COVID-19 Time Series Forecasting (India)

**Objective**: The objective of this project is to analyse the time-series of SARS-CoV-2/Covid-19 confirmed and death cases and study the trends to forecast expected future trends in India.

About Covid-19: What is corona virus Corona viruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered coronavirus causes coronavirus disease COVID-19. COVID-19 is the infectious disease caused by the most recently discovered corona virus. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019.

Summary: I collected data about the confirmed and deaths till October and predicted the same for November, and then compared actual data to the predicted one.

The following libraries were used:

#libraries  
library(openxlsx) #xsls file

## Warning: package 'openxlsx' was built under R version 4.0.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.3

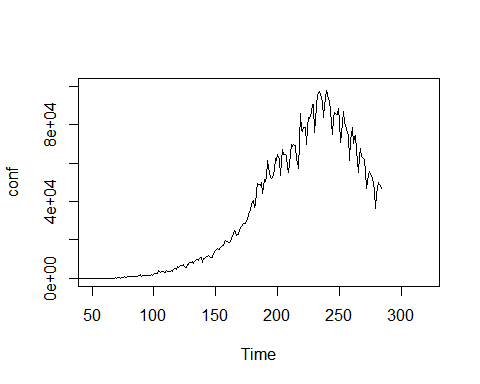
Raw data is imported and manipulated as per requirements. We ensure that the datatype in the imported file is interpreted exactly as what it was before, or at least what we want it to be interpreted as.

#processing raw data  
oct\_data=read.xlsx("datasets/conf\_new\_oct.xlsx",sheet = 1,startRow = 1,colNames = T,rowNames = F,detectDates = T)  
nov\_data=read.xlsx("datasets/conf\_new\_nov.xlsx",sheet = 1,startRow = 1,colNames = T,rowNames = F,detectDates = T)  
oct\_data[,1]<-as.Date(oct\_data[,1])  
class(oct\_data[,1])

## [1] "Date"

For now we do not use data\_nov as it will be later used to compare to the model

conf<-as.ts(oct\_data$Confirmed)  
plot.ts(conf,xlim=c(50,320),ylim=c(0,100000))



The data begins from 22-Jan-2020 and the x axis is the number of days after the date. The values start to show trends after about 60 days so the graph is cropped.

summary(conf)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 436.2 10897.0 28817.2 56517.5 97894.0

Now we are going to fit an Autoregressive model for the data, store predicted values and standard error values. Standard error values will be used to create a confidence interval of 95%.

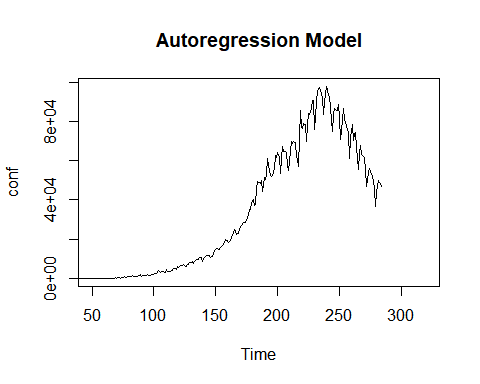
ar\_conf<-arima(conf,order=c(1,0,0)) #autoregression model  
ar\_pred<-predict(ar\_conf,n.ahead=30)$pred  
ar\_se<-predict(ar\_conf,n.ahead=30)$se

CONFIRMED CASES PREDICTION

The following few graphs show how data of confirmed cases can be used to predict for next month and comparison of the same is shown with the actual data of November.

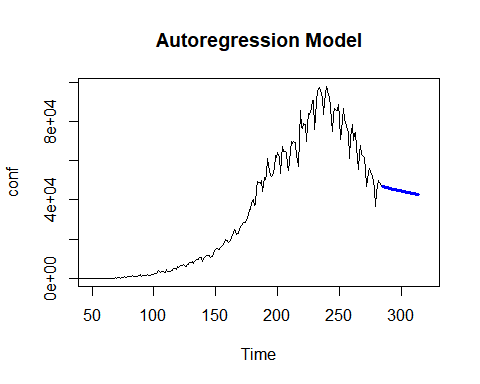
Actual data of confirmed cases till October end.

plot.ts(conf,xlim=c(50,320),main="Autoregression Model")



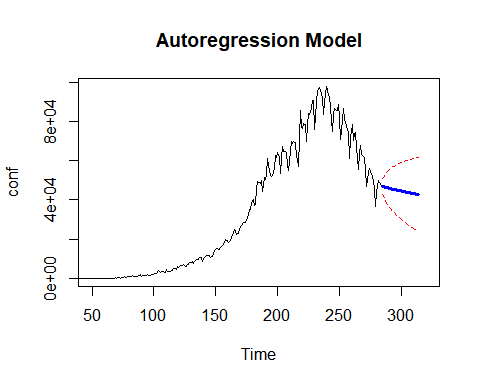
Predicted values are added to the graph.

plot.ts(conf,xlim=c(50,320),main="Autoregression Model")  
points(ar\_pred,type='l',col="blue",lwd=3)



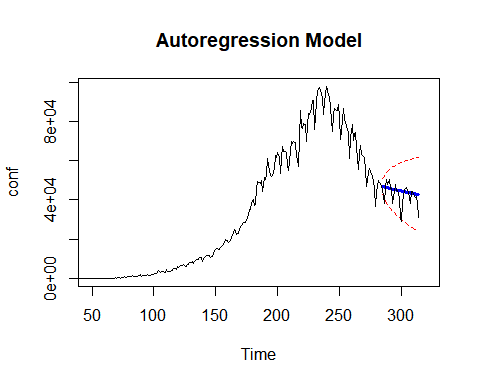
Confidence interval of 95% is represented by the area bordered by the red dotted-line. The basis is standard error and for farther dates, there is scope for more standard errir than that of closer dates.

plot.ts(conf,xlim=c(50,320),main="Autoregression Model")  
points(ar\_pred,type='l',col="blue",lwd=3)  
points(ar\_pred-ar\_se,type='l',col="red",lty=2); points(ar\_pred+ar\_se,type='l',col="red",lty=2)



Finally we’re comparing the actual data of November with that of the predicted region of values.

plot.ts(conf,xlim=c(50,320),main="Autoregression Model")  
points(ar\_pred,type='l',col="blue",lwd=3)  
points(ar\_pred-ar\_se,type='l',col="red",lty=2); points(ar\_pred+ar\_se,type='l',col="red",lty=2)  
points(nov\_data,type='l')



Removing all the data that we don’t need to deal with further.

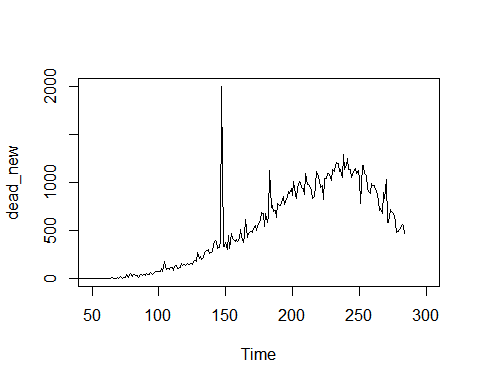
rm(list=ls(pattern='^conf|^ar|new$|data$'))

DEATHS PREDICTION

Importing data for deaths.

oct\_data2=read.xlsx("datasets/dead\_new\_oct.xlsx",sheet = 1,startRow = 1,colNames = T,rowNames = F,detectDates = T)  
nov\_data2=read.xlsx("datasets/dead\_new\_nov.xlsx",sheet = 1,startRow = 1,colNames = T,rowNames = F,detectDates = T)  
oct\_data2[,1]<-as.Date(oct\_data2[,1])

dead\_new<-as.ts(oct\_data2$Deaths)  
plot(dead\_new,ylim=c(0,2000),xlim=c(50,300))

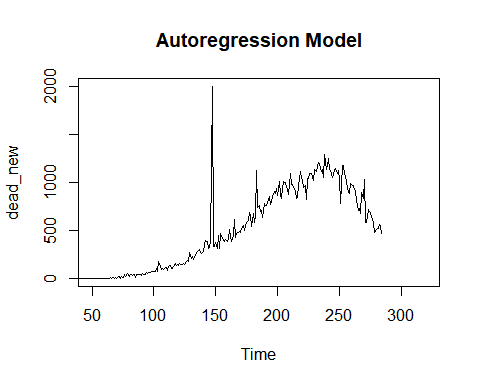


ar\_dead<-arima(dead\_new,order=c(1,0,0))  
ar\_pred<-predict(ar\_dead,n.ahead=30)$pred  
ar\_se<-predict(ar\_dead,n.ahead=30)$se

The following few graphs show how data of deaths can be used to predict for next month and comparison of the same is shown with the actual data of November.

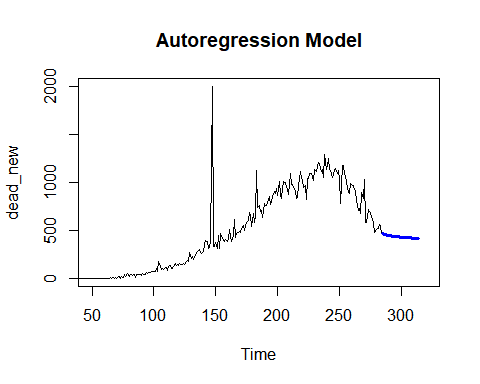
Actual data of deaths till October end.

plot.ts(dead\_new,xlim=c(50,320),main="Autoregression Model")



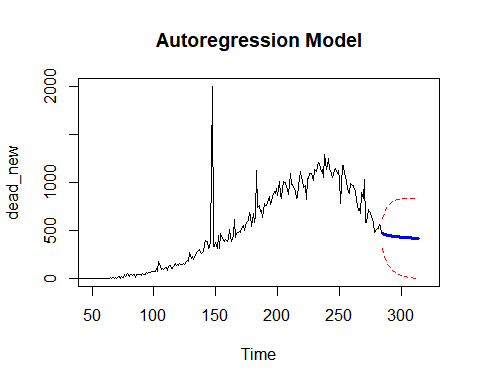
Predicted values are added to the graph.

plot.ts(dead\_new,xlim=c(50,320),main="Autoregression Model")  
points(ar\_pred,type='l',col="blue",lwd=3)



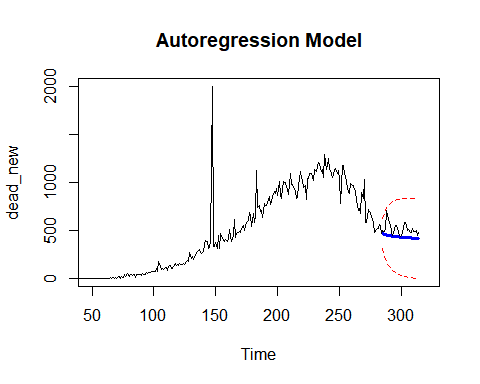
Confidence interval of 95% is represented by the area bordered by the red dotted-line. The basis is standard error and for farther dates, there is scope for more standard errir than that of closer dates.

plot.ts(dead\_new,xlim=c(50,320),main="Autoregression Model")  
points(ar\_pred,type='l',col="blue",lwd=3)  
points(ar\_pred-ar\_se,type='l',col="red",lty=2); points(ar\_pred+ar\_se,type='l',col="red",lty=2)



Finally we’re comparing the actual data of November with that of the predicted region of values.

plot.ts(dead\_new,xlim=c(50,320),main="Autoregression Model")  
points(ar\_pred,type='l',col="blue",lwd=3)  
points(ar\_pred-ar\_se,type='l',col="red",lty=2); points(ar\_pred+ar\_se,type='l',col="red",lty=2)  
points(nov\_data2,type='l')



rm(list=ls(pattern='^conf|^ar|new$|data2$'))