Time series analysis of Mobile application user statistics

Jonathan Tonglet – R0827509

Advanced Time Series Analysis- D0M63B

Professor : Christophe Croux

Presentation of the dataset

- Dataset published on Kaggle in 2018 *
- User statistics for an unnamed mobile app
- Hourly data during a week (169 observations)
- Starts on the 22/12/2018 and ends on the 29/12/2018

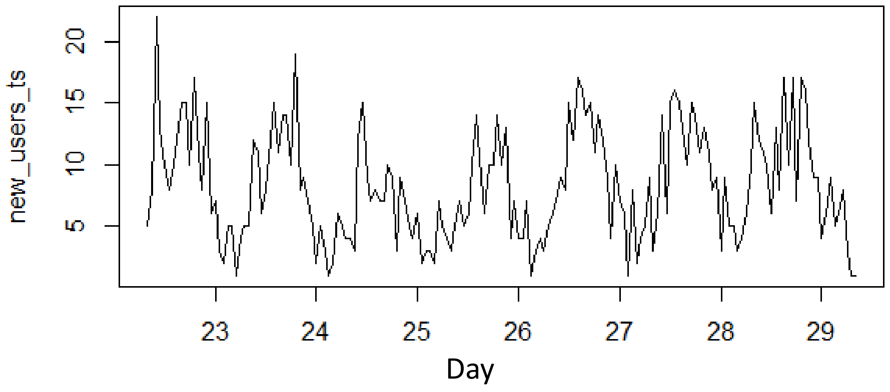
- In this analysis, we focus on 2 time series
  - Number of new users per hour
  - Number of active users per hour

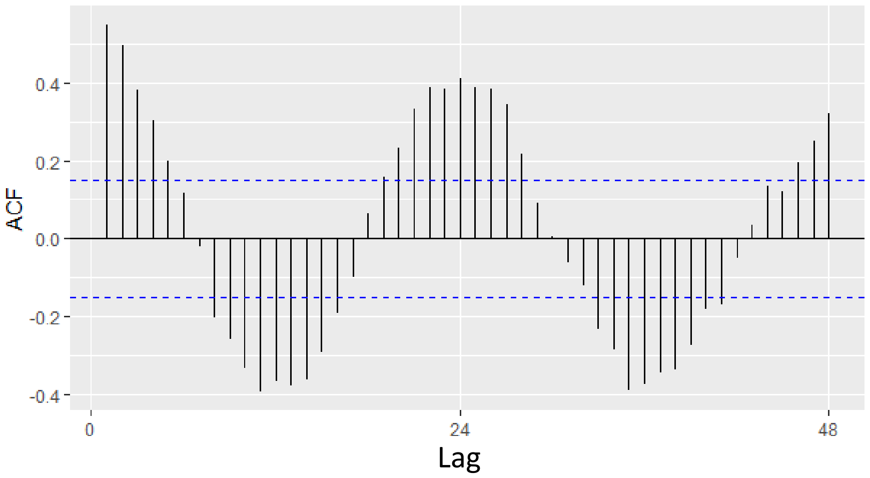
- Interesting business questions to study
  - Do new users arrive at specific time periods during the day/ during the week?
  - Does the number of active users at a time ***t-1*** have a impact on new users at time ***t*** (recommendations, app sharing, …)?

* Beer, W.(2018).*Mobile application users statistics* (Version 4). Retrieved from : https://www.kaggle.com/wolfgangb33r/usercount?select=app.csv

Univariate TS analysis – New users

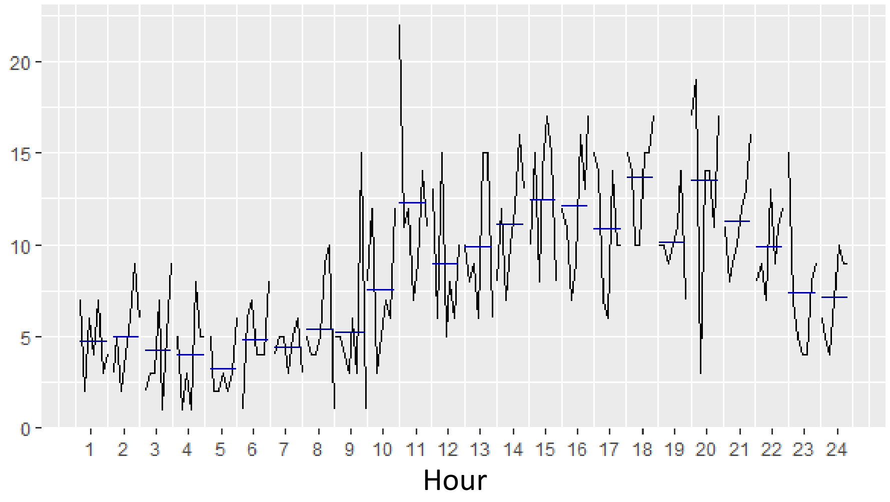
A) Analysis of the TS

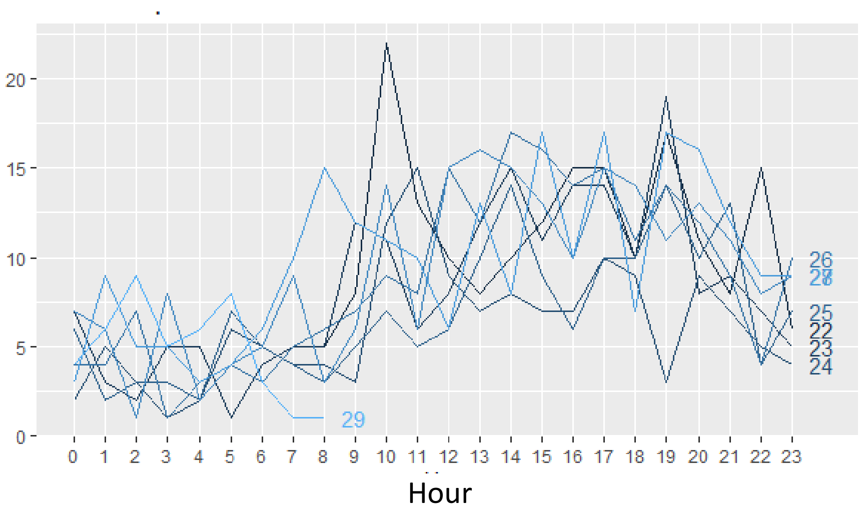





```
        Box-Ljung test

data:  new_users_ts
X-squared = 306.36, df = 15, p-value < 2.2e-16
```

```
data:  new_users_ts
ADF(1) = -3.8438, p-value = 0.003115
alternative hypothesis: true delta is less than 0
sample estimates:
     delta
-0.2777697
```





Comments

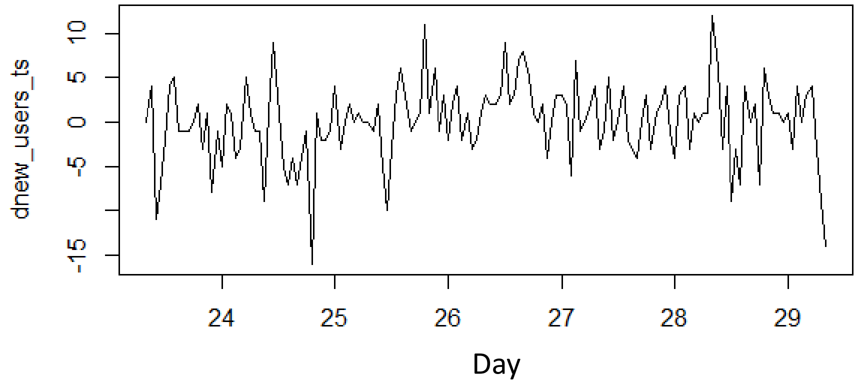
- Test results : the TS is not stationary and is not white noise (critical value = 0.05 *)
- Clear seasonal (**daily**) pattern
- No major differences between the 7 days of the week
- Month and seasonal plots confirm the seasonal pattern
- Peak in new users in the afternoon

→ Go into seasonal differences of **lag 24**

* The same critical value will be applied for all subsequent tests in the report

Univariate TS analysis – New users

B) Analysis of the TS in seasonal differences

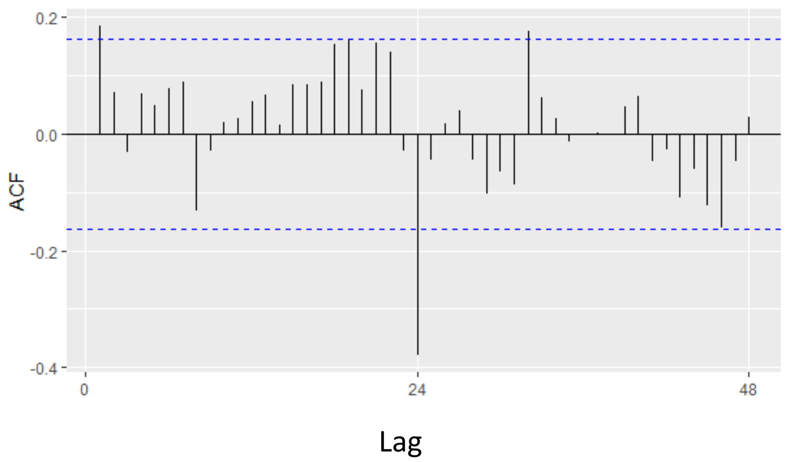


Comments

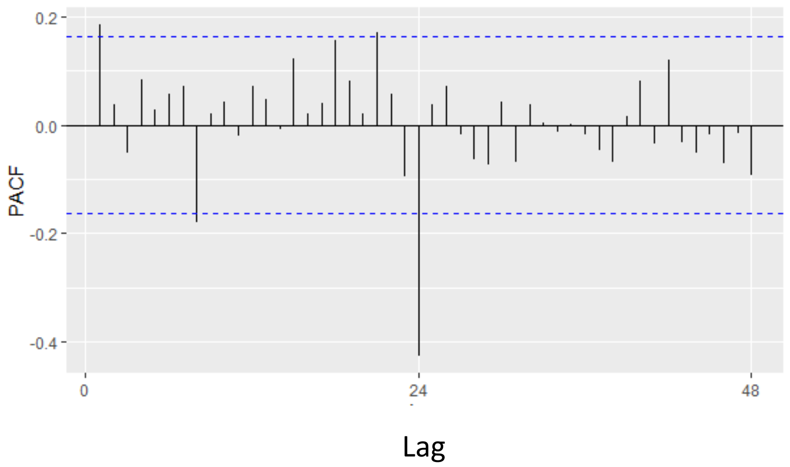
- ADF test result: the TS in seasonal differences is stationary
- Ljung-Box test : the TS seems to be white noise

```
        Box-Ljung test

data:  dnew_users_ts
X-squared = 14.739, df = 15, p-value = 0.4703
```

```
        ADF test

data:  dnew_users_ts
ADF(0) = -8.9585, p-value = 7.4e-13
alternative hypothesis: true delta is less than 0
sample estimates:
   delta
-0.80174
```



- Correlograms show that there still exists a seasonal effect at lag 24
- The TS is not purely white noise
- We will use SARIMA models
- Different possible models :
  - SARIMA(1,0,1)(1,0,1)
  - SARIMA(1,0,1)(1,0,0)
  - SARIMA(1,0,1)(0,0,1)



Univariate TS analysis – SARIMA models

Model 1 – Estimation & Validation SARIMA(1,0,1)(1,0,1)	Model 2 – Estimation & Validation SARIMA(1,0,1)(0,0,1)

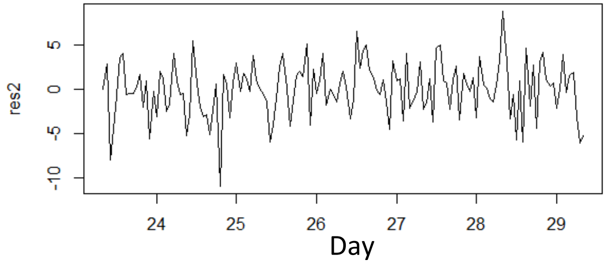
```
arima(x = snew_users_ts, order = c(1, 0, 1), seasonal = c(1, 0, 1))

Coefficients:
         ar1      ma1     sar1     sma1  intercept
      0.9774  -0.8875  -0.1262  -0.8235    -0.0004
s.e.  0.0307   0.0526   0.1590   0.3198     0.3871

sigma^2 estimated as 10.38:  log likelihood = -390.37,  aic = 792.74
```

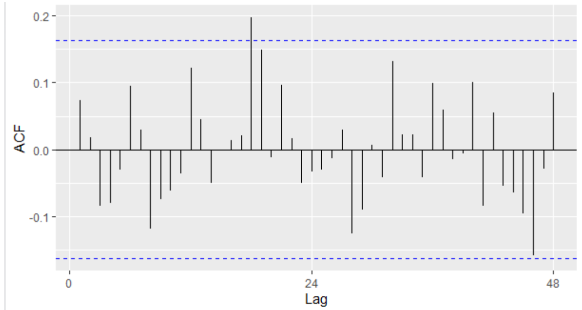
- Sar1 coefficient & intercept are not significant
- We move to more simple models
    - Model 2 : SARIMA(1,0,1)(0,0,1)
    - Model 3 : SARIMA(1,0,1)(1,0,0)

```
arima(x = snew_users_ts, order = c(1, 0, 1), seasonal = c(0, 0, 1))

Coefficients:
         ar1      ma1     sma1  intercept
      0.9786  -0.8962  -0.9997    -0.0120
s.e.  0.0327   0.0490   0.2852     0.3706

sigma^2 estimated as 9.342:  log likelihood = -390.77,  aic = 791.55
```

- All coefficients are significant except the intercept

Analysis of the residuals






```
        Box-Ljung test

data:  res2
X-squared = 11.478, df = 15, p-value = 0.718
```

- From the graphs and the test, we conclude that the residuals are white noise
- The model is validated

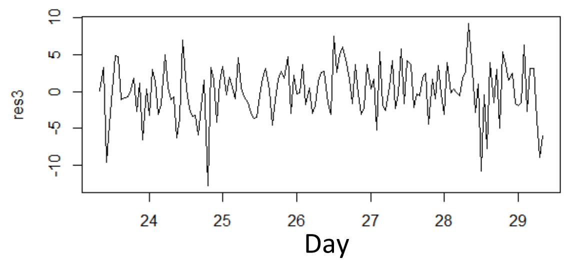
Univariate TS analysis – SARIMA models

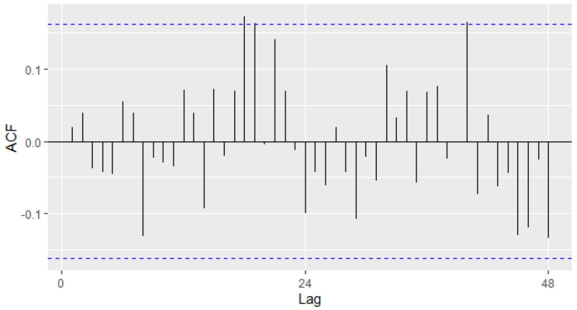
Model 3 – Estimation & Validation SARIMA(1,0,1)(1,0,0)	Models 2 & 3 Comparison

```
arima(x = snew_users_ts, order = c(1, 0, 1), seasonal = c(1, 0, 0))

Coefficients:
         ar1      ma1     sar1  intercept
      0.8822  -0.7078  -0.5685    -0.0260
s.e.  0.1761   0.2799   0.0744     0.5205

sigma^2 estimated as 13.16:  log likelihood = -397.36,  aic = 804.72
```

- All coefficients are significant except the intercept

Analysis of the residuals






```
        Box-Ljung test

data:  res3
X-squared = 8.2977, df = 15, p-value = 0.9113
```

- From the graphs and the test, we conclude that the residuals are white noise
- The model is validated

	AIC
Model 2	791,5
Model 3	804,7

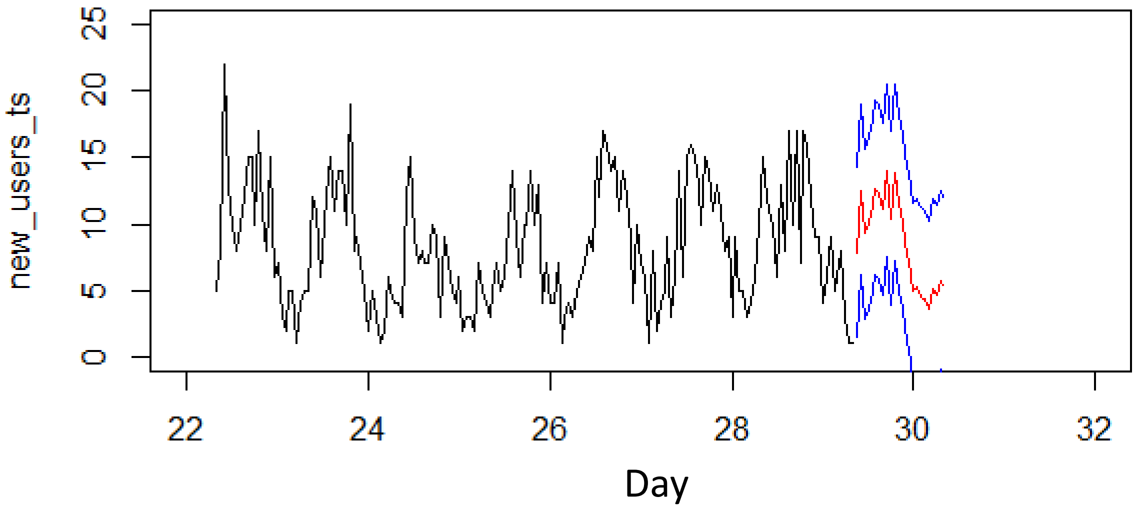
	BIC
Model 2	806,4
Model 3	819,6

- Model 2 is preferred because it has the lowest AIC & BIC values
- We will now make forecasts based on the two models

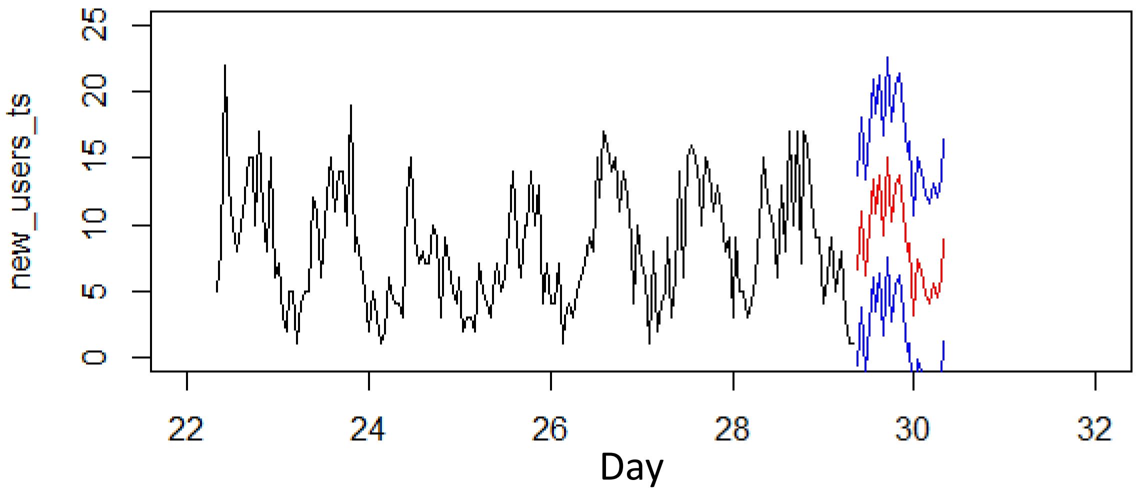
Univariate TS analysis – Forecasting

- Forecasting for the next 24 hours (29/12/18 at 10 AM – 30/12/18 at 10 AM)

Model 2 – Forecast
SARIMA(1,0,1)(0,0,1)



Model 3 – Forecast
SARIMA(1,0,1)(1,0,0)



Comments

- Both forecasts are very similar
- MAE almost equal (expanding window approach)
- Prediction intervals are close to the forecasted values
- Predict a peak of new users in the 29th afternoon
- Predict a low number of new users during the night
- Globally respect the seasonal pattern of previous days

	MAE
Model 2	215
Model 3	216

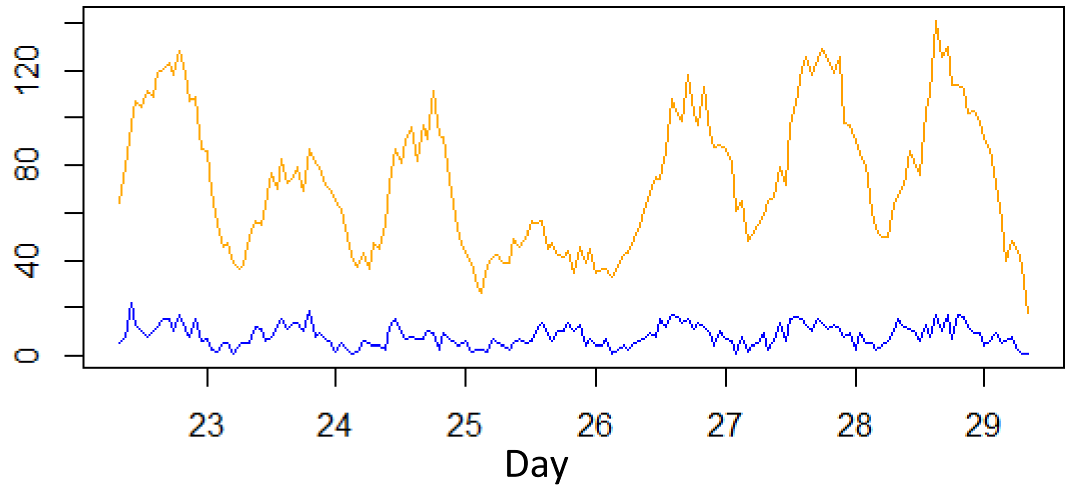
Diebold-Mariano test

```
        Diebold-Mariano Test

data:  error2.herror3.h
DM = -0.61028, Forecast horizon = 1, Loss function power = 1, p-value =
0.5436
alternative hypothesis: two.sided
```

We reject that the forecasting performances of model 2 and model 3 are significantly different

Multivariate TS Analysis – Active users

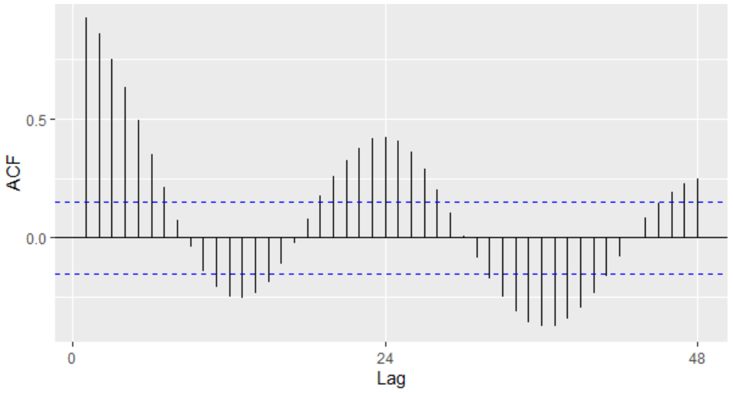
A) Introducing a new TS : number of active users



- Blue : Number of new users
- Orange : Number of active users

```
         ADF test

data:  active_users_ts
ADF(3) = -3.8439, p-value = 0.003113
alternative hypothesis: true delta is less than 0
sample estimates:
     delta
-0.1114991
```

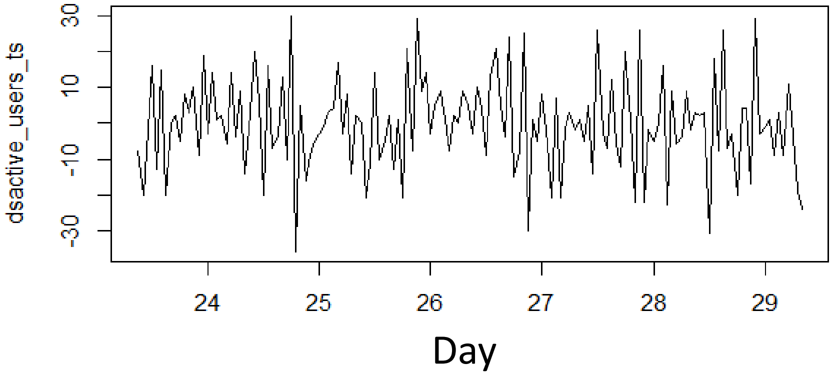
```
        Box-Ljung test

data:  active_users_ts
X-squared = 568.25, df = 15, p-value < 2.2e-16
```



- Test results : the TS is not stationary and is not white noise
- It has a seasonal pattern **and** a trend

→ We go in differences and in seasonal differences (lag = 24)

B) Number of active users in difference & seasonal differences




```
         ADF test

data:  dsactive_users_ts
ADF(0) = -16.956, p-value < 2.2e-16
alternative hypothesis: true delta is less than 0
sample estimates:
     delta
-1.391063
```

```
        Box-Ljung test

data:  dsactive_users_ts
X-squared = 33.196, df = 15, p-value = 0.0044
```

- ADF test: the TS is now stationary
- Ljung-Box test : TS is not white noise

→ We can now create a multivariate model

Multivariate TS Analysis – ADLM (2)

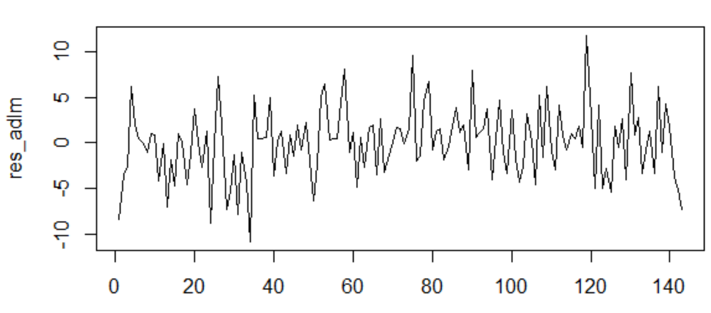
- sY = number of new users in seasonal difference
- dsX = number of active users in difference and seasonal difference

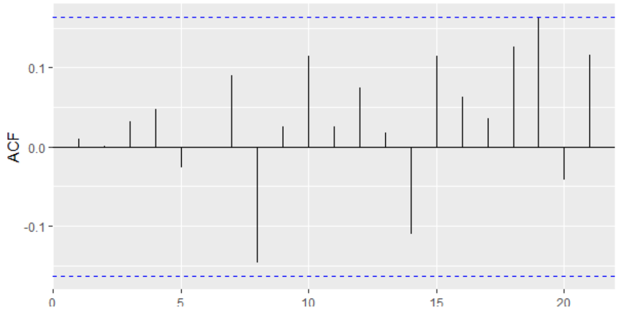
```
lm(formula = sY.0 ~ dsX.1 + dsX.2 + sY.1 + sY.2)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.8365  -2.6627   0.2428   1.8854  11.7861 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02066    0.32673  -0.063  0.94968    
dsX.1        0.16262    0.02701   6.020 1.49e-08 ***
dsX.2        0.03391    0.03011   1.126  0.26204    
sY.1         0.23773    0.08558   2.778  0.00623 ** 
sY.2         0.01694    0.08192   0.207  0.83645    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.903 on 138 degrees of freedom
Multiple R-squared:  0.2432,	Adjusted R-squared:  0.2213 
F-statistic: 11.09 on 4 and 138 DF,  p-value: 7.906e-08
```

Comments

- dsX1 and sY1 coefficients are significant
- The other coefficients are not significant
- R-squared : 24% of the variance is explained by the independent variables
- The variables are jointly significant (F-statistic)

Analysis of the residuals





Hour

Lag

```
Box-Ljung test

data:  res_adlm
X-squared = 12.367, df = 15, p-value = 0.651
```

- From the graphs and the test, we can conclude that the residuals are white noise
- The model is validated

Multivariate TS Analysis – ADLM (2)

Test for Granger causality

```
Analysis of Variance Table

Model 1: sY.0 ~ dsX.1 + dsX.2 + sY.1 + sY.2
Model 2: sY.0 ~ sY.1 + sY.2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    138 2102.2
2    140 2668.2 -2      -566 18.578 7.168e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We strongly reject that there is no Granger causality
- The number of active users (in difference) provides incremental predictive power to predict the number of new users (in difference)

Conclusion

This analysis helped us to get more insights in the data of the mobile app.
We have discovered that :

- New users tend to arrive during the afternoon.
- The same pattern is repeated every day of the week with almost no differences.
- The time series is quite stable over time. Forecasted values have narrow prediction intervals.
- The number of new users at time ***t-1*** and the number of active users at time ***t-1*** can help us explain the number of new users at time ***t***. However, we have to be careful with this assumption and can’t conclude that there is a direct causality.
- Peak time for new users corresponds to peak time for active users. Interesting information for the social media/marketing strategy of the company.