prices of houses decrease, which then increases the demand for houses. That is, houses become less expensive once sales are lowest while houses become more costly once sales are at their highest. Thus our goal is to forecast future house sales from the given history in our data set to help determine the optimal time to purchase or sell a home.

In order to forecast, I first plotted the time series without the last 10 values and checked for trend, seasonality, or changes in variance. I then applied a Box-Cox transformation to make the data stationary and to stabilize the variance. Next I applied differencing at lags 1 and 12 since there was a trend and seasonal component. Then I plotted the ACF and PACF to identify potential models for the data: SARIMA $(0, 0, 0)$ x $(2, 0, 1)_{12}$, SARIMA $(0, 0, 0)$ x $(4, 0, 1)_{12}$, SARIMA $(2, 0, 0)$ x $(2, 0, 1)_{12}$. After checking causality and invertibility, plotting their residuals, performing diagnostic checks, and comparing AIC, I concluded that Model 3 was the most suitable for forecasting. Model 3 was able to forecast successfully since the true values were within the confidence intervals of the forecasted values.

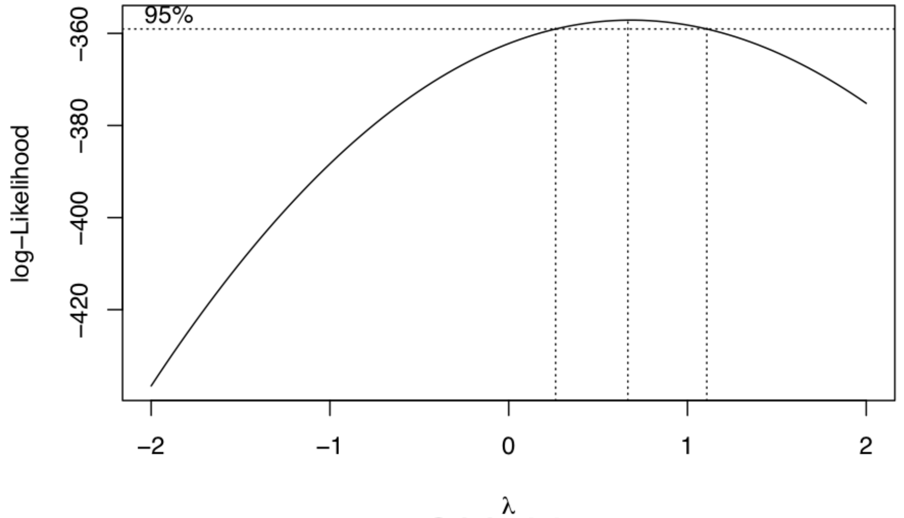
The data set can be found on datamarket.com and was provided by the Time Series Data Library. All statistical analysis on the project was performed using RStudio.

Analysis

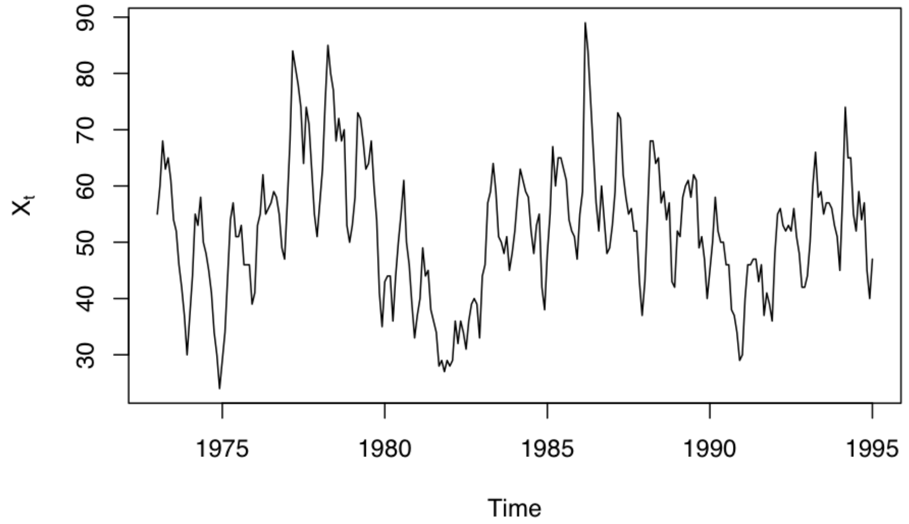
First we start by plotting the time series.

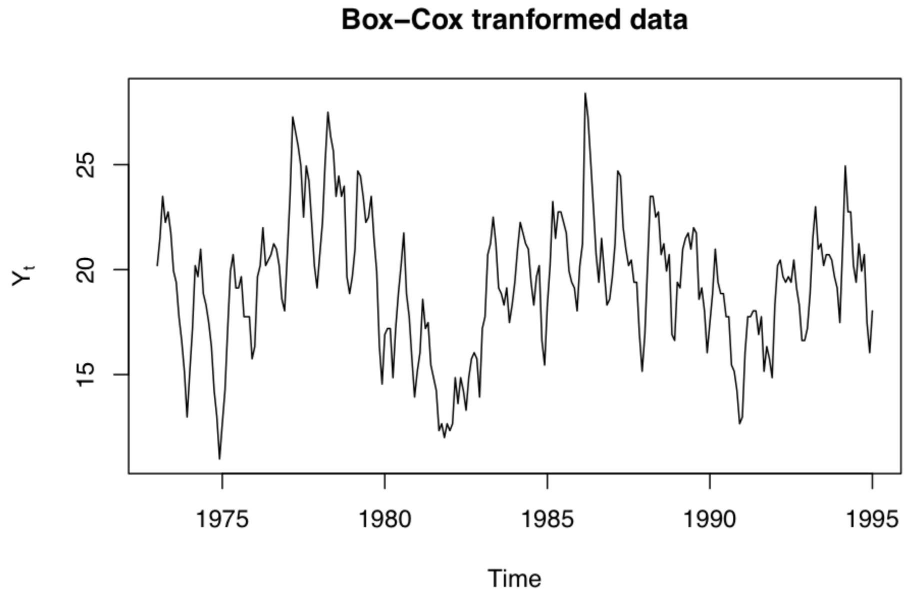


We observe that there is a cyclical increasing and decreasing trend in the data, a strong seasonal component, and volatility of the variance over time displayed by the changing range of values across the different time intervals. Thus we conclude that our data is not normally distributed and we must perform the Box-Cox transformation.



Original data





```
#Compare Var
var(houses)
```

```
## [1] 145.2358
```

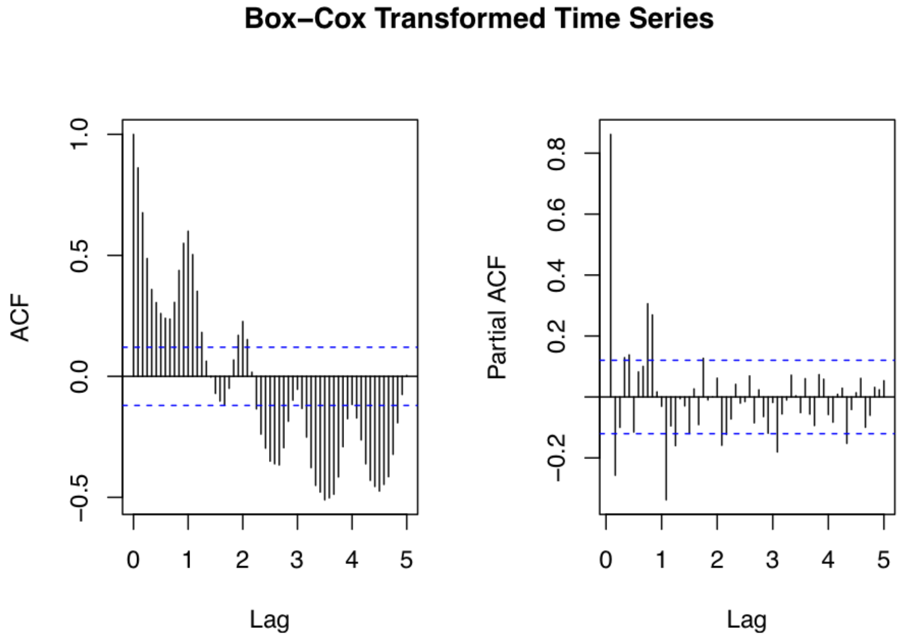
```
var(houses.bc)
```

```
## [1] 10.51813
```

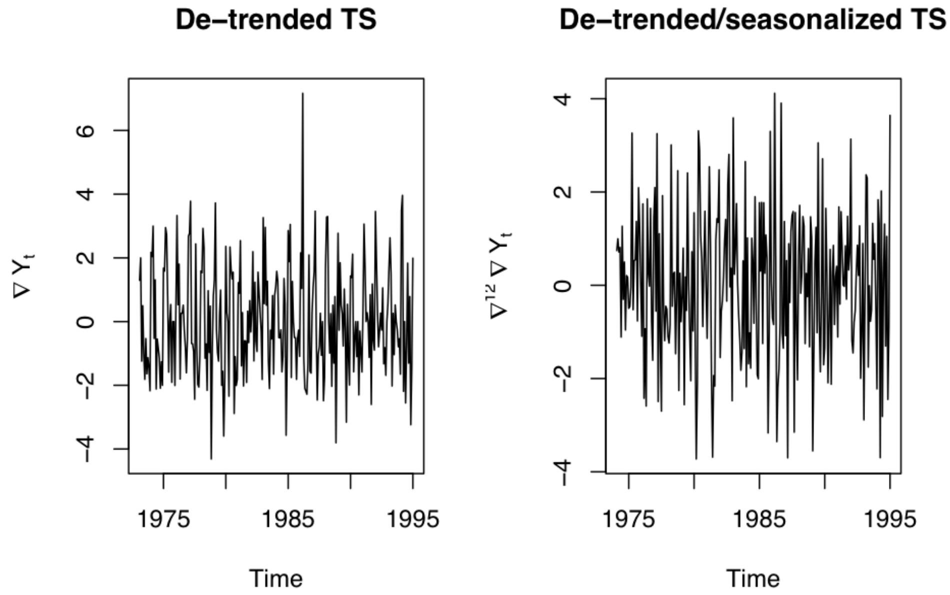
Time series model building and forecasting is typically done under the assumption that the data is normally distributed. Since the data is not normally distributed, I applied a Box-Cox transformation to find the optimal λ, transform the data, and re-plot the time series. The dashed vertical lines in the plot are the 95% confidence interval for the true λ value in the Box-Cox transformation. Since the interval does not include $\lambda = 0$, then the Box-Cox transformation to stabilize the variance is

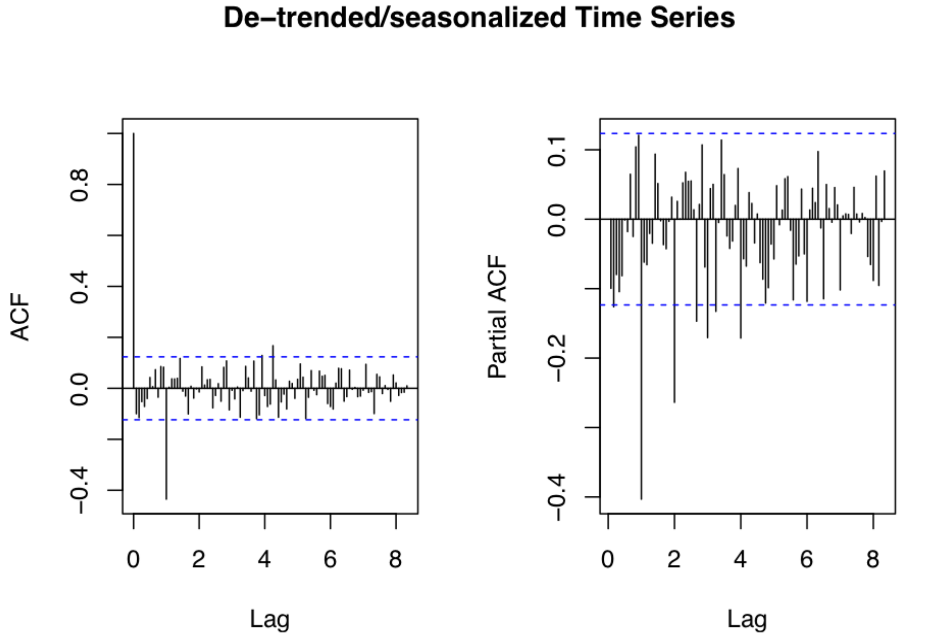
$$Y_t = \frac{1}{\lambda}(X_t^\lambda - 1);$$

By plotting the graph and calculating the variance we see that the Box-Cox transformation has lowered the variance and lessened the volatility.



Then I plotted the ACF and PACF of the transformed series to determine possible models to forecast our data. However, when plotting the ACF and PACF I noticed a cycle in the transformed data as there were significant correlations every twelve lags. This implies that the period of the seasonal component is 12, that is $d = 12$.





I used differencing first at lag 1 to remove the trend component and then second at lag 12 to remove the seasonal component. After using Box-Cox transformation and differencing at lags

1 and 12, the time series is now stationary.

Now we can use the ACF and PACF to fit possible models for our data. I observed that there is a seasonal component since there are spikes at lags $l = 12n$ where $l = 12n$, $n \in \mathbb{N}$. The ACF cuts off after lag 12 so we can assume SMA(1). The PACF cuts off after lag 4 so we can assume SAR(4). We also consider SAR(2) since the lags observed after lag 2 are not as large as the lags prior meaning that they may just be from noise.

Next we examine the lags 1 through 11 to determine the order of MA and AR. The ACF cuts off after lag 0 therefore we must have MA(0). The PACF decays quickly so we will consider AR(0). Another possible order of AR may be AR(2) as the ACF may be tailing off while the PACF cuts off after lag 2.

Thus we consider the following models:

- 1. SARIMA $(0,0,0)$ x $(2,0,1)_{12}$
  - AIC $= 1245.06$
  - $\nabla_{12}\nabla X_t = (X_t - X_{t-12})(X_t - X_{t-1})Z_t$
  - $(X_t + 1.0058X_{t-12} + 0.3863X_{t-24})\nabla_{12}\nabla X_t = (Z_t - 0.2581Z_{t-12})$
- 2. SARIMA $(0,0,0)$ x $(4,0,1)_{12}$
  - AIC $= 1195.66$
  - $\nabla_{12}\nabla X_t = (X_t - X_{t-12})(X_t - X_{t-1})Z_t$
  - $(X_t - 1.5084X_{t-12} + 0.5952X_{t-24} + 0.2843X_{t-36} - 0.3710X_{t-48})\nabla_{12}\nabla X_t = (Z_t - 0.9864Z_{t-12})$
- 3. SARIMA $(2,0,0)$ x $(2,0,1)_{12}$
  - AIC $= 918.39$
  - $\nabla_{12}\nabla X_t = (X_t - X_{t-12})(X_t - X_{t-1})Z_t$
  - $(X_t - 0.8225X_{t-1} - 0.0435X_{t-2})(X_t - 0.3374X_{t-12} - 0.2941X_{t-24})\nabla_{12}\nabla X_t = (Z_t + 0.0548Z_{t-12})$

We check to see whether our models are causal and invertible by seeing if the roots of the polynomials are outside of the unit circle.

```
## [1] 1.301838+0.945452i 1.301838-0.945452i
## [1] 3.874467+0i
```

The SAR and SMA polynomial roots for Model 1 are outside of the unit circle. Therefore Model 1 is both causal and invertible.

```
## [1]  1.0001054-0.0000000i -1.8115651+0.0000000i  0.7888835-0.9302691i
## [4]  0.7888835+0.9302691i
```

```
## [1] 1.013788+0i
```

Some SAR polynomial roots for Model 2 are inside of the unit circle while some are outisde of the unit circe. SMA polynomial roots are outside of the unit circle. Therefore Model 2 is not causal but it is invertible.

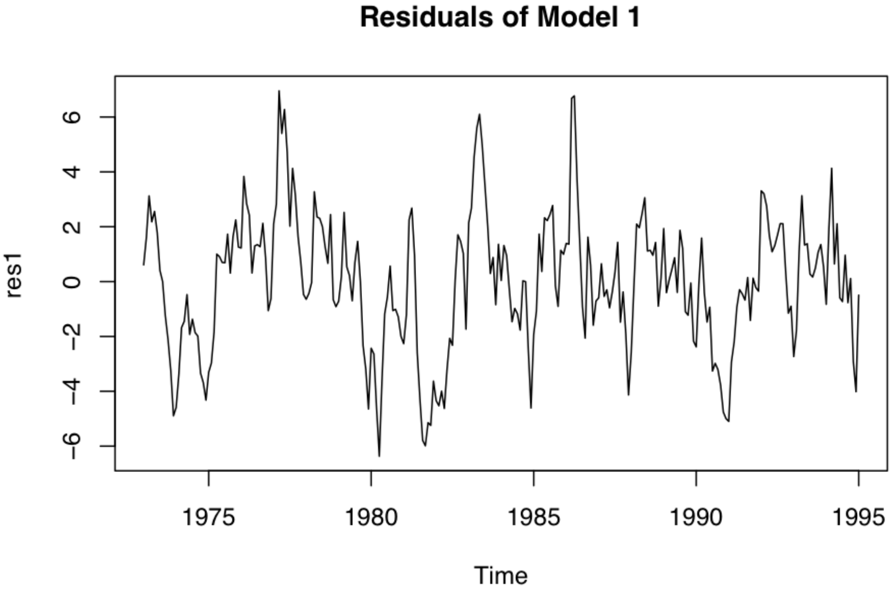
```
## [1]   1.14631+0i -20.05436+0i
```

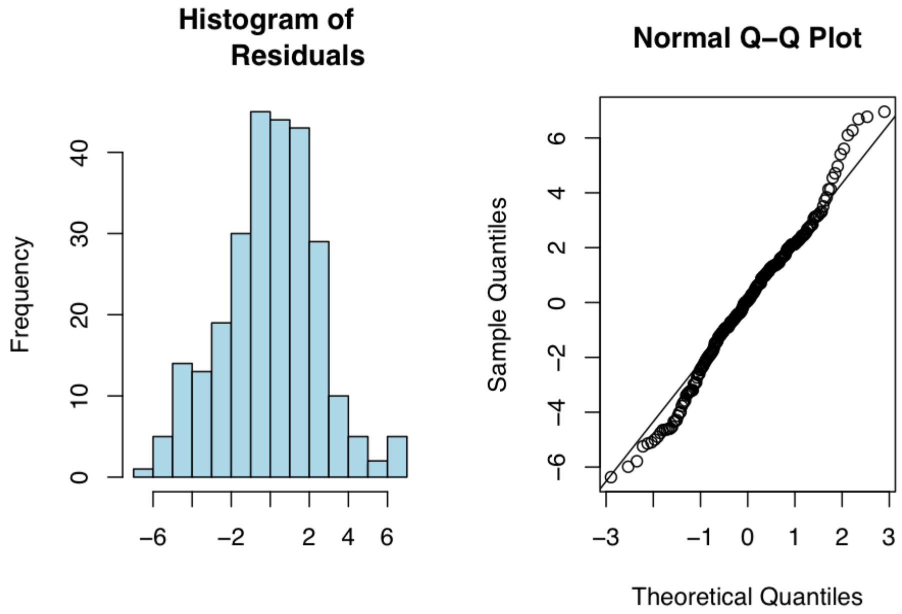
```
## [1]  1.357509-0i -2.504738+0i
```

```
## [1] -18.24818+0i
```

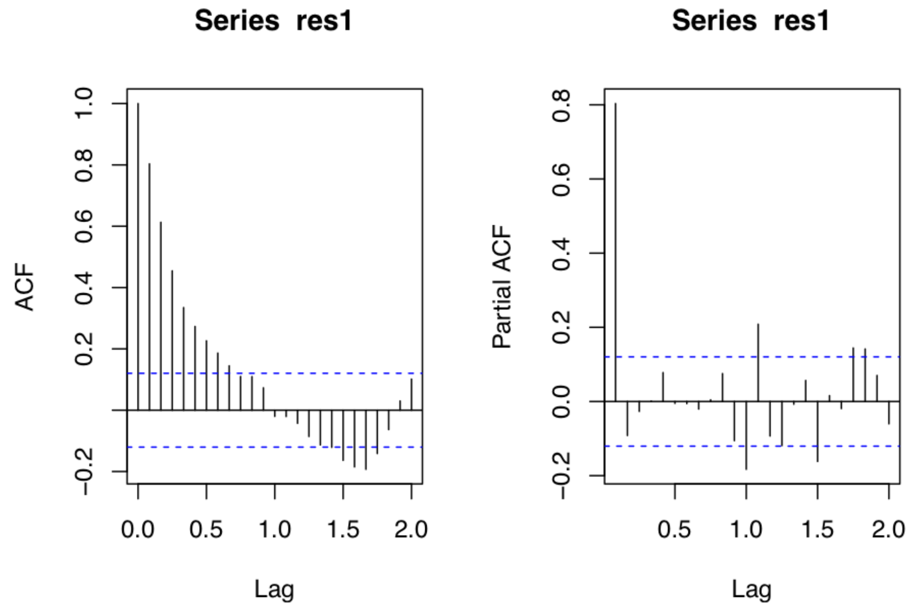
The AR, SAR, and SMA polynomial roots are all outside of the unit circle therefore Model 3 is both causal and invertible.

Next we analyze the residuals and perform diagnostic checks for each model.

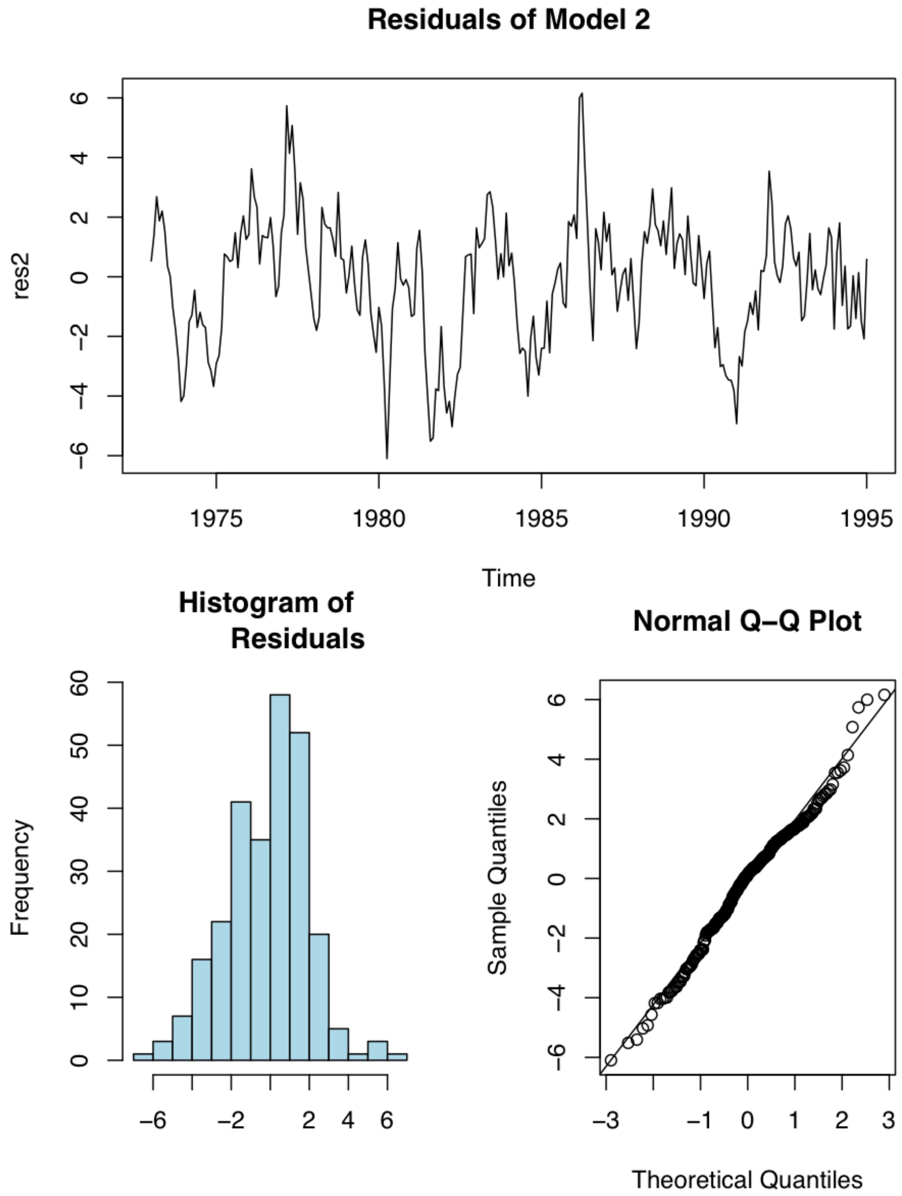




```
## [1] -0.02993651
## [1] 6.042932
##
##  Box-Pierce test
##
## data:  res1
## X-squared = 416.9, df = 16, p-value < 2.2e-16
##
##  Box-Ljung test
##
## data:  res1
## X-squared = 424.35, df = 16, p-value < 2.2e-16
##
##  Box-Ljung test
##
## data:  res1^2
## X-squared = 188.79, df = 16, p-value < 2.2e-16
##
##  Shapiro-Wilk normality test
##
## data:  res1
## W = 0.98895, p-value = 0.04085
```



We see that the plot of the residuals seems to display a bit of a trend, change of variance, and seasonality. The plot of the histogram is symmetric and the QQ plot seems to be normally distributed. However the first model does not pass the Box-Pierce, Ljung Box, Mcleod-Li, and Shapiro-Wilkes test since the p-values are smaller than .05. The ACF and PACF of the residuals are not within the confidence intervals so they do not resemble white noise. Since Model 1 does not pass most of the diagnostic checks and the plots do not resemble white noise, we cannot use Model 1 to forecast our data. Thus we omit our first model as a possible candidate.



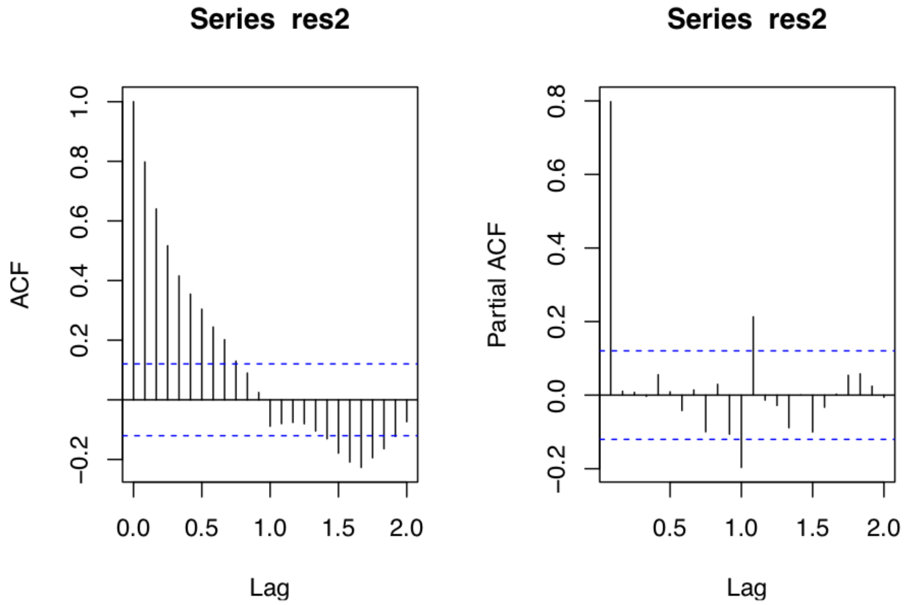
```
## [1] -0.1600864
## [1] 4.354718
##
##  Box-Pierce test
```

```
##
## data:  res2
## X-squared = 494.73, df = 16, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  res2
## X-squared = 504.26, df = 16, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  res2^2
## X-squared = 156.06, df = 16, p-value < 2.2e-16
```

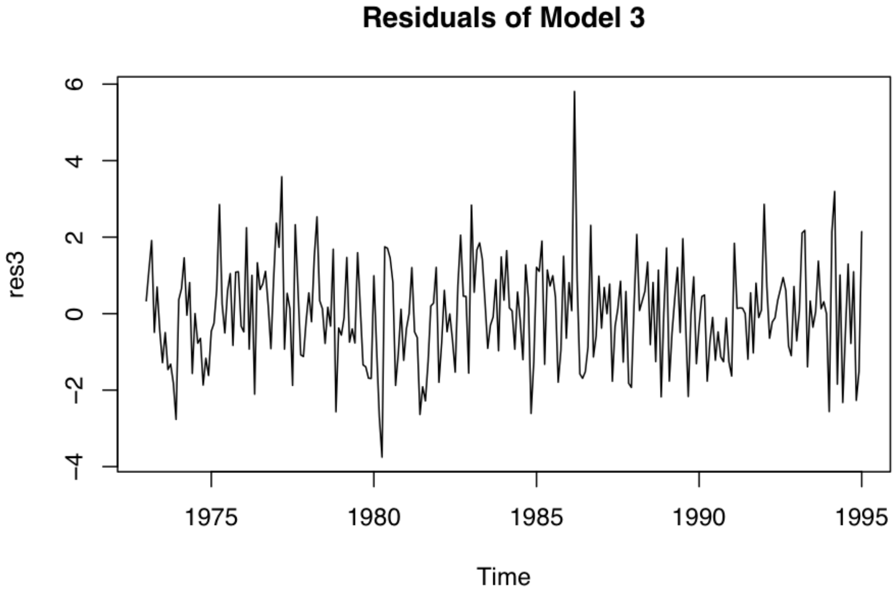
```
##
##  Shapiro-Wilk normality test
##
## data:  res2
## W = 0.98667, p-value = 0.01462
```

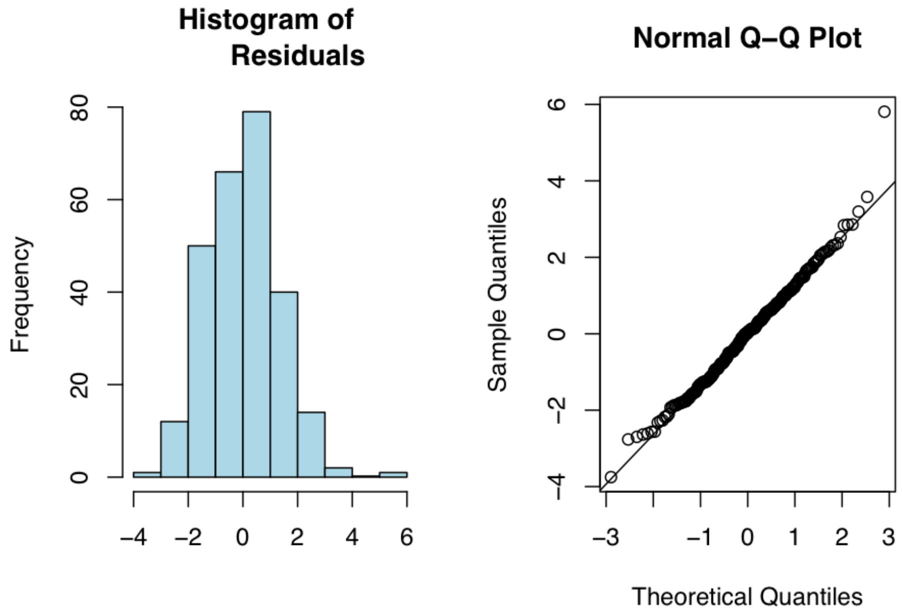


Plotting the residuals we see a bit of a trend, change of variance, and seasonality. The histogram is not as symmetric and but the QQ plot appears to be normally distributed. Model 2 does not pass the Box-Pierce, Ljung-Box, Shapiro Wilk, and Mcleod-Li test since p-values are less than .05.

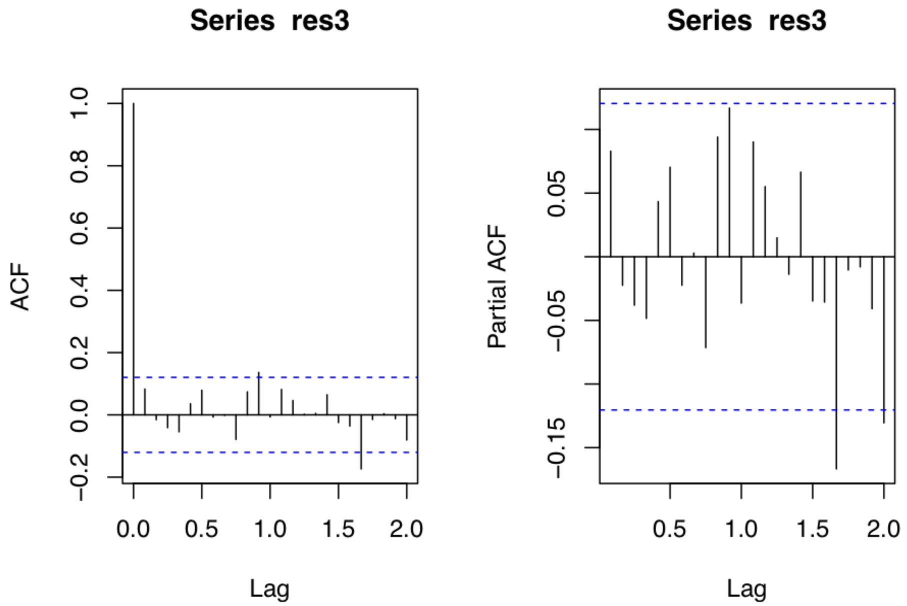
The lags of both the ACF and PACF extend beyond the confidence intervals and do not resemble white noise.

Although the QQ plot appears to be normally distributed, Model 2 does not pass the diagnostic checks and it is not causal. Since causality is necessary to forecast future values, we omit our second model as a possible candidate.





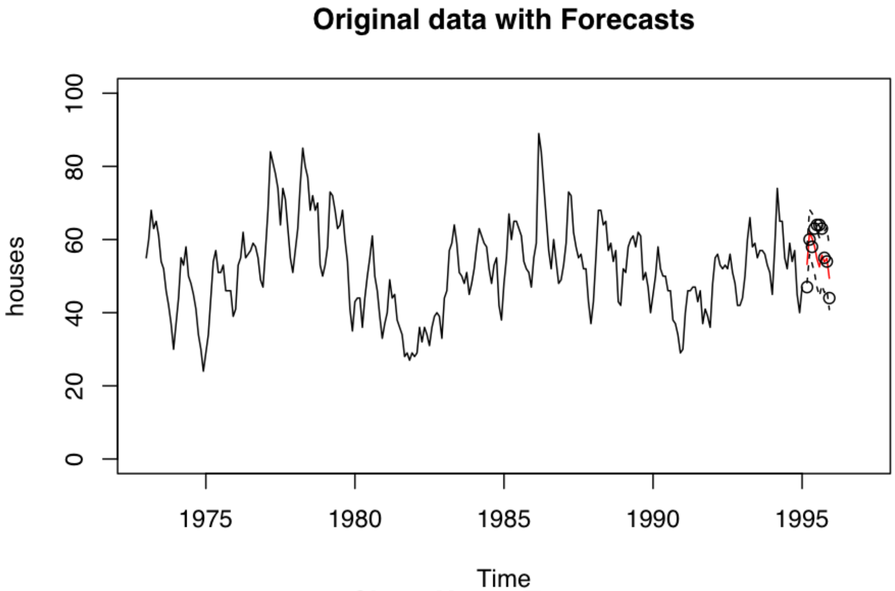
```
## [1] -0.0008290847
## [1] 1.747655
##
##  Box-Pierce test
##
## data:  res3
## X-squared = 15.479, df = 14, p-value = 0.3462
##
##  Box-Ljung test
##
## data:  res3
## X-squared = 16.121, df = 14, p-value = 0.3061
##
##  Box-Ljung test
##
## data:  res3^2
## X-squared = 6.3433, df = 16, p-value = 0.9839
##
##  Shapiro-Wilk normality test
##
## data:  res3
## W = 0.9897, p-value = 0.05726
```

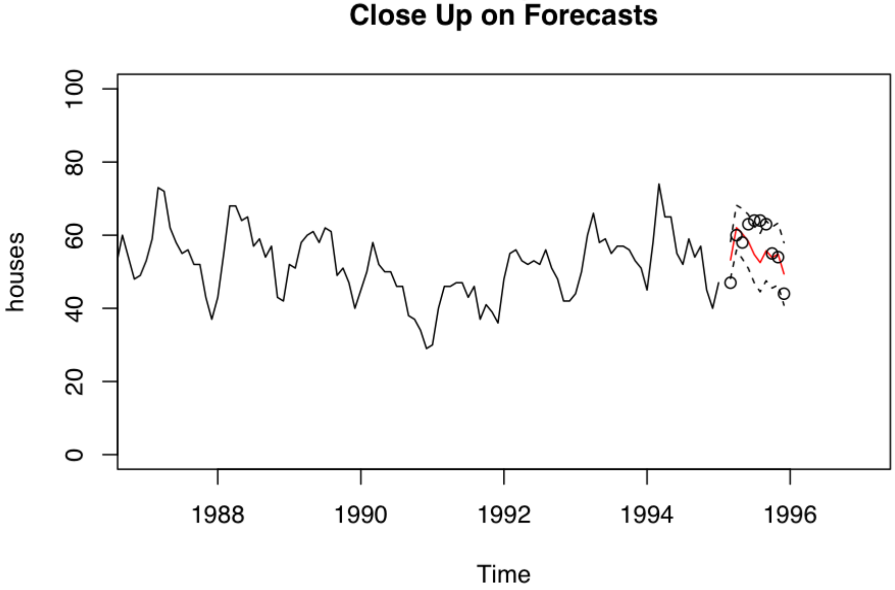


Residuals for Model 3 display no trend, change in variance, nor seasonal component. Both our histogram and QQ plot appear to be normally distributed. Performing diagnostic checks, Model 3 does not pass Box-Pierce nor Ljung-Box but it does pass the Shapiro Wilk, and Mcleod-Li test with p-values greater than .05.

The ACF extends slightly beyond the confidence interval around lag 1.7 but aside from this resemble white noise.

In addition, Model 3 has the lowest AIC of the three models and it is both causal and invertible. Thus we proceed with this model and conclude that Model 3 is the best model to use for forecasting.





I forecasted 10 points with Model 3 along with confidence intervals. The original points are dots and the forecasted points are a red line. Since the true values are within the confidence intervals of the forecasted values, it is clear that Model 3 accurately forecasts the data.

Conclusion

Model 3. SARIMA $(2,0,0)$ x $(2,0,1)_{12}$

- $(X_t - 0.8225X_{t-1} - 0.0435X_{t-2})(X_t - 0.3374X_{t-12} - 0.2941X_{t-24})\nabla_{12}\nabla X_t = (Z_t + 0.0548Z_{t-12})$

Model 3 was used to forecast future sales of new one family households since it had the lowest AIC, the residuals resembled white noise, and it passed most of the diagnostic checks. We see that it has successfully forecasted values close to the true values of the data set. Thus our model met Box-Jenkins assumptions for forecasting time series and our goal was achieved.

Acknowledgements

I would like to express my gratitude to a former classmate Annika Tan and the TA Nhan Huynh for providing me assistance with endeavors throughout this project. I would also like to thank my professor Raya Feldman for providing information and equipping me with the necessary tools to complete this project.

References

- Data Set Link

Appendix

```r
#Import Data
houses.csv = read.table("monthly-sales-of-new-onefamily-h.csv"
                        , sep=",", header=FALSE, skip=1, nrows=265 )
#Create TS & Plot TS
houses = ts(houses.csv[,2], start = c(1973,1), frequency = 12)
ts.plot(houses,main = "Original data")

#BC Transformation on data
library(MASS)
t = 1:length(houses)
fit = lm(houses ~ t)
bcTransform = boxcox(houses ~ t,plotit = TRUE)

lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
houses.bc = (1/lambda)*(houses^lambda-1)

#Plot transformed TS
ts.plot(houses,main = "Original data",ylab = expression(X[t]))
ts.plot(houses.bc,main = "Box-Cox tranformed data", ylab = expression(Y[t]))


#Compare Var
var(houses)
var(houses.bc)


#Plotting ACF/PACF of transformed data
{op = par(mfrow = c(1,2))
  acf(houses.bc,lag.max = 60,main = "")
  pacf(houses.bc,lag.max = 60,main = "")
  title("Box-Cox Transformed Time Series", line = -1, outer=TRUE)
  par(op)}


#Differencing data at lag 1
y1 = diff(houses.bc, 1)
plot(y1,main = "De-trended Time Series",ylab = expression(nabla~Y[t]))

#Differencing data at lag 12
y12 = diff(y1, 12)
ts.plot(y12,main = "De-trended/seasonalized Time Series",ylab
        = expression(nabla^{12}~nabla~Y[t]))

#Plot ACF/PACF of differenced data
acf(y12,lag.max = 100,main = "")
pacf(y12,lag.max = 100,main = "")
title("De-trended/seasonalized Time Series")

#Candidate Models
fit1<-arima(houses.bc, order=c(0,0,0), seasonal=list(order=c(2,0,1), period=12))
```

```r
fit2<-arima(houses.bc, order=c(0,0,0), seasonal=list(order=c(4,0,1), period=12))

fit3<-arima(houses.bc, order=c(2,0,0), seasonal=list(order=c(2,0,1), period=12))


#Causal/Invertible Fit1
polyroot(c(1, -1.0058, 0.3863  )) #SAR2 for Model 1
polyroot(c(1, -0.2581 )) #SMA1 for Model 1


#Causal/Invertible Fit2
polyroot(c(1, -1.5084, 0.5952, 0.2843,-0.3710 )) #SAR4 for Model 2 ||
#Not stationary bc ar root inside unit circle
polyroot(c( 1, -0.9864)) #SMA1 for Model 2

#Causal/Invertible Fit3
polyroot(c(1, -0.8225,  -0.0435)) #AR2 for Model 3
polyroot(c(1, -0.3374,-0.2941)) #SAR2 for Model 3
polyroot(c(1, 0.0548 )) #SMA1 for Model 3



#Residuals of first model
res1<-residuals(fit1)
plot(res1, main="Residuals of Model 1")

{op = par(mfrow = c(1,2))
  hist(res1, col="light blue", xlab="", main="Histogram of
       Residuals")
  qqnorm(res1);qqline(res1)
  par(op)}
mean(res1)
var(res1)

#Diagnostic checking for Model 1
Box.test(res1,lag=16, type = c("Box-Pierce"), fitdf=0)
Box.test(res1,lag=16, type = "Ljung", fitdf=0)
Box.test(res1^2, lag=16, type = "Ljung", fitdf=0)
#McLeod-Li/Quadratic correlation test
shapiro.test(res1)

{op = par(mfrow = c(1,2))
  acf(res1)
  pacf(res1)
  par(op)}



#Residuals of second model
res2<-residuals(fit2)
plot(res2, main="Residuals of Model 2")
{op = par(mfrow = c(1,2))
  hist(res2, col="light blue", xlab="", main="Histogram of
       Residuals")
  qqnorm(res2);qqline(res2)
  par(op)}
mean(res2)
var(res2)

#Diagnostic checking for Model 2
Box.test(res2,lag=16, type = c("Box-Pierce"), fitdf=0)
Box.test(res2,lag=16, type = "Ljung", fitdf=0)
Box.test(res2^2,lag=16, type = "Ljung", fitdf=0)
shapiro.test(res2)

{op = par(mfrow = c(1,2))
  acf(res2)
  pacf(res2)
  par(op)}



#Residuals of third model
res3<-residuals(fit3)
plot(res3, main="Residuals of Model 3")
{op = par(mfrow = c(1,2))
  hist(res3, col="light blue", xlab="", main="Histogram of
       Residuals")
  qqnorm(res3);qqline(res3)
  par(op)}

mean(res3)
var(res3)

#Diagnostic checking for Model 3
Box.test(res3, lag=16,type = c("Box-Pierce"), fitdf=2)
Box.test(res3, lag=16,type = "Ljung", fitdf=2)
Box.test(res3^2, lag=16,type = "Ljung", fitdf=0)
shapiro.test(res3)
{op = par(mfrow = c(1,2))
  acf(res3)
  pacf(res3)
  par(op)}



#True Points
x.true = c("1995-02","1995-03","1995-04","1995-05","1995-06","1995-07",
           "1995-08","1995-09","1995-10","1995-11")
y.true=c(47,60,58,63,64,64,63,55,54,44)

#Forecasting data points and CI using Model 3
pred <- predict(fit3, n.ahead=10)
newpred<-(2/3*pred$pred+1)^3/2
{ts.plot(houses,xlim=c(1973,1997), ylim=c(0,100), main =
           "Original data with Forecasts")
```

```
space=(1995-1973)/length(houses)

index=2:11*space
indextoadd=1995+index
points(indextoadd,pred$pred,type="l",col="red")

lines(indextoadd,pred$pred-1.96*pred$se,lty="dashed")
lines(indextoadd,pred$pred+1.96*pred$se,lty="dashed")
points(indextoadd, y.true)
points(x.true, y.true, pch = "*", col = "blue")}
```