Sainyukta Gujar

22060641026

Time Series Analysis Minitab Assignment

**Defining the Question:**

In the question a dataset is provided containing the number of Covid positive cases and number of deaths due to Covid in India from 24th of March, 2020 to 11th May, 2020. It is expected to perform a time series analysis on a dataset, which includes Smoothing the dataset, ACF and PACF study and transformations, if required.

**Analysis of the dataset:**

Tools used:

The ACF tool is used to plot the autocorrelation function and PACF tool is used to plot the partial autocorrelation function. This tools are useful in determining the order of the ARIMA (Auto Regressive Integrated Moving Average) models. Also the descriptive statistics and graphical summary tools are used to perform EDA.

Overview:

A time series is actually a set of observations measured at specified (usually equal) time interval. Here, the interest lies in smoothing the dataset and drawing insights about ARIMA (p, d, q) models using the ACF and PACF plots.

In the dataset there are 49 observations and 2 variables (number of Covid cases and number of deaths) and another column is Date. The observations are measured along a specified and equal time interval and thus the dataset is a time series.

Moving forward, a smoothing technique among the three smoothing techniques (Exponential Smoothing, Holts Smoothing and Holts-Winter Smoothing) has to be applied on the dataset, to conclude which technique gives the best result. The ACF and PACF plots need to be plotted to identify the number of lags and hence decide the order of the ARIMA model. ACF is used to determine the order of MA model as it cuts off after lag q and PACF is used to determine the order of AR model as it cuts off after lag p. Appropriate transformation methods are expected to be used if required.

Softwares used:

The softwares used for the analysis are MS Excel and Minitab. The data for number of Covid cases and number of deaths is cumulative in nature. Hence MS-Excel is used to de cumulate the dataset and then use it for analysis. Minitab is used for smoothing the dataset, plotting ACF and PACF, fitting the ARIMA model and for transformations.

Outputs and Graphs:

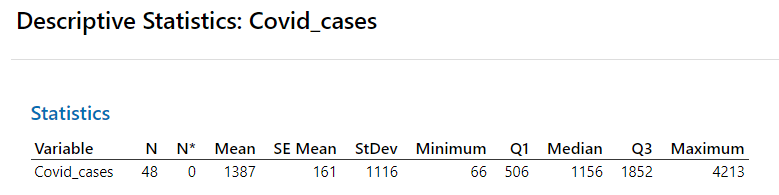


Fig. (1)

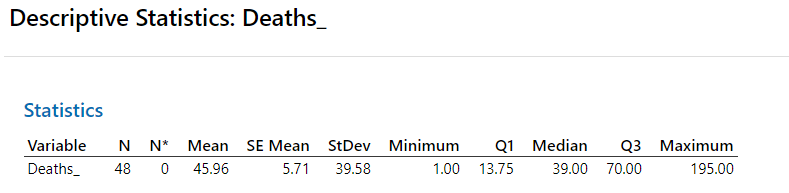


Fig. (2)

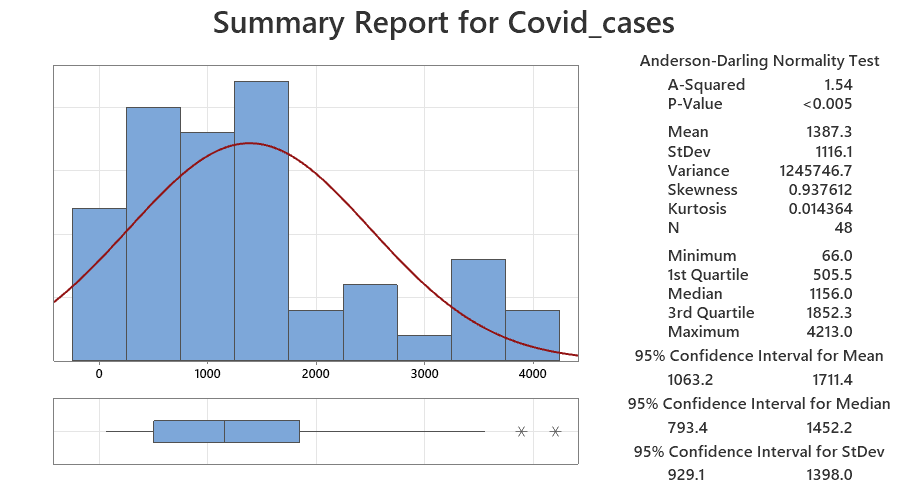


Fig. (3)

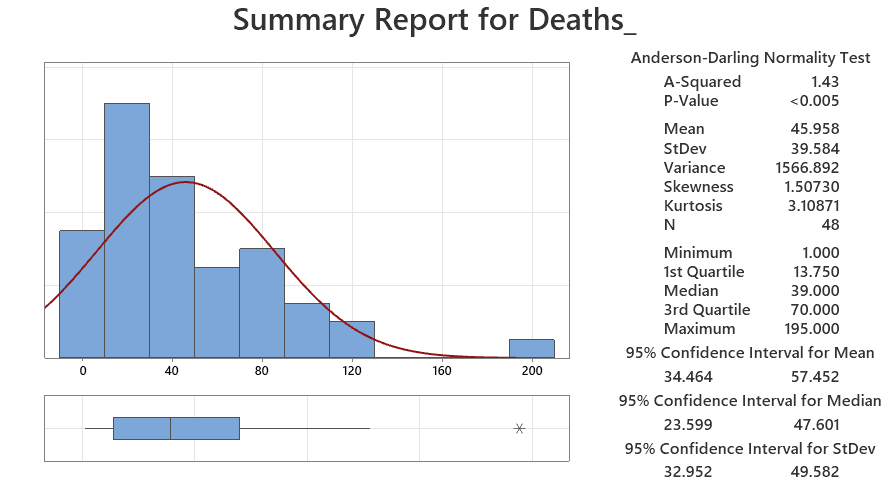


Fig. (4)

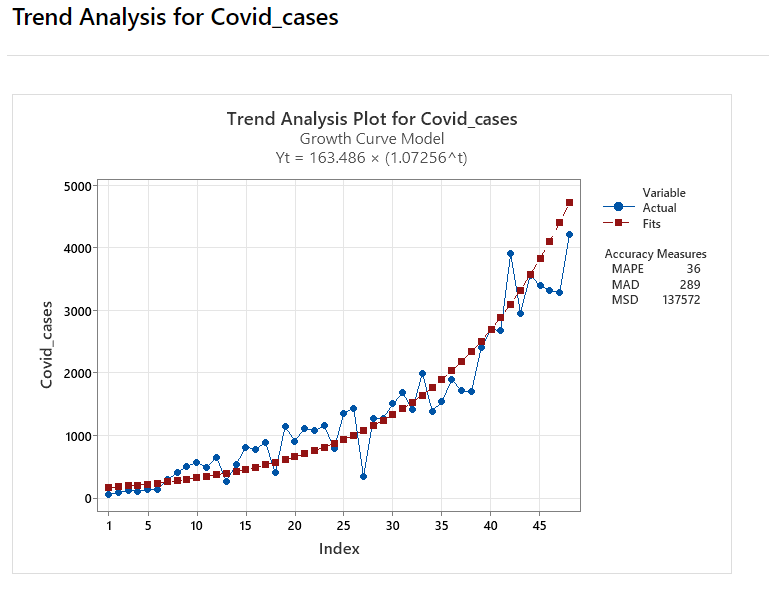


Fig. (5)

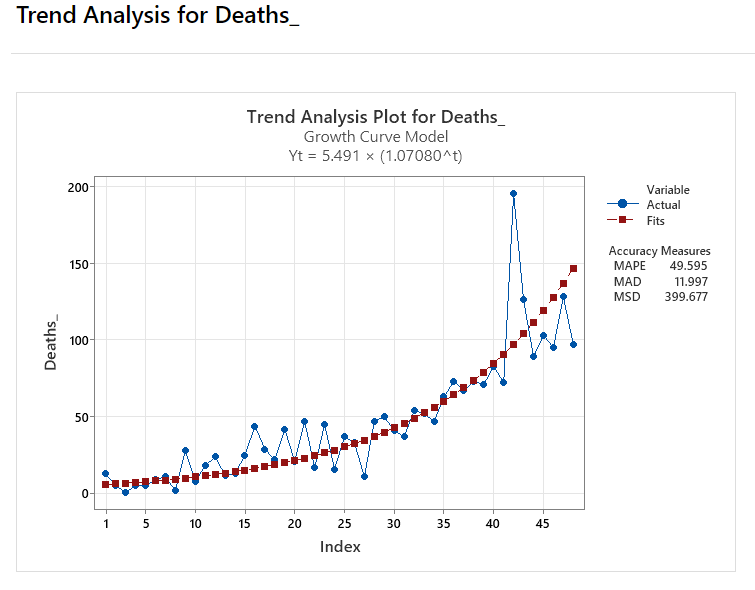


Fig. (6)

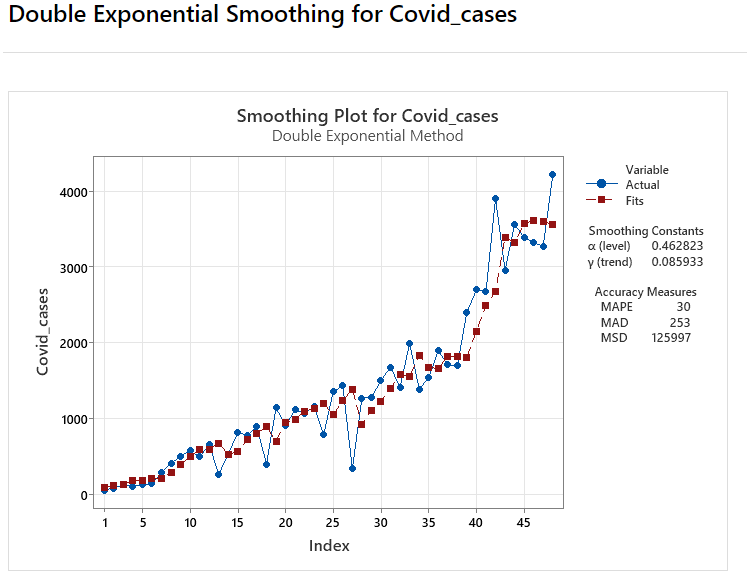


Fig. (7)

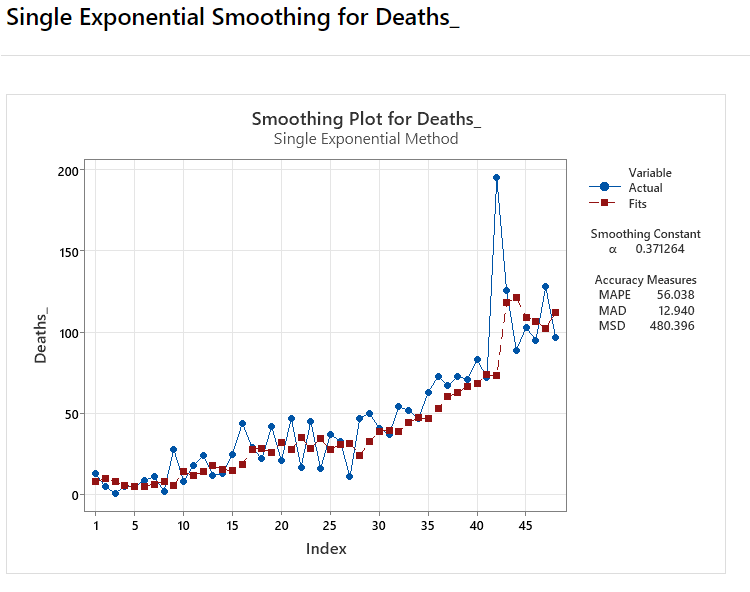


Fig. (8)

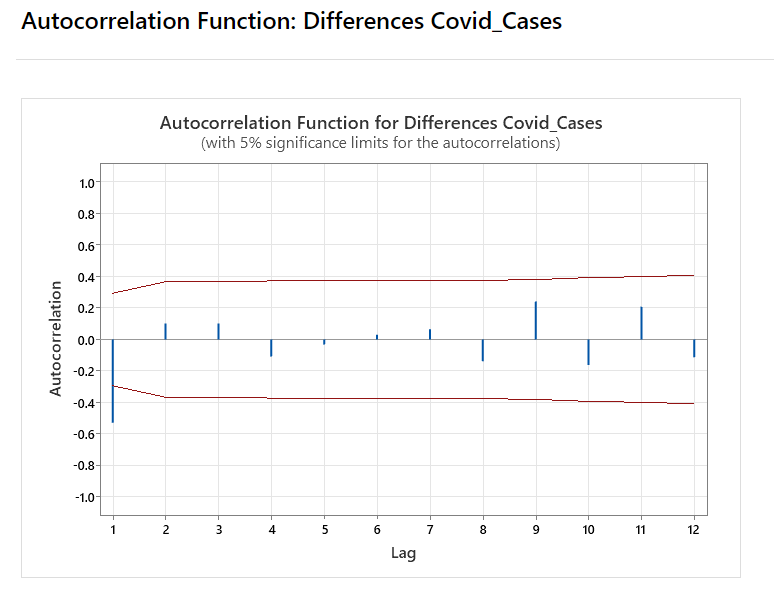


Fig. (9)

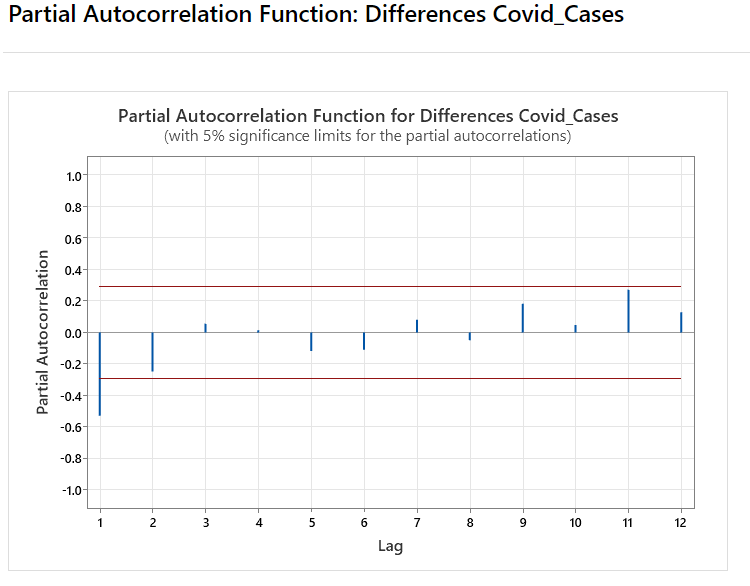


Fig. (10)

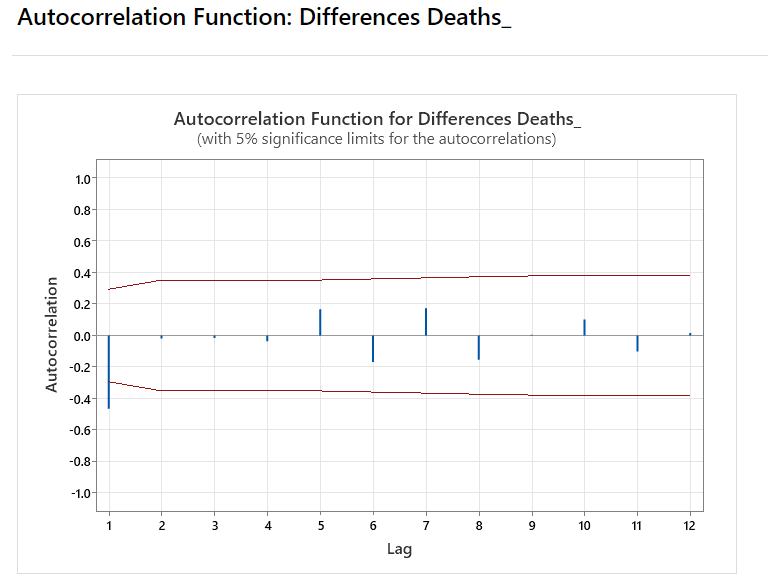


Fig. (11)

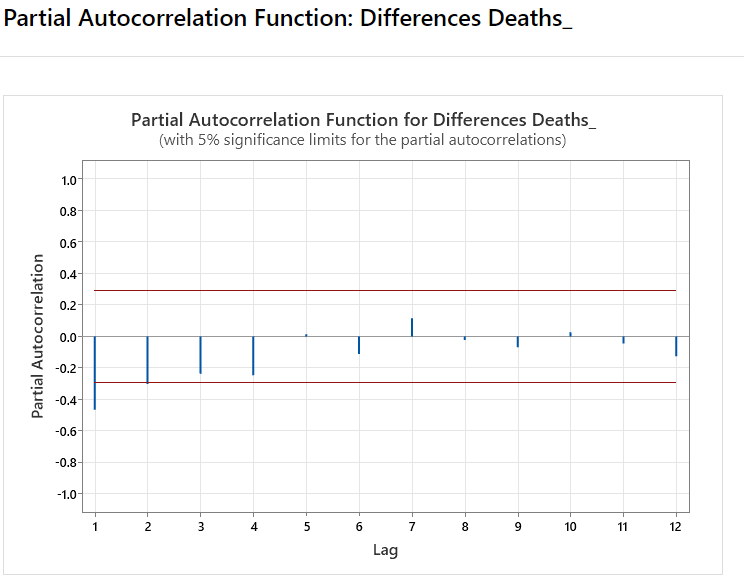


Fig. (12)

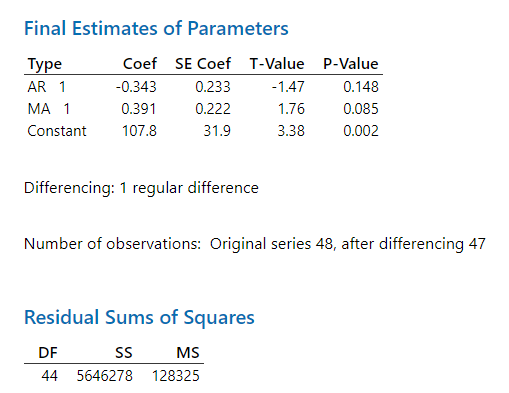


Fig. (13)

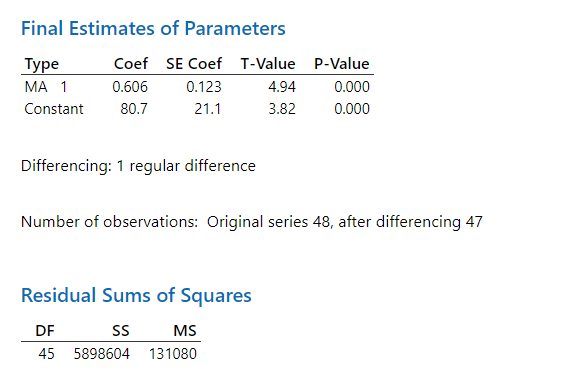


Fig. (14)

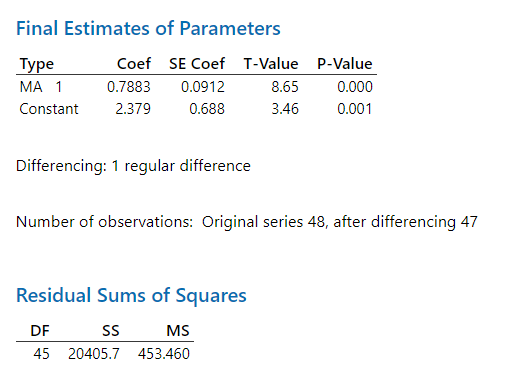


Fig. (15)

Exploratory Data Analysis:

From the Descriptive Stats function in the Stats tab on Minitab, the descriptive statistics for both the columns (number of Covid cases and number of deaths can be) obtained. The first row is deleted from the dataset as there arises a need to decummulate the data and the first row remains empty since the preceding observation is subtracted from the given observation.

* For Covid cases it can be observed that in total there are 48 observations without any missing observation. The mean number of Covid cases over the given period is 1387 and the standard deviation is 1116. Minimum number of Covid positive cases is 66 whereas the maximum number of Covid positive cases in a day is 4213. Also it is observed that the data is positively skewed and platykurtic. (Ref. Fig. (1) and Fig. (3))
* For deaths it can be observed that in total there are 48 observations without any missing observation. The mean number of deaths over the given period is 46 and the standard deviation is 39.58. Minimum number of deaths due to Covid is 1 whereas the maximum number of deaths due to Covid in a day is 195. Also it is observed that the data is positively skewed and mesokurtic. (Ref. Fig. (2) and Fig. (4))

**Interpretation:**

To start with trend analysis is performed for the number of Covid cases and it is observed that the MAPE value for the growth curve model is 36 and is the least when compared with the MAPE values of linear model and Quadratic model (64 and 44 respectively). Hence concluded that Exponential model is best fit for data with the equation as Yt =163.486\*(1.07256^t). (Ref. Fig. (5)).

Further, trend analysis is performed for the number of deaths and it is observed that the MAPE value for the growth curve model is 49.595 and is the least when compared with the MAPE values of linear model and Quadratic model (69.753 and 63.156 respectively). Hence concluded that Exponential model is best fit for data with the equation as Yt =5.491\*(1.07080^t). (Ref. Fig. (6))

Then, a double exponential smoothing technique is used for the number of Covid cases and it can be observed that α and γ values are closer to zero when compared with the two other techniques (single exponential smoothing technique uses α=0.568 and triple exponential smoothing technique uses α=0.2 and MAPE=36) suggesting that the data is smoothened to a very good extend. (Ref. Fig. (7))

Moving to the number of deaths the single exponential smoothing technique is used and the value of α is very closer to zero when compared with the two other techniques (double exponential smoothing technique uses α=0.67 and triple exponential smoothing technique uses α=0.2 and MAPE=58.046) suggesting that the data is smoothened to a very good extend. (Ref. Fig. (8))

Since the dataset is non stationary, the next step is to use differencing. In the case of number of Covid cases it can be seen that when plot of ACF and PACF is plotted for differences a lag of 1 is obtained. (Ref. Fig. (9) and Fig. (10)).

Also, in the case of number of deaths it can be seen that when plot of ACF and PACF is plotted for differences a lag of 1 is obtained. (Ref. Fig. (11) and Fig. (12)).

Lastly, an ARIMA (1, 1, 1) is fitted in the case of number of Covid cases since the ACF and PACF plots of differences for number of Covid cases indicate a lag of 1 for both the AR and MA models. However it can be observed that though the MS value appears to be less, the p-values indicate that the model is not statistically significant and hence an ARIMA (0, 1, 1) is fitted which has a MS value nearly as the same as the ARIMA (1, 1, 1) model but whose p values are statistically significant. (Ref. Fig. (13) and Fig. (14)).

For the case of number of deaths ARIMA (0, 1, 1) is fitted since the ACF and PACF plots of differences for number of deaths indicate a lag of 1 for both the AR and MA models. Also, it can be noted that the MS value appears to be very less and the p-values indicate that the model is statistically significant. (Ref. Fig. (15))

**Conclusion:**

From the trend analysis it can be concluded that both number of Covid cases and number of deaths follow an exponential growth curve indicating that both the numbers are expected to rise exponentially.

Single Exponential Smoothing technique is appropriate in case of number of deaths and Double Exponential Smoothing technique is appropriate in case of number of Covid cases.

From the ACF and PACF plots for differences it is concluded that ARIMA(1,1,1) model is best fit for number of Covid cases when value of MS is taken into consideration, but ARIMA(0,1,1) is a more statistically significant model. On the other hand, for the number of deaths, ARIMA(0,1,1) is the best fit model since it has p values that are statistically significant and lest MS value.