**DEATH RATES FORECAST IN INDIA DUE TO COVID-19**

**By D.Pavithra,**

**21-PST-025**

**M.Sc Statistics (I Year)**

**Loyola College**

**ABSTRACT:**

COVID-19, a novel coronavirus, is currently a major worldwide threat. It has infected more than a million people globally leading to hundred-thousands of deaths. In such grave circumstances, it is very important to predict the future death numbers to support prevention of the disease and aid in the healthcare service preparation. The most important factor in preventing the spread of the Virus locally is to empower the citizens with the right information. Following that notion, I have developed a model and fitted it for forecasting the number of deaths due to COVID-19 in India. The graph plotted using the time series analysis model indicates an upward trend in the curve, showing an exponential increase in the no. of deaths in the future.

1. **INTRODUCTION:**

The pandemic of COVID-19 started from Wuhan, China in December 2019. The spread was slow initially, and the first case was reported in Kerala, India on 30th Jan 2020. The number of cases were very slow till March and it started increasing rapidly by March 2020. Eventually, some of the cases turned out as Deaths and the first death in India was reported on 13th March 2020.

At present, there is neither a treatment nor a vaccination for the COVID-19 infection. The only option left is to predict the future cases and take necessary precautions against it.

Modelling the disease and providing forecast for future no. of deaths may help the medical professionals to work in an efficient and precautious way. I’ve used Auto-regressive Model (AR Model) which is more realistic and has least AIC and Mean Squared Error compared to other models. I have employed the best AR Model that fits my dataset and have predicted the no. of deaths for next 45 days (till July 30th, 2020) using python.

1. **METHODOLOGY**
2. **Datasets:**

Deaths cases due to COVID-19 in India along with the date of occurrence have been collected from the official website of Kaggle. The first death reported in India was on 13th March 2020. So I’ve taken the dataset from 13th March to 16th June to build predictive model.

**Source** – [www.kaggle.come](http://www.kaggle.come)

1. **Model Development**

An autoregressive (AR) model predicts future behavior based on past behavior. It’s used for forecasting when there is some correlation between values in a time series and the values that precede and succeed them. AR Model is the best prediction model for forecasting a time series.

**Model**

An AR(p) model is an autoregressive model where specific lagged values of yt are used as [predictor variables](https://www.statisticshowto.com/independent-variable-definition/#Predictor). Lags are where results from one time period affect following periods.

The value for “p” is called the *order*.

**The AR(p) model is defined by the equation:**  
yt = δ + φ1yt-1 + φ2yt-2 + … + φpyt-1 + At  
Where:

* yt-1, yt-2…yt-p are the past series values (lags),
* At is white noise (i.e. randomness)

1. **PYTHON**
2. **Packages Used:**

* **pandas** - Since my datasets are in tabular form, I imported the package ‘pandas’ to organize and manipulate my data. And also, pandas.plotting was used to plot autocorrelation.
* **numpy** - Data analysis is all about playing around numbers. Since most of my datasets are numbers, I imported the package ‘numpy’ to manipulate numerical data.
* **matplotlib.pyplot** – It is easy to understand data when it is visually represented. I imported the package ‘matplotlib’ to visualize 2D plots of array like bar diagram, scatter plot etc.
* **statsmodels -**  It is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. I imported AR model and plotted partial autocorrelation plot using this package.

1. **Code:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as pyplt

from pandas.plotting import autocorrelation\_plot

from statsmodels.graphics.tsaplots import plot\_pacf

from statsmodels.tsa.ar\_model import AR

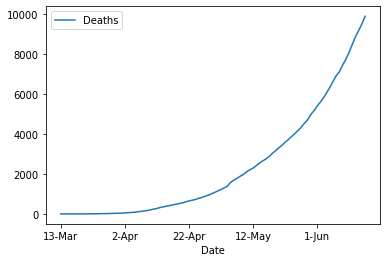
from sklearn.metrics import mean\_squared\_error

df = pd.read\_csv('Covid\_final.csv',header=0, parse\_dates=[0],index\_col=0,squeeze=True)

**#NUMBER OF DEATHS – PLOT**

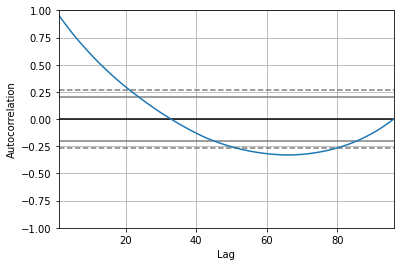
df.plot()

pyplt.show()



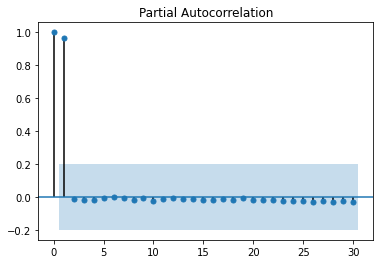
**#Auto-Correlation Plot**

autocorrelation\_plot(df)

****

**#Partial-Autocorrelation Plot**

plot\_pacf(df,lags=30)



**#SPLITTING THE DATA INTO 80% Train and 20% Test**

X = df.values

train = X[0:77]

test = X[77:]

**#AR MODEL - Fitting**

model\_ar = AR(train)

model\_ar\_fit = model\_ar.fit()

predictions = model\_ar\_fit.predict(start=77,end=139)

**#Plotting – Test and Predicted Values**

pyplt.plot(test)

pyplt.plot(predictions,color='red')

**#To check the Goodness of the Fit**

mean\_squared\_error(test,predictions)

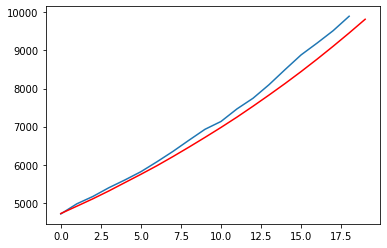
**#Forecasting the No. Of Deaths for next 45 Days**

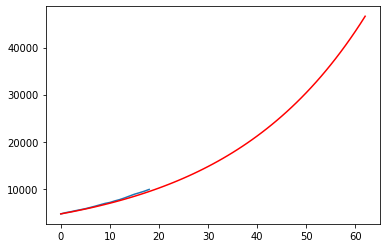
pd.DataFrame(date,predictions)

**4) RESULT & CONCLUSION:**

The no. of deaths for next 45 days is listed as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Predicted No. Of Deaths** | **Date2** | **Predicted No. Of Deaths3** |
| 15th June 2020 | 9111.05 | 8th July 2020 | 21233.36 |
| 16th June 2020 | 9460.06 | 9th July 2020 | 22013.8 |
| 17th June 2020 | 9820.66 | 10th July 2020 | 22822.17 |
| 18th June 2020 | 10195.63 | 11th July 2020 | 23659.3 |
| 19th June 2020 | 10584.58 | 12th July 2020 | 24526.08 |
| 20th June 2020 | 10986.37 | 13th July 2020 | 25423.62 |
| 21st June 2020 | 11401.84 | 14th July 2020 | 26353.17 |
| 22nd June 2020 | 11832.53 | 15th July 2020 | 27315.85 |
| 23rd June 2020 | 12279.32 | 16th July 2020 | 28312.69 |
| 24th June 2020 | 12741.79 | 17th July 2020 | 29344.91 |
| 25th June 2020 | 13219.91 | 18th July 2020 | 30413.87 |
| 26th June 2020 | 13715.17 | 19th July 2020 | 31520.91 |
| 27th June 2020 | 14228.64 | 20th July 2020 | 32667.3 |
| 28th June 2020 | 14760.4 | 21st July 2020 | 33854.39 |
| 29th June 2020 | 15310.64 | 22nd July 2020 | 35083.7 |
| 30th June 2020 | 15880.27 | 23rd July 2020 | 36356.76 |
| 1st July 2020 | 16470.49 | 24th July 2020 | 37675.11 |
| 2nd July 2020 | 17081.94 | 25th July 2020 | 39040.3 |
| 3rd July 2020 | 17714.88 | 26th July 2020 | 40454.01 |
| 4th July 2020 | 18370.09 | 27th July 2020 | 41918.02 |
| 5th July 2020 | 19048.76 | 28th July 2020 | 43434.11 |
| 6th July 2020 | 19751.77 | 29th July 2020 | 45004.09 |
| 7th July 2020 | 20479.72 | 30th July 2020 | 46629.88 |





Blue Line – Actual no. of Deaths

Red Line – Predicted no. Of deaths

The above plot shows the time series graph of the no. of deaths due to COVID-19 from June 15th – July 30th. An **increasing trend** is shown by the Red line which depicts the predicted no. of deaths in future, hence, **suggesting a rise in deaths due to COVID-19.**

As per the model forecast, the death cases are expected to greatly rise in the coming days. The time series analysis shows an exponential enhancement in the no. of deaths. And also, by the **end of July 30**, if the situation seems to be the same, the **no. of death cases will reach 50k**. However, social distancing and sanitizing can help in controlling the cases and the situation.