Problem 1: A Song of Ice and Fire (9 percent)

Workday 2

Danny Moncada monca016

March 26, 2020

```
suppressWarnings(suppressPackageStartupMessages({
  library(TSA)
  library(ggplot2)
  library(dplyr)
  library(forecast)
  library(tseries)
}))
```

Part 0: Data Cleaning

Wikipedia web visit (Sessions per day) was counted. Data were collected from 11.29.2015 to 11.28.2016 – 366 days in total.

Question 1

Please load the data and plot it.

What do you think about this time series? Is it likely to be stationary? Is there any spikes? Is there a trend? Does it have any missing data? (You don't need to answer these questions in the answer sheet, but developing a habit of thinking over these questions would be very helpful for future time series analysis in real-world practices.)

Hints:

1. Please specify `header=F`

```
# Read in the text file as table
sif_df <- read.table("Wiki_A_Song_of_Ice_and_Fire_web_visit.txt", header = F)

# Summary table of sif table
summary(sif_df)
```

```
##        V1
##  Min.   : 5221
##  1st Qu.: 6535
##  Median : 7274
##  Mean   : 7653
##  3rd Qu.: 8280
##  Max.   :20101
```

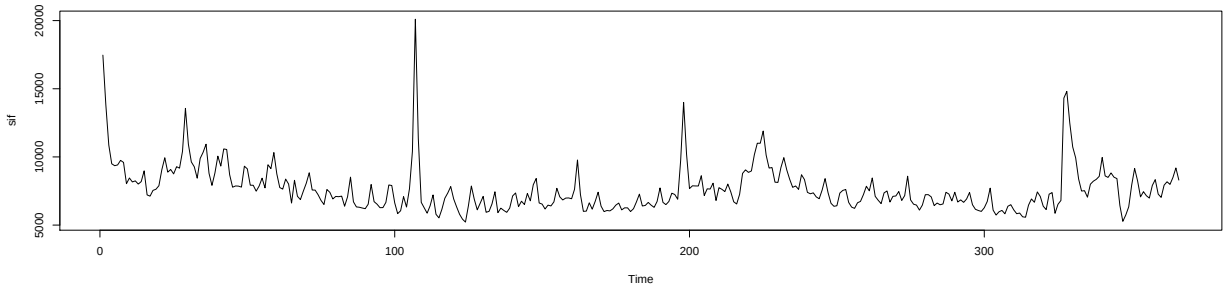
```r
# the time series should be called "sif"
sif = ts(sif_df, start = 1, end = 366)

# ADF test to check stationarity
adf.test(sif)
```

```
## Warning in adf.test(sif): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  sif
## Dickey-Fuller = -4.509, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

```r
# Plot the data
ts.plot(sif)
```



* Is it likely to be stationary? Per the ADF test (p-value = 0.01), the process is indeed stationary.

- Is there any spikes? There are four spikes based on the plot.
- Is there a trend? There seems to be a seasonal trend... which is strange.
- Does it have any missing data? Per the summary table, no NA values... so not likely to have missing data.

Question 2

Please split the time series into a training set and a testing set. We are interested in forecsting the web visit volume in the next 7 days, so please define the training set as the first 359 data points, and the testing set as the last 7 data points.

Hints:

1. Please use the `ts()` function in the `tseries` package to change the numeric data into a Time-Series
2. For the training set, `start=1`, and `end=359`
3. For the testing set, `start=360`, and `end=366`

```r
# please write your code below
sif_train = ts(sif[1:359], start = 1, end = 359)
sif_test = ts(sif[360:366], start = 360, end = 366)
```

Part 1: ARIMA alone

Question 1

Now let's see how ARIMA model alone performs on this dataset. For simplicity, please run an `auto.arima()` to select the parameters. What's the selected model? Are there any significant coefficients?

Hints:

- Roughly a coefficient is significant if its magnitude is at least twice as large as its standard error.

```r
# please write your code below
arima0 = auto.arima(sif_train)
arima0
```

```
## Series: sif_train
## ARIMA(2,1,1)
##
## Coefficients:
##          ar1      ar2      ma1
##       0.7736  -0.2261  -0.9163
## s.e.  0.0647   0.0580   0.0447
##
## sigma^2 estimated as 1331249:  log likelihood=-3031.15
## AIC=6070.31   AICc=6070.42   BIC=6085.83
```

- The selected model, based on the auto.arima() output is is an ARIMA(2,1,1).
- There are NO significant coefficients based on the output.

Question 2

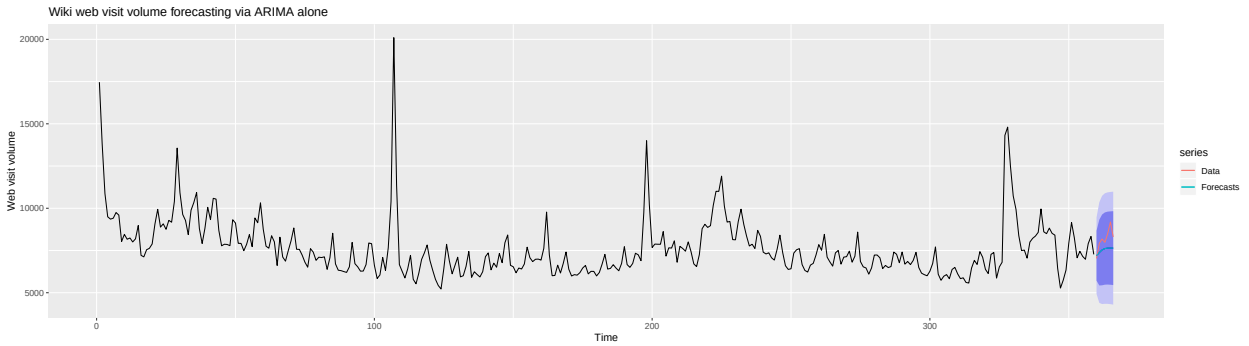
Please forecast the web visit volume in the next 7 days. Plot the forecasts, the raw data, and the 80% and 95% prediction intervals. What do you think of the performance of the forecasting? Our lowest expectation is that at least the 95% prediction inverval should cover the true data. Our highest expectation is that the forecasts highly aligh with the true data.

Hints:

1. You may borrow our code from the previous lecture (US consumption, CO2, or the bitcoin example).

```r
# Forecast on the testing set
arima0_forecast = forecast(arima0, h = length(sif_test))

# Plot the data
autoplot(arima0_forecast) +
  autolayer(sif_test, series="Data") +
  autolayer(arima0_forecast$mean, series="Forecasts") +
  labs(title="Wiki web visit volume forecasting via ARIMA alone", y="Web visit volume")
```




```
arima0_forecast_mu = forecast(arima0, h = length(sif_test))$mean
#The forecasts are acceptable, but not perfect, as we can see that it does not capture the trend.
```

Question 3

Please Calculate the RMSE of the forecasts. This RMSE will be used as a benchmark for comparison below.

```
# Report the RMSE
rmse_arima = accuracy(arima0_forecast, sif_test)[2,2]
rmse_arima
```

```
## [1] 795.7763
```

Part 2: Linear model alone

Question 1

Create two regressors, one is t, the other is t^2. You may call them `t1` and `t2`. We will use these two variables for linear regression. Later on you may try polynomials with a higher degree.

Hints:

1. For the testing set, t starts from 360.

```
# please write your code below
t = 1:366
t1 = t[1:359]
t2 = t1^2

t1_test = t[360:366]
t2_test = t1_test^2
```

Question 2

On the training set, run a linear regression using `sif` against both `t1` and `t2`. Are all coefficients significant?

Hints:

1. For the testing set, t starts from 360.

```
# please write your code below
lm0 = lm(sif_train ~ t1 + t2)
summary(lm0)
```

```
##
## Call:
## lm(formula = sif_train ~ t1 + t2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2710.7  -922.0  -356.4   462.6 12495.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.439e+03  2.540e+02  37.157  < 2e-16 ***
## t1          -2.290e+01  3.259e+00  -7.027 1.08e-11 ***
## t2           5.392e-02  8.766e-03   6.151 2.07e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1595 on 356 degrees of freedom
## Multiple R-squared:  0.1365, Adjusted R-squared:  0.1316
## F-statistic: 28.13 on 2 and 356 DF,  p-value: 4.555e-12
```

Question 3

Please extrapolate the results to the testing sets, and illustrate them.

Hints:

1. In the `forecast()` function, you need to specify `newdata=`.
2. `newdata` should have column names match those in the previous question.
3. Please also convert the forecast values into `ts()`.

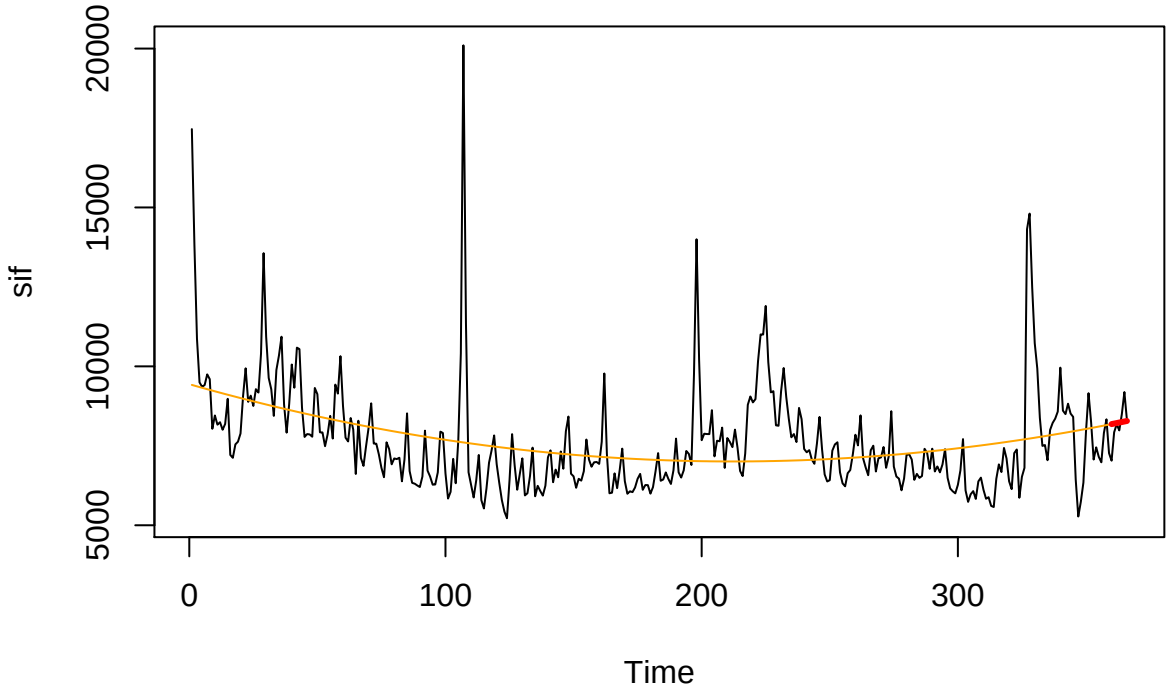
```
# please write your code below
X_test = data.frame(t1 = t1_test, t2 = t2_test)

lm0_forecast = ts(forecast(lm0, newdata = X_test, h = 7))

lm0_forecast_mu = lm0_forecast$mean

lm0_forecast_mu = ts(lm0_forecast_mu,start = 360,end = 366)

ts.plot(sif)
lines(c(lm0$fitted.values),col="orange")
lines(c(rep(NA,length(sif_train)),lm0_forecast_mu),col="red",lwd="3")
```



Question 4

Please calculate the RMSE on the testing set for the linear model above.

Hints:

```
# please write your code below
rmse_lm = accuracy(lm0_forecast, sif_test)[2,2]
rmse_lm
```

```
## [1] 583.4326
```

Part 3: Linear model PLUS Arima

Question 1

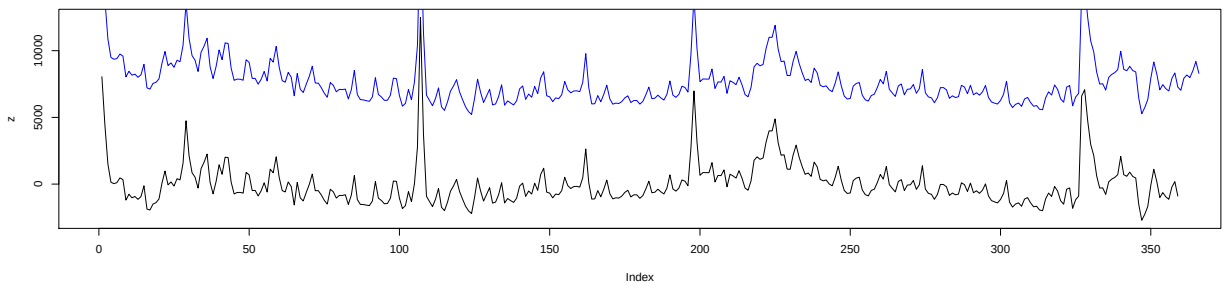
Extract the residuals of the linear model above. Consider it as a new time series for Arima. Plot it and compare the curve with the one in Part 0.

Hints:

1. You can use `$residuals` to extract residuals from a linear model.

```r
# please write your code below
z = lm0_forecast$residuals

par(mfrow = c(1, 1))
plot(z, type = "l")
lines(sif, col = "blue")
```



Question 2

Run an `auto.arima()` on the residuals, and calculate the forecasts on the testing set. What's the model selected by `auto.arima()`?

Hints:

1. Similar to Question 1 and 2 in Part 1.

```r
# please write your code below
arima1 = auto.arima(lm0_forecast$residuals)
arima1
```

```
## Series: lm0_forecast$residuals
## ARIMA(1,0,1) with zero mean
##
## Coefficients:
##          ar1     ma1
##       0.5654  0.2816
## s.e.  0.0629  0.0704
##
## sigma^2 estimated as 1300166:  log likelihood=-3035.78
## AIC=6077.56   AICc=6077.63   BIC=6089.21
```

```r
z_forecast = forecast(arima1, h = length(sif_test))

z_forecast_mu = z_forecast$mean
```

- The model selected by the auto.arima() output is an ARIMA(1,0,1).

Question 3

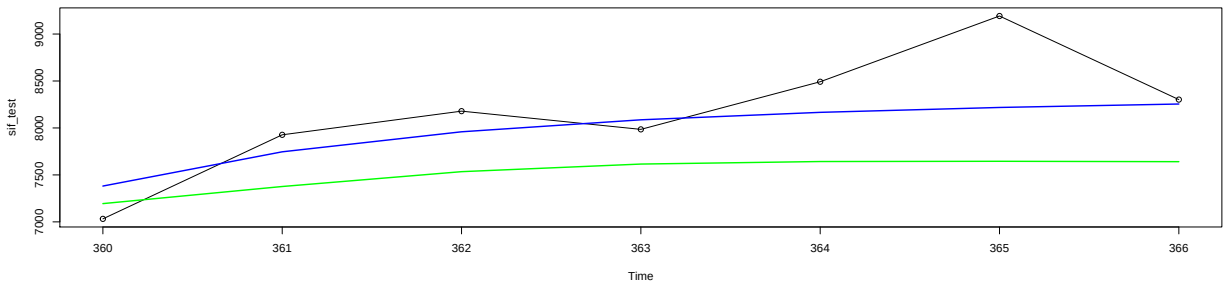
Now please calculate the forecasts combining both the linear model and the ARIMA, and illustrate the results from all three models (i.e., arima alone, linear model alone, linear model plus arima)

Hints:

1. Add the predicted residuals to the predicted values by the linear model

```
# please write your code below
y_forecast = lm0_forecast_mu + z_forecast_mu

plot(sif_test, type = "o")
lines(arima0_forecast$mean, col = "green", lw = "2")
lines(lm0_forecast$mean, col = "red", lw = "2")
lines(y_forecast, col = "blue", lw = "2")
```



Question 4

Now calculate the RMSE on the testing set for the linear model PLUS arima above. How do you compare the RMSE with the one given by arima or the linear model?

Hints:

```
# please write your code below
rmse_lm_plus_arima = accuracy(y_forecast, sif_test)[1,2]
rmse_lm_plus_arima
```

```
## [1] 425.9883
```

Done!

Congratulations!