# G0O00a: Statistical Data Analysis
# Academic year 2019-2020: Project 2

Clément Cerovecki, Sebastiaan Höppner

This homework involves a regression analysis on the Bike Sharing Dataset. You are provided with a training set on which you have to fit two linear regression models, and a test set for which you have to make predictions. The training set contains the counts of rental bikes for both casual and registered users during 2011 with the corresponding weather information. The test set contains the same data during 2012.

The regressors are:

- `season`: 1:winter, 2:spring, 3:summer, 4:fall
- `holiday`: is 1 when the day is a holiday and 0 otherwise.
- `workingday`: is 1 when the day is neither a weekend nor a holiday, and 0 otherwise.
- `weathersituation`:
  1. Clear, Few clouds, Partly cloudy, Partly cloudy
  2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- `temperature`: Normalized temperature in Celsius (i.e., all values are transformed to [0,1]).
- `feelingtemperature`: Normalized feeling temperature in Celsius.
- `humidity`: Normalized humidity.
- `windspeed`: Normalized wind speed.

There are two response variables (which should not be used as predictor variables):

- `casual`: count of casual users
- `registered`: count of registered users

Each group of 1, 2 or 3 students chooses the zipped folder as determined according to the group division on Toledo. For example, group 1 will use the training and test sets (i.e. train.csv and test.csv) in "Group1.zip".

You answer the questions by carrying out an appropriate analysis with R. The presentation of the results, including figures, is made in a written report of maximum 6 pages (12pt font). Only report results and interpretations, do not recite theory from the course! Only include the relevant figures.

One single folder containing your report and R script should be uploaded on Toledo at the latest on **22 June 2020**. This project is graded on 5 points.

**Exercise: linear regression**

The goal is to construct two linear regression models based on the training set: one for predicting the number of casual users, one for predicting the number of registered users. The performance of each model is evaluated on the test set.

Construct good linear models by selecting and/or transforming variables while making use of the appropriate techniques. Be creative! For example, you could select different regressors and/or transformations depending on the response variable (i.e. `casual` or `registered`) that you try to predict. However, do not use PCR or ridge regression in this analysis.

**Hint**: (1) Remember, if you transform a variable in the training set, then you must apply the exact same transformation on the corresponding variable in the test set (using only statistics that are based on the training set). Otherwise, your models will fail to accurately predict the reponse variables of the test set. (2) Remember, you are not allowed to use any "information" based on the test set. The test set must be regarded as new, unseen data. For example, it is not allowed to standardize a variable based on the empirical mean and standard deviation of that variable in the test set. Instead, you must estimate and use the corresponding statistics of the training set.

Check if your models satisfy the Gauss-Markov conditions. If you do not succeed in meeting all these conditions for one or both of your models, then explain the shortcomings of your model(s).

Which insights do you gain from your models? Which variables have an influence on the count of casual users and the count of registered users?

Predict the number of casual and registered users based on the data in the test set. Report and discuss the performance of the models on the test set by using the following performance measures:

- Root mean squared error (RMSE)
- Mean absolute error (MAE)
- R-squared ($R^2$)
- Adjusted R-squared ($R^2_{adj}$)
- Pearson correlation between predicted response values and observered response values.

Again, be creative in the construction of your models! Try to obtain good performance results on the test set.

Good luck!