Abstract

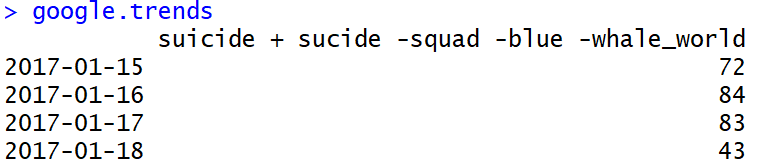
This project assignment is about timeseries and forecasting using Arima. We extract data from google trends based on the term suicide and terms related to it from 2017-01-15 to 2017-03-30. We then apply ARIMA model on it to the to forecast the values for the next 19 days. We then compare it with the actual values of the search results to see the accuracy of the forecast. We check for the stationarity of the series and then use acf and pacf to decide p,d and q values of our arima model. We also apply auto arima and compare our forecast results with the real results. We are basically recreating the analysis performed by the John W. Ayers to see whether there was significant increase in the search results related to the suicide query.

Keywords: [Timeseries, forecast, ARIMA, prediction, 13 reasons why, acg, pacf, stationarity]

**Methodology and Analysis**

Using gtrends function in r we collect the data related to suicide from the time period 2017-January-15 to 2017-March-30.

google.trends = gtrends(c("suicide + sucide -squad -blue -whale"), gprop = "web", time = "2017-01-15 2017-03-30")[[1]]



We get the number of interest over time relative to the highest point on the chart for the given region and time. A value of 100 is a peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.

We then plot the timeseries of the trends:

A screenshot of a cell phone

Description generated with high confidence

Here we can see the data for 75 days, let us now check for the stationary of the data.

We can check for the stationary using the Augmented Dickey-Fuller Test.

adf.test(timeseries, alternative = "stationary", k=0)

A screenshot of a cell phone

Description generated with very high confidence

We get the p-value of 0.01 which is less than 0.05 hence we can reject the null hypothesis. We need to go with the alternative hypothesis that the time series is stationary.

We then make acf plots. Autocorrelation is the linear dependence of a variable with itself at two points in time. For stationary processes, autocorrelation between any two observations only depends on the time lag h between them.

ts\_acf <- acf(timeseries)

A screenshot of a cell phone

Description generated with very high confidence

Looking the act plot we keep the MA term as 0.

Let us plot pacf. Partial autocorrelation is the autocorrelation between yt and yt–h after removing any linear dependence on y1, y2, ..., yt–h+1.

ts\_pacf <- pacf(timeseries)

A screenshot of a cell phone

Description generated with very high confidence

Looking at the PACF graph we keep the AR term as 3.

We finally keep arima values as p = 3, d = 1, q = 0

fit <- arima(timeseries, c(3,1,0))

We then use the predict function in r to find the forecast of the next 19 days and then plot it.

pred <- predict(fit, n.ahead = 19)

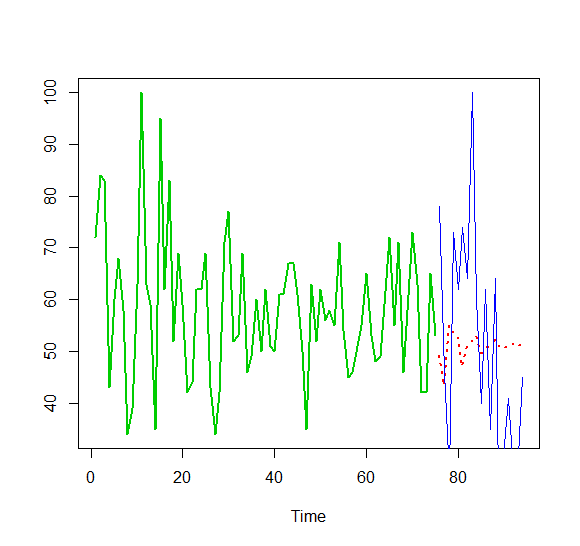
ts.plot(timeseries,pred$pred, lty = c(1,3), col=c(3,2), lwd = 2)

A picture containing screenshot

Description generated with high confidence

We then compare it with the real values of suicide related google trends results using:

lines(timeseries1, col = "blue")



We can see the actual values in blue line and see how the actual results seem amplified version of the forecast.

Here we can see the analysis performed by the John W. Ayers was actually true there is significant increase in the search results related to the suicide query.

We compare the actual results and the forecast using a table:

A picture containing wall

Description generated with high confidence

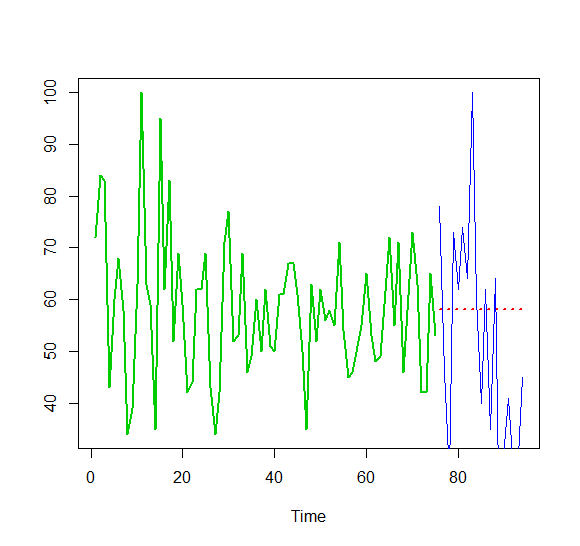
Let us calculate the error rate by differencing the forecast and actual results.

error <- sqrt((final$actual-final$prediction)^2)

error <- sum(error)

We get the error rate of 352.42, which is calculated by squaring the differencing and then taking the square root of it.

Let us perform the same analysis using auto arima,



We get the error rate of 361.04 which is worse than what we calculated using our interpretations looking at the acf and pacf graphs.

We also try other combinations of ARIMA model there is the R file and also compare the aic and bic values.

**Conclusion**

We have learned the implementation of time series and calculated the future values using arima forecasting. We have used the acf and pacf plots to determine the p,d and q values and then used the arima function to calculate the forecast. Looking at the pacf graph we got the value of AR as 3 and looking at the acf graph we got the value of MA as 0.

After using the arima model we got the predicted values and after comparing with the real values we can see the analysis performed by the John W. Ayers was actually true there is significant increase in the search results related to the suicide query.

Comparing it with different p, d and q values we develop a better understanding of arima model.

References

1) Econometric Modeler. (n.d.). Retrieved from <https://www.mathworks.com/help/econ/autocorrelation-and-partial-autocorrelation.html>

2) Srivastava, T., Choudhary, A., Dar, P., Shaikh, F., & Ray, S. (2018, April 20). A Complete Tutorial on Time Series Modeling in R. Retrieved from <https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/>

3) Jones, R. (2017, July 31). Research Finds Disturbing Suicide Search Trends Following Release of Netflix's 13 Reasons Why. Retrieved from https://gizmodo.com/research-finds-disturbing-suicide-search-trends-followi-1797398484