FORECASTING TIME SERIES HOMEWORK 2

TEAM G

Vratul Kapur
Paul Jacques-Mignault
Sheena Miles
Ashley O'Mahony
Stavros Tsentemeidis
Karl Westphal
Irune Maury Arrue

LOAD DATA

To begin with, we load the Coca-Cola dataset, which consists of quarterly data, in our working directory. At this stage split the data into **train** – **test**, in order to validate our models based on actual predictions, on unseen data.

```
y<-data[,2][1:95]        # TRAINING SET
y_test<-data[,2][96:107]  # TEST SET
```

CHECKING STATIONARITY

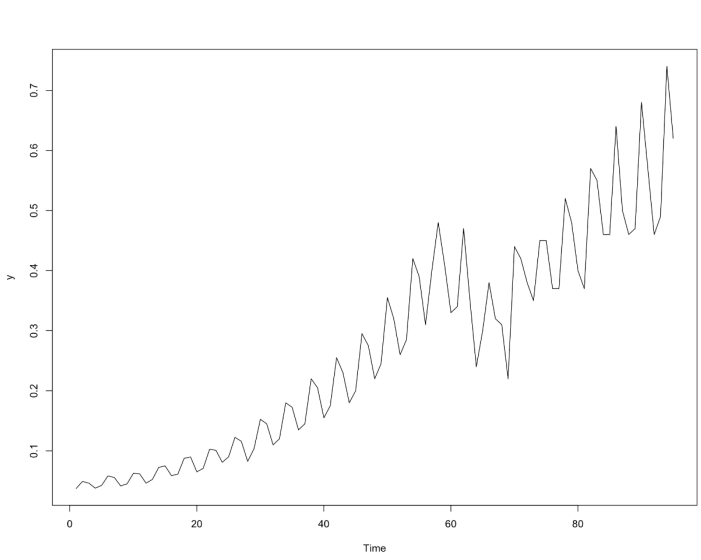
By plotting our timeseries, it is noticed that our data is not stationary both in the **mean** (not constant) as well as, in the **variance** (not constant either). In order to verify our observation, we will use the formal test. Also, as our data is about quarters, we think that there might be seasonality, so we also use the formal test for seasonal differences. In order to do so we set **s = 4**, as again our data is quarterly.

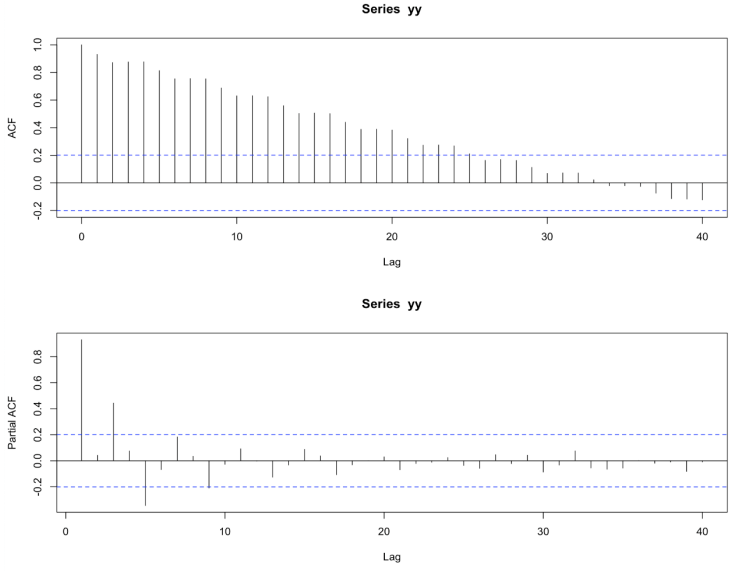
```
s=4
nsdiffs(y,m=s,test=c("ocsb"))      # seasonal differences
ndiffs(y, alpha=0.05, test=c("adf")) # regular differences
yy <- log(y)
```

PLOT ACF & PACF

Based on the above tests we see that our model should take into account **1 regular and 1 seasonal difference**, in order to deal with the lack of stationarity in our data. We also take the **logs**, to fix the variance. Moving on with our analysis, we need to plot the **PACF** and the **ACF** graphs to detect spikes. In order to do so, we define the number of lags to be equally to 40, which corresponds to 10 years as our data is quarterly once more.

```
ts.plot(y)
nlags=40
par(mfrow=c(2,1))
acf(yy,nlags)
pacf(yy,nlags)
```

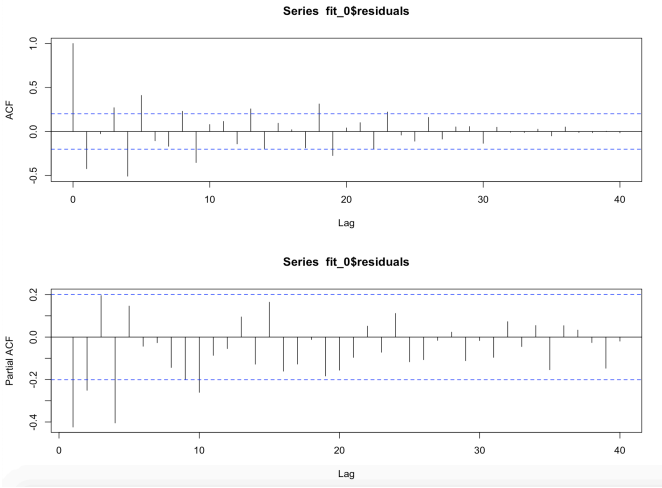




START MODELLING NON-SEASONAL PART WITH AR

```
fit_0<-arima(y,order=c(0,1,0),seasonal=list(order=c(0,1,0),period=s))

par(mfrow=c(2,1))
acf(fit_0$residuals,nlags)
pacf(fit_0$residuals,nlags)
ndiffs(fit_0$residuals, alpha=0.05, test=c("adf")) # regular differences?
nsdiffs(fit_0$residuals, m=s,test=c("ocsb")) # seasonal differences?
```



We can start modelling from the **non-seasonal** part cause it seems more "strong" at this point in time. In our first fit, we define the number of differences for both to be equal to 1, as already seen before. By looking at the **PACF** (simpler) we are trying to detect an **AR** model. Based on the graph, we choose to start with an **AR (3)** as 4 corresponds on a seasonal lag, and anything after that will include the 4th one.

```
fit_1<-arima(yy, order=c(3,1,0),seasonal=list(order=c(0,1,0),period=s))
fit_1
```

By looking at the summary of the model we can see that lag 2 and 3 are not significant so we re-tune our model.

```
Call:
arima(x = yy, order = c(3, 1, 0), seasonal = list(order = c(0, 1, 0), period = s))

Coefficients:
          ar1      ar2     ar3
      -0.5239  -0.1166  0.1978
s.e.   0.1027   0.1157  0.1020

sigma^2 estimated as 0.01149:  log likelihood = 73.03,  aic = -138.05
```

```
fit_2<-arima(yy, order=c(1,1,0),seasonal=list(order=c(0,1,0),period=s))
fit_2
```

```
Call:
arima(x = yy, order = c(1, 1, 0), seasonal = list(order = c(0, 1, 0), period = s))

Coefficients:
          ar1
      -0.4616
s.e.   0.0926

sigma^2 estimated as 0.0127:  log likelihood = 68.65,  aic = -133.31
```

Having concluded that our lags are significant at a 95% level, we can now have a look at our model's residuals. By observing the plot, we can see that our data is close to stationary (which we assume), with the exception of some outliers. Moving on, we check our residuals for WN in order to determine how good our model is.

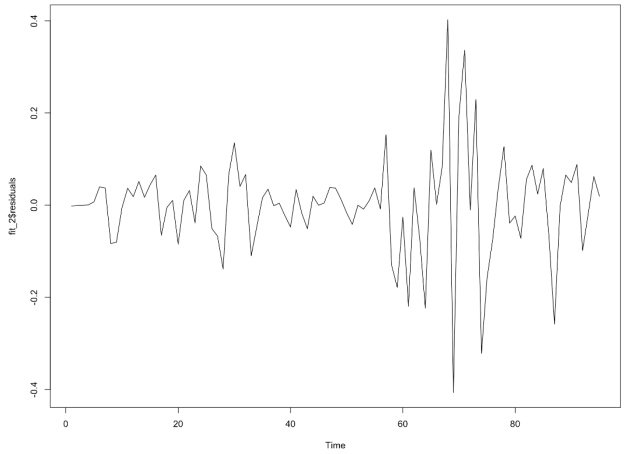
```
ts.plot(fit_2$residuals)
par(mfrow=c(2,1))
acf(fit_2$residuals, nlags)
pacf(fit_2$residuals, nlags)
Box.test(fit_2$residuals, lag=40)
shapiro.test(fit_2$residuals)
```

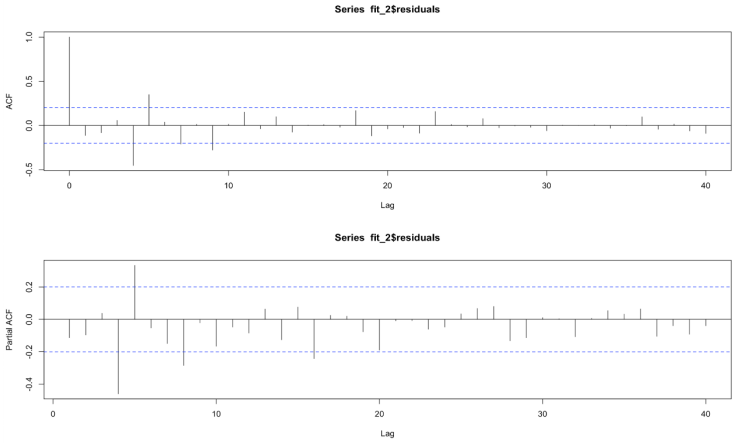
```
        Box-Pierce test

data:  fit_2$residuals
X-squared = 59.517, df = 40, p-value = 0.02413
```

```
        Shapiro-Wilk normality test

data:  fit_2$residuals
W = 0.90171, p-value = 2.852e-06
```





Our residuals are not WN and not normal, so our job is not done yet. Let's move on to the seasonal part of the model. By looking at the PACF plot, we can see that lag 16 is nearly out of bounds. So we are adding a SAR(4) in our model.

```
fit_3<-arima(yy,order=c(1,1,0),seasonal=list(order=c(4,1,0),period=s))
fit_3
```

```
Call:
arima(x = yy, order = c(1, 1, 0), seasonal = list(order = c(4, 1, 0), period = s))

Coefficients:
          ar1     sar1     sar2     sar3     sar4
      -0.3308  -0.6790  -0.4092  -0.2979  -0.1519
s.e.   0.1064   0.1087   0.1230   0.1236   0.1024

sigma^2 estimated as 0.008629:  log likelihood = 84.91,  aic = -157.83
```

By looking at the summary, we can tell that all our lags are not significant so we retune.

```
fit_3a<-arima(yy,order=c(1,1,0),seasonal=list(order=c(3,1,0),period=s))
fit_3a
```

```
Call:
arima(x = yy, order = c(1, 1, 0), seasonal = list(order = c(3, 1, 0), period = s))

Coefficients:
          ar1     sar1     sar2     sar3
      -0.3472  -0.6386  -0.3496  -0.1925
s.e.   0.1040   0.1065   0.1167   0.1024

sigma^2 estimated as 0.008881:  log likelihood = 83.84,  aic = -157.67
```

By looking at the summary again, we can tell that all our lags are not significant. So we re-tune and move on checking our residuals. To do that we keep the first 2 lags of the seasonal part and only the first one from the non-seasonal.

```
fit_4<-arima(yy,order=c(1,1,0),seasonal=list(order=c(2,1,0),period=s))
fit_4
```

```
Call:
arima(x = yy, order = c(1, 1, 0), seasonal = list(order = c(2, 1, 0), period = s))

Coefficients:
          ar1     sar1     sar2
      -0.3568  -0.5879  -0.2310
s.e.   0.1053   0.1055   0.1005

sigma^2 estimated as 0.009274:  log likelihood = 82.12,  aic = -156.25
```

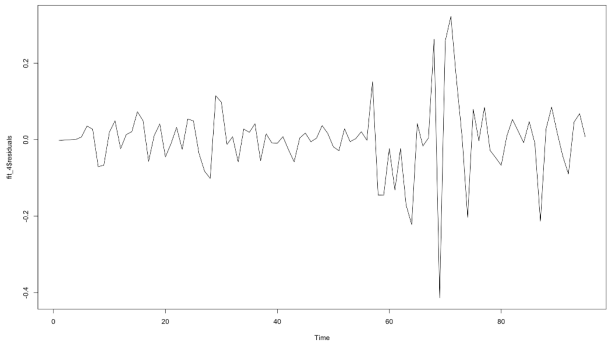
```
ts.plot(fit_4$residuals)
par(mfrow=c(2,1))
acf(fit_4$residuals,nlags)
pacf(fit_4$residuals,nlags)
Box.test(fit_4$residuals,lag=20)
shapiro.test(fit_4$residuals)
```

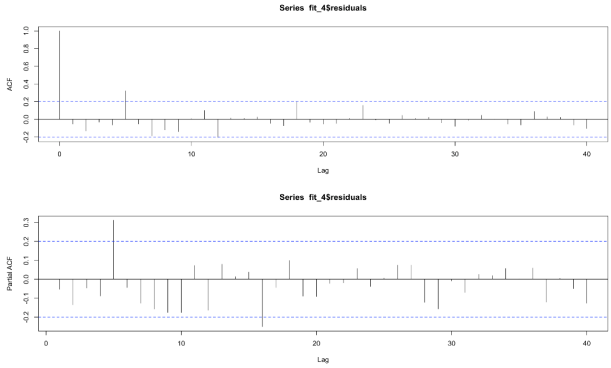
```
        Box-Pierce test

data:  fit_4$residuals
X-squared = 29.021, df = 20, p-value = 0.08735
```

```
        Shapiro-Wilk normality test

data:  fit_4$residuals
W = 0.87419, p-value = 1.856e-07
```



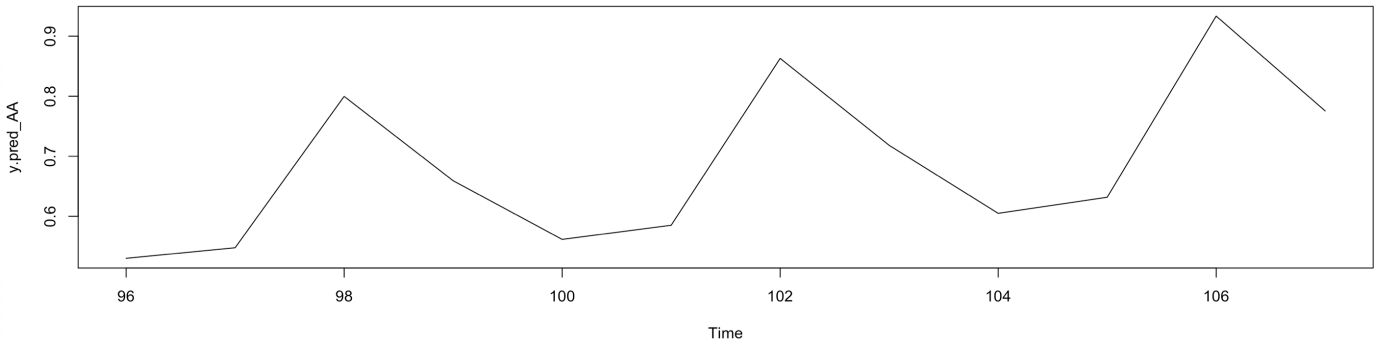


Based on the graph but also on the Box - Ljung Test we can tell that our residuals are WN, which indicates that our model is sufficiently good to be considered one of our possible models. As a next step, let's apply predictions to our model.

```
y.pred_A<-predict(fit_4,n.ahead=12)
y.pred_A$pred
y.pred_A$se
```

After that, it is time to undo the *logarithmic transformation* that was applied to fix the non-stationarity in the variance.

```
y.pred_AA <- exp(y.pred_A$pred)
y.pred_AA
ts.plot(y.pred_AA)
```



Finally, even though our residuals are not normal, a 95% confidence interval is a good approximation. Based on that we calculate the following:

```
quantile(fit_4$residuals, probs=c(0.025,0.975)) # 95% confidence interval
```

```
      2.5%      97.5%
-0.2098218  0.2226874
```

START MODELLING NON-SEASONAL PART WITH MA

We can start modelling from the non-seasonal part cause it seems more "strong" at this point in time. In our first fit, we define the number of differences for both to be equal to 1, as already seen before. By looking at the ACF (first plots at the beginning of the doc again) we are trying to detect an MA model. Based on the graph, we choose to start with an MA (3) as 4 corresponds on a seasonal lag, and anything after that will include the 4th one.

```
fit_10<-arima (yy, order=c(0,1,3), seasonal=list(order=c(0,1,0),period=s))
fit_10
```

```
Call:
arima(x = yy, order = c(0, 1, 3), seasonal = list(order = c(0, 1, 0), period = s))

Coefficients:
          ma1     ma2      ma3
      -0.7477  0.4704  -0.6144
s.e.   0.0875  0.1186   0.0903

sigma^2 estimated as 0.01026:  log likelihood = 77.27,  aic = -146.55
```

Having concluded that our lags are significant at a 95% level, we can now have a look at our model's residuals.

By observing the plot we can see that our data is close to stationary (which we assume), with the exception of some outliers. Moving on, we check our residuals for WN in order to determine how good our model is.

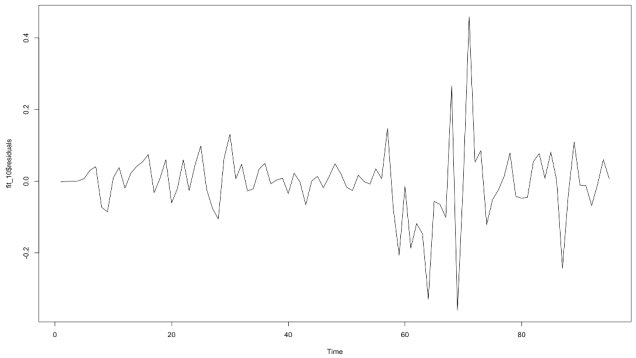
```
ts.plot(fit_10$residuals)
par(mfrow=c(2,1))
acf(fit_10$residuals,nlags)
pacf(fit_10$residuals,nlags)
Box.test(fit_10$residuals,lag=20)
shapiro.test(fit_10$residuals)
```

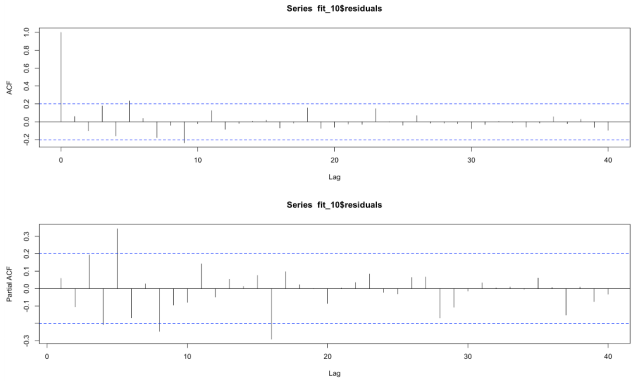
```
        Box-Pierce test

data:  fit_10$residuals
X-squared = 26.35, df = 20, p-value = 0.1546
```

```
        Shapiro-Wilk normality test

data:  fit_10$residuals
W = 0.85417, p-value = 3.155e-08
```





Our residuals are WN, but our job is not done yet. Let's move on to the seasonal part of the model. By looking at the PACF plot, we can see that lag 16 is out of bounds. So we are adding a SMA(4) in our model.

```
fit_30<-arima(yy,order=c(0,1,3),seasonal=list(order=c(4,1,0),period=s))
fit_30
```

```
Call:
arima(x = yy, order = c(0, 1, 3), seasonal = list(order = c(4, 1, 0), period = s))

Coefficients:
          ma1      ma2     ma3     sar1     sar2     sar3     sar4
      -0.4523  -0.0926  0.2108  -0.7023  -0.5136  -0.3315  -0.1433
s.e.   0.1223   0.1158  0.1892   0.1099   0.1373   0.1223   0.1025

sigma^2 estimated as 0.008218:  log likelihood = 86.96,  aic = -157.92
```

By looking at the summary, we can tell that some lags are not significant from both parts of the formula. So, we re-tune our model and move on checking residuals.

```
fit_40<-arima(yy,order=c(0,1,1),seasonal=list(order=c(3,1,0),period=s))
fit_40
```

```
Call:
arima(x = yy, order = c(0, 1, 1), seasonal = list(order = c(3, 1, 0), period = s))

Coefficients:
          ma1     sar1     sar2     sar3
      -0.4300  -0.6276  -0.3743  -0.2152
s.e.   0.0992   0.1065   0.1141   0.1012

sigma^2 estimated as 0.008568:  log likelihood = 85.34,  aic = -160.69
```

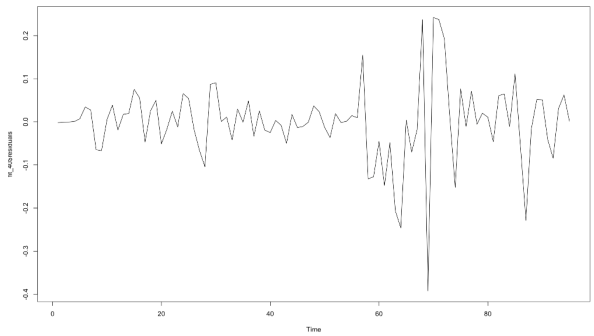
```
ts.plot(fit_40$residuals)
par(mfrow=c(2,1))
acf(fit_40$residuals,nlags)
pacf(fit_40$residuals,nlags)
Box.test(fit_40$residuals,lag=20)
shapiro.test(fit_40$residuals)
```

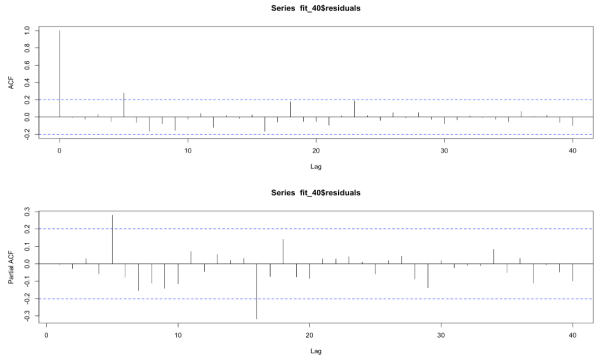
```
        Box-Pierce test

data:  fit_40$residuals
X-squared = 21.93, df = 20, p-value = 0.3443
```

```
        Shapiro-Wilk normality test

data:  fit_40$residuals
W = 0.88829, p-value = 7.172e-07
```



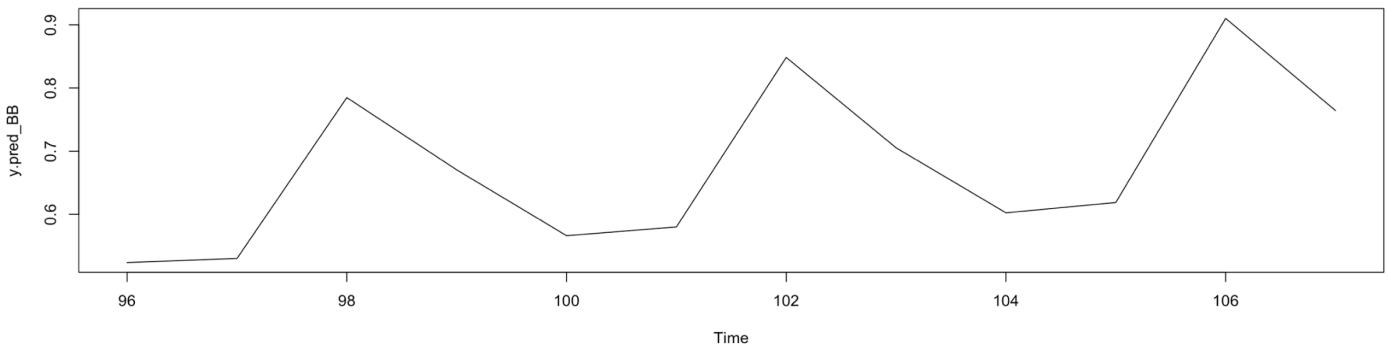


Based on the graph but also on the Box - Ljung Test we can tell that our residuals are WN, which indicates that our model is sufficiently good to be considered one of our possible models. As a next step, let's apply predictions to our model.

```
y.pred_B<-predict(fit_40,n.ahead=12)
y.pred_B$pred
y.pred_B$se
ts.plot(y.pred_B$pred)
```

After that it is time to undo the logarithmic transformation that was applied to fix the non-stationarity in the variance.

```
y.pred_BB <- exp(y.pred_B$pred)
y.pred_BB
ts.plot(y.pred_BB)
```



Finally, even though our residuals are not normal, a 95% confidence interval is a good approximation. Based on that we calculate the following:

```
quantile(fit_40$residuals, probs=c(0.025,0.975)) # 95% confidence interval
```

```
      2.5%      97.5%
-0.2206733  0.2217790
```

START MODELLING SEASONAL PART WITH AR & MA

We can start modelling from the seasonal part. In our first fit, we define the number of differences for both to be equal to 1, as already seen before. By looking at the ACF (first plots at the beginning of the doc again) we are trying to detect an MA model. Based on the graph, we choose to start with an SMA(2) as 8 corresponds on the last seasonal lag out of limits.

```
fit_100<-arima(yy, order=c(0,1,0),seasonal=list(order=c(0,1,2),period=s))
fit_100
```

```
Call:
arima(x = yy, order = c(0, 1, 0), seasonal = list(order = c(0, 1, 2), period = s))

Coefficients:
          sma1    sma2
      -0.8326  0.0508
s.e.   0.1002  0.0940

sigma^2 estimated as 0.009234:  log likelihood = 81.13,  aic = -156.25
```

By looking at the summary of the model we can see that lag 1 is the only *significant* so we re-tune our model.

```
fit_200<-arima(yy, order=c(0,1,0),seasonal=list(order=c(0,1,1),period=s))
fit_200
```

```
Call:
arima(x = yy, order = c(0, 1, 0), seasonal = list(order = c(0, 1, 1), period = s))

Coefficients:
          sma1
      -0.7959
s.e.   0.0698

sigma^2 estimated as 0.00926:  log likelihood = 80.98,  aic = -157.96
```

Having concluded that our lags are significant at a 95% level, we can now have a look at our model's residuals.

By observing the plot we can see that our data is close to stationary (which we assume), with the exception of some outliers. Moving on, we check our residuals for WN in order to determine how good our model is.

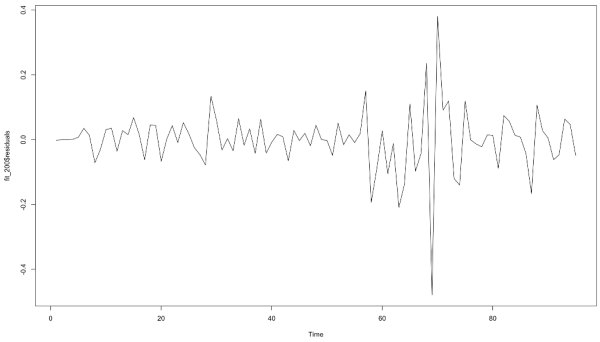
```
ts.plot(fit_200$residuals)
par(mfrow=c(2,1))
acf(fit_200$residuals,nlags)
pacf(fit_200$residuals,nlags)
Box.test(fit_200$residuals,lag=20)
shapiro.test(fit_200$residuals)
```

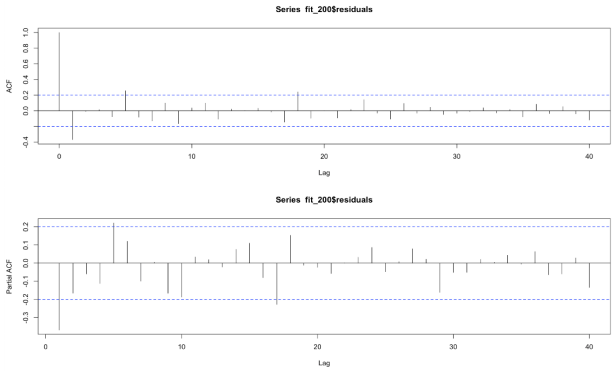
```
        Box-Pierce test

data:  fit_200$residuals
X-squared = 36.273, df = 20, p-value = 0.01428
```

```
        Shapiro-Wilk normality test

data:  fit_200$residuals
W = 0.86011, p-value = 5.25e-08
```





Our residuals are not WN, but our job is not done yet as we observe that there are lags off the limits. So now we move on to the non-seasonal part of the model. By looking at the PACF plot, we can see that lag 1 is nearly out of bounds. So we are adding a AR(1) in our model. But, looking at the ACF we see lag 1 also out of bounds. So we are adding a MA(1) in to another model. The reason why we stay at lag 1 is again the fact that if we include any lag after lag 4, seasonality will be included.

```
fit_300<-arima(yy,order=c(1,1,0),seasonal=list(order=c(0,1,1),period=s))
fit_300
```

```
Call:
arima(x = yy, order = c(1, 1, 0), seasonal = list(order = c(0, 1, 1), period = s))

Coefficients:
          ar1     sma1
      -0.3728  -0.8246
s.e.   0.0994   0.0799

sigma^2 estimated as 0.007937:  log likelihood = 87.59,  aic = -169.17
```

```
fit_3000<-arima(yy,order=c(0,1,1),seasonal=list(order=c(0,1,1),period=s))
fit_3000
```

```
Call:
arima(x = yy, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = s))

Coefficients:
          ma1     sma1
      -0.4218  -0.8058
s.e.   0.0894   0.0763

sigma^2 estimated as 0.007752:  log likelihood = 88.77,  aic = -171.54
```

By looking at the summaries, we can tell that all our lags are significant. So, we move on checking our residuals for both models.

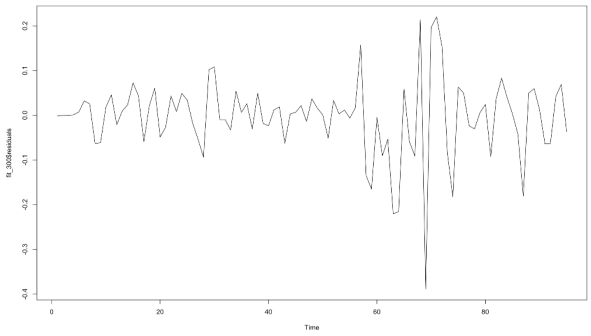
```
ts.plot(fit_300$residuals)
par(mfrow=c(2,1))
acf(fit_300$residuals,nlags)
pacf(fit_300$residuals,nlags)
Box.test(fit_300$residuals,lag=20)
shapiro.test(fit_300$residuals)
```

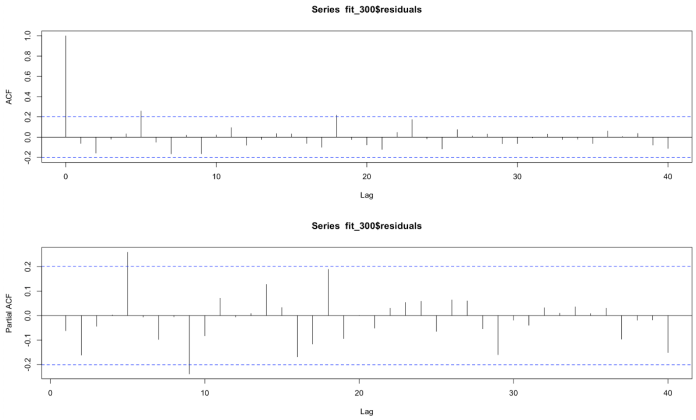
```
        Box-Pierce test

data:  fit_300$residuals
X-squared = 22.766, df = 20, p-value = 0.3004
```

```
        Shapiro-Wilk normality test

data:  fit_300$residuals
W = 0.90861, p-value = 6.041e-06
```





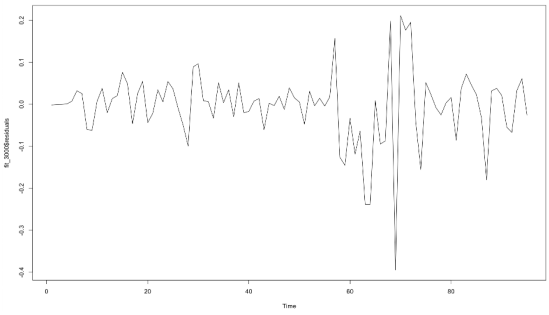

```
ts.plot(fit_3000$residuals)
par(mfrow=c(2,1))
acf(fit_3000$residuals,nlags)
pacf(fit_3000$residuals,nlags)
Box.test(fit_3000$residuals,lag=20)
shapiro.test(fit_3000$residuals)
```

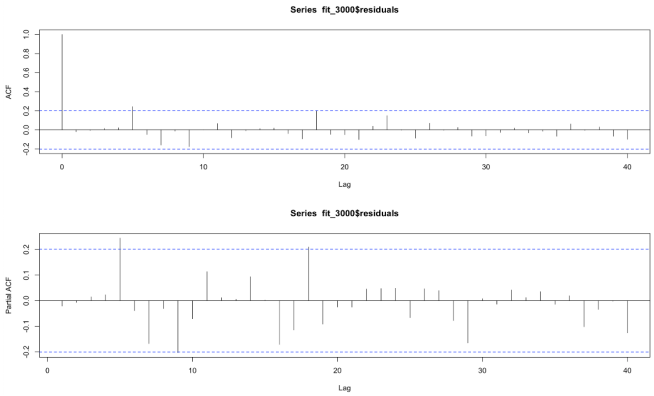
```
        Box-Pierce test

data:  fit_3000$residuals
X-squared = 17.614, df = 20, p-value = 0.6128
```

```
        Shapiro-Wilk normality test

data:  fit_3000$residuals
W = 0.88593, p-value = 5.682e-07
```



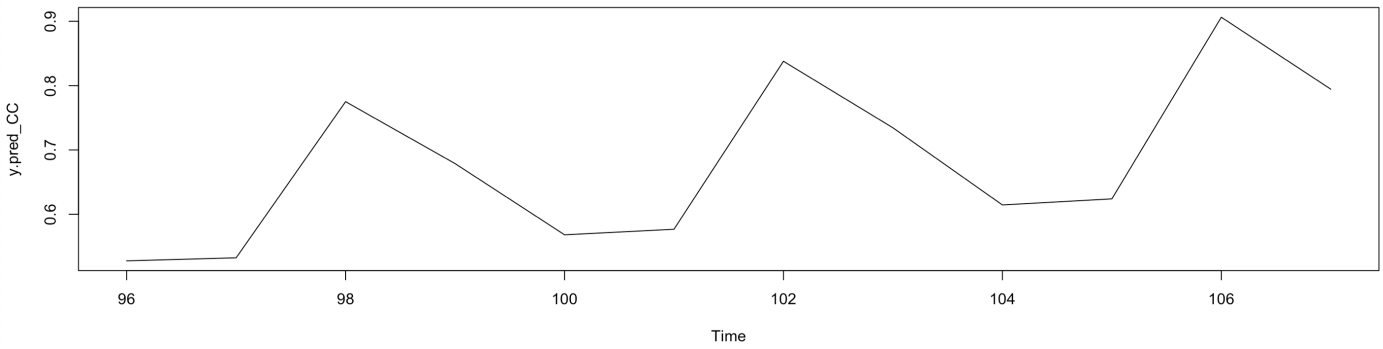


Based on the graphs, but also on the Box - Ljung Test we can tell that our residuals are WN, which indicates that our models are sufficiently good to be considered one of our possible models. As a next step, let's apply predictions to our models.

```
y.pred_C<-predict(fit_300,n.ahead=12)
y.pred_C$pred
y.pred_C$se
```

After that it is time to undo the logarithmic transformation that was applied to fix the non-stationarity in the variance.

```
y.pred_CC <- exp(y.pred_C$pred)
ts.plot(y.pred_CC)
```



Finally, even though our residuals are not normal, a 95% confidence interval is a good approximation. Based on that we calculate the following:

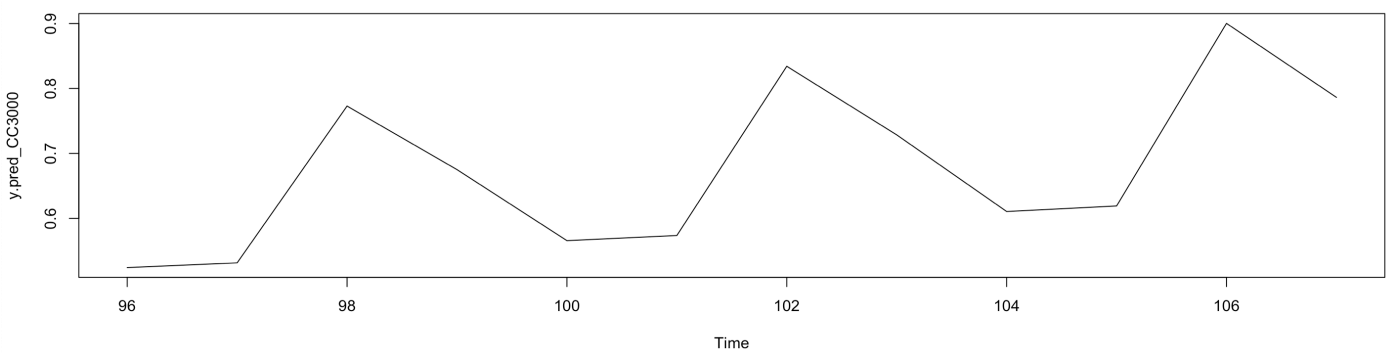
```
quantile(fit_300$residuals, probs=c(0.025,0.975)) # 95% confidence interval
```

```
      2.5%      97.5% 
-0.2039979  0.1827766 
```

```
y.pred_C3000<-predict(fit_3000,n.ahead=12)
y.pred_C3000$pred
y.pred_C3000$se
```

After that it is time to undo the logarithmic transformation that was applied to fix the non-stationarity in the variance.

```
y.pred_CC3000 <- exp(y.pred_C3000$pred)
ts.plot(y.pred_CC3000)
```



Finally, even though our residuals are not normal, a 95% confidence interval is a good approximation. Based on that we calculate the following:

```
quantile(fit_3000$residuals, probs=c(0.025,0.975)) # 95% confidence interval
```

```
      2.5%      97.5%
-0.2176682  0.1884963
```

COMPARE PREDICTION PERFORMANCE OF THE 4 MODELS

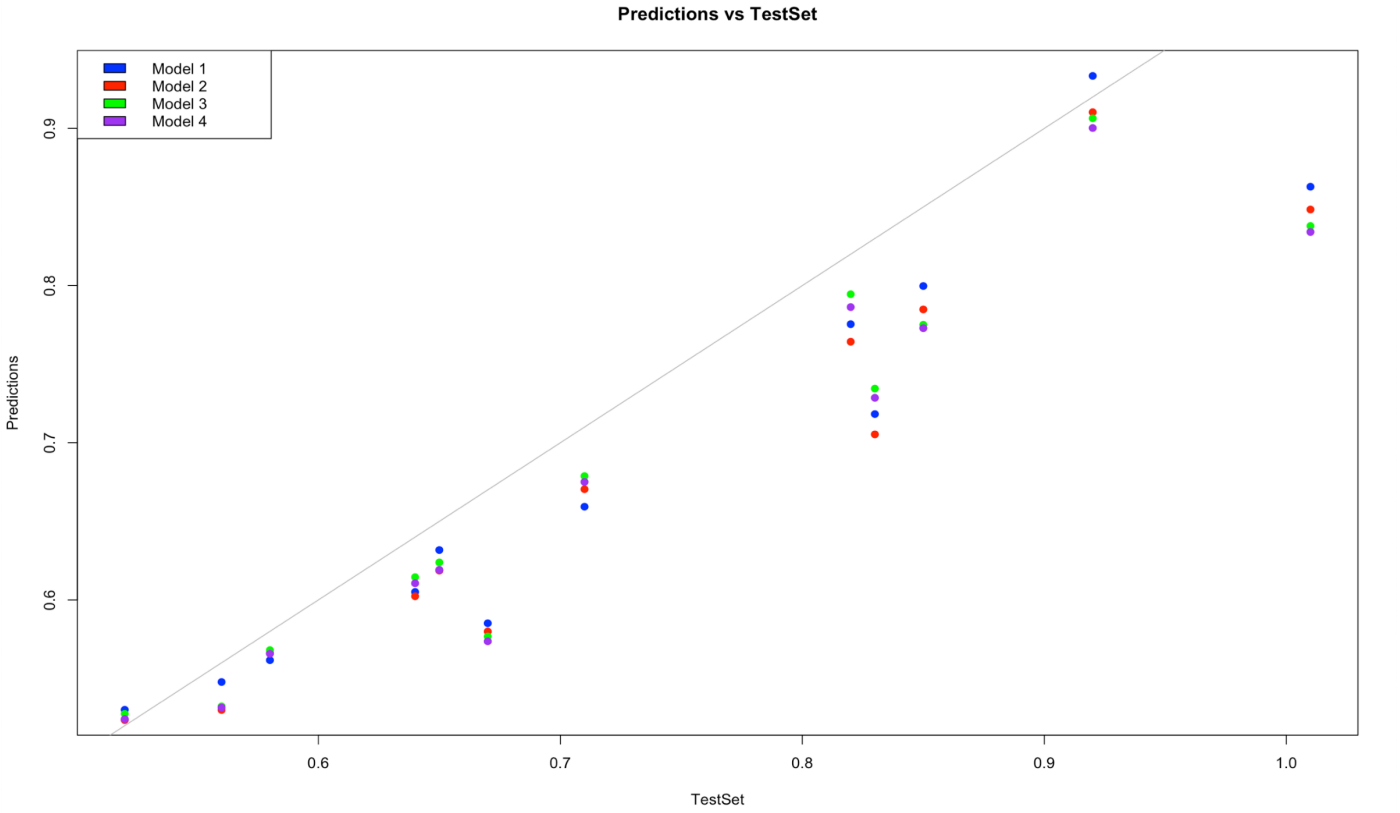
Model 1: fit_4<-arima(yy,order=c(1,1,0),seasonal=list(order=c(2,1,0),period=s))
Model 2: fit_40<-arima(yy,order=c(0,1,1),seasonal=list(order=c(3,1,0),period=s))
Model 3: fit_300<-arima(yy,order=c(1,1,0),seasonal=list(order=c(0,1,1),period=s))
Model 4: fit_3000<-arima(yy,order=c(0,1,1),seasonal=list(order=c(0,1,1),period=s))

```
y_total <- cbind(y.pred_AA,y.pred_BB,y.pred_CC,y.pred_CC3000,y_test)
colnames(y_total) <- c('First Model','Second Model','Third Model','Fourth Model','Test Values')
y_total
```

First Model	Second Model	Third Model	Fourth Model	Test Values
0.5301319	0.5235411	0.5275311	0.5242119	0.52
0.5477592	0.5299681	0.5322404	0.5315240	0.56
0.7996619	0.7848048	0.7750533	0.7728656	0.85
0.6592771	0.6704094	0.6787926	0.6750631	0.71
0.5616499	0.5661631	0.5680925	0.5657483	0.58
0.5851433	0.5798966	0.5767045	0.5736398	0.67
0.8628425	0.8483696	0.8378766	0.8341042	1.01
0.7182143	0.7053065	0.7344416	0.7285523	0.83
0.6050459	0.6023296	0.6144700	0.6105759	0.64
0.6317292	0.6186254	0.6238592	0.6190926	0.65
0.9333448	0.9102410	0.9063461	0.9001952	0.92
0.7754170	0.7642560	0.7944717	0.7862798	0.82

```
plot(y_test, y.pred_AA, type = 'p', col = 'blue', cex = 1, pch=19,
   main="Predictions vs TestSet",
   ylab="Predictions", xlab='TestSet')
points(y_test, y.pred_BB, type = 'p', col = 'red', cex=1, pch=19)
points(y_test, y.pred_CC, type = 'p', col = 'green', cex =1, pch=19)
points(y_test, y.pred_CC3000, type = 'p', col = 'purple', cex=1, pch=19)

legend("topleft",
    c("Model 1","Model 2", "Model 3",'Model 4'),
    fill=c("blue","red",'green', 'purple')
)
abline(a=0, b=1, col = 'grey')
```




```
accuracy(y.pred_AA, y_test)
accuracy(y.pred_BB, y_test)
accuracy(y.pred_CC, y_test)
accuracy(y.pred_CC3000, y_test)

print(paste0('The MAPE of Model 1 is ',accuracy(y.pred_AA, y_test)[5]))
print(paste0('The MAPE of Model 2 is ',accuracy(y.pred_BB, y_test)[5]))
print(paste0('The MAPE of Model 3 is ',accuracy(y.pred_CC, y_test)[5]))
print(paste0('The MAPE of Model 4 is ',accuracy(y.pred_CC3000, y_test)[5]))
```

```
"The MAPE of Model 1 is 6.35230267583398"
"The MAPE of Model 2 is 7.06015825342512"
'The MAPE of Model 3 is 6.39656425833563"
'The MAPE of Model 4 is 6.83209122285819"
```

We also calculate different performance metrics, and we use the MAPE as defined. Based on the results mentioned below, the model that performs the best is the first one, Model 1 with the following formula:

fit_4<-arima(yy,order=c(1,1,0),seasonal=list(order=c(2,1,0),period=s))