IC 272: DATA SCIENCE - III

LAB ASSIGNMENT – VI

Auto-regression

Student's Name: Arpit Singh **Mobile No: 6265104315**

Roll Number: B20084 **Branch: CSE**

1 a.

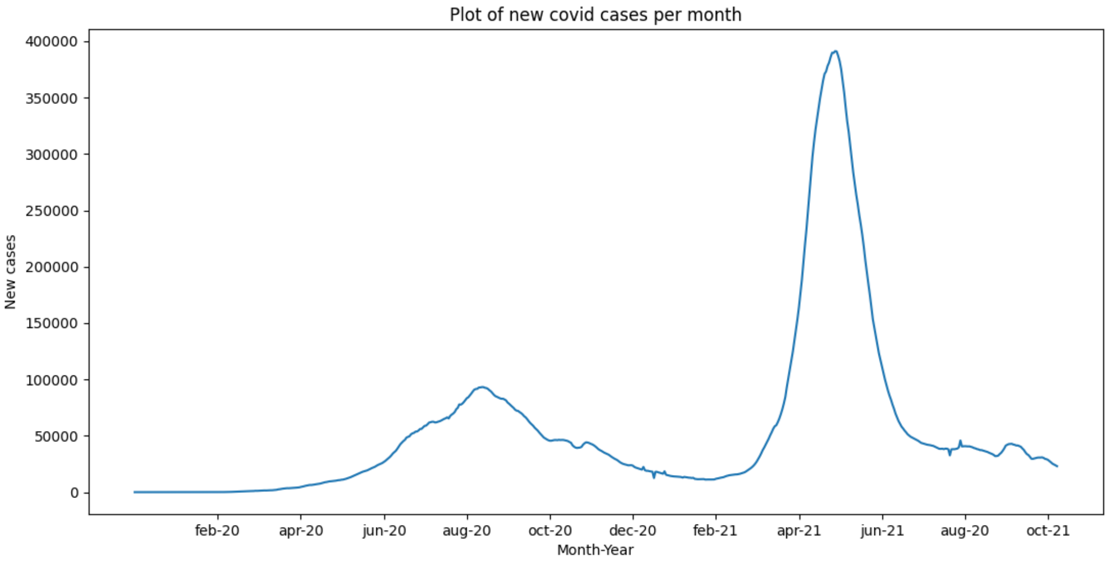


Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. The days one after the other have similar power consumption.
2. The reason behind the similarity is because the graph is continuous so the consequent days have similar no. of confirmed cases.
3. First wave was started at jun 20 and subsided around oct 20 and peaked on Aug and second waves was shorter as compared to first in duration but in cases it was much larger and happened around apr to jun 21.

b. The value of the Pearson's correlation coefficient is 0.999.

Inferences:

1. From the value of Pearson’s correlation coefficient, we can infer that the degree of correlation between the two time sequences is highly correlated.
2. We generally expect observations (here number of COVID-19 cases) on days one after the other to be similar. This is because the Pearson’s correlation coefficient is almost 1 so it should be similar as it is highly correlated.
3. They are highly correlated and consequent days have similar covid cases because the coefficient of correlation is almost equal to 1, that's the reason behind the above two inferences.

c.

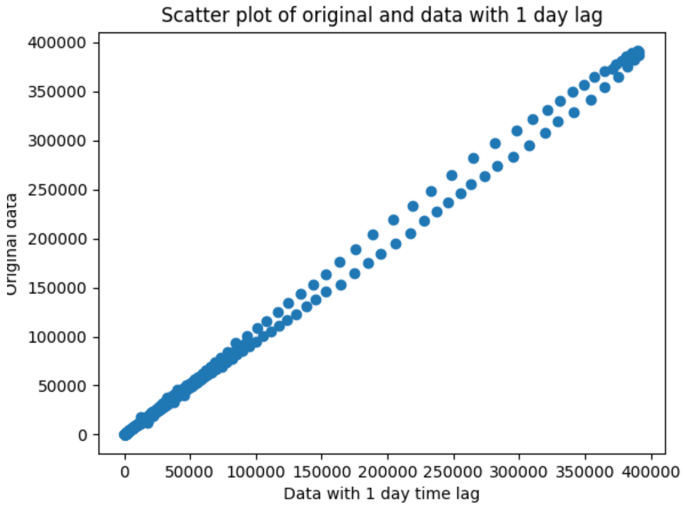


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of the spread of data points, correlation between the two sequences is very high.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson’s correlation coefficient calculated in 1b.
3. As it can be observed that the curve is around the diagonal of the plot so its correlation must be near 1 as calculated and inferred.

d.

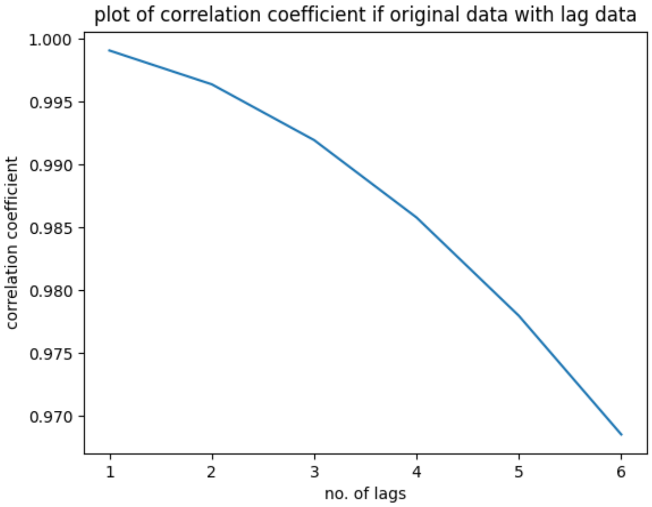


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Correlation coefficient value decreases with respect to increase in lags in time sequence.
2. As the no. of cases is similar to the days which are near to each other, increasing the lag increases the deviation causing decrease in correlation.

e.

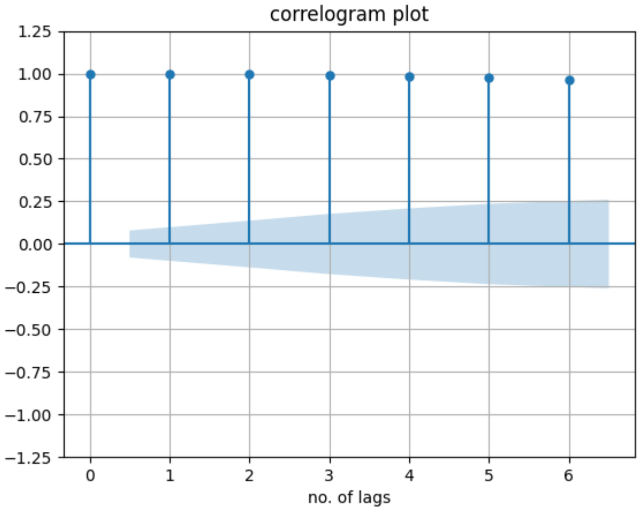


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Correlation coefficient value decreases with respect to increase in no. of lags in time sequence.
2. As the no. of cases is similar to the days which are near to each other, increasing the lag increases the deviation causing decrease in correlation.

2

a. The coefficients obtained from the AR model are:- Wo = 59.955, W1 = 1.037, W2 = 0.262, W3 = 0.028, W4 = -0.175, W5 = -0.152.

b. i.

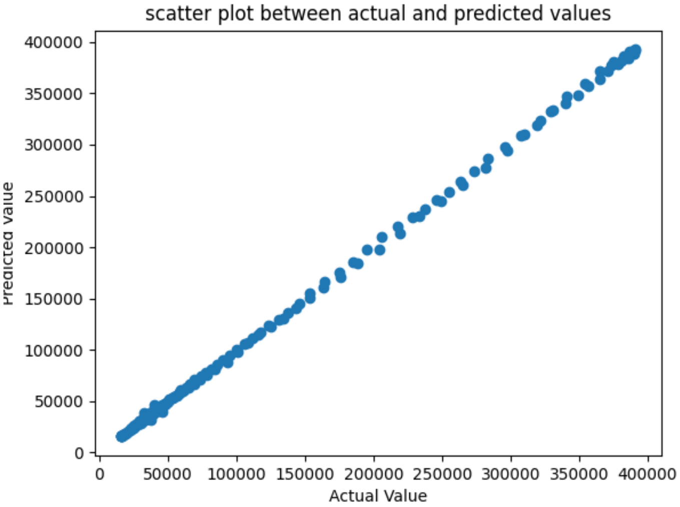
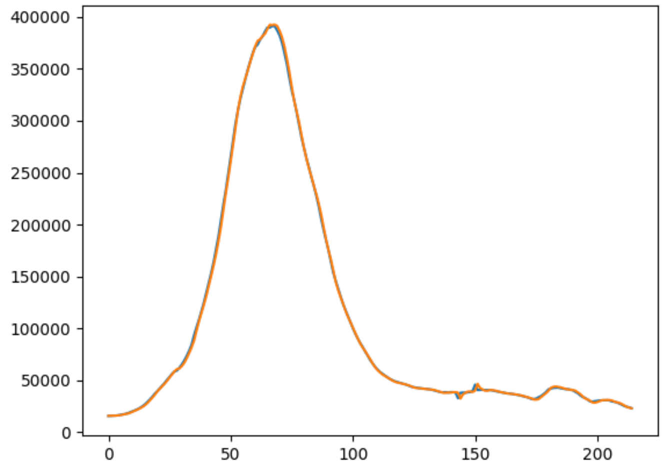


Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, The nature of the correlation between the two sequences is highly correlated and its value is almost 1.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. Our model predicts correctly, as the correlation is almost equal to 1, and it also reflects the observation observed before.

ii.



Inferences:

1. From the plot of predicted test data time sequence vs. original test data sequence we can observe that our model is reliable as it gives almost identical predictions. So it will be reliable for future predictions.

iii.

The RMSE(\%) and MAPE between predicted power consumed for test data and original values for test data are 1.825, 1.575 respectively.

Inferences:

1. From the value of RMSE(\%) and MAPE value we can observe that our model is reliable and accurate.
2. The reason behind the above inference is the value of RMSE(\%) and MAPE. it is below 2 which shows how reliable the data is.

3

Table 1 RMSE (%) and MAPE between predicted and original data values wrt lags in time sequence

Lag value	RMSE (%)	MAPE
1	**5.373**	**3.446**
5	**1.825**	**1.575**
10	**1.685**	**1.519**
15	**1.612**	**1.496**
25	**1.703**	**1.535**

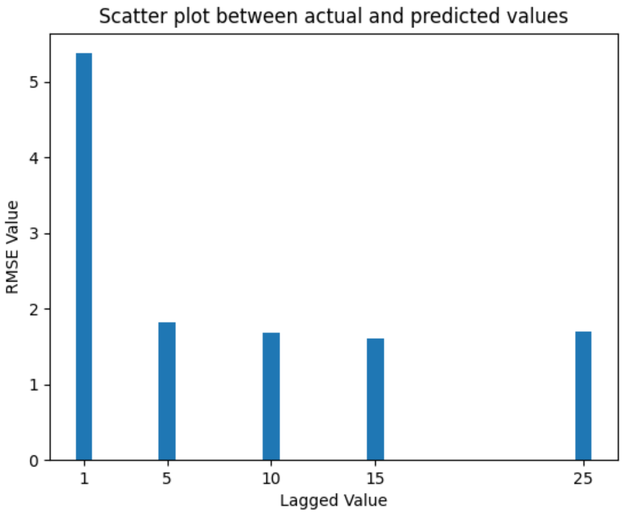


Figure 7 RMSE(%) vs. time lag

Inferences:

1. RMSE(%) decreases quickly from 1 to 5 but after that it decreases gradually almost constantly with an increase in lag value.
2. The reason for the above inference is the complex model needed to fit our data more accurately. So when the lag is increased from 1 to 5 the accuracy improves significantly but after that the increase is gradual.

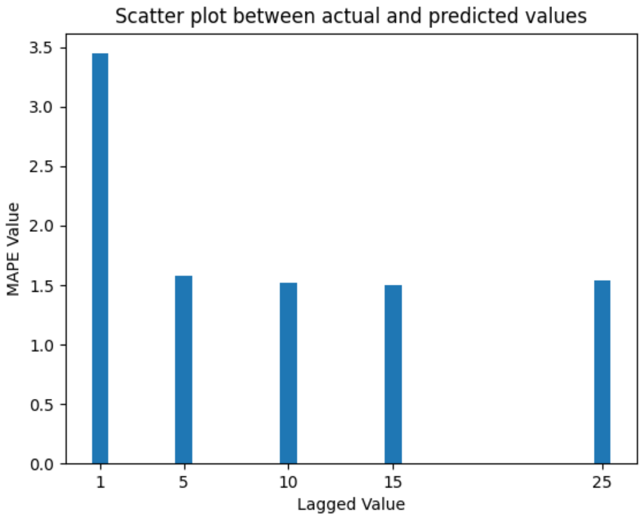


Figure 8 MAPE vs. time lag

Inferences:

1. MAPE decreases quickly from 1 to 5 but after that it decreases gradually almost constantly with an increase in lag value.
2. The reason for the above inference is the complex model needed to fit our data more accurately. So when the lag is increased from 1 to 5 the accuracy improves significantly but after that the increase is gradual.

4

The heuristic value for the optimal number of lags is 77.

The RMSE(%) and MAPE value between test data time sequence and original test data sequence are 1.759, 2.026 respectively.

Inferences:

1. No, Based upon the RMSE(%) and MAPE value, heuristics for calculating the optimal number of lags doesn’t improve the prediction accuracy of the model
2. The reason behind Inference 1 can be seen that the RMSE (%) for lag value 10 is less than that of the optimal lag value and we can also see that as the observations are made for every day AR(77) doesn't make sense than that of a lag of around one day.
3. The prediction accuracy obtained without heuristic is less as compared to with heuristic for calculating optimal lag with respect to RMSE (%) and MAPE values.