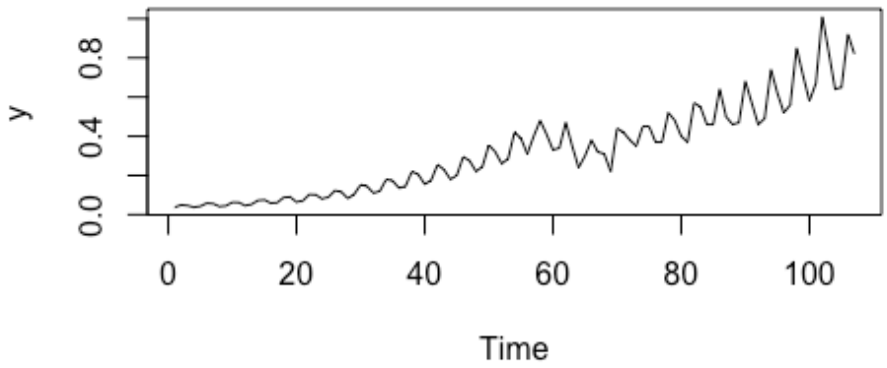
Forecasting Time Series: Coca Cola Earnings

By **Mohamed Khanafer**

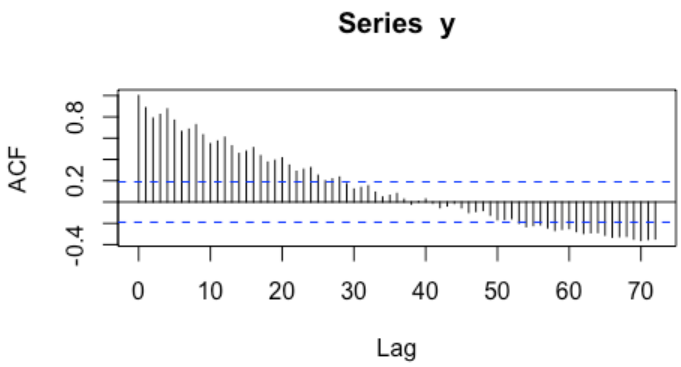
Table of Content

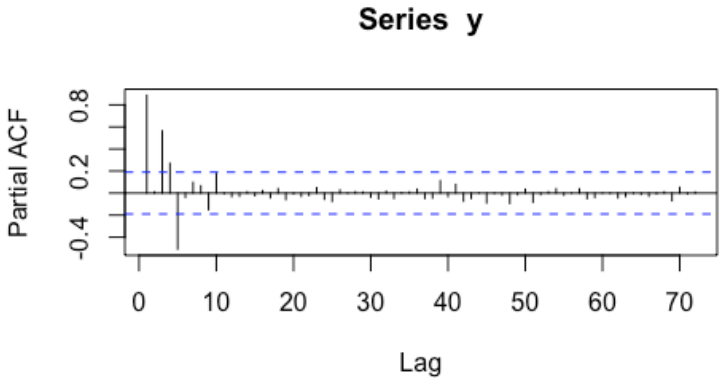
Part I: *Finding linear time series models, using the Box-Jenkins methodology, for the quarterly earnings per share of Coca-Cola*

1. Plotting the graph, ACF and PACF of our data



Observation: we clearly see that the data on hand does not exhibit stationarity nor in the mean nor in the variance. We will confirm those hypotheses with the tests. We also see clear seasonality in the data with up and downs repeated over the period. This is coupled with an upward trend, and it is worth mentioning that we could consider using the logarithms of the data to remove the non-constant variance we are dealing with. We will consider this option later on. We now turn to the ACF and PACF:





Observation: we can confirm the non-stationarity of the data here with the decline to 0 in the ACF. Running the formal.

2. Looking at the tests for the required differences

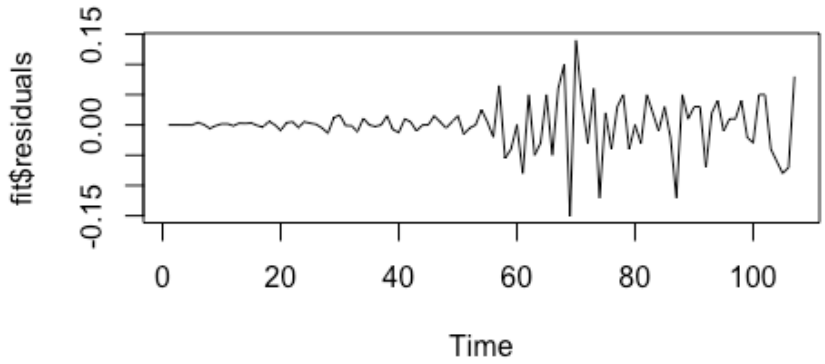
To check how many differences are needed for making our data stationary, we compute the Dickey-Fuller test to check for the needed regular differences. The result is that we need 1 regular difference. We know that we are dealing with quarterly data and thus have an S

parameter of 4. And, we use also the test to check for seasonal differences needed and we see that one seasonal difference is also needed.
Thus, we perform those transformations using d=1 and D=1 in the following form:

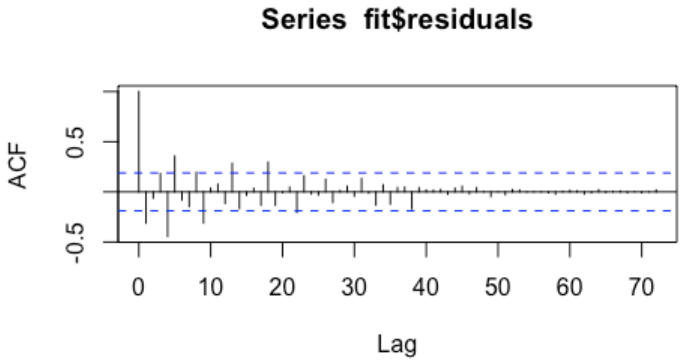
arima(y,order=c(0,1,0),seasonal=list(order=c(0,1,0),period=s))

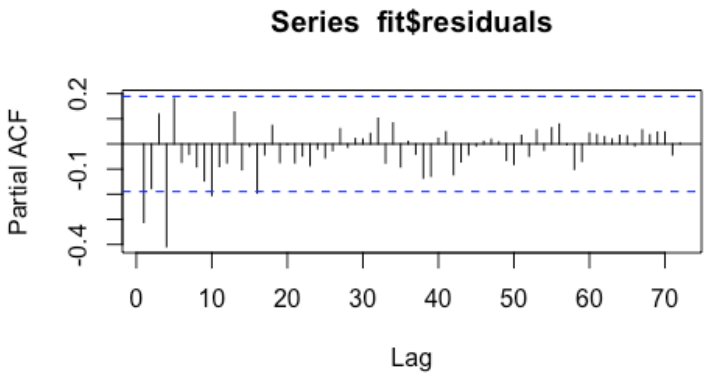
and plot our data again.

3. Plotting the data after the transformations



Observation: we see that at least the mean is stationary following the transformations and running the two tests again, we see that we do not need any further regular nor seasonal differences. This is why we could consider the logarithms later on.





Observation: now that we are dealing with stationary data, we can start looking for the parameters p,q, P, and Q.

From the above ACF and PACF, we could propose various models:

arima(y,order=c(0,1,0),seasonal=list(order=c(1,1,0),period=s))
arima(y,order=c(0,1,0),seasonal=list(order=c(0,1,1),period=s))
arima(y,order=c(0,1,0),seasonal=list(order=c(0,1,2),period=s))
arima(y,order=c(0,1,18),seasonal=list(order=c(0,1,0),period=s))

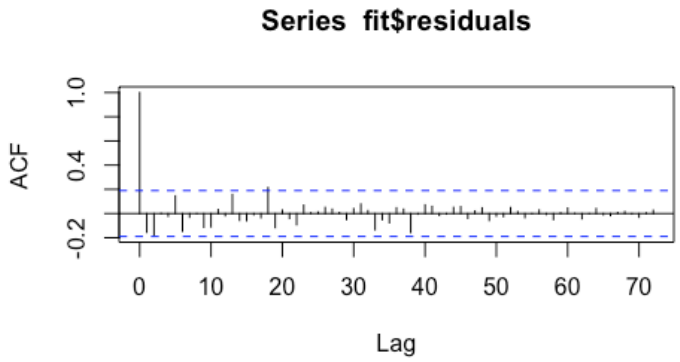
Other models could be proposed but we decided to start with the simpler ones and build on them. So, we start with the Seasonal Autoregressive of order 1 given that we have lag 4 is clearly out of bounds and then we will build on top of that.

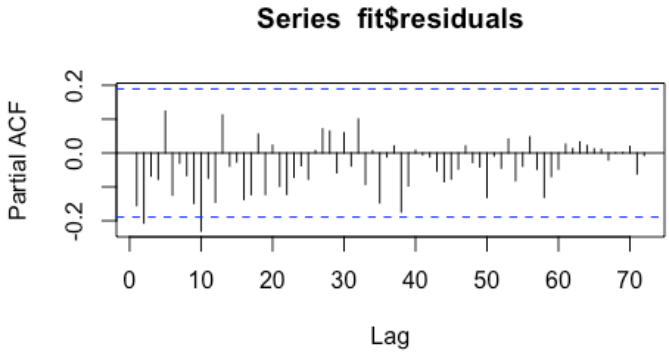
4. Estimating the first model

We fit the following model:

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

We get as a coefficient sar1 = -0.4842 with a standard error of 0.0899 and can thus conclude this parameter is significant.





Observation: we still have lags out of bound so maybe we could add something to the model, but first we checked to see if we had white noise in the residuals to see if this could be a viable model.

So, using the Box Test, we get a p-value = 0.3962 and thus we can use this model to predict our data.

Building on the 1st model: because we see that lag 2 is out of bound in the PACF, we estimated this model too:

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

And we got the coefficient as:

	ar1	ar2	sar1
	-0.1912	-0.2093	-0.4702
s.e.	0.1030	0.0996	0.1003

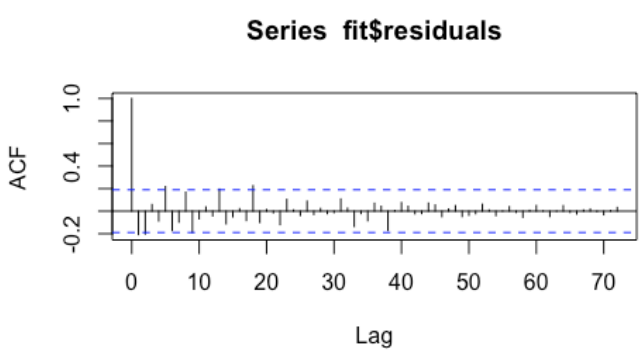
We concluded that even if the parameters ar1 and ar2 are almost non significant, we could still try to use this model because its residuals were also white noise (with a p-value of 0.53).

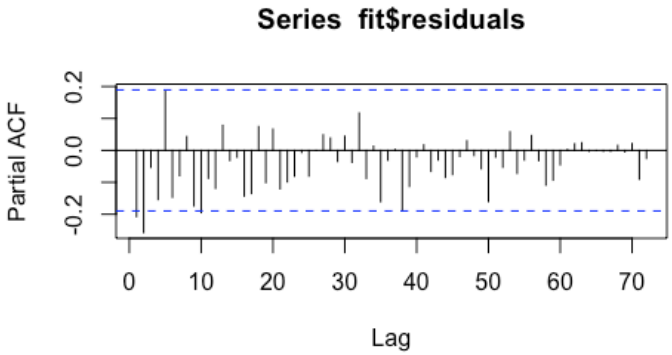
5. Estimating the second model

We fit the following model:

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

We get as a coefficient sma1 = -0.4242 with a standard error of 0.0812 and can thus conclude this parameter is significant.





Observation: we still have lags out of bound so maybe we could add something to the model, but first we checked to see if we had white noise in the residuals to see if this could be a viable model.

So, using the Box Test, we get a p-value = 0.009833 and thus we cannot use this model to predict our data. We decide thus to add the second lag in the PACF to see if the model could turn out to be significant.

Building on the 2nd model: because we see that lag 2 is out of bound in the PACF, we estimated this model too:

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

And we got the coefficient as:

	Coefficients:		
	ar1	ar2	sma1
	-0.2467	-0.2960	-0.4994
s.e.	0.0960	0.1071	0.0947

We concluded that all the parameters are significant. And using the Box test, we get a p-value of 0.2545 and thus we can use this model.

Part II: *Comparing all the models in terms of forecasting*

1. Calculating the MAPE of the 3 models using the Recursive Scheme

Given that for a time series we cannot randomly split the data in order not to destroy its structure, we will divide it into in sample and out-of-sample parts. Then we apply a recursive or a rolling scheme to have a N point ahead predictions. This thus allow us to compare real values and predictive values in the test sample.

Since the MAPE is the better measure in terms of comprehension, it is the one we used to compare the different models. So, here are the results using the Recursive Method:

	N=1	N=2	N=3	N=4
Model 1 (010|110)	6.374052	8.61975	8.989462	9.024862
Model 2 (210|110)	5.563923	7.953592	8.838176	8.523163
Model 3 (210|011)	5.59801	7.871255	8.369848	7.796061

Observation: we can directly see that the errors of the 3 models are quite similar. We also see that the error increases with the time horizon, which is an expected behavior. However, if we were to choose one of them, we would probably take Model 3 given that it is able to slightly better predict the 2,3, and 4 steps ahead. Maybe if someone is really interested in the first quarter estimate, he could choose the Model 2 but the difference in error is really minimal.

2. Calculating the MAPE of the 3 models using the Rolling Scheme

Given that we do not have a structural change in the data, we do not expect the 2 schemes to give us different results. But we used it to make sure our hypothesis was correct:

	N=1	N=2	N=3	N=4
Model 1 (010|110)	6.370019	8.611569	8.984889	9.021739
Model 2 (210|110)	5.56574	7.950915	8.834667	8.520701
Model 3 (210|011)	5.611824	7.882132	8.374048	7.794272

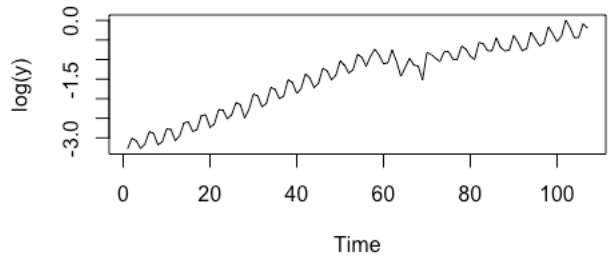
Observation: we indeed have identical measures and a similar conclusion as using the Recursive Method.

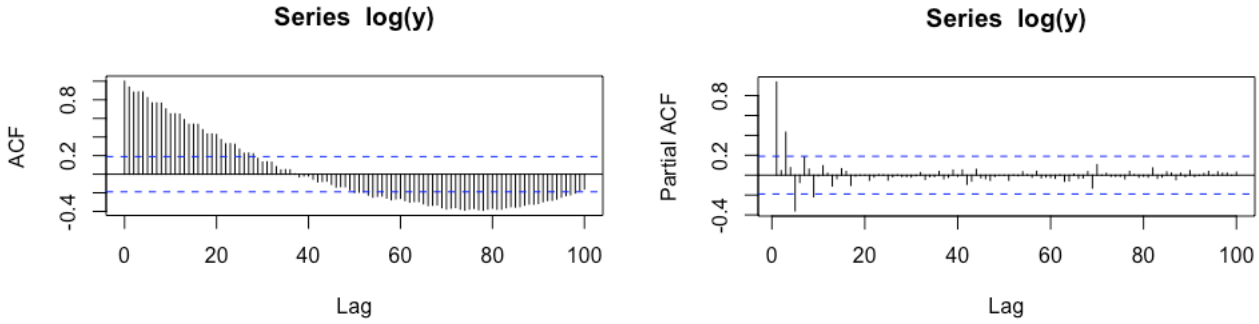
Part III: *Experimenting with Logarithms*

As illustrated in Part I, the variance of the data was not stationary. And because we know it is financial data, it could help to take the logarithms of the data and try building models. This is what we did here to see if the performance of the models would improve when we take this into consideration.

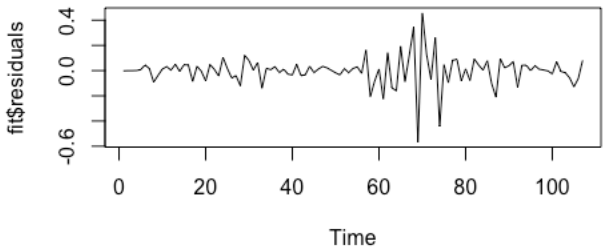
1. Building 2 models with logarithms

We followed the same procedure as used previously but this time using log(y). We had almost a similar plot for the logs and ACF and PACF:

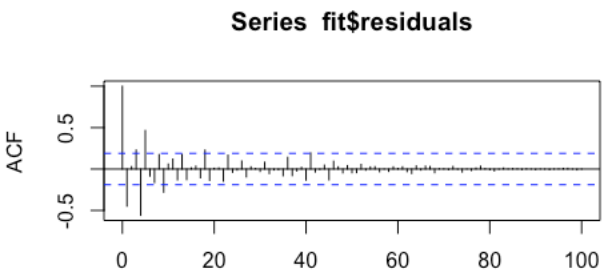


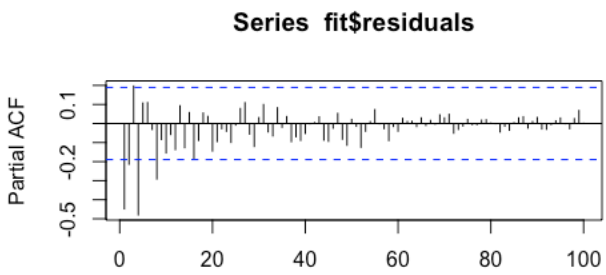


Observation: we see a similar behavior as in the normal data. Doing the tests, we get that we also need a regular and a seasonal transformation so here d=1 and D=1 as well. And we get, after the transformations:



Observation: the data looks indeed more stationary in the variance as compared with the normal data without the logarithms. However, we still see a cluster of volatility present. And we get the ACF and PACF as follows:





Observation: based on these ACF and PACF, we built the following models:

Model 4:

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

With coefficients:

	ar1	sar1
	-0.3143	-0.4845
s.e.	0.0975	0.0887

And a p-value for the Box test of 0.1597.

Model 5:

```
fit<-arima(log(y),order=c(1,1,0),seasonal=list(order=c(2,1,0),period=s))
```

With coefficients:

	ar1	sar1	sar2
	-0.3372	-0.5932	-0.2260
s.e.	0.0996	0.1000	0.0954

And a p-value for the Box test of 0.1231.

2. Forecasting the 2 logarithms models

After undoing the logarithm transformation using exp in the function to score the models, we get the 2 following MAPE scores for the 2 models (using the Recursive Scheme):

	N=1	N=2	N=3	N=4
Model 4 (110|110)	5.456378	7.52264	9.145756	9.275793
Model 5 (110|210)	5.457136	7.043966	8.055772	7.958494

Observation: when compared to the first 3 models developed earlier, we can see that they are very similar in terms of errors. But if we had to choose one, it would be the Model 5 using logarithms for the slightly better score.