Subject

Exercise A We consider n observations $y_1, \ldots, y_n$ of a variable and n vectors x_i ($t(x_i) = (x_{i1}, \ldots, x_{ik})$) where we have the observations of k variables.
For $i \in \{1, \ldots, n\}$, we assume that y_i is an observation of Y_i with

$$Y_i \sim \mathcal{N}(t(x_i).\beta, \sigma_i^2)$$

where β is a vector of dimension k ($t(\beta) = (\beta_1, \ldots, \beta_k)$).
We assume that the variable Y_i are independent.
Let consider the following populations :

- $I_1 = \{1, \ldots, n_1\}$ indices associated to the first population with n_1 elements
- $I_2 = \{n_1 + 1, \ldots, n_1 + n_2\}$ indices associated to the second population with n_2 elements
- ....
- $I_p = \{n_1 + \ldots + n_{p-1} + 1, \ldots, n\}$ indices association to the population p with n_p elements

We assume that if $i \in I_j$, $\sigma_i^2 = j.\sigma^2$.

We want to estimate β and σ^2 thanks to maximum likelihood.

1. What is the expression of $f_{Y_i}(y_i)$, where f_{Y_i} denotes the density function of Y_i?
2. Show that $\hat{\beta}$ and $\hat{\sigma}^2$ are solution of:

$$\begin{cases} \sum_{l=1}^{p} \frac{1}{l} \sum_{i \in I_l} (y_i - t(x_i)\beta)^2 = n\sigma^2 \\ \forall j = 1, \ldots, k, \ \sum_{l=1}^{p} \frac{1}{l} \sum_{i \in I_l} (y_i - t(x_i)\beta).x_{ij} = 0 \end{cases}$$

3. Prove that the previous system is similar to:

$$\begin{cases} ||A(Y - X\beta)||^2 = n\sigma^2 \\ t(X)A^2(Y - X\beta) = 0 \end{cases}$$

   where Y and X are the classical matrices involved in linear model, and A is a diagonal matrix whose elements are $\frac{1}{\sqrt{l}}$ for i such that $i \in I_l$

4. We assume that $t(X)A^2X$ is invertible. Give the expression of $\hat{\beta}$ and $\hat{\sigma}^2$.
5. Prove that $n\hat{\sigma}^2 = ||V||^2$ where V is a centered gaussian vector.
6. Deduce that $\mathbb{E}(||V||^2)$ is the sum of the diagonal elements of the covariance matrix of V.
7. Prove that $\frac{n\hat{\sigma}^2}{n-k}$ is an unbiased estimator of σ^2.

8. Let denote by X_l the matrix with n_l rows and k columns composed of the rows of X associated to the indices of I_l. And Y_l is the vector with n_l elements associated to the elements of I_l for Y. Let denote $\hat{\beta}_l = (t(X_l).X_l)^{-1}.t(X_l).Y_l$. Prove that $\hat{\beta}_l$ is an unbiased estimator of β.

Exercise B Let consider the model

$$y_i = \beta_1 + \beta_2.x_i + \beta_3.\sqrt{x_i} + \varepsilon_i \quad for\ i \in \{1, \ldots, n\},$$

The variables ε_i are gaussian, independent with expectation 0 and variance σ^2.

1. Write the matricial equation with $t(\beta) = (\beta_1, \beta_2, \beta_3)$.
2. For a dataset, with $n = 100$, we find :

$$t(X).X = \begin{pmatrix} ? & ? & 222 \\ ? & 3767 & ? \\ ? & 1408 & 544 \end{pmatrix} \quad t(X).Y = \begin{pmatrix} -569 \\ -4505 \\ -1610 \end{pmatrix} \quad t(Y).Y = 651900$$

   What are the values of '¿

3. What is the mean of the x_i?
4. What are the estimations of β_1, β_2 and β_3?
5. What is the estimation of σ^2?
6. Compute a confidence interval for β_2 with level 95%.
7. Perform the test $\beta_2 = 0$ with respect $\beta_2 \neq 0$ with level 10%.
8. Compute the coefficient of determination.
9. Compute a confidence interval for y_{n+1} at level 95%, knowing $x_{n+1} = 49$.
10. Compute a confidence interval for y_{n+1} at level 95%, knowing $x_{n+1} = 25$.
11. Which one is the biggest one? Why?

Exercise C Let consider the files data1 and data2.
For data1, we want to explain the development while in data2, we want to explain the Grade.
By considering the nature of the variables, create several models and give the best one.