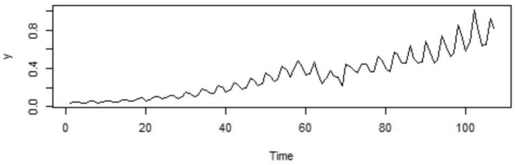
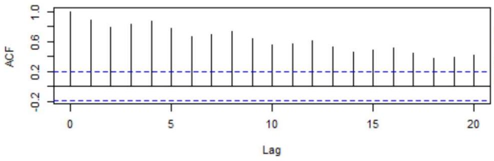
Forecasting Time Series - Assignment 2

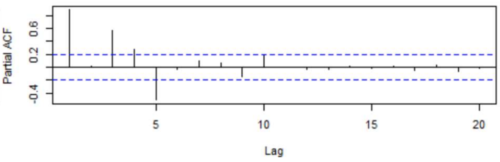
Group C

In this assignment we are analyzing Coca Cola's quarterly earnings and trying to create valid models to forecast future earnings. We created 5 models but will only explain the processes for models 1, 3, and 5.

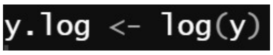
We will begin by plotting the data. It appears non stationary in the mean, as well as the variance (systematic change in variance as it grows over time, in a heteroscedastic manner). We also see a slow decay in the ACF, indicating it is not stationary. There also appears to be seasonality behaviour, which makes sense as we are analyzing quarterly earnings which may depend on the season.





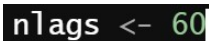


We perform the NDIFFS and NSDIFFS tests and confirm that both 1 regular transformation and 1 seasonal transformation are required. S = 4 because we are talking about quarterly data, which repeats every 4 periods. Because it is non-stationary in the variance, we should do a logarithmic transformation.



```
y.log <- log(y)
```

We now plot again and see that it is now stationary in the variance, unlike before. We check NDIFFS and NSDIFFS again and still see that 1 transformation for each is required. We also check the ACF/PACF and still see non-stationarity.

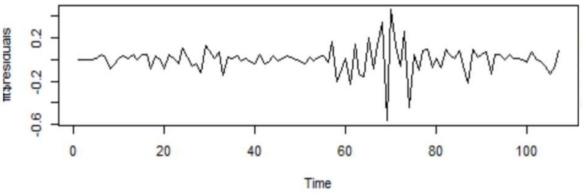


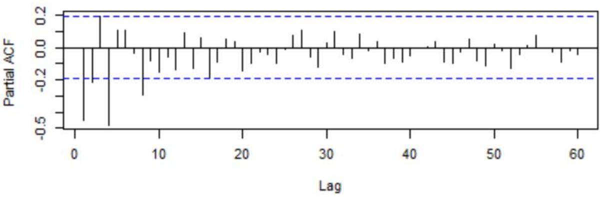
```
nlags <- 60
```

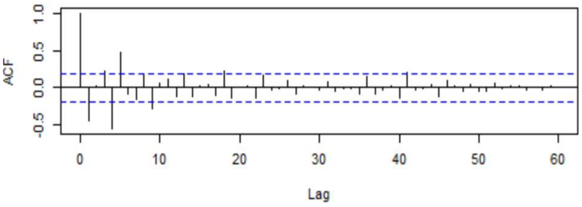
We will put nlags as 60 as this reflects 15 years of data which is reasonable.

<u>Model 1</u>

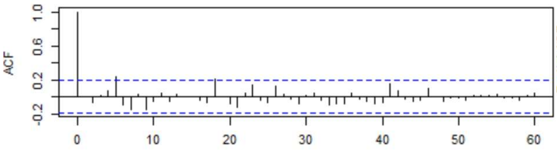
We will start with a SARIMA(0, 1, 0) x (0, 1, 0) 4 model. This is a good starting point as we know that 1 transformation is required for both.

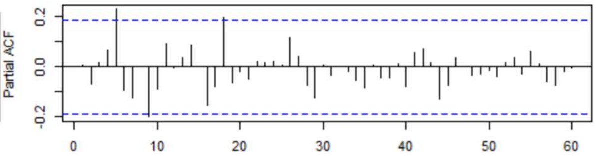




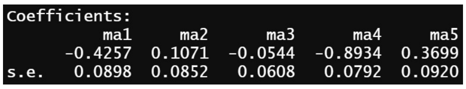


The ACF appears to be slightly simpler to analyze so we will start there by incorporating lag 5 and thus having a SARIMA(0, 1, 5) x (0, 1, 0) 4 model.





Lags 5, 9, and 18 appear to potentially be out of bounds in the PACF. MA5 is significant at a 95% confidence level because the absolute value of estimator/S.E. is greater than 1.96.

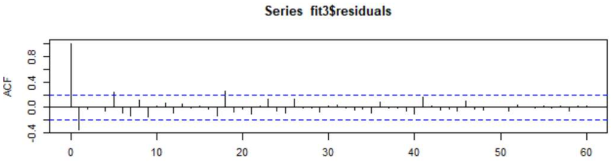


```
Coefficients:
          ma1     ma2      ma3      ma4     ma5
      -0.4257  0.1071  -0.0544  -0.8934  0.3699
s.e.   0.0898  0.0852   0.0608   0.0792  0.0920
```

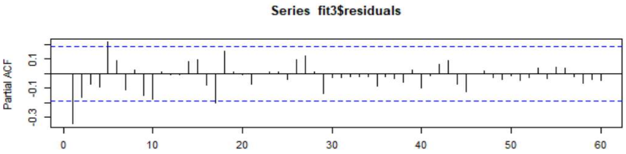
The Box-Test gives a p-value of 0.4739 which signals that we fail to reject the null hypothesis and our data is uncorrelated, and thus WN. Therefore, our linear model in the first moment cannot be improved and it is a valid model.

Model 3

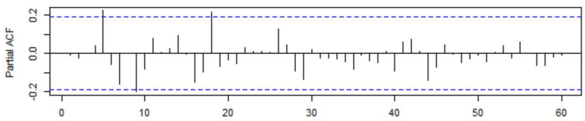
As we are aware that our initial test advised that we should do 1 regular transformation and 1 seasonal transformation. So we start with SARIMA(0, 1, 0) x (0, 1, 0) 4. We did this already in the prior model but this time we will focus on the seasonality, as ACF lag 4 is very out of bounds, and thus run a SARIMA(0, 1, 0) x (0, 1, 1) 4 and see the corresponding ACF/PACF below.

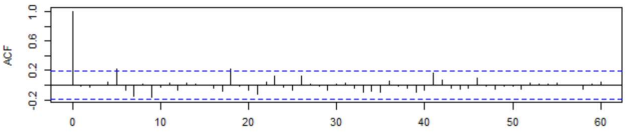


Looking at PACF lag 1 is quite out of bounds so we do one more adjustment by using a SARIMA(1, 1, 0) x (0, 1, 1) 4.



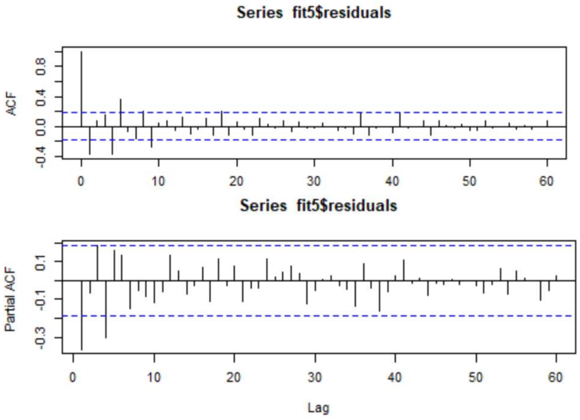
SARIMA(1, 1, 0) x (0, 1, 1) 4 is our final model. Looking at the ACF/PACF, all graphs are within bounds. The box test gives a p-value 0.527 and the data is not correlated so it is WN and a valid linear model..



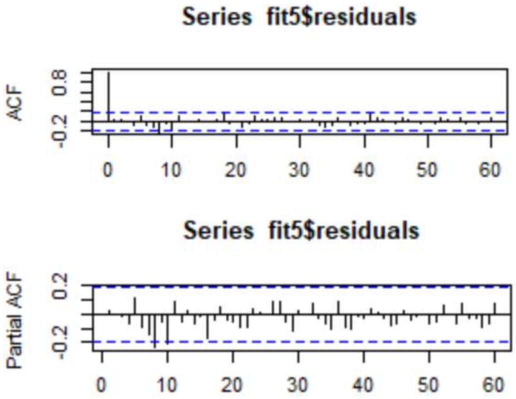


Model 5

The regular and seasonal parts can have interesting effects on each other. This time we began by using a SARIMA(5, 0, 0) x (0, 0, 0) 4 model due to our initial results where the AR(5) was out of bounds.



The resulting ACF indicates lag 4 is out of bounds so we will try with a SARIMA(5,0,0) x (1,0,0) 4

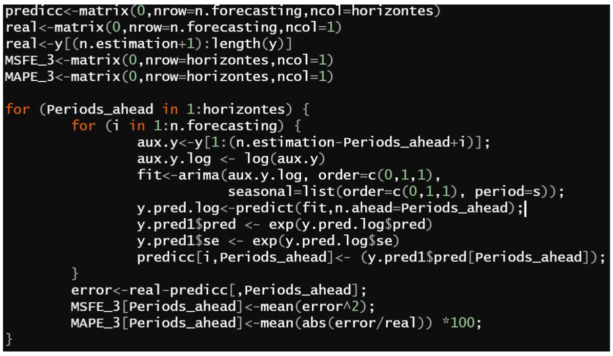


Our ACF is simpler and has lag 8 out of bounds so we will do the SARIMA(5, 0, 0) x (1, 0, 2) 4. We also receive an error stating that a transformation is required, and this makes sense, so we take a seasonal transformation as well.

We end up with a SARIMA(5,0,0) x (1, 1, 2) 4 model that has all lags within bounds and a p-value of 0.9783 is received by the Box-Test. This model is WN and cannot be improved further. It is a valid linear model.

Error Calculations

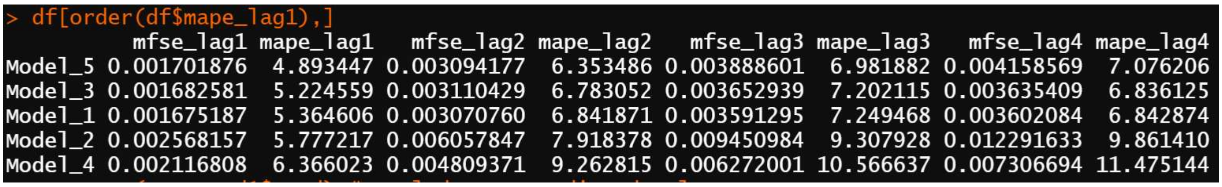
We calculated MFSE and MAPE scores using code shown in class. Here is a sample of code used.



```
predicc<-matrix(0,nrow=n.forecasting,ncol=horizontes)
real<-matrix(0,nrow=n.forecasting,ncol=1)
real<-y[(n.estimation+1):length(y)]
MSFE_3<-matrix(0,nrow=horizontes,ncol=1)
MAPE_3<-matrix(0,nrow=horizontes,ncol=1)

for (Periods_ahead in 1:horizontes) {
        for (i in 1:n.forecasting) {
                aux.y<-y[1:(n.estimation-Periods_ahead+i)];
                aux.y.log <- log(aux.y)
                fit<-arima(aux.y.log, order=c(0,1,1),
                           seasonal=list(order=c(0,1,1), period=s));
                y.pred.log<-predict(fit,n.ahead=Periods_ahead);
                y.pred1$pred <- exp(y.pred.log$pred)
                y.pred1$se <- exp(y.pred.log$se)
                predicc[i,Periods_ahead]<- (y.pred1$pred[Periods_ahead]);
        }
        error<-real-predicc[,Periods_ahead];
        MSFE_3[Periods_ahead]<-mean(error^2);
        MAPE_3[Periods_ahead]<-mean(abs(error/real)) *100;
}
```

Below are the results ordered by model.

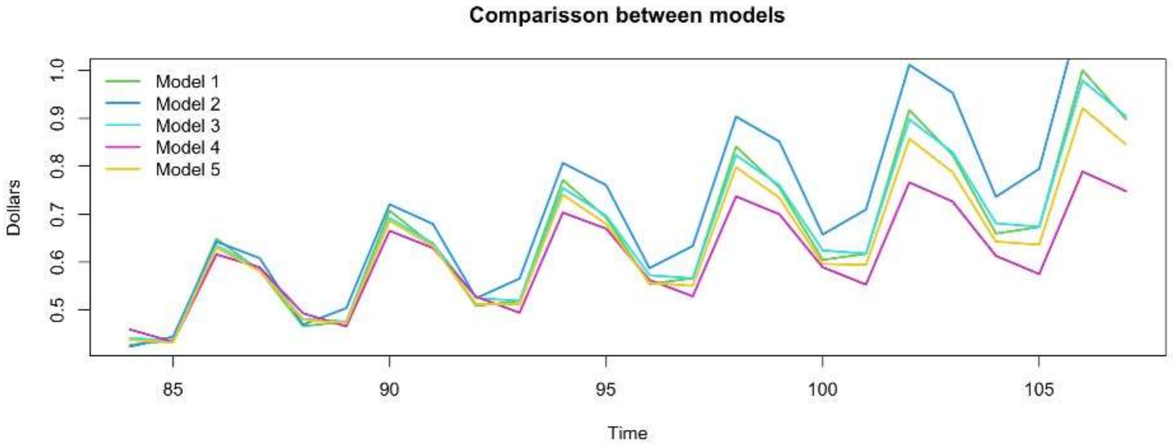


```
> df[order(df$mape_lag1),]
         mfse_lag1 mape_lag1   mfse_lag2 mape_lag2   mfse_lag3 mape_lag3   mfse_lag4 mape_lag4
Model_5 0.001701876  4.893447 0.003094177  6.353486 0.003888601  6.981882 0.004158569  7.076206
Model_3 0.001682581  5.224559 0.003110429  6.783052 0.003652939  7.202115 0.003635409  6.836125
Model_1 0.001675187  5.364606 0.003070760  6.841871 0.003591295  7.249468 0.003602084  6.842874
Model_2 0.002568157  5.777217 0.006057847  7.918378 0.009450984  9.307928 0.012291633  9.861410
Model_4 0.002116808  6.366023 0.004809371  9.262815 0.006272001 10.566637 0.007306694 11.475144
```

MAPE uses absolute values to calculate the mean percentage error for the forecast. We consider the Mean Absolute Percentage Error (MAPE) to be the most relevant and we see that Model 5 has the lowest score here for Lags 1-3. However, it’s worth noting that for Lag 4, Model 5 performed worse than Model 3. We also analyzed Lags 5-8 and Model 5’s performance worsened then.

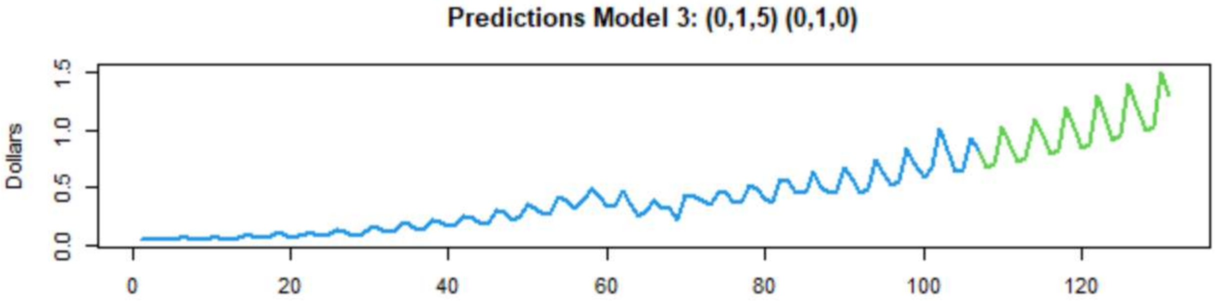
The result is that we consider both Model 5 and Model 3 to be best. Model 3 might be more appropriate for forecasting further into the future due to its lower MAPE score when forecasting farther ahead, while Model 5 might be better for short-term forecasts. Using both can be useful for creating a more holistic view.

Here is a graphical representation of differences between them all. Worth noting is that Models 3 and 5 (best performing models) tend to have higher predictions (measured in Dollars on the y-axis) in comparison than Model 4, which scored lowest with MFSE/MAPE scoring.



Comparisson between models

Additionally, we also looked at predictions generated by Model 3. We can see that sales are expected to continue to rise while continuing to exhibit cyclical behaviour.



Predictions Model 3: (0,1,5) (0,1,0)