# Model Reuse with Bike Rental Station Data

**Authors:**

1. **Arun Bala Subramaniyan, M.S. Industrial Engineering, Arizona State University.**

2. **Dr. Rong Pan, Associate Professor of Industrial Engineering, Arizona State University.**

## Introduction and Motivation

Bike Rental Stations are a good business in places with large number of tourists and also used for routine work. In this project, the bike rental station located in Valencia, the third largest city of Spain is considered. The bike rental company would like to predict the number of bikes available in each station three hours in advance.

The uses for such prediction are

1) A user plans to rent (or return) a bike in 3 hours time and wants to choose a bike station which is not empty (or full).

2) The company would like to avoid situations where a station is empty or full

The predictions depend on location, time, weather conditions and other factors. Also, the profile of bike availability in the station is learnt from historical information.

## Methodology

The idea is to reuse the models learnt from the old stations (station 1 to station 200) and predict the number of bikes in the new stations (station 201 to station 275).

## Model Extraction

There are 7 base models and 6 trained models available for each of the 200 stations. As the deployment data for stations 201 to 275 is given, these 13 models can be used for predicting the number of bikes in the stations 201 to 275. The model with less Mean Absolute Error (MAE) is selected as the best model for the station. This process is continued for selecting the best model for all the new stations (201 to 275). The R software package is used for this purpose. The extracted best models for the new stations are stored in .csv file.

In some cases, the prediction result is negative or it exceeds the maximum limit of the bikes that can be parked in a station. This can be overcome by adding a constraint such that whenever the result is negative, the value is reset to zero and whenever the result exceeds the maximum limit, the value is reset to the number of docks at the particular station. So, this helps in reducing the error value.

## Prediction

Using the extracted models, the number of bikes at the new stations is predicted. The same constraints are applied to avoid negative values and over fitting. The R software is used for predicting the number of bikes. The results of this prediction are stored in .csv file.

## Other Methods tried for prediction

Instead of reusing the trained models, new models were built with the given deployment data for stations 201 to 275. Some of the methods used are given below.

## Ordinary Least Squares Method

After collecting and cleaning the data, the first model was built using all the regressors under consideration. A thorough analysis of this full model, including residual analysis and multicollinearity check was done. The best subset regression was also tried. The normal probability plot showed deviations at the upper and lower tails. The residual plots (deleted residuals versus the fitted values) for some stations showed a significant double bow pattern, violating the assumption of constant variance. Also, there were some outliers observed in the plot of residual versus observations.

To explore about the outliers, the values of ordinary residuals, studentized residuals, leverage (HI1), Cook's distance, DFFIT were collected. Though some outliers were observed, no influential points were noticed which has been confirmed from the cook's distance. Since the reason for the unusual observations were not explicit, these observations were not removed and hence, included for modelling. As the number of outliers were less, removing these observations did not improve the model significantly. Finally, the MAE value obtained for this model was 2.724.

## Poisson Regression

In order to improve the model further and make it useful for prediction, the Poisson Regression analysis was done. The reason for choosing the Poisson regression was that the response variable involved counting the number of bikes, which was discrete. Also, the log link was particularly attractive for Poisson regression as it ensures that all of the predicted values of the response variable will be nonnegative. The MAE for this model was 3.068.

## Zero Inflated Poisson (ZIP) Regression

As there was a possible evidence of excess of zeroes when examining the response data, there aroused a doubt that the zeroes might be inflated. So, in order to solve this problem, the Zero Inflated Poisson Regression analysis was used. The MAE for this model turned out to be 2.774.

## Transformation and Polynomial Regression

Since the data involved historical information, the logarithmic transformation of the regressor variables was done, interaction terms were added and regressed against the response. The model improved slightly but the MAE for the models without transformation was better than for the transformed model.

## Results and Discussions

Finally, the idea of reusing the linear models built for the old stations seems to be better than building the models for the given deployment data. This is obvious because, the old models were obtained from training dataset which had data for two years, but the deployment data is just for one month. Also, rounding the values affects the MAE significantly. Initially, the MAE increased after rounding the values. But, at the later stages of model building, the MAE decreased with rounding the results. The best MAE obtained after rounding for the small test challenge is 2.502.

## References

[1] Data source: http://reframe-d2k.org/Challenge

[2] Textbook: "Introduction to Linear Regression Analysis"- Douglas C. Montgomery, Elizabeth A. Peck, Geoffrey Vining.

[3] Website: http://www.ats.ucla.edu/stat/r/dae/zipoisson.htm

[4] Website: http://www.ats.ucla.edu/stat/r/dae/zinbreg.htm