Time series modeling of ODI scores

Rahul.P.R

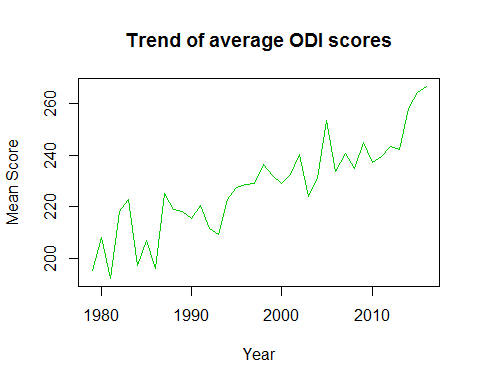
3 July 2017

The average of all the matches conducted in a particular year was used to obtain the mean and median for the same. The trend was plotted for median score and it was understood that structural changes initiated by ICC have majorly pushed the median scores over the years.

In the initial data there were 47 years considered however now for time series modeling only 38 years are considered eventually leaving out years where the number of matches played were less than 20. The years are 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978 . 2017, the current year was removed and lets try predicting the same!

Now, lets understand the time series. Before that its wise to understand if the time series is stationary/non-stationary. ARIMA(Auto regression integrated moving average) model is applicable only on a stationary series. To understand the data series in picture and decide on stationary/non stationary lets do both descriptive as well as statistic analyss

1. Descriptive

A stationary series typically doesnt have the mean across the years increasing i.e., in simple terms there would be no visible trend.

The year-on-year plot above clearly shows that there is a considerable increase in the average score over the years (requires a diff function) and also variation over time which will eventually require a log function i.e., to neglect the effect of trend and variation and making it stationary. Though seasonality would be present in the data it is anyways taken care in the ARIMA model itself

1. Augemented Dickey Fuller test

Though there is some indecisiveness/statistical evidence with the descriptive plots the above named test is conclusive. Initially, running the test directly on the raw data without any transformation and then applying the same post-transformation or stationarizing the time series using log and diff

##   
## Augmented Dickey-Fuller Test  
##   
## data: data\_for\_acf\_pacf  
## Dickey-Fuller = -5.7495, Lag order = 0, p-value = 0.01  
## alternative hypothesis: stationary

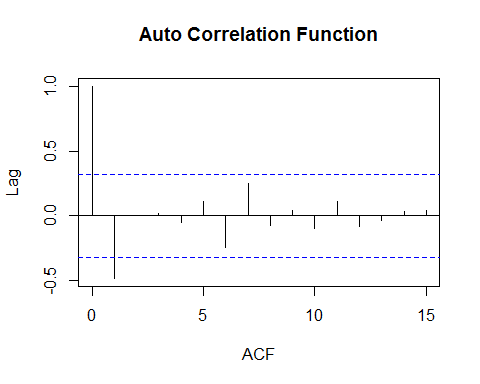
The time series considered here is already stationary without transoframtion or operation such as log/diff since the the null hypothesis is rejected. The agumented test statistic is typically a negative number, the more the negative it is the stronger the rejection of null hypothesis. Now, with diff operation alone i.e., removing the trend component, the test works as

##   
## Augmented Dickey-Fuller Test  
##   
## data: diff(data\_for\_acf\_pacf)[-which(is.na(diff(data\_for\_acf\_pacf)))]  
## Dickey-Fuller = -9.7899, Lag order = 0, p-value = 0.01  
## alternative hypothesis: stationary

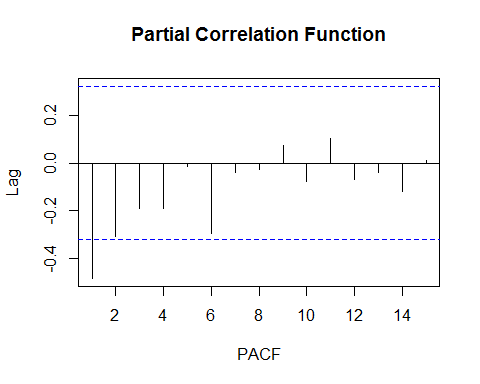
Now, the test statistic has become more negative as seen above i.e., decreasing from -5.7495 to -9.7899 and now the rejection of null hypothesis is even stronger. Since variation was witnessed we are incorporating the log function as well and now the test statistic looks like

##   
## Augmented Dickey-Fuller Test  
##   
## data: diff(log(data\_for\_acf\_pacf))[-which(is.na(diff(log(data\_for\_acf\_pacf))))]  
## Dickey-Fuller = -9.9682, Lag order = 0, p-value = 0.01  
## alternative hypothesis: stationary

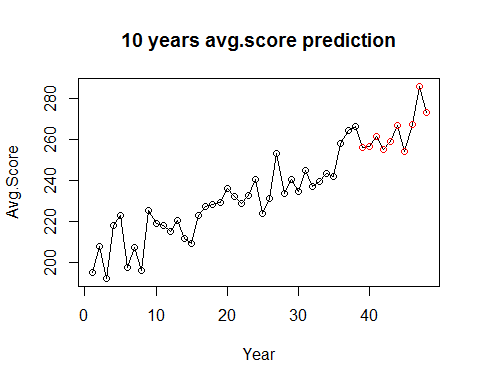
The value decrease even more to -9.9682 from -9.7899. Though there is no significant decrease as observed earlier we still incorporate the same to compeletely staionarize since variation along the time series was visible from the descriptive plots

The series now is stationary enough to do any kind of time series modeling. Lets find the parameters for the ARIMA model using ACF (Auto Correlation function) and PACF (Partial correlation function)

In ACF curve it cuts off the blue dotted line at 2 and hence the value of **p** is 2 i.e., the arima model would require two MA coeffecients MA(2). Now, to the PACF curve to understand the number of auto regression components



The PACF curve cuts off in the first value and hence the value of **q** or the number of regression components required for the ARIMA model is 1 i.e., it is a AR(1) model. Since the series was transformed to stationary using diff the value of **d** is 1. The parameters (p,d,q), where p is the number of auto regressive coefficients, d is the degree of differencing, q is the number of moving average coefficients which in this cases takes the form (2,1,1).Since difference is incorporated into the parameters already in the form of d we fit the model using the only remaining log transformation. Now predicting the next two years



The accuracy of the model however could be tested by predicting for 2016 by considering data only upto 2015. Also, model diagnostics is important and the same would be coverd in detail in the preceding article/blog