Msc Stat. for Smart Data
2022-2023
matthieu.marbac-lourdelle@ensai.fr

Parallel Computing
Lab Report

ENSAI
École nationale de la statistique et de l'analyse de l'information

---

Context

We want to implement in R, a procedure for selecting the variables in the logit model (without using the `glm` function implemented in R). The procedure for selecting the variables is a stepwise search (with the backward or forward direction) which optimizes the prediction error estimated by cross-validation (leave-one-out).

Instructions You can do this project by group of two students. All your scripts should be developed in a R project. All your experiments must be reproducible. You have to write a PDF report. For each task, your report should presents:

- The names of the scripts which are related to.
- Few sentences which explains the functions.
- One or two sentences of conclusion.
- The names of the html files produced by your code profiling (if relevant).

Your report and the ZIP of your project should be send on Moodle for October, 14th (one sending per group of students).

About the logit model

Let $(X_1^\top, Y_1), \dots, (X_n^\top, Y_n)$ be observed independent copies of the random vector

$$(X^\top, Y) \text{ with } X \in \mathbb{R}^d \text{ and } Y \in \{0, 1\}.$$

The distribution of Y given $X = x$ is assumed to be a logit model, such that

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{e^{x^\top \beta}}{1 + e^{x^\top \beta}} \text{ and } \mathbb{P}(Y = 0 \mid X = x) = 1 - \mathbb{P}(Y = 1 \mid X = x),$$

where $\beta \in \mathbb{R}^p$ is the vector of the model parameters.

For prediction model, the procedures of selection of variables reduce the variance of the estimators. When p is large, it is convenient to assume that only r coefficients of β are not zero (with $r << p$). This leads to select a subset of the covariates as relevant for the prediction of Y. We said that a variable is relevant to predict Y if its coefficient is not zero.

Tasks

1. In this part, you have to use basic instructions (like loops). Don't use any optimization techniques.
   (a) Implement the function `rlogit` which generate observations from a logit model.
   (b) Implement the function `basic.mle` which takes as input argument the subset of the relevant variables and the variable to predict. This function returns the maximum likelihood estimator. This estimator is given by a Newton-Raphson algorithm that you have to implement (or any other algorithm of optimization). Detail the formulas of your algorithm in the report.
   (c) Implement the function `basic.cv` which takes as input argument the sample and the set of the relevant variables. This function returns the estimator of the error of prediction obtained by cross-validation for any subset of covariates by using the MLE of the model.

(d) Implement the function `basic.modelcomparison` which takes as input argument the sample and a set of competing models. This function returns the best model (*i.e.,* the subset of the relevant variables) and its estimator of the prediction error.

(e) Implement the function `basic.modelselection` which takes as input argument the sample and the direction (backward or forward). This function returns the best model (*i.e.,* the subset of the relevant variables) and its estimator of the prediction error.

2. Manage the exceptions. Give, in your report, a list of the exceptions you have considered.

3. Code profiling and code improvement (without parallel computing). Explain which functions must be optimized and propose different versions of these functions. Finally, justify which solution should be considered.

4. Improve your code by considering parallel computing. You can consider that the user works on Linux. Justify your choice.

5. Illustrate the consistency of the procedure of model selection by a reproducible numerical experiment.

6. (Bonus) Illustrate the consistency of the procedure of model selection by a reproducible numerical experiment where the data generation is performed on different CPU. Discuss the case where this approach improve the code of Task 5.