Advanced examination

For this exam, you are expected to upload a single pdf containing everything that you want the professor to read. Make sure your code is also available. Exercise 2 should under no circumstances span over more than 20 pages. Make sure that only informative outputs are printed.

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 $p+1$ rows.
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 happens if we add constraints on β?

1. Let us consider 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 criterion under those constrains 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 data file.

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.

(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:

In this exercise, you are expected to use what you learnt in ASML class to take into account the specificities of each one of those 2 datasets.

1. Consider the dataset data_advanced.RData . Construct different models to explain the response variable Y.
   Apply a method to determine which constructed model 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.