Bocconi University - academic year 2023-2024

20592 - Probability and Statistics

(Computational Statistics: Filippo Ascolani)

Final project

Deadline: January 11, 2023.

1 The dataset

Consider the "SAheart" dataset used in the first practice session (you can find the txt file with the data in the folder dedicated to the final project). It consists of a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. In particular, you have:

1. A binary response variable "chd", which is equal to 1 if the individual suffered from a coronary heart disease.
2. Other 9 covariates of interest.

The goal is to assess the impact of the covariates (e.g. cholesterol) on the occurrence of a coronary heart disease. A more elaborate description is given in the txt file named "Dataset_explanation".

2 Probit regression

Probit regression is an alternative Generalized Linear Model for binary data. In particular, given p covariates and n observations, the model reads

- **Random component**: $Y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}(\mu_i)$, with $i = 1, \dots, n$.

- **Systematic component**: $\eta_i = X_i^\top \beta$.
- **Link function**: $\Phi^{-1}(\mu_i) = \eta_i$, where Φ is the cumulative distribution of a standard Gaussian distribution.

The likelihood can be written as:

$$L(\beta\,;\, Y, X) = \prod_{i=1}^{n} \left[\Phi(\eta_i)\right]^{Y_i} \left[1 - \Phi(\eta_i)\right]^{1-Y_i},$$

which is analytically intractable. Both in classical and Bayesian formulation of the model, computational approximations are needed to perform statistical inference.

3 Assignment

You have to:

1. Create a Python script which implements the Fisher scoring algorithm for Probit regression (see the next Section for details on the model).
2. Create a Python script which implements a Markov Chain Monte Carlo method (e.g. random walk Metropolis-Hastings) for a Bayesian probit regression.
3. Apply the two scripts to the "SAheart" dataset and write a report which discusses the findings.

In particular the report should include:

- A description of the two algorithms.
- The analytical derivation of all the quantities used in the algorithms (updating equations, acceptance probabilities, ecc.)
- Basic diagnostics for the convergence of the algorithms.
- A discussion of the prior distributions chosen in point 2.
- A final discussion on the fitted models (interpretation of the parameters, ecc.), in particular regarding the role of cholesterol.

The report should be around $3/5$ **pages.**

Remark 1: Python libraries for numerical computing (e.g. Numpy) are allowed, but the algorithms should be implemented from scratch.

Remark 2: the report should **not** include lines of code.

Remark 3: I should be able to run the scripts on my computer, without any additional information. Failure to reproduce the results on the report will lead to **zero points** awarded.