Estimating bike-share trips : Dublin City

Maxim Le Cloerec

August 2019

1 Data selection

Nidhin George's previous work [1] has provided a R table containing an observation set on Dublin self-service bikes. Among these observations is a set of interesting information in the study of the activity of the self-service bicycles of Dublin, such as : the number of a station, its name and address, its number of bike stands, as well as the number of stands available at a certain date and at a certain time. In order to estimate the number of bike journeys per day in Dublin, the first step was to create a new variable containing the number of bikes available for each observation of our table. This was done through a simple calculation of subtracting the total number of stands from a station minus the number of stands available. As a result of this calculation, the variable contained some errors, such as available bikes under 0, which is obviously impossible. All the observations containing these errors, more than a hundred, have been removed from the table.

Observations being collected every 10 minutes, it is possible to obtain a maximum of 144 observations per day for a station. The article writing by Cyril Médard de Chardon and Geoffrey Caruso [2] indicates that it is advisable to keep for the study the days comprising at least 95 per cent of these 144 observations. The Dublin bikes only run from 5:00 am to 0:30 am, this represents 26 10-minute intervals where Dublin bikes will not be used, except those returned. The choice was therefore to reduce the 144 daily observations to 144 minus 26, so 118 observations. 95 per cent of 118 being equal to 112, only the days for which the 100 stations in Dublin had at least 112 observations were kept, only 74 days out of the initial 426 days.

2 Calculation

As given in the article cited above, the first step consisted in calculating the differences in bikes available between each period of $\Delta = 10$ minutes, for each station. To do this, the reduced table of 74 days was sorted by date, per station,

then per hour and the calculations were simply done by subtracting the available bikes value from an observation to its previous observation.

Now having all the difference in available bikes, a function named overall() was created to calculate the total sum of interactions for each of the 74 days. The result is then available in a table called result_table. It is also possible to get the interactions of a day only via the function unique_date() which takes a day as a parameter.

The number of bike journeys is calculated via the following equation :

$$journeys = \frac{|interactions - rebalancing| + collisions}{2}$$

Not having the OD (origin-destination) data to calculate the rebalancing, it was necessary to find an alternative to estimate this rebalancing. After thinking about how to tackle this problem, all in a limited time, the idea was to use the station clusters obtained during the previous work of Nidhin George [1] to estimate this rebalancing. Indeed, the previous works tended to show that stations are opposed in their behavior. For example, for cluster 3 stations, the number of stands available is constant and high from 0:00 am to 5:00 am and begins to reduce to reach its lowest from 9:00 am to 3:00 pm to finally regain its state of the morning. This is the total opposite for cluster 4. We can therefore imagine that cluster 4 stations which have a lot of bicycles between 9:00 am and 3:00 am are used to rebalance the stations of cluster 3 which have few bits on this period.

The goal was therefore to find in each of the clusters extreme values likely to be rebalancing. In order to detect the most extreme values, it was necessary to reduce the table by not selecting the observations for which the difference in bikes was 0. The study of extreme values is therefore done on 384 505 observations, around 38 per cent of the initial table. Indeed, the number of extreme values with the full data was too great and low difference in bikes such as 3 could be considered as extreme values.

For each of the extreme values obtained afterwards, we subtract this value from the average of the cluster at the given time, to estimate the number of rebalancing. For example, if a difference for a cluster 3 station is 12 at 1:00 am while the average for a cluster 3 station at 1:00 am is 2, we can imagine that the number of rebalance bikes will be 12-2 = 10, even if it is only proper to a personal and unscientific reflection. We were also able to calculate the number of collisions per day, simply adding up the differences over a whole day. This sum should normally be 0 if each trip has arrived at its destination.

Therefore, we obtained the number of interactions X_Δ, the rebalancing R_Δ, the collisions C_Δ, to now calculate the number of bike journeys per day. Note that the standard deviation for our estimated rebalancing is 650, which

remains quite important, with values sometimes very low, which can be explained by important dates : 26 rebalancing on Christmas Day against 1 065 on average. Collision figures seem to be more in adequation with an average of 12 per day, which may include bike losses or maintenance. However, some figures seem abnormal, like -275 and +716 for example, and make the analysis more complicated.

3 Regression

The Médard de Chardon and Caruso document [2] proposes 4 different models to evaluate. Through this section, we will discuss these 4 models, called the individual aggregated model, the combined day aggregated model, the interval aggregation model and the finally the station aggregation model. For each of the models, the aim will be to analyze the R exits as well as to verify the different assumptions on the residues, such as linearity, the normality of the residues, homoscedasticity and the independence of residual terms errors in order to evaluate these different models.

3.1 Individual day aggregated model (DAM)

The first model is based on the hypothesis of the relation between the number of paths and the number of interactions divided by 2. It is given by the following formula :

$$T_d = 0 + \beta_1 T_{x\Delta d} + \beta_2 T_{x\Delta d}^2 + \varepsilon$$

In order to evaluate the R regression, here are graphical representations allowing to validate or not the assumptions :

We used to check the linear relationship assumptions with the residuals vs fitted graph. A horizontal line, without distinct patterns would be an indication for a linear relationship, and that does not looks good here. The normal Q-Q plot is used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line and that is a good point here. Scale-Location is used to check the homoscedasticity. Horizontal line with equally spread points is a good indication of homoscedasticity. This is not the case in our example, where we have kind of heteroscedasticity problem. We identify influential cases with the residuals vs leverage graph, that is extreme values that might influence the regression results when included or excluded from the analysis.

Here, only 2 out of 4 criteria are satisfied, which does not validate the model. In addition, the values that could be influential on the analysis were removed and this did not improve the result.

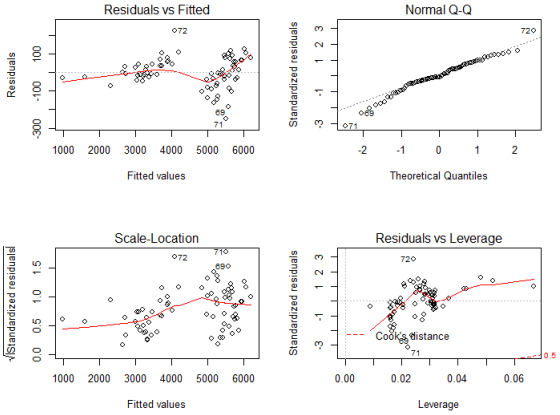

Figure 1: DAM assumptions

One way to satisfy these assumptions may be to modify one of the predictive variables by its napierian logarithm or its square root. After several regression tests, the following model was found :

$$T_d = 0 + \beta_1 log(T_{x\Delta d}) + \beta_2 T_{x\Delta d}^2 + \varepsilon$$

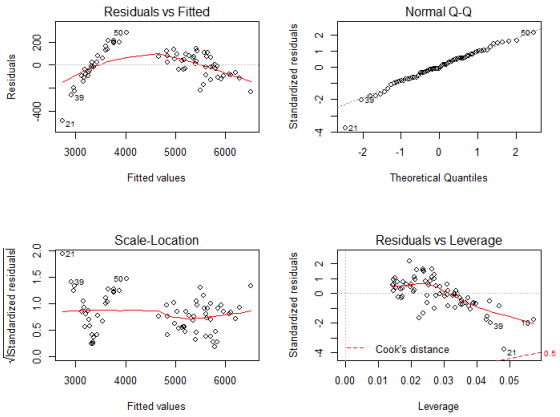

Figure 2: Modify DAM assumptions

Here, the normal Q-Q plot, the scale-location and the residuals vs. leverage all seem to be satisfied. In spite of that some doubts could be emitted concerning the linearity which seems to follow a slight pattern. However, the linearity test

(raintest) satisfied the hypothesis of linearity, as well as the other tests on the other assumptions (shapiro.test, bptest, dwtest).

For this analysis, 3 influential values were previously detected and subsequently removed. Among these 3 values were very low rebalancing values probably related to Christmas periods, as well as a collision value of 716 which is difficult to explain. Here is the R output of this analysis :

```
Call:
lm(formula = test$T ~ 0 + log1p(test$Tx) + test$Tx2)

Residuals:
    Min      1Q  Median      3Q     Max
-485.71  -84.23   -7.70   77.70  283.11

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
log1p(test$Tx) 3.000e+02  4.498e+00   66.71   <2e-16 ***
test$Tx2       7.205e-05  1.190e-06   60.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 131.8 on 69 degrees of freedom
Multiple R-squared:  0.9993,    Adjusted R-squared:  0.9992
F-statistic: 4.642e+04 on 2 and 69 DF,  p-value: < 2.2e-16
```

Figure 3: DAM output

With $\beta_1 = 3.000\text{e+}02$, $\beta_2 = 7.205\text{e-}05$, and also the error RMSE/$\overline{T_d}$ = 0.028 which is good.

3.2 Combined day aggregated model (cDAM)

The second model is based on the same hypothesis as the previous model, but here we also use a the active stations parameter in order to normalize the density of activity. It is given by the following formula :

$$T_d = 0 + \beta_1 T_{x\Delta d} + \beta_2 \frac{T_{x\Delta d}^2}{A_{x\Delta d}} + \varepsilon$$

In order to evaluate the R regression, here are statistical tests results allowing to validate or not the assumptions :

To satisfy every assumptions, p-values of the Rainbow test, Shapiro test, and Durbin-Watson test must be over 5 per cent, and p-value of the Breusch-Pagan test under 5 per cent. Here, only two assumptions are validated, including the normality of residuals and the independent errors.

Other tests have been attempted, using the methods previously mentioned, such as the use of the napierian logarithm, the square root or even the suppression of extreme values. However, for each of the models, none satisfied the four assumptions and therefore none was reliable.

```
> raintest(cDAM)

        Rainbow test

data:  cDAM
Rain = 2.7531, df1 = 37, df2 = 34, p-value = 0.001814

> shapiro.test(resid(cDAM))

        Shapiro-Wilk normality test

data:  resid(cDAM)
W = 0.97369, p-value = 0.1251

> bptest(cDAM)

        studentized Breusch-Pagan test

data:  cDAM
BP = 2.8143, df = 2, p-value = 0.2448

> dwtest(cDAM)

        Durbin-Watson test

data:  cDAM
DW = 1.7159, p-value = 0.09812
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 4: cDAM tests output

3.3 Interval aggregation model (IAM)

The interval aggregation model use the interval of times in order to estimate the number of interactions rather than the number of daily trips. It sums all interactions to get the total number of interactions for every interval duration. Here, the active stations variable is also use to normalize the density :

$$i_{\Delta dt} = 0 + \beta_1 x_{\Delta dt} + \beta_2 A_{x\Delta dt} + \beta_3 A^2_{x\Delta dt} + \varepsilon$$

In order to evaluate the R regression, here are graphical representations allowing to validate or not the assumptions :

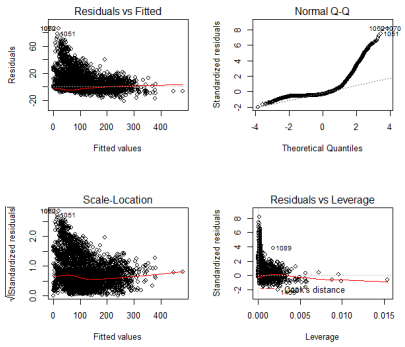

Figure 5: IAM assumptions

Since this model uses the number of interactions and stations used for each time period, the intervals from 0:30 am to 5:00 am have been removed to keep only those where the BSS is really running. Here, the graphic representations do not satisfy the assumptions. By using the napierian logarithm on the variable $x_{\Delta dt}$, the graphical representations below seem more promising. Unfortunately, the statistical tests performed subsequently do not validate the model.

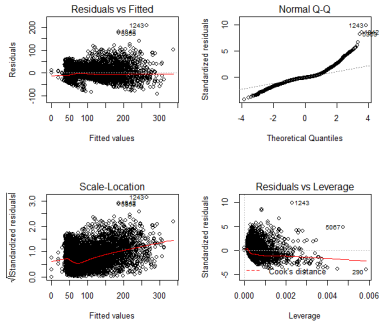

Figure 6: Modify IAM assumptions

3.4 Station aggregation model (SAM)

The station aggregation model use the frequency of station us throughout the day and not the number of active stations in the BSS, as the previous ones. Here, variable A gives the total number of time slots where a station was used. The model is given by the following formula :

$$i_{\Delta sd} = 0 + \beta_1 x_{\Delta sd} + \beta_2 A_{x\Delta sd} + \beta_3 A^2_{x\Delta sd} + \varepsilon$$

In order to evaluate the R regression, here are graphical representations allowing to validate or not the assumptions :

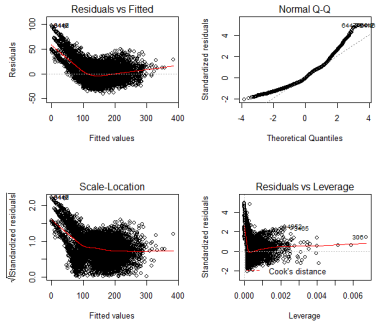

Figure 7: SAM assumptions

Once again, the graphic representations do not satisfy the assumptions. By using the square root on every explanatory variables, the graphical representations below seem more promising. Unfortunately, the statistical tests performed subsequently do not validate the model.

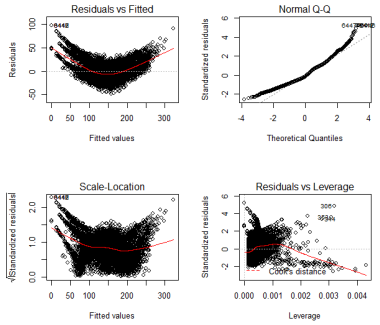

Figure 8: Modify SAM assumptions

3.5 Conclusion

We conclude that three of the four models could not be valid because of the lack of satisfaction of the assumptions. Only the first model estimating the number of journeys per day concluding. It also has good results since its RMSE/$\overline{T_d}$ is a little under 0.03 while that of Cyril Médard de Chardon and Geoffrey Caruso [2] for DAM is 0.06 for a $\Delta = 10$. It could be that the first model is the best since uses explanatory variables using personal rebalancing estimates. Other models using these explanatory variables with data variables, such as active station variables A_Δ, may be responsible for poor results in assumptions.

The future goal would eventually be to be able to determine a more complex algorithm that could better estimate the rebalancing, which seems to be the key to the problem here, since we do not have any OD data available for Dublin Bikes. It may be possible to use more methods to validate assumptions, such as the White's robust standard error for homoscedasticity, and not just use simple techniques like napierian logarithms or square root.

4 Appendix

https://github.com/mackesim/dublinbikes

References

[1] G. Nidhin. *Implementation and Comparison of Strategies to Predict the Availability of Dublin Bikes*. 2018.

[2] C. Medard de Chardon and G. Caruso. *Estimating bike-share trips using station level data*. 2015.