Aston Business School
Birmingham

PORTFOLIO TASK 5 – ARIMA MODELS

Introduction

A dataset on UK's Covid – 19 cases which includes number of cases and number of deaths per day from 1 January 2020 up to 14 June 2020 is provided. In this task, the number of cases for the next seven days (15 – 21 June 2020) is to be predicted.

To estimate an appropriate ARIMA Model for forecasting the number of cases, Box Jenkins Methodology in SPSS software will be used. The variables in the dataset include date, day, month, year, cases, deaths, countries and territories, country territory code, population data of 2018 and continent. Ensure that the data is ready for analysis, i.e., it should not have missing values and it should be in the right format.

Analysis

Step 1: Identification of data

There was one case (in excel), where the value of number of cases was negative. It was replaced by the average of previous four days' number of cases as it cannot be negative. The next step is to plot the number of cases.

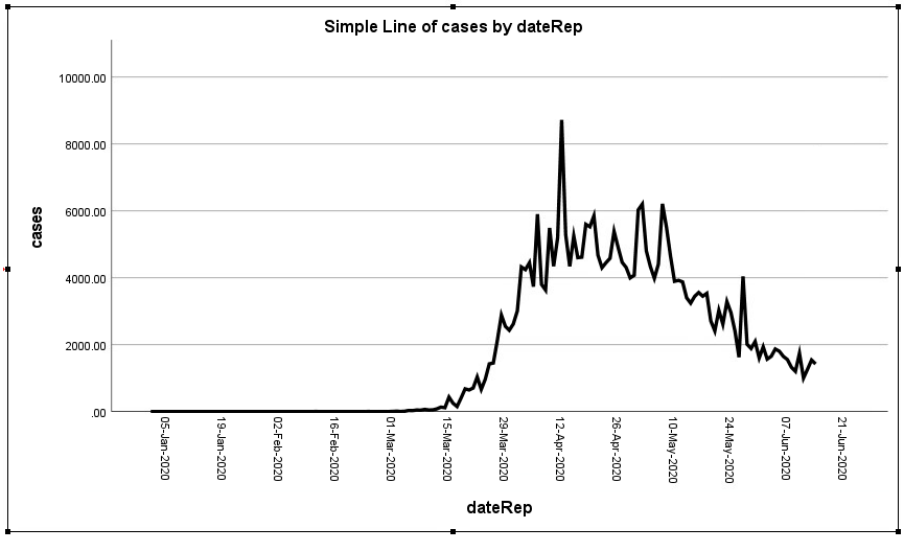

Fig: Line chart of cases over time

From the graph, it can be observed that first there is a sharp increase in the number of cases and then there is a downward trend. It is clear that the data is non stationary. To build an

ARIMA model, that data must be converted to stationary, which can be done by calculating the first difference.

Considering the time, from where the cases are regular. A graph is plotted again to see the variation over time.

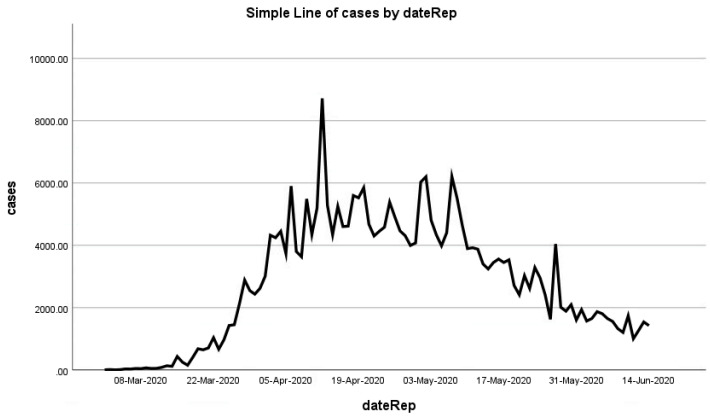

Fig: Line chart showing cases from 01-Mar-20

In this graph, a sinusoidal trend is observed.

For this data, the below ACF and PACF plots are obtained.

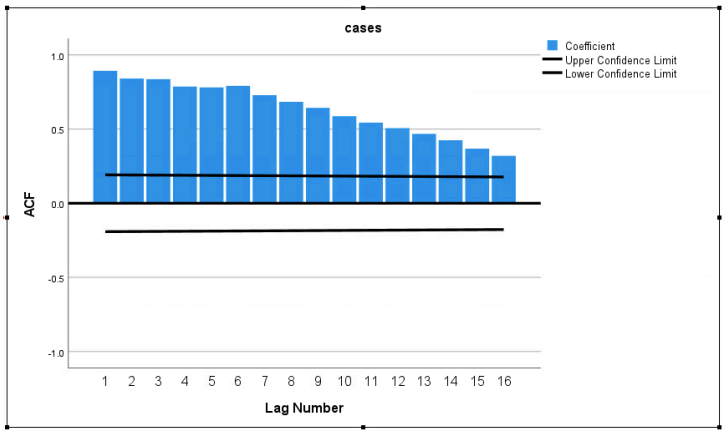

Fig: ACF Plot for initial data

In the ACF plot there is a gradual decay and most of the spikes are significant.

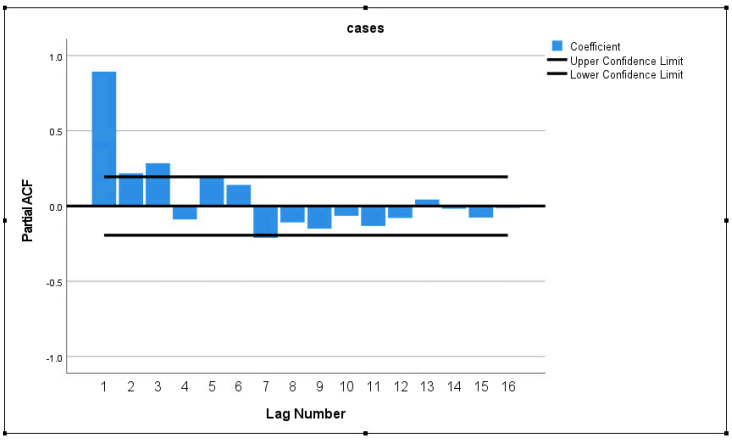

Fig: PACF Plot for initial data

In PACF plot, there is one spike which is significant. From the plots above, the non-stationarity of the data is proved.

Now for any further analysis, the data ought to be converted to stationary data which can be done by calculating the differenced data. This operation can directly be performed in the SPSS by checking the 'difference' option in 'Autocorrelation' of Forecasting option.
The ACF and PACF plots for the differenced data are shown below.

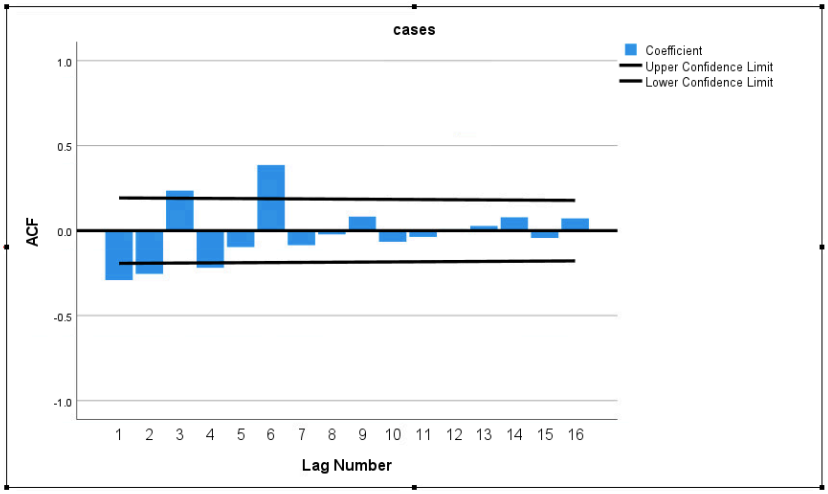

Fig: ACF plot for first differenced data

There is one slightly significant lag in ACF plot at t = 5. There are 5 lags in total.

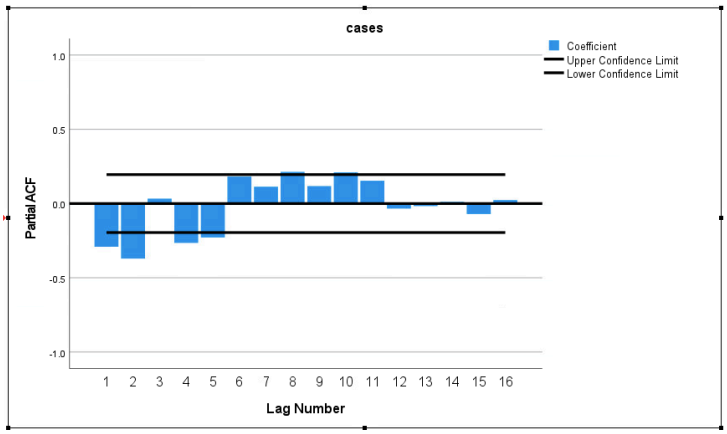

Fig: PACF plot for first differenced data

There are no significant lags in the PACF plot. There are 6 lags in total.

The general form of ARMA model that shows if a series is stationary or not is ARIMA Model, represented as ARIMA (p, d, q). Here, p is time period at which there is a final lag in PACF plot, q is number of lags in ACF plot and d is the number of times difference is calculated to convert a non-stationary data into stationary.

Therefore, from the above information and plots, the ARIMA model in this case is represented as ARIMA (6, 1, 5).

Step 2: Estimation of parameters of model

The ARIMA model is built using the 'Forecasting' option in the SPSS. The Model summary is shown below.

Model Description

			Model Type
Model ID	cases	Model_1	ARIMA(6,1,5)

Model Summary

Model Fit

					Percentile						
Fit Statistic	Mean	SE	Minimum	Maximum	5	10	25	50	75	90	95
Stationary R-squared	.422	.	.422	.422	.422	.422	.422	.422	.422	.422	.422
R-squared	.888	.	.888	.888	.888	.888	.888	.888	.888	.888	.888
RMSE	680.804	.	680.804	680.804	680.804	680.804	680.804	680.804	680.804	680.804	680.804
MAPE	21.653	.	21.653	21.653	21.653	21.653	21.653	21.653	21.653	21.653	21.653
MaxAPE	166.436	.	166.436	166.436	166.436	166.436	166.436	166.436	166.436	166.436	166.436
MAE	446.335	.	446.335	446.335	446.335	446.335	446.335	446.335	446.335	446.335	446.335
MaxAE	2800.245	.	2800.245	2800.245	2800.245	2800.245	2800.245	2800.245	2800.245	2800.245	2800.245
Normalized BIC	13.534	.	13.534	13.534	13.534	13.534	13.534	13.534	13.534	13.534	13.534

Model Statistics

		Model Fit statistics				Ljung-Box Q(18)			
Model	Number of Predictors	Stationary R-squared	R-squared	RMSE	MAE	Statistics	DF	Sig.	Number of Outliers
cases-Model_1	0	.422	.888	680.804	446.335	2.822	7	.901	0

Fig: Table showing model summary of ARIMA (6,1,5)

The Model parameters are shown in the image below. It provides information on different lags (main parameter being their significance). In the AR part there are 6 lags. In the MA part there are 5 lags.

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	R-squared	RMSE	MAE	Ljung-Box Q(18): Statistics	DF	Sig.	Number of Outliers
cases-Model_1	0	.422	.888	680.804	446.335	2.822	7	.901	0

ARIMA Model Parameters

					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	AR	Lag 1	.103	.326	.316	.753
				Lag 2	.218	.260	.839	.404
				Lag 3	-.017	.239	-.069	.945
				Lag 4	-.112	.237	-.474	.636
				Lag 5	-.133	.189	-.702	.484
				Lag 6	.362	.121	2.984	.004
			Difference		1			
			MA	Lag 1	.641	.342	1.874	.064
				Lag 2	.434	.366	1.187	.238
				Lag 3	-.371	.369	-1.006	.317
				Lag 4	-.043	.319	-.134	.894
				Lag 5	-.132	.282	-.470	.640

Fig: Table showing model statistics and parameters of ARIMA (6,1,5)

In the residual plots below, all the lags are between the confidence intervals, this indicates that the model is performing (forecasting) well enough.

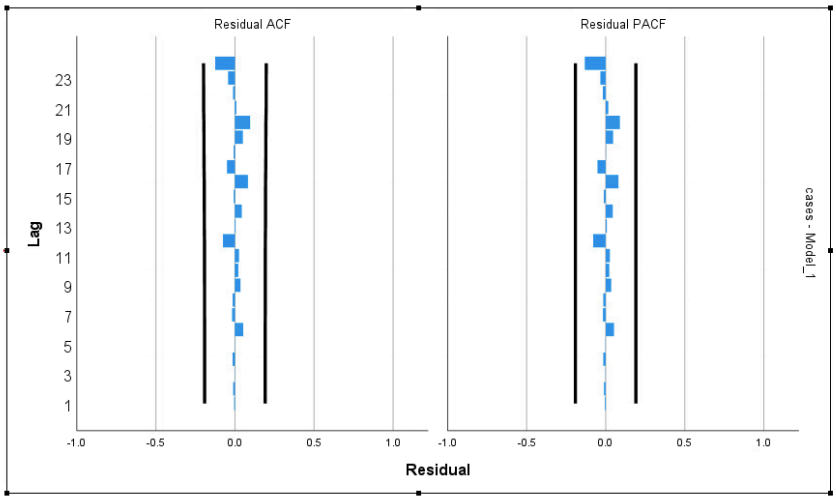

Fig: Graph showing residual ACF and PACF for ARIMA (6,1,5)

Step 3: Diagnostic Checking

The Model's adequacy can be checked by observing the 'Ljung Box' in Model Statistics. The Mean Absolute Error (MAE) value is 446.335. The sig. value should be more than 5% for the model to be adequate.

To find the most parsimonious model, generally, the most insignificant lag should be removed. But in SPSS, we cannot choose the lag we wish to remove. As the highest value of insignificance is observed in AR model, remove a lag in the AR model first. It is important to ensure less MAE value.

After removing a lag, the following values are obtained for ARIMA (5,1,5). The MAE value is 451.652. There are still few insignificant values which can be removed, the highest value is observed in MA model this time.

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	R-squared	RMSE	MAE	Ljung-Box Q(18): Statistics	DF	Sig.	Number of Outliers
cases-Model_1	0	.376	.879	703.862	451.652	12.912	8	.115	0

ARIMA Model Parameters

					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	AR	Lag 1	-.556	.245	-2.272	.025
				Lag 2	-.068	.241	-.282	.779
				Lag 3	-.110	.207	-.531	.597
				Lag 4	-.192	.182	-1.054	.294
				Lag 5	-.554	.134	-4.139	<.001
			Difference		1			
			MA	Lag 1	-.009	32.335	.000	1.000
				Lag 2	.494	32.052	.015	.988
				Lag 3	-.133	16.038	-.008	.993
				Lag 4	-.107	11.695	-.009	.993
				Lag 5	-.471	15.172	-.031	.975

Fig: Table showing model statistics for ARIMA (5,1,5)

Removing a lag from the MA model, the following model is obtained. The model for ARIMA (5,1,4) is shown below.

Model Description

Model ID			Model Type
Model ID	cases	Model_1	ARIMA(5,1,4)

Model Summary

Model Fit

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile: 5	10	25	50	75	90	95
Stationary R-squared	.413	.	.413	.413	.413	.413	.413	.413	.413	.413	.413
R-squared	.886	.	.886	.886	.886	.886	.886	.886	.886	.886	.886
RMSE	678.802	.	678.802	678.802	678.802	678.802	678.802	678.802	678.802	678.802	678.802
MAPE	21.314	.	21.314	21.314	21.314	21.314	21.314	21.314	21.314	21.314	21.314
MaxAPE	166.300	.	166.300	166.300	166.300	166.300	166.300	166.300	166.300	166.300	166.300
MAE	439.472	.	439.472	439.472	439.472	439.472	439.472	439.472	439.472	439.472	439.472
MaxAE	2997.246	.	2997.246	2997.246	2997.246	2997.246	2997.246	2997.246	2997.246	2997.246	2997.246
Normalized BIC	13.440	.	13.440	13.440	13.440	13.440	13.440	13.440	13.440	13.440	13.440

Fig: Image showing model statistics of ARIMA (5,1,4)

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	R-squared	RMSE	MAE	Ljung-Box Q(18): Statistics	DF	Sig.	Number of Outliers
cases-Model_1	0	.413	.886	678.802	439.472	4.604	9	.867	0

ARIMA Model Parameters

					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	AR	Lag 1	.966	.238	4.061	<.001
				Lag 2	-.775	.271	-2.855	.005
				Lag 3	.668	.228	2.927	.004
				Lag 4	-.587	.133	-4.409	<.001
				Lag 5	.339	.119	2.850	.005
			Difference		1			
			MA	Lag 1	1.513	.246	6.143	<.001
				Lag 2	-1.048	.409	-2.565	.012
				Lag 3	.630	.390	1.617	.109
				Lag 4	-.393	.221	-1.776	.079

Fig: Image showing model statistics of ARIMA (5,1,4)

In this case, all the values are significant except one. The MAE has decreased to 439.472. The. sig value is 0.867. On removing the final lag from the MA model, the MAE keeps increasing. Therefore, ARIMA (5,1,4) is the parsimonious model.

Step 4: Forecasting

As the satisfactory model is found, the values can now be predicted. To forecast the values from 15 June 2020 to 21 June 2020, add these dates in the data and set range in 'select cases' for forecasting in SPSS as it cannot forecast for the data not mentioned in the dataset. After running the model, the following forecast values are obtained.

Forecast

Model		167	168	169	170	171	172	173
cases-Model_1	Forecast	1370.16	1344.27	1096.37	1065.19	1202.70	1190.50	1188.02
	UCL	2717.10	2822.70	2598.42	2666.82	2841.96	2930.66	3255.74
	LCL	23.21	-134.16	-405.68	-536.45	-436.57	-549.66	-879.69

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

Fig: Table showing the forecasted values

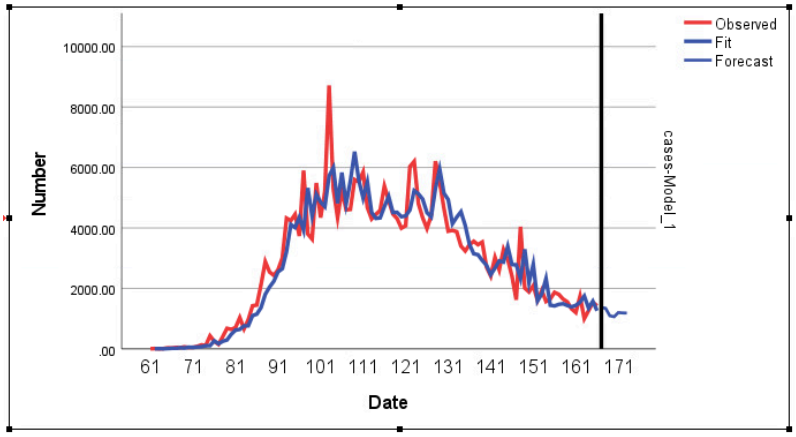

Fig: Graph showing the predicted values

Conclusion

ARIMA models are the combination of of AR(p) and MA(q) processes called autoregressive moving average (ARMA (p, q)). These are used to forecast the number of covid cases in this case using Box-Jenkins Methodology. The most important step is to identify the type of data as it decides the next steps of estimating and testing the model. Finally, ARIMA (5,1,4) is found to be the best model with less MAE and good adequacy.