G0O01: Statistical models and data-analysis
Academic year 2016-2017: Project 2

This homework must be made **individually**. It consists of two parts, both of which involve an analysis of the cars dataset. This dataset contains the fuel consumption and emission data of cars, from 2000 to 2013. A description of the variables can be found on page 3. Note that the variables $manufacturer$, *model* and *description* should not be included in any of your models.

The first part involves a regression analysis on the cars dataset with $co2$ as response variable and (a subset of) the other variables as predictors. The second part consists of a classification analysis on the *euro_standard* of the cars.

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 12 pages (12pt font), excluding the appendix. The appendix must include the R-code used to carry out the analysis. Only report results and interpretations, do not recite theory from the course!

Deliver the report by email to Wannes Van den Bossche on **5 June 2017** at the latest (w.vandenbossche@kuleuven.be). This project is quoted on 8 points.

First step: create your personal dataset
Read in the dataset and execute the following R-code:

```
set.seed(0012345)
selVec <- c(sample(1:dim(X)[1],1000))
XTrain <- X[selVec,]
XTest <- X[-selVec,]
```

In this code, 0012345 should be replaced by your student number and `X` is the given data set. Now you have a unique training data set `XTrain` of 1000 observations and a test set `XTest` of 500 observations. Do not forget this step!

Exercise 1: linear regression

1. Perform an explorative analysis on your training data and remove observations that could have a strong influence on the outcome of your analysis.

2. Construct a good linear model (with $co2$ as response variariable) by selecting and/or transforming variables by making use of the appropriate techniques. Do not use PCR or ridge regession in this analysis. Check whether your model satisfies the Gauss-Markov conditions. If you do not succeed in meeting all these conditions explain the shortcomings of your model.

Exercise 2: classification

1. Execute the following R-code:

   ```
   XTrain[XTrain$euro_standard <= 4,"euro_standard"] = 0
   XTrain[XTrain$euro_standard >= 5,"euro_standard"] = 1
   XTest[XTest$euro_standard <= 4,"euro_standard"] = 0
   XTest[XTest$euro_standard >= 5,"euro_standard"] = 1
   ```

   This code turns the variable *euro_standard* into a binary factor which can be interpreted as old (*euro_standard*=0) and new (*euro_standard*=1) Euro standards.

2. Fit a logistic model that uses the variables (except *euro_standard*) in `XTrain` to predict *euro_standard* of the cars. Select variables with the appropriate techniques. Interpret the result.

3. What is the apparent error rate APER of your final model? And what is the error rate when the final model is applied to the test data `XTest`?

4. Carry out linear and quadratic discriminant analysis for the classes formed by *euro_standard* using the continuous regressors in your final logistic model. Compare their performance to that of the logistic model.

Good luck!

Variable	Description
manufacturer	Car manufacturer or importer.
model	Car model.
description	Further details on the car model.
euro_standard	Euro Standard to which the record applies.
transmission_type	Transmission type. Either Automatic or Manual.
engine_capacity	Engine capacity in cubic centimetres (cc).
fuel_type	Fuel type this car uses, e.g. Diesel, Petrol, Electric, Hybrid, etc.
urban_metric	Fuel comsumption in urban conditions in liters per 100 Kilometers (l/100 Km).
extra_urban_metric	Fuel comsumption in extra-urban conditions in liters per 100 Kilometers (l/100 Km).
combined_metric	Combined fuel comsumption: average of the urban and extra-urban tests, weighted by the distances covered in each part, in liters per 100 Kilometers (l/100 Km).
noise_level	External noise emitted by a car shown in decibels
co2	CO2 emissions in grammes per kilometre (g/km).
co_emissions	Carbon monoxide emissions in milligrammes per kilometre (mg/km).
nox_emissions	Nitrogen oxides emissions in milligrammes per kilometre (mg/km).