G0O02a: Statistical Data Analysis: Project 3

Prof. Mia Hubert

May 2022

The project must be made **individually**. It consists of the analysis of a cars data set. This data set contains the fuel consumption and emission data of cars from 2000 to 2013.

You answer the questions by performing an appropriate analysis with `R`. The discussion of the results and the necessary figures are reported in a written text that consists of a maximum of 5 pages (12pt font size). This file should be named "LastName_FirstName_Project3.pdf". Only report results and interpretations, do not repeat theory from the course! Include the figures when they are discussed, not at the end of the report. Additionally a separate file with the full `R` script should be provided. It should be named "LastName_FirstName_Project3.R".

Upload the documents on Toledo no later than **June 15, 23h**. This project is graded on 3.5 points.

The variables of the cars data set are:

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 centimeters (cc).
fuel_type	Fuel type this car uses, Diesel, Petrol or Hybrid.
urban_metric	Fuel consumption in urban conditions in liters per 100 Kilometers (l/100 Km).
extra_urban_metric	Fuel consumption in extra-urban conditions in liters per 100 Kilometers (l/100 Km).
combined_metric	Combined fuel consumption: 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	CO_2 emissions in grammes per kilometer (g/km).
co_emissions	Carbon monoxide emissions in milligrammes per kilometer (mg/km).
nox_emissions	Nitrogen oxides emissions in milligrammes per kilometer (mg/km).

Note that the variables `manufacturer`, `model` and `description` should not be included in any of your models.

Read in the data set and execute the following R-code:

```
cars_data <- read.table("Project3_data.txt", sep="", header=T)
cars_data$euro_standard <- as.factor(cars_data$euro_standard)
cars_data$transmission_type <- as.factor(cars_data$transmission_type)
cars_data$fuel_type <- as.factor(cars_data$fuel_type)
set.seed(0012345)
data_ind <- sample.int(n=nrow(cars_data), size=500, replace=F)
mydata <- cars_data[data_ind, ]
```

In this code, 0012345 should be replaced by your student number.

Consider **all** variables from `mydata` (except `manufacturer`, `model` and `description`). The goal is to predict the CO_2 emissions of cars from the other variables.

1. Consider the linear regression model containing all observed predictor variables. Do not transform the predictors nor the response. Interaction terms or higher order terms do not need to be included. Describe the shortcomings of this model, e.g.
   - Does it suffer from multicollinearity?
   - Does it violate the Gaussian-Markov conditions?
   - Are some predictors not significant?
2. Study whether a transformation of the response variable can improve the model.
3. Study whether a selection of predictor variables can improve the model.
4. If the variable `euro_standard` is still in your model, test whether all slopes corresponding to that variable are simultaneously equal to zero or not. If `euro_standard` is not in your model, perform the same test for `fuel_type`. If both predictors are not included, you can skip this question.
5. Compute the 95% confidence interval for β_1.
6. Compute the 99% prediction interval for a car with `euro_standard` = 4, `transmission_type` = "Manual", `engine_capacity` = 2196, `fuel_type` = "Petrol", `urban_metric` = 9.2, `extra_urban_metric` = 5.6, `combined_metric` = 6.9, `noise_level` = 72, `co_emissions` = 273.5, `nox_emissions` = 43.