Advanced examination

For this exam, you are expected to submit a pdf containing all your answers, numerical results, graphics. Your code should also be available in its original format (Rmd, R or other). Make sure your reasoning is clearly explained at all steps. All exercises have similar weights in your final mark. Your total work should not span over more than 25 pages.

Exercise 1:

Let us consider a multiple regression framework with p explanatory variables such that:

$$\mathbb{Y} = \mathbb{X}.\beta + \mathbb{U}$$

where $\mathbb{U}$ is the noise vector such that :

$$\mathbb{U} \sim \mathcal{N}(0, \sigma^2.I_n)$$

with I_n the identity matrix with n rows and columns and β is a column vector with $p+1$ columns.

We know that if $\mathbb{X}'\mathbb{X}$ is invertible, then the least square estimator for β is

$$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}.\mathbb{X}'\mathbb{Y}$$

But now, what happends if we add constraints on β?

1. Let us consider the constraints given by $R.\beta = r$ where R is a $q \times (p+1)$ matrix, $q < p+1$, q being also the rank of R.
   Prove that the solution of the least square problem under those constraints is:
   $$\hat{\beta}_c = \hat{\beta} + (\mathbb{X}'\mathbb{X})^{-1}.R'(R.(\mathbb{X}'\mathbb{X})^{-1}.R')^{-1}(r - R.\hat{\beta})$$
2. Let us consider the dataset Ozone.txt. We consider Y as being the concentration in Ozone (maxO3) and all the other numerical variables are the explanatory variables, except obs that should be deleted. Using the above results:
   (a) Determine the model involving all the explanatory variables.
   (b) Determine the model obtained with the constraint : $\beta_{T9} + \beta_{T12} + \beta_{T15} = 0$ where β_{T9} for instance represents the coefficient associated to the explanatory variable $T9$.
   (c) Compare the two models.

Exercise 2:

1. Consider the dataset data_advanced and construct different models to explain the response variable Y.
   Apply a method to determine which one of these models is the best one on this dataset. Try to explain what you obtain.
2. Do the same with the observations associated to the real-world data on PM10 pollution un Rouen area, observations that are available in the VSURF package.

Exercise 3: Write a function to perform cross-validation error to compare the linear model computed thanks to ordinary least-square and CART algorithm.

Exercise 4: Consider the article title 'MTGAUE' and focus on the independence test. Explain what is done, why, the difficulties and the advantages.