**Indian Institute of Information Technology-Dharwad**

**Name:** Modem Praveen **Roll No:**17BCS035 **Mail:** modemiiit@gmail.com

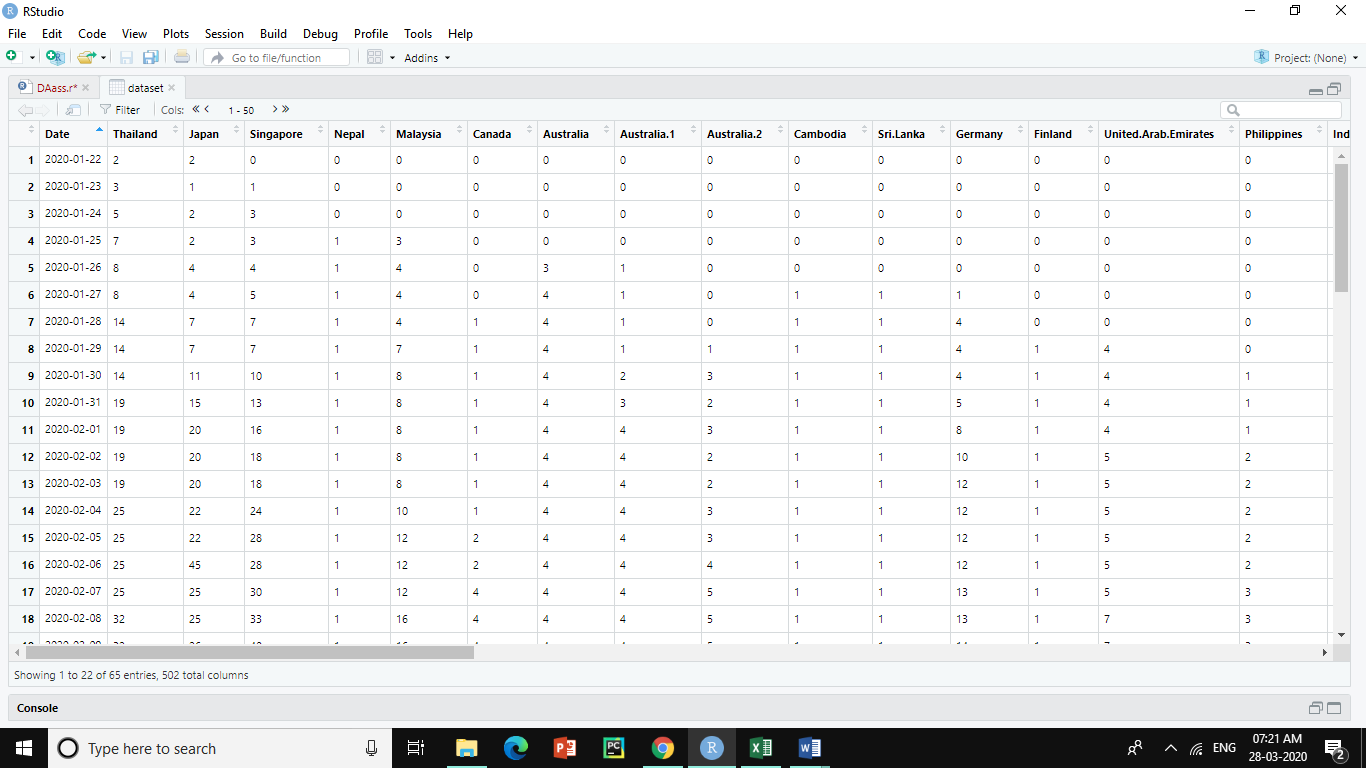
**Data Analytics**

**Time series analysis on Corona virus infected cases**

**Files Url :** <https://github.com/modem0011/Covid-19-analysis>

This dataset consists of 50 columns,1st column has Dates (22 Jan,2020 – 26 Jan ,2020) and other columns has country wise Corona infected cases.

**Data Cleaning and Manipulating**: Data cleaning and Data manipulation is done in a convenient way.



**Selecting Data Set**

dataset<-read.csv(file.choose())

Download the dataset from above URL ( given in page 1) and enter above code in R Studio.It will ask which file to select . Select “cleaned\_file.csv “ from downloaded folder.

**Describing data**

install.packages(“Hmisc”)

library(“Hmisc”)

describe(dataset)

To know about data, enter above code and check output. It describes whole dataset.

**Note:** Try “summary(dataset)” , “ str(dataset) ” for clear understanding about dataset.(“str ” shows structure of dataset.)

**There are many countries are there in dataset so we can focus on some countries which we want to predict and analyse.**

**Here I am focusing on India, Italy, Norway, Israel, Iraq, Spain, Brazil, Belgium, Sweden.**

**Select columns using $ symbol**

**Example:**

dataset$Date

dataset$India

dataset$Italy

**Note**: There are several methods for selecting columns. Above is one of the method used in this project

**Define format of date is format by using below code.**

install.packages(“lubridate”)

library(lubridate)

dataset2$Date<-dmy(dataset$Date)

**Visualising**

par(mfrow=c(2,2))

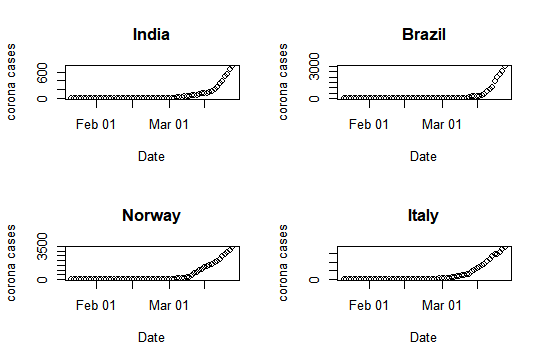
plot(dataset$Date,dataset$India,xlab = "Date",ylab = "corona cases",main = "India")

plot(dataset$Date,dataset$Brazil,xlab = "Date",ylab = "corona cases",main = "Brazil")

plot(dataset$Date,dataset$Norway,xlab = "Date",ylab = "corona cases",main="Norway")

plot(dataset$Date,dataset$Italy,xlab = "Date",ylab = "corona cases",main = "Italy")

**Note: with par function we plot n plots in a single screen . Here we plotted as (2,2) matrix**



All above plots are looking little similar but some variations are there .If we see Y-axis ranges we can notice changes.

**Time series object with the data**

**Time series** is a **series** of data points in which each data point is associated with a timestamp. ... The data for the **time series** is stored in an **R object** called **time**-**series object**. It is also a **R** data **object** like a vector or data frame. The **time series object** is created by using the ts() function.

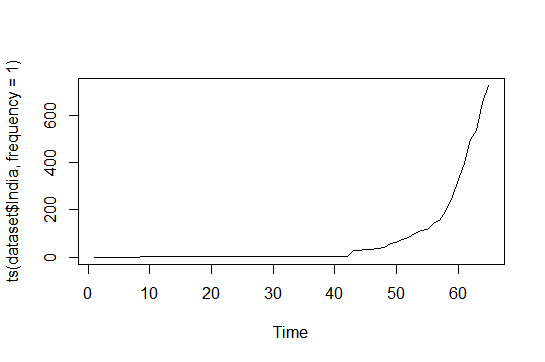
**Code:**

ts(dataset$India)

ts(dataset)

par(mfrow=c(1,1))

plot(ts(dataset$India))



**Finding Mean**

This dataset is in the form of cumulative sums of corona cases in different countries. So finding monthly or yearly mean values is not possible.

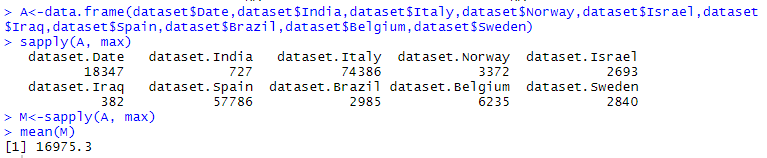
But we can find mean value of corona cases in selected countries.

**Code:**

A<-data.frame (dataset$Date, dataset$India, dataset$Italy, dataset$Norway, dataset$Israel, dataset$Iraq, dataset$Spain, dataset$Brazil, dataset$Belgium, dataset$Sweden)

B<-sapply(A, max)

mean(B)



**Above mean value says in an average in every country 16975.3 people got infected by Corona Virus.**

**Boxplots of a corona cases in a span of 2 months (approx)**

par(mfrow=c(2,3))

boxplot(dataset$India,main="India",col="red")

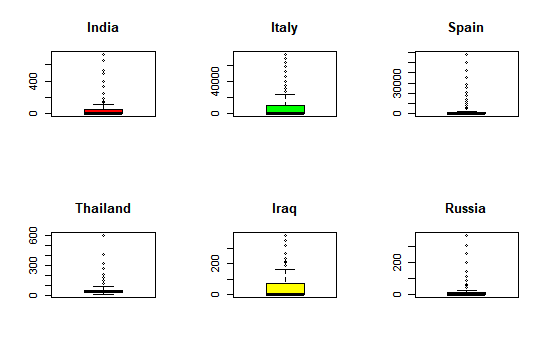
boxplot(dataset$Italy,main="Italy",col="green")

boxplot(dataset$Spain,main="Spain",col="pink")

boxplot(dataset$Thailand,main="Thailand",col="blue")

boxplot(dataset$Iraq,main="Iraq",col="yellow")

boxplot(dataset$Russia,main="Russia",col="skyblue")



**Dots in plots represents outliers**

|  |
| --- |
| **Note:**    Stl function required atleast 2 periods . But here we have only 1 period (dates).To overcome this problem we have to collect yearly data . As we know corona started spreading recently .So there is no possibility of collecting more data.  Only 1 seasonality we have .That too not periodic.It’s additive.  **Residuals:** |

**Model for the data using the HoltWinters method**

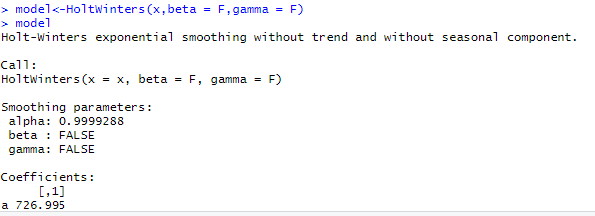
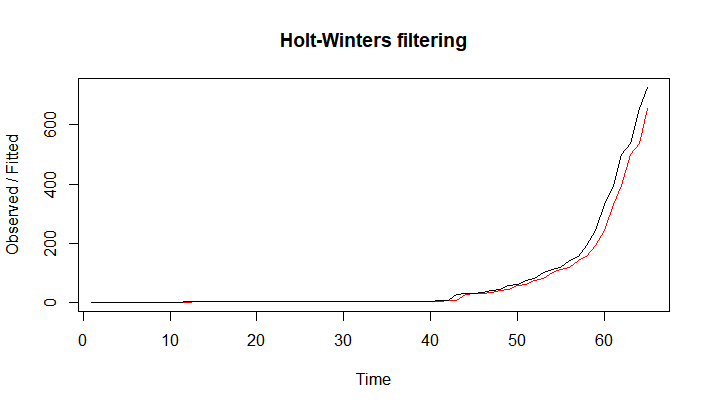
x<-ts(dataset$India)

model<-HoltWinters(x,beta = F,gamma = F)

model

plot(model)

Here I mentioned beta and gamma values as False because those are trend and seasonality coefficients .As we know our data don’t have proper seasonality .Which I mentioned in page:7. so I assigned them as False and by default it took alpha as 0.9999288 for above data.



**Visualisation of HoltWinters model for different countries**

x1<-ts(dataset$India)

x2<-ts(dataset$Italy)

x3<-ts(dataset$Norway)

x4<-ts(dataset$Sweden)

model1<-HoltWinters(x1,beta = F,gamma = F)

model2<-HoltWinters(x2,beta = F,gamma = F)

model3<-HoltWinters(x3,beta = F,gamma = F)

model4<-HoltWinters(x4,beta = F,gamma = F)

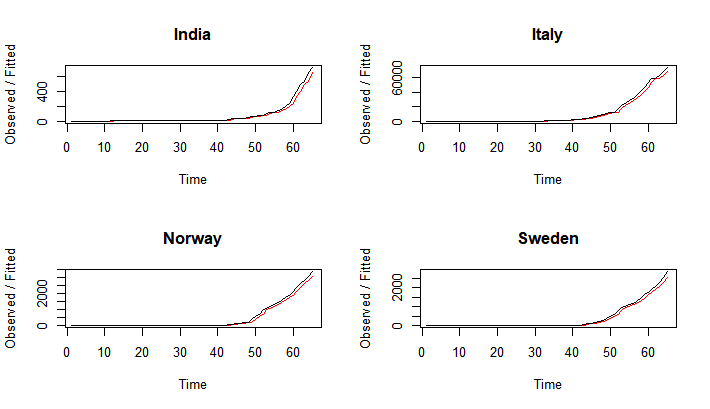
par(mfrow=c(2,2))

plot(model1,main = "India")

plot(model2,main="Italy")

plot(model3,main = "Norway")

plot(model4,main="Sweden")

****

**Observed and Fitted lines are plotted above**

**Forecasting country wise next 4 days’ corona cases.**

x1<-ts(dataset$India)

x2<-ts(dataset$Norway)

x3<-ts(dataset$Sweden)

x4<-ts(dataset$Italy)

x5<-ts(dataset$Brazil)

x6<-ts(dataset$Belgium)

x7<-ts(dataset$Israel)

x8<-ts(dataset$Iraq)

x9<-ts(dataset$Spain)

x10<-ts(dataset$Malaysia)

model1<-HoltWinters(x1,beta = F,gamma = F)

model2<-HoltWinters(x2,beta = F,gamma = F)

model3<-HoltWinters(x3,beta = F,gamma = F)

model4<-HoltWinters(x4,beta = F,gamma = F)

model5<-HoltWinters(x5,beta = F,gamma = F)

model6<-HoltWinters(x6,beta = F,gamma = F)

model7<-HoltWinters(x7,beta = F,gamma = F)

model8<-HoltWinters(x8,beta = F,gamma = F)

model9<-HoltWinters(x9,beta = F,gamma = F)

model10<-HoltWinters(x10,beta = F,gamma = F)

library(forecast)

forecast(model1,4)

forecast(model2,4)

forecast(model3,4)

forecast(model4,4)

forecast(model5,4)

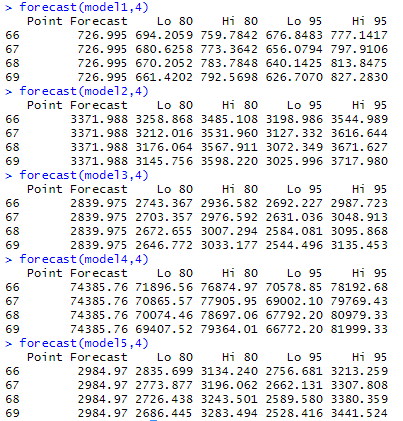
forecast(model6,4)

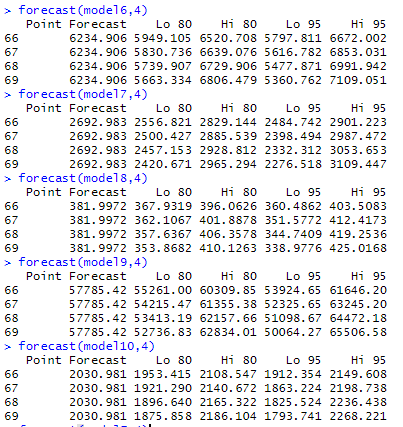
forecast(model7,4)

forecast(model8,4)

forecast(model9,4)

forecast(model10,4)





Time series started from 1. It considered 22nd Jan 2020 as 1 and so on…According to that we trained model from 1 to 65 (22nd Jan2020 to 26th March 2020) and we Predicted Corona cases for 66,67,68,69 (27th March 2020 – 30th March 2020 )

**Visualising Predictions**

par(mfrow=c(3,3))

plot(forecast(model1,4),col = "red",main = "India")

plot(forecast(model2,4),col="green",main = "Norway")

plot(forecast(model3,4),col="blue",main="Sweden")

plot(forecast(model4,4),col="brown",main = "Italy")

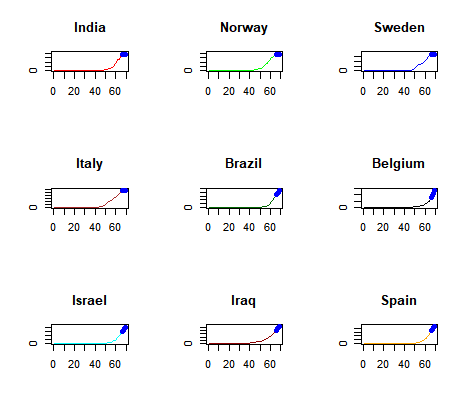
plot(forecast(model5,4),col="dark green",main="Brazil")

plot(forecast(model6,4),col = "black",main = "Belgium")

plot(forecast(model7,4),col="cyan",main="Israel")

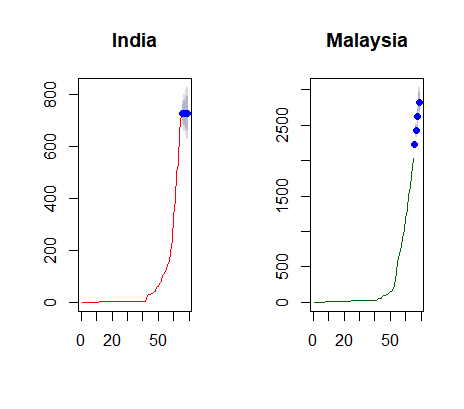
plot(forecast(model8,4),col = "dark red",main="Iraq")

plot(forecast(model9,4),col="orange",main = "Spain")



**Thick Blue area in plot is predicted area**.

**Clear view of plot for India and Malaysia.**

 **Above 4 blue dots are predicted values using HoltWinters model**

**Actual Corona Cases**

Above what we saw is Predicted values and below values are Actual values.

Today is 29 march ,2020 . But I predicted till March 30 so here I am comparing with 27th ,28th actual and predicted values.

**Source:** [CoronaWorldmeter](https://www.worldometers.info/coronavirus/)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| India | Norway | Sweden | Italy | Brazil | Belgium | Israel | Iraq | Spain | Malaysia |
| 887 | 3771 | 3069 | 86498 | 3417 | 7284 | 3035 | 458 | 65719 | 2161 |
| 987 | 4015 | 3447 | 92472 | 3904 | 9134 | 3619 | 506 | 73235 | 2320 |

We can compare 66,67(27 march 2020,28 march 2020) predicted values(page: 11) with these above values.

**Visualizing Actual VS Predicted**

par(mfrow=c(2,2))

plot(c(66,67),c(887,987),xlim = c(66,67),ylim = c(100,1500),"b",col="green", ylab="Actual vs Predict",main = "India" )

lines(c(66,67),c(777,797),"b",col="red")

plot(c(66,67),c(3771,4015),xlim = c(66,67),ylim = c(1000,5000),"b",col="green", ylab="Actual vs Predict",main = "Norway" )

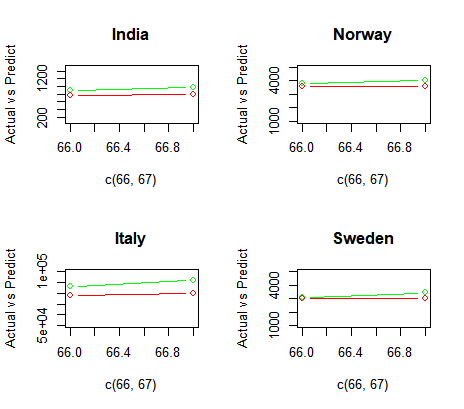
lines(c(66,67),c(3544,3616),"b",col="red")

plot(c(66,67),c(86498,92472),xlim = c(66,67),ylim = c(50000,100000),"b",col="green", ylab="Actual vs Predict",main = "Italy" )

lines(c(66,67),c(78192,79769),"b",col="red")

plot(c(66,67),c(3069,3447),xlim = c(66,67),ylim = c(1000,5000),"b",col="green", ylab="Actual vs Predict",main = "Sweden" )

lines(c(66,67),c(2987,3048),"b",col="red" )



**Green is actual and Red is predicted line.**

* For this particular dataset model prediction is very poor because of not giving beta and gamma coefficients(seasonality and Trend coefficients).
* I already mentioned reason in page 7.

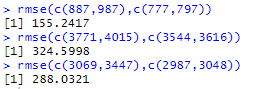
**RMSE**

**library(Metrics)**

rmse(c(887,987),c(777,797)) # india

rmse(c(3771,4015),c(3544,3616)) #Norway

rmse(c(3069,3447),c(2987,3048)) #Sweden



…………………………………………………………………………………………………………………..

**Changing alpha Value and Comparing OutPuts**

model1<-HoltWinters(x1,beta = F,gamma = F)

model1

a=forecast(model1,2)

a

model1<-HoltWinters(x1,beta = F,gamma = F,alpha=0.7)

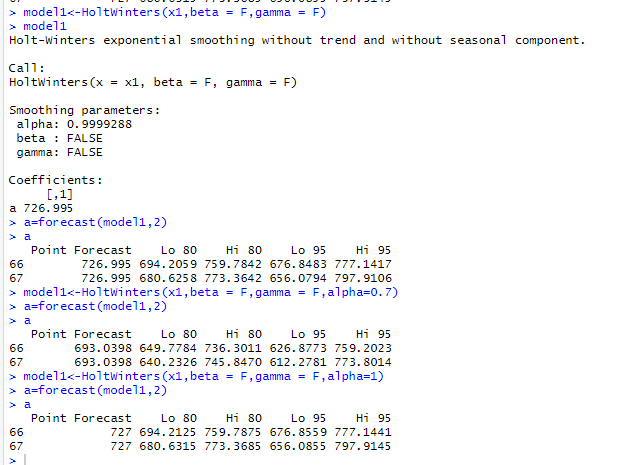
a=forecast(model1,2)

a

model1<-HoltWinters(x1,beta = F,gamma = F,alpha=1)

a=forecast(model1,2)

a

****

In above code we gave 3 different alpha values one is default (0.9999288) and other two are 0.7 and 1

Before we found rmse values using default alpha predictions .But here if we give alpha value 0.7 . It is predicting far way value than actual value . so rmse will be higher

If we change alpha value >0.9999288 then predicted value is going closer to actual Value. So rmse will be lower than before alpha values

**Model for the data using the ARIMA method**

x1<-ts(dataset$India)

x2<-ts(dataset$Norway)

x3<-ts(dataset$Sweden)

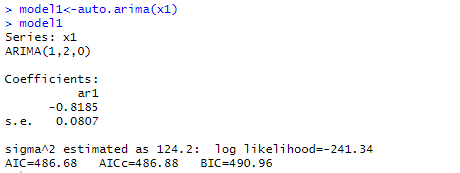
x4<-ts(dataset$Italy)

x5<-ts(dataset$Brazil)

x6<-ts(dataset$Belgium) model1<-auto.arima(x1) # India

x7<-ts(dataset$Israel) model

x8<-ts(dataset$Iraq)

x9<-ts(dataset$Spain)

x10<-ts(dataset$Malaysia)

model1<-auto.arima(x1)

model2<-auto.arima(x2)

model3<-auto.arima(x3)

model4<-auto.arima(x4)

model5<-auto.arima(x5)

model6<-auto.arima(x6)

model7<-auto.arima(x7)

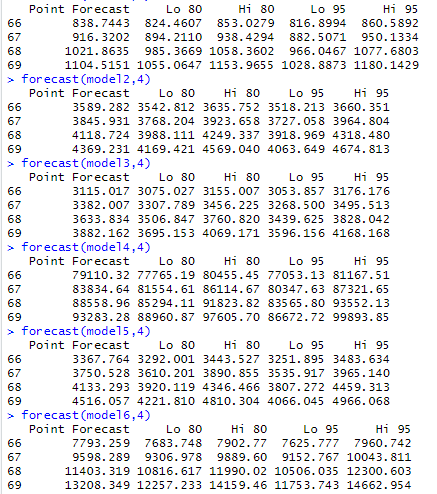
model8<-auto.arima(x8)

model9<-auto.arima(x9)

model10<-auto.arima(x10)

**Here we are using ARIMA model.We didn’t mention order so by default it is taking 1,2,0 values.**

**Predicting Corona cases using ARIMA Model**

****

forecast(model1,4)

forecast(model2,4)

forecast(model3,4)

forecast(model4,4)

forecast(model5,4)

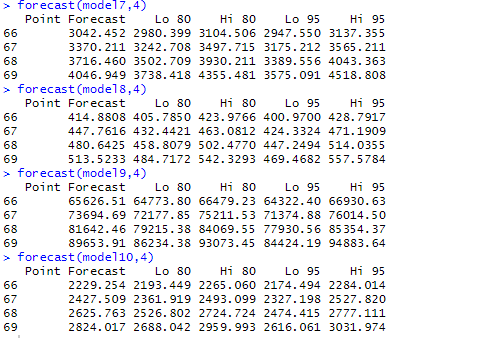
forecast(model6,4)

forecast(model7,4)

forecast(model8,4)

forecast(model9,4)

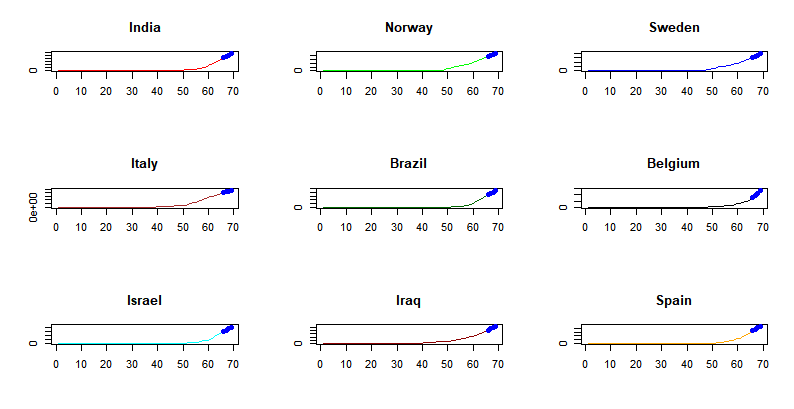
forecast(model10,4)

****

Time series started from 1 it considered 22nd Jan 2020 as 1 and so on…According to that we trained model from 1 to 65 (22nd Jan2020 to 26th March 2020) and we Predicted Corona cases for 66,67,68,69 (27th March 2020 – 30th March 2020 )

**Example**: For India (model 1) 27th -30th march it predicted 860,950,1077,1180 corona cases in India.

**Visualization**

****

**Actual Corona Cases**

Above what we saw is Predicted values and below values are Actual values .

Today is 29 march ,2020 but I predicted till March 30 so here I am comparing with 27th ,28th actual and predicted values.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| India | Norway | Sweden | Italy | Brazil | Belgium | Israel | Iraq | Spain | Malaysia |
| 887 | 3771 | 3069 | 86498 | 3417 | 7284 | 3035 | 458 | 65719 | 2161 |
| 987 | 4015 | 3447 | 92472 | 3904 | 9134 | 3619 | 506 | 73235 | 2320 |

**Predicted Corona Cases**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| India | Norway | Sweden | Italy | Brazil | Belgium | Israel | Iraq | Spain | Malaysia |
| 860 | 3660 | 3176 | 81167 | 3483 | 7960 | 3137 | 428 | 66930 | 2284 |
| 950 | 3964 | 3495 | 87321 | 3965 | 10043 | 3565 | 471 | 76014 | 2527 |

From ARIMA Model we are getting closer predictions.

**Actual vs Predicted**

par(mfrow=c(2,2))

plot(c(66,67),c(887,987),xlim = c(66,67),ylim = c(100,1500),"b",col="green", ylab="Actual vs Predict",main = "India" )

lines(c(66,67),c(860,950),"b",col="red")

plot(c(66,67),c(3771,4015),xlim = c(66,67),ylim = c(1000,5000),"b",col="green", ylab="Actual vs Predict",main = "Norway" )

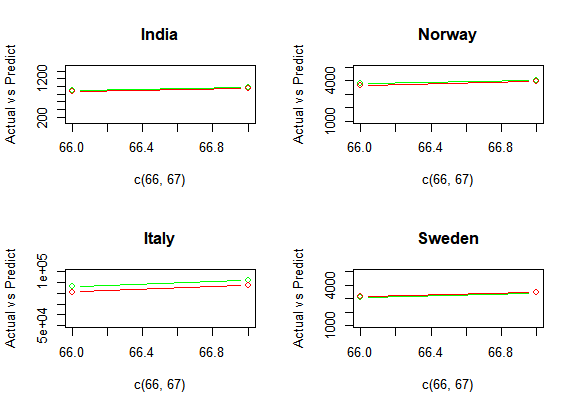
lines(c(66,67),c(3660,3964),"b",col="red")

plot(c(66,67),c(86498,92472),xlim = c(66,67),ylim = c(50000,100000),"b",col="green", ylab="Actual vs Predict",main = "Italy" )

lines(c(66,67),c(81167,87321),"b",col="red")

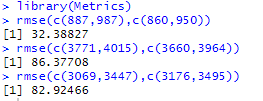
plot(c(66,67),c(3069,3447),xlim = c(66,67),ylim = c(1000,5000),"b",col="green", ylab="Actual vs Predict",main = "Sweden" )

lines(c(66,67),c(3176,3495),"b",col="red" )

****

**From plots we can see our model is almost perfectly predicting**

**RMSE**



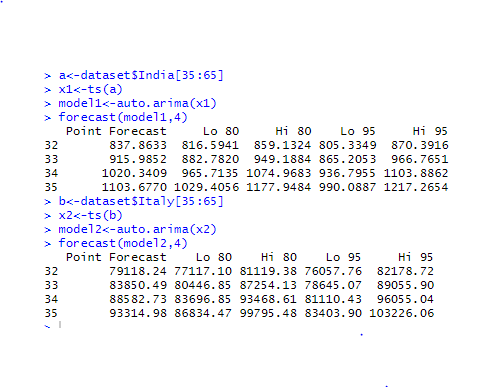
library(Metrics)

rmse(c(887,987),c(860,950)) # India

rmse(c(3771,4015),c(3660,3964)) #Norway

rmse(c(3069,3447),c(3176,3495)) # Sweden

**Cleaning little more**



a<-dataset$India[35:65]

x1<-ts(a)

model1<-auto.arima(x1)

forecast(model1,4)

b<-dataset$Italy[35:65]

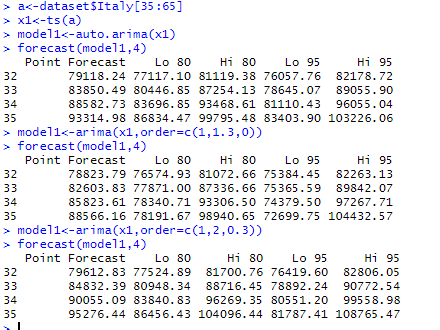
x2<-ts(b)

model2<-auto.arima(x2)

forecast(model2,4)

If we see our dataset in initial dates, we have very less corona infected people ( 0 or in single digit ). Actually those columns are affecting our accuracy.So I took past month corona cases and I predicted values. It gave closer predictions than before .

**Changing Order values**



a<-dataset$Italy[35:65]

x1<-ts(a)

model1<-auto.arima(x1)

forecast(model1,4)

model1<-arima(x1,order=c(1,1.3,0))

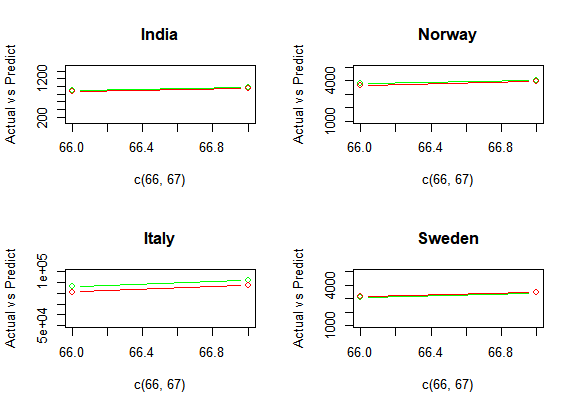
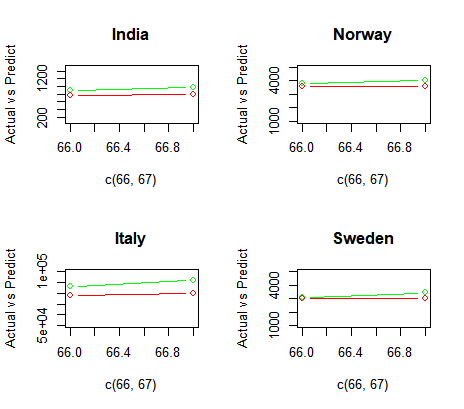
forecast(model1,4)

model1<-arima(x1,order=c(1,2,0.3))

forecast(model1,4)

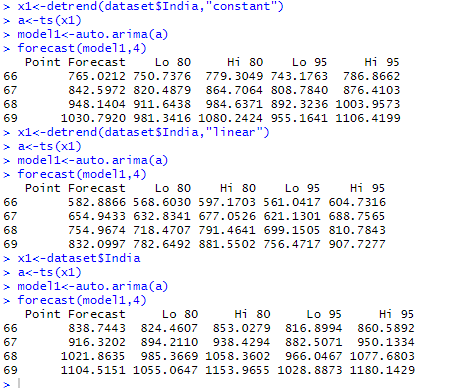
If we change order values(p,q,d) we are getting good predictions. But above output screenshot we can see that auto.arima is taking order automatically.Internally it will try with all possible permutations of order values and it will select best order values which gives less AIC and BIC values.As we know here we are predicting only 4 upcoming corona cases values so concluding auto.arima predictions are giving less accuracy than other orders is wrong way according to my observations. So I think taking auto.arima is good for this dataset.

**Holtwinters VS Arima**

****

Above plots are clearly saying that ARIMA model is giving closer predictions when compared with Holtwinters model. For my dataset Holtwinters model is not suitable.Because as we saw in page:7 we don’t have proper seasonality and because of that we can’t give beta and gamma values .But prediction will depend on those 2 coefficients too.so we are getting low accuracy for it. Here I’m concluding ARIMA is best method for this dataset than Holtwinters method.

**Detrending**

x1<-detrend(dataset$India,"constant")

a<-ts(x1)

model1<-auto.arima(a)

forecast(model1,4)

x1<-detrend(dataset$India,"linear")

a<-ts(x1)

model1<-auto.arima(a)

forecast(model1,4)

x1<-dataset$India

a<-ts(x1)

model1<-auto.arima(a)

forecast(model1,4)

Trending is removing fluctuations in data and in case if we detrend our model then fluctuations will increase and we will get bad predictions as shown in above output.

**Future predictions**

x1<-ts(dataset$India[20:65])

x2<-ts(dataset$Italy[20:65])

x3<-ts(dataset$Sweden[20:65])

x4<-ts(dataset$Norway[20:65])

x5<-ts(dataset$Spain[20:65])

model1<-auto.arima(x1)

model2<-auto.arima(x2)

model3<-auto.arima(x3)

model4<-auto.arima(x4)

model5<-auto.arima(x5)

forecast(model1,7)

forecast(model2,7)

forecast(model3,7)

forecast(model4,7)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Country | 27  Mar,2020 | 28  Mar,2020 | 29  Mar,2020 | 30  Mar,2020 | 31  Mar,2020 | 1  April,2020 | 2  April,2020 |
| India | 864 | 956 | 1088 | 1195 | 1327 | 1443 | 1577 |
| Italy | 81584 | 88035 | 94581 | 101263 | 108086 | 115046 | 122138 |
| sweden | 3197 | 3560 | 3937 | 4327 | 4728 | 5140 | 5560 |
| Norway | 3676 | 3992 | 4363 | 4742 | 5128 | 5535 | 5953 |
| spain | 67195 | 76487 | 86111 | 95951 | 105996 | 116226 | 126626 |

forecast(model5,7)

**Source :** [CoronaWorldmeter](https://www.worldometers.info/coronavirus/)

**Predictive Analysis**

* According to our predictions spreading of virus is very fast .
* So if it continuous then the world will fall in danger.

**Prescriptive analysis**

* Wash your hands regularly for 20 seconds, with soap and water or alcohol-based hand rub
* Cover your nose and mouth with a disposable tissue or flexed elbow when you cough or sneeze
* Avoid close contact (1 meter or 3 feet) with people who are unwell
* Stay home and self-isolate from others in the household if you feel unwell
* Don't Touch your eyes, nose, or mouth if your hands are not clean