CSC 425 Homework 3
Sarah Cummings

Problem 2)

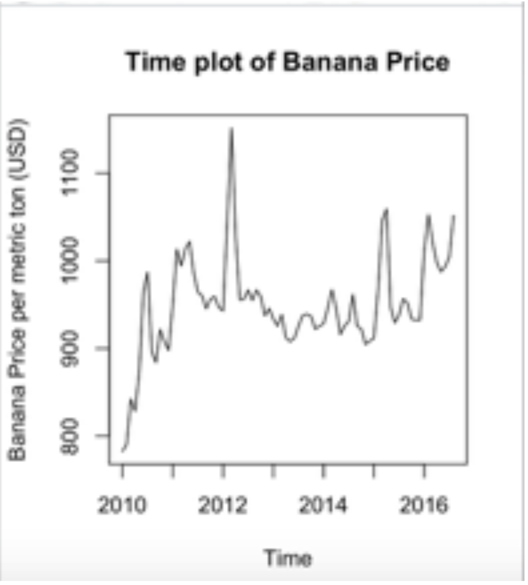
a. Explain why the MA(q) models are regarded as finite memory models
In an MA(q) model, points that are more than q units apart are uncorrelated. The memory is hence "finite" because past a certain point, the past data has no affect on the present and future data.

b. Discuss the behaviors of the PAFT function for MA(q) and AR(p) process
For a MA(q) model, the autocovaraice function cuts off beyond lag q. For an AR(p) model, the auto covariance function decays exponentially without cutting off.

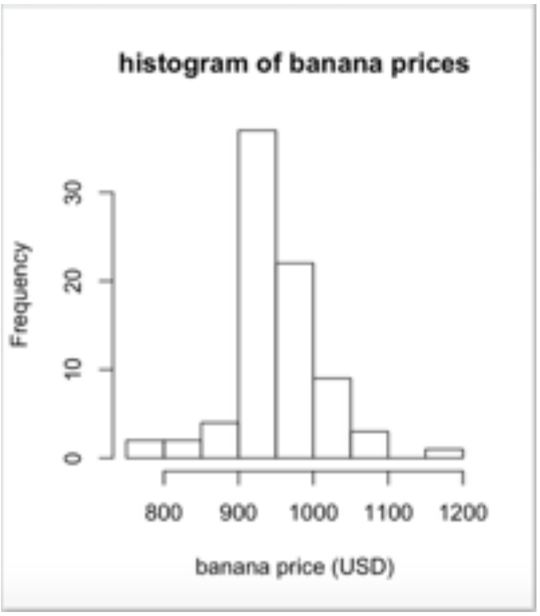
c. Given a timeseries of 100 observations with the following auto correlations:
p1=-0.49, p2=0.31, p3=0.21, p4-0.11 and |pk|<0.01 for k>4. What model would you pick for this series?
I would pick an AR model, given the quick decay to zero. Most likely AR(4).

Problem 3)

a. Time plot for the price data, and discussion of trends and variability

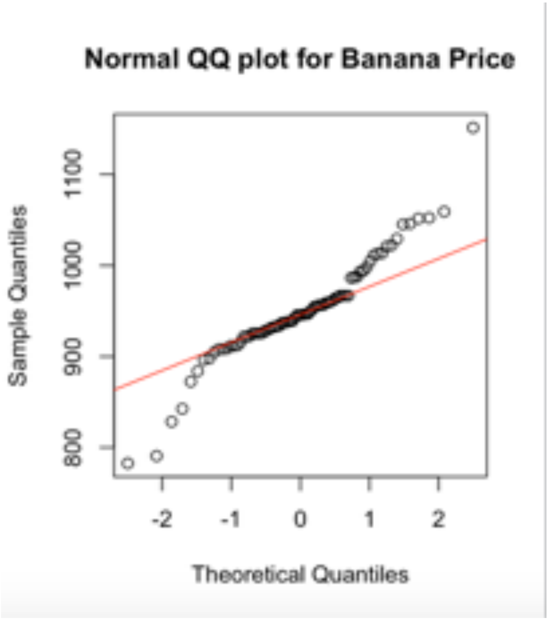


The time plot shows that the banana price ranges from <$800 to almost $1200. Most values appear to be in the $900 to $1050 range, though the pattern is inconsistent. In general, it looks like the price is increasing.



b. Analyze distribution of prices plus discussion

As seen in the histogram, the price for bananas is not normally distributed. We have the following



**summary statistics: Min: 782.69, Max 1151.43, 1. Quartile: 925.41, 3. Quartile: 966.85, Mean: 950.277250
Median: 945.650000**

Looking at the normal qq plot, we see the distribution of banana price has heavy tails—our data has more extreme values than would be expected for a normal distribution.

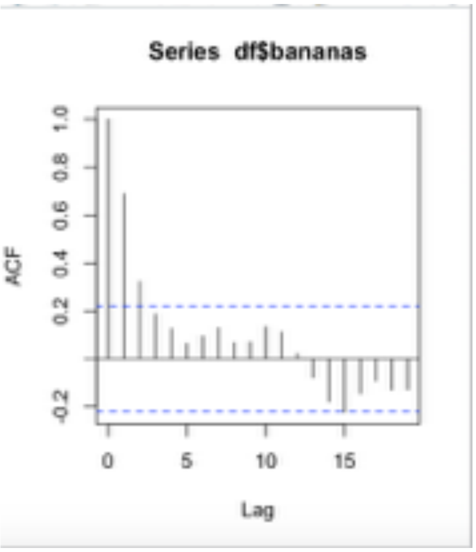
c. Evaluate if price has stationary behavior with analysis

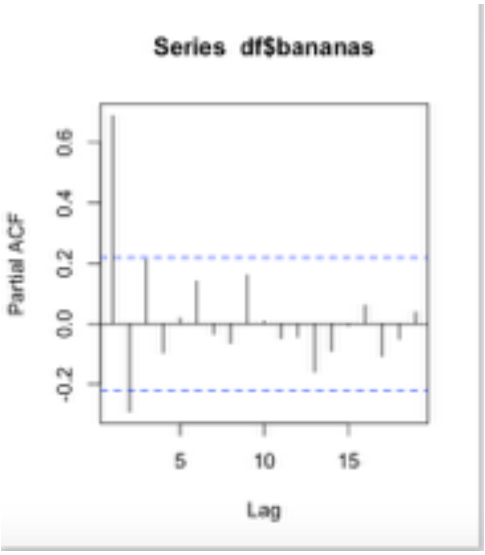
Given the small p values of both of our Box Ljung tests, we reject the null hypothesis of independence and conclude the change rates are serially correlated.

To analyze for stationary behavior, we will use the box Ljung test, and later confirm with ACF and PACF tests

```
Box-Ljung test
data:  priceTS
X-squared = 51.176, df = 3, p-value = 4.487e-11

Box-Ljung test
data:  priceTS
X-squared = 53.686, df = 6, p-value = 8.536e-10
```

d. Analysis of ACF and PACF plots. Does the process show clear AR behavior or MA behavior?



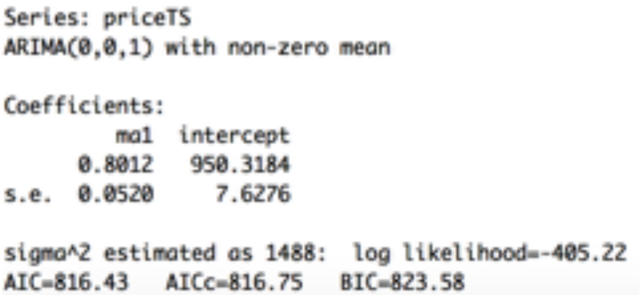


Looking at the plots, this function seems to show more MA behavior than AR. Rather than having a continual decay to zero in the ACF function, it decays to lag 5 and then goes back up. The PACF decays toward zero as is typical for the MA process.

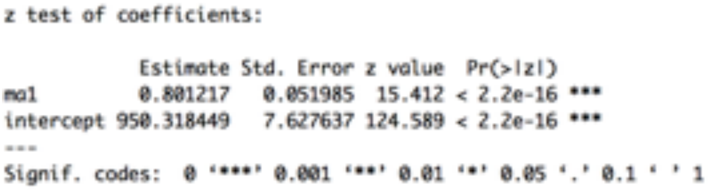
e. Fit an analyze a MA model for the data

I tried an MA(1) and an MA(2) for this data and they both seemed to be fairly suitable

MA(1):

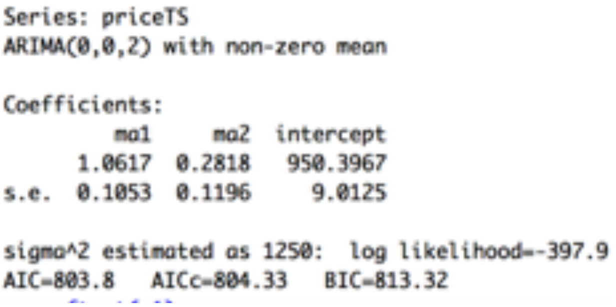


```
Series: priceTS
ARIMA(0,0,1) with non-zero mean

Coefficients:
         ma1  intercept
      0.8012   950.3184
s.e.  0.0520     7.6276

sigma^2 estimated as 1488:  log likelihood=-405.22
AIC=816.43   AICc=816.75   BIC=823.58
```

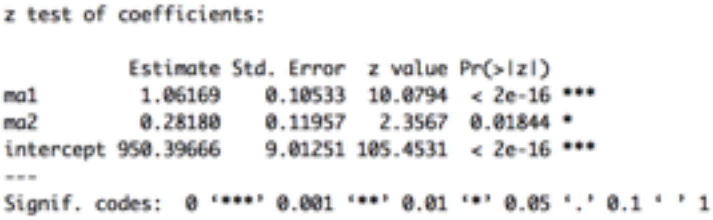


```
z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)    
ma1        0.801217   0.051985  15.412 < 2.2e-16 ***
intercept 950.318449   7.627637 124.589 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MA(2):



```
Series: priceTS
ARIMA(0,0,2) with non-zero mean

Coefficients:
         ma1     ma2  intercept
      1.0617  0.2818   950.3967
s.e.  0.1053  0.1196     9.0125

sigma^2 estimated as 1250:  log likelihood=-397.9
AIC=803.8   AICc=804.33   BIC=813.32
```



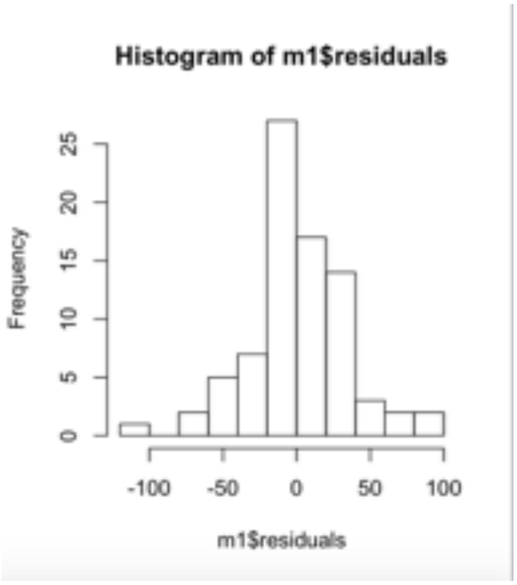
```
z test of coefficients:

           Estimate Std. Error  z value Pr(>|z|)    
ma1         1.06169    0.10533  10.0794  < 2e-16 ***
ma2         0.28180    0.11957   2.3567  0.01844 *  
intercept 950.39666    9.01251 105.4531  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

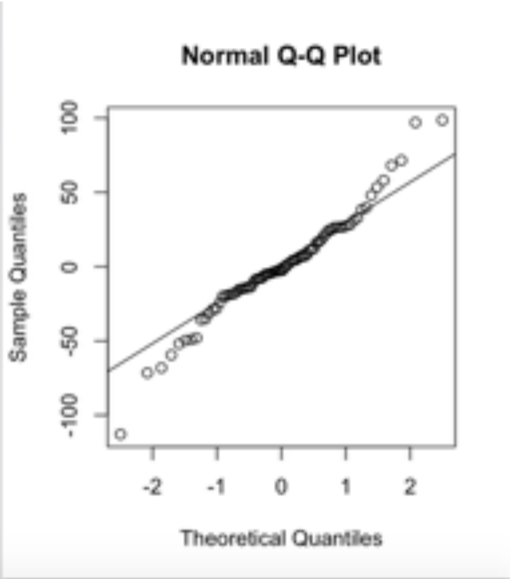
Residual analysis for MA(2):

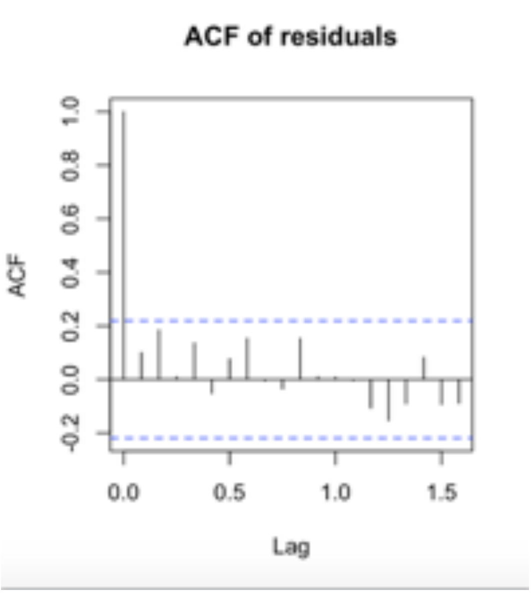
Box-Ljung test
data: m1$residuals
X-squared = 6.0767, df = 4, p-value = 0.1935

Box-Ljung test
data: m1$residuals
X-squared = 10.52, df = 8, p-value = 0.2304

Box-Ljung test
data: m1$residuals
X-squared = 10.538, df = 10, p-value = 0.3947

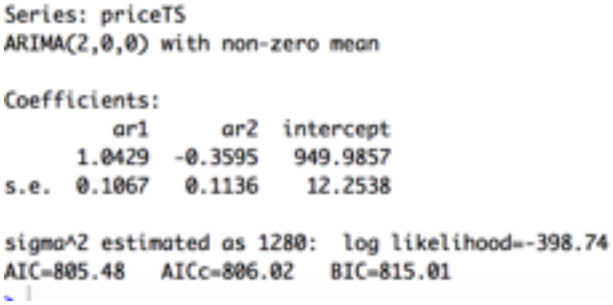




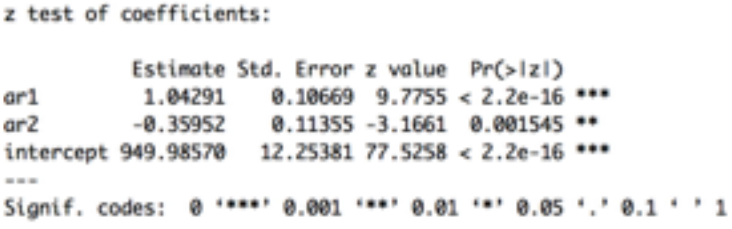


Looking at the output for our coefficient tests, both models have all significant parameters. I would go with the MA(2) between the two of them, for a better fit. The ACF values of residuals are all small and statistically different from zero. The LB Box test for residuals unfortunately does not reject the null hypothesis of independence. This model provides a fairly good fit but it could be better.

f. Fit an AR model to the data and analyze



```
Series: priceTS
ARIMA(2,0,0) with non-zero mean

Coefficients:
         ar1      ar2  intercept
      1.0429  -0.3595   949.9857
s.e.  0.1067   0.1136    12.2538

sigma^2 estimated as 1280:  log likelihood=-398.74
AIC=805.48   AICc=806.02   BIC=815.01
```

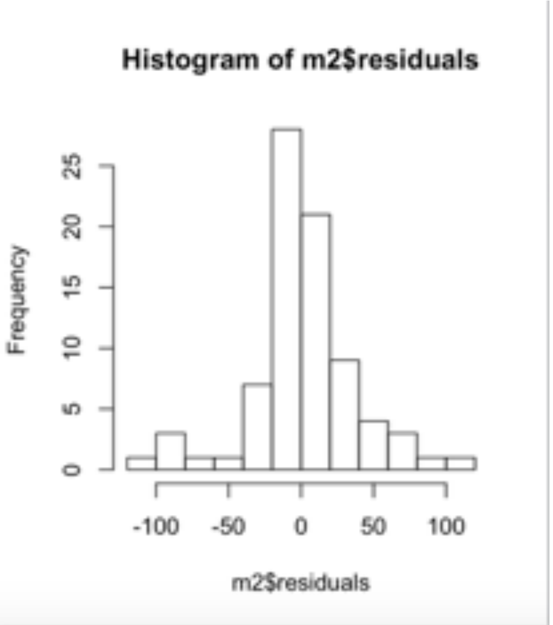


```
z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)    
ar1         1.04291    0.10669  9.7755 < 2.2e-16 ***
ar2        -0.35952    0.11355 -3.1661  0.001545 ** 
intercept 949.98570   12.25381 77.5258 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

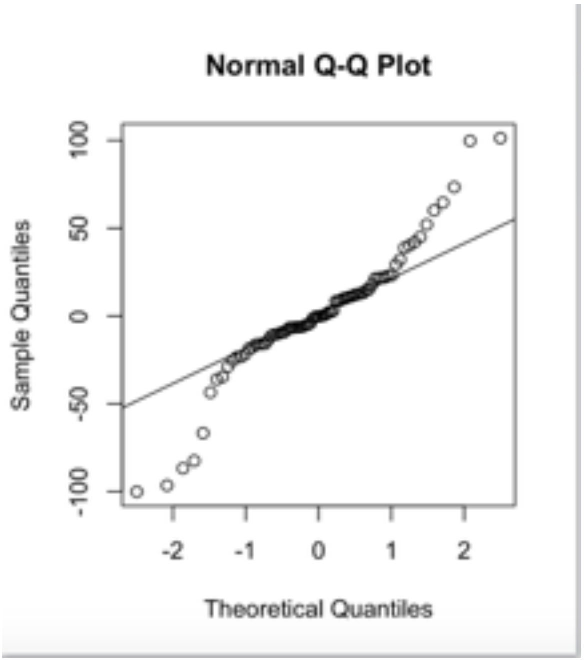
Residual Analysis:

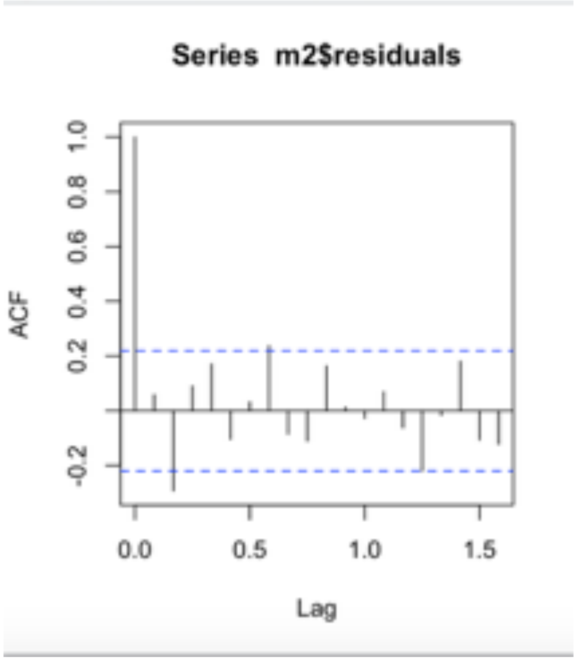
Box-Ljung test
data: m2$residuals
X-squared = 11.75, df = 4, p-value = 0.01931

Box-Ljung test
data: m2$residuals
X-squared = 21.168, df = 8, p-value = 0.006713



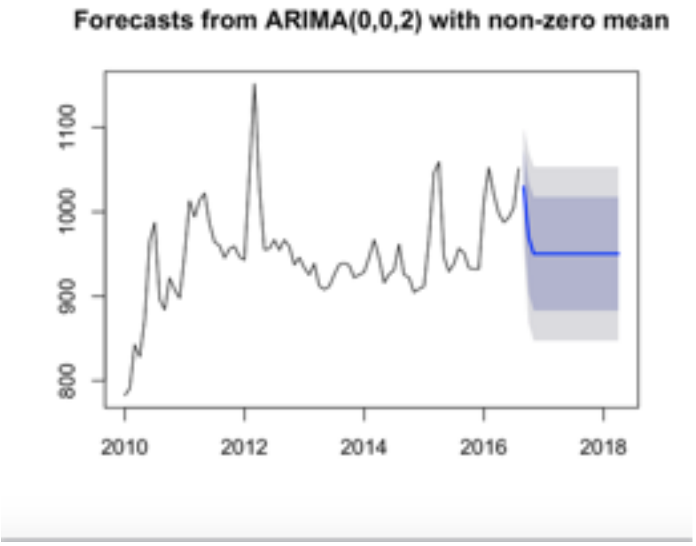
```
Box-Ljung test
data:  m2$residuals
X-squared = 21.258, df = 10, p-value = 0.01936
```

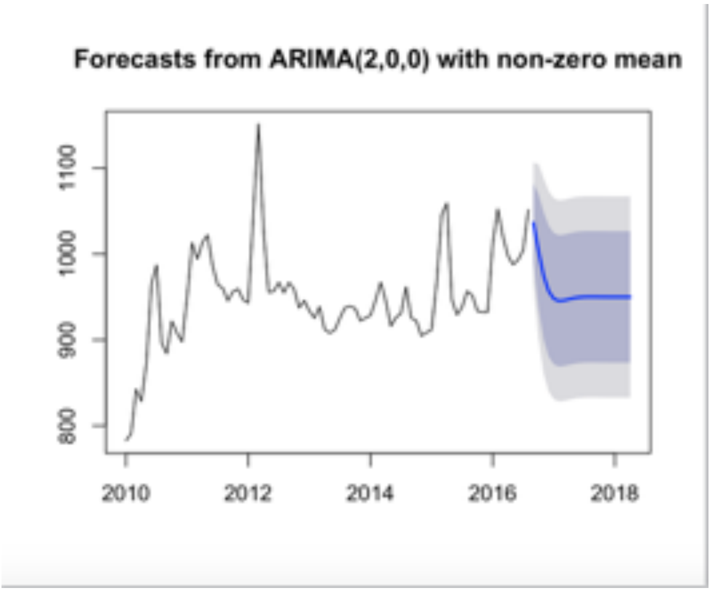




For the AR model, all the parameters are significant and the LB box tests rejects the null hypothesis of independence. Unfortunately the residuals are not normally distributed, and the residual at lag two is significantly different from zero. This model provides an ok fit but could be better.

g. Forecasts for both models and comparison of behavior: MA model left, AR model at right



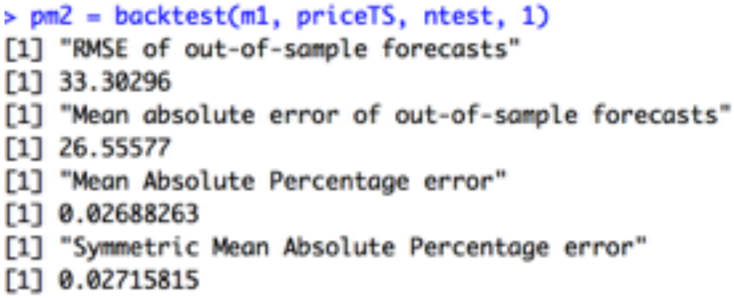


The forecasts for these models are very similar, though the width for the prediction interval of the AR model is larger. In general the behavior is the same, though the 2017 point is little lower in the AR model.

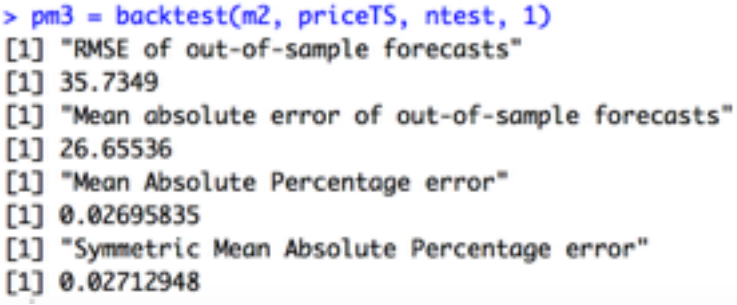
h. What model is selected using the BIC criterion?

The MA. The BIC is lower (813.32) for the MA than for the AR(815.01). With autoarima, though, the best model according to BIC is ARIMA(1,0,1)

i. Backtesting procedure using 80% for training and 20% for testing for each model, with comparison of out-of-sample accuracy of the forecasts for the two models,.



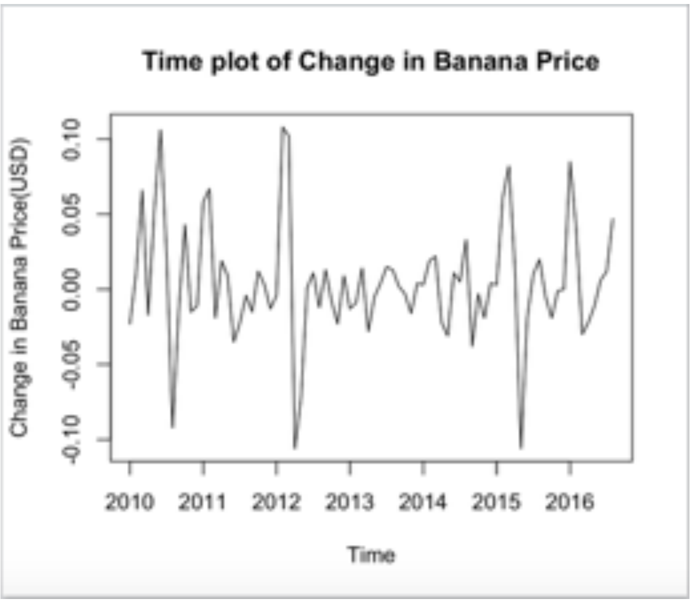
```
> pm2 = backtest(m1, priceTS, ntest, 1)
[1] "RMSE of out-of-sample forecasts"
[1] 33.30296
[1] "Mean absolute error of out-of-sample forecasts"
[1] 26.55577
[1] "Mean Absolute Percentage error"
[1] 0.02688263
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.02715815
```



```
> pm3 = backtest(m2, priceTS, ntest, 1)
[1] "RMSE of out-of-sample forecasts"
[1] 35.7349
[1] "Mean absolute error of out-of-sample forecasts"
[1] 26.65536
[1] "Mean Absolute Percentage error"
[1] 0.02695835
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.02712948
```

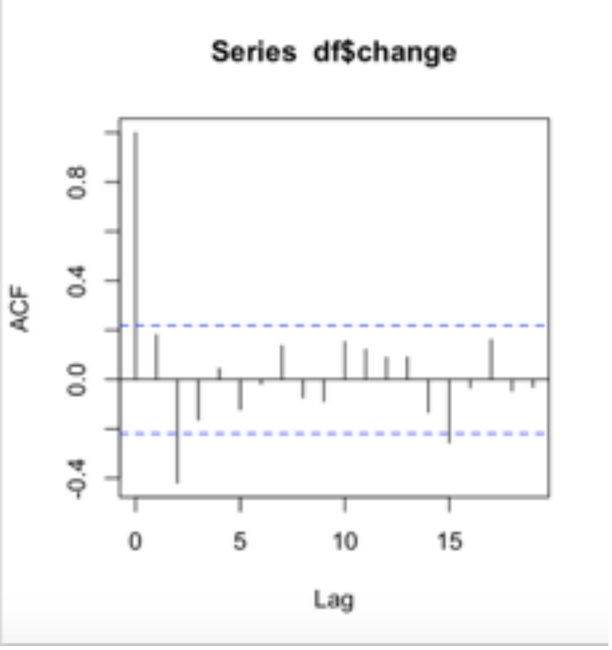
Our MA model (backtest above left) has a smaller RMSE for out-of sample forecasts than the AR model, at 33.30296 versus the 35.7349 for the AR model. The MA model also has a smaller mead absolute error than the AR model.

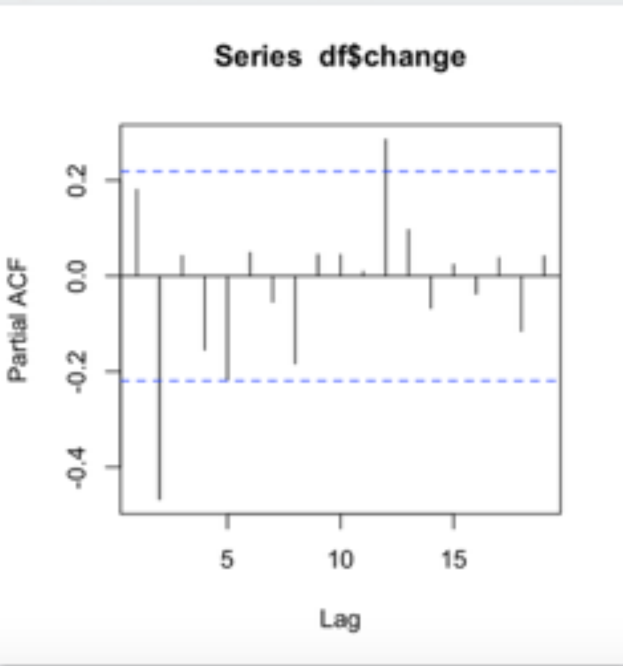
j. Time plot for price change



The change in banana price ranges from -0.10 to 0.10. It appears as though the mean is at 0, with consistent variability throughout.

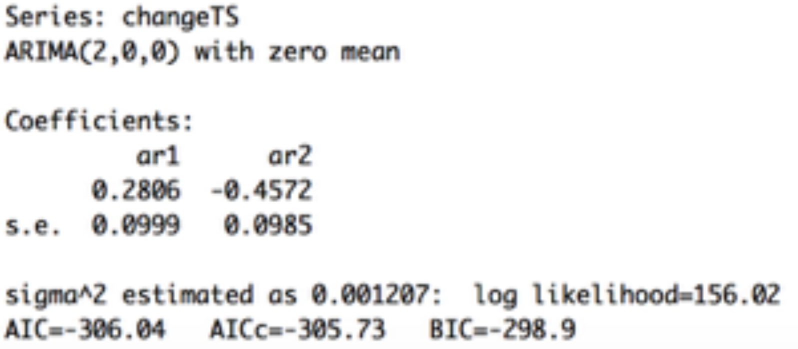
k. Analysis of PACF and ACF plots



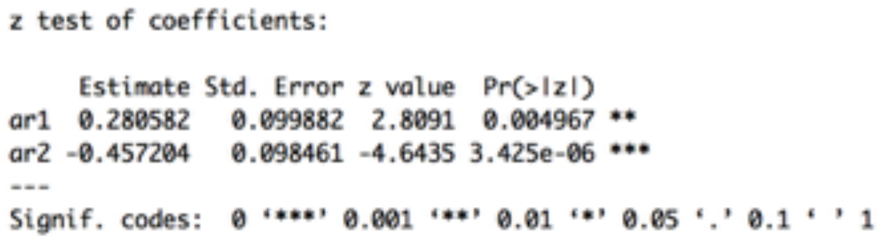


Looking at the ACF, it looks like the model decays towards zero, although lag 2 and lag 15 are significantly different from zero. The PACFT has lag 12 and lag 1 significantly different than zero. The plots don’t clearly show which model would be best.

l. AR model for the price change data. Remove any parameter that is not significant, using the FIXED option in the Arima() function. Write down the final AR model.



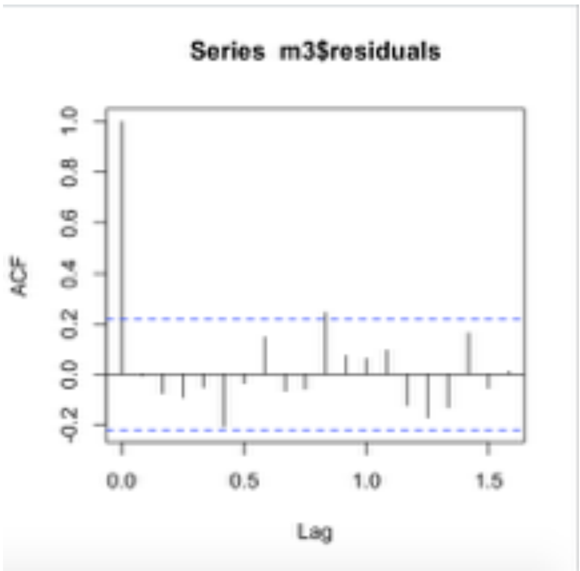
```
Series: changeTS
ARIMA(2,0,0) with zero mean

Coefficients:
         ar1      ar2
      0.2806  -0.4572
s.e.  0.0999   0.0985

sigma^2 estimated as 0.001207:  log likelihood=156.02
AIC=-306.04   AICc=-305.73   BIC=-298.9
```

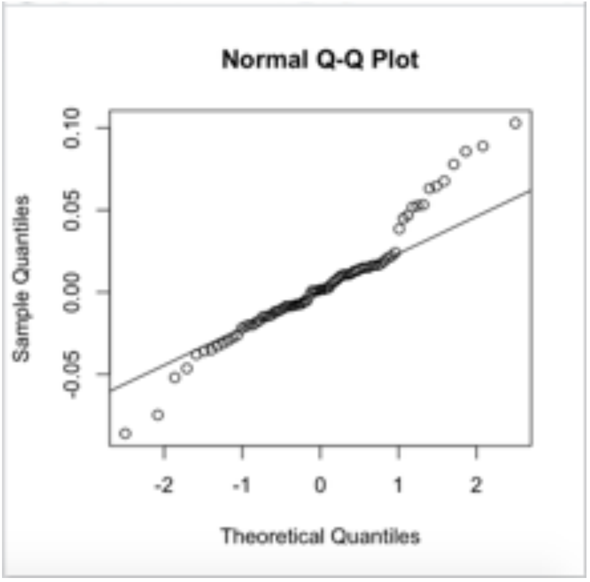


```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1  0.280582   0.099882  2.8091  0.004967 **
ar2 -0.457204   0.098461 -4.6435 3.425e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Final AR model: $\mathbf{X_t=0.2806\,(X_{t-1}) - 0.4572(X_{t-2}) + a_t}$

m. Analysis of Residuals





Box-Ljung test

data: m3$residuals

X-squared = 5.3013, df = 4, p-value = 0.2578

Box-Ljung test

data: m3$residuals

X-squared = 13.549, df = 8, p-value = 0.09429

Box-Ljung test

```
data:  m3$residuals

X-squared = 14.441, df = 10, p-value = 0.1538
```

For this AR model, all the parameters are significant. Unfortunately the residuals are not normally distributed, and there are some residuals significantly different from zero. This model provides an ok fit, but could be better.

Code:

```
library(fBasics)
library(tseries)
library(zoo)
library(forecast)
library(lmtest)


#set working directory
setwd("/Users/sarahcummings/Documents/csc425")
df<-read.csv('banana_price.csv')
head(df)


#create two timeseries objects
priceTS<-ts(df$bananas,start = c(2010,1),end = c(2016,8), frequency = 12)
changeTS<-ts(df$change,start = c(2010,1),end = c(2016,8), frequency = 12)

#timeplot of Price TS
plot(priceTS,ylab='Banana Price per metric ton (USD)', main= "Time plot of Banana
        Price")
#distibution of price histogram
hist(df$bananas,xlab = 'banana price (USD)', main= "histogram of banana prices")

basicStats(df$bananas)
qqnorm(df$bananas, main = 'Normal QQ plot for Banana Price')
qqline(df$bananas, col = 2)

#analysis of stationary behavior
```

```
acf(df$bananas)
pacf(df$bananas)

#Box test
Box.test(priceTS,lag=3,type='Ljung')
Box.test(priceTS,lag=6,type='Ljung')

# apply a automated order selection procedure
auto.arima(priceTS, stationary=T, seasonal=T)


#Fit an MA model
m1= Arima(priceTS, order=c(0,0,2), method='ML', include.mean=T)
m1
# T-tests on coefficients
coeftest(m1)

# RESIDUAL ANALYSIS
Box.test(m1$residuals,lag=6,type='Ljung', fitdf=2)
Box.test(m1$residuals,lag=10,type='Ljung', fitdf=2)
Box.test(m1$residuals,lag=12,type='Ljung', fitdf=2)
acf(m1$residuals)

hist(m1$residuals)
qqnorm(m1$residuals)
qqline(m1$residuals)

acf(m1$resid, main="ACF of residuals")

#Fit an AR model
m2=Arima(priceTS, order=c(2,0,0))
m2
#Coeff test
coeftest(m2)

# RESIDUAL ANALYSIS
Box.test(m2$residuals,lag=6,type='Ljung', fitdf=2)
Box.test(m2$residuals,lag=10,type='Ljung', fitdf=2)
Box.test(m2$residuals,lag=12,type='Ljung', fitdf=2)
acf(m2$residuals)

hist(m2$residuals)
qqnorm(m2$residuals)
qqline(m2$residuals)

#Forecast
```

```
forecast.Arima(m1, h=10)
plot(forecast.Arima(m1, h=20))

forecast.Arima(m2, h=10)
plot(forecast.Arima(m2, h=20))

auto.arima(priceTS, max.P=8, max.Q=8, ic="bic")

#Backtesting
ntest=round(length(priceTS)*0.8)
source("backtest.R")
pm2 = backtest(m1, priceTS, ntest, 1)
pm3 = backtest(m2, priceTS, ntest, 1)

#Plot for change ts
plot(changeTS,ylab='Change in Banana Price(USD)', main= "Time plot of Change in
        Banana Price")

#Analysis of ACF and PACF for change
acf(df$change)
pacf(df$change)

m3<-Arima(changeTS, order=c(2,0,0), include.mean=F)
m3
coeftest(m3)

# RESIDUAL ANALYSIS
Box.test(m3$residuals,lag=6,type='Ljung', fitdf=2)
Box.test(m3$residuals,lag=10,type='Ljung', fitdf=2)
Box.test(m3$residuals,lag=12,type='Ljung', fitdf=2)
acf(m3$residuals)

hist(m3$residuals)
qqnorm(m3$residuals)
qqline(m3$residuals)
```