STAT 656: Bayesian Data Analysis
Fall 2023
Homework 4

Note: For each question, no credit will be given unless work is shown.

Gibbs sampler for the probit model

35 points

Consider N pairs of observations (x_i, y_i), where $x_i \in \Re^d$ and $y_i \in \{0, 1\}$. These are generated as follows:

$$z_i \sim N(x_i\beta, 1), \qquad y_i = \mathbb{1}(z_i > 0), \qquad i \in 1, \ldots, N.$$

This is thus the probit model of Lecture 16-17. Write X, Y, and Z for the collection of x_i, y_i and z_i's We will place a $N(0, \Sigma_b)$ prior on $\beta \in \Re^d$, and draw samples from the posterior $p(\beta, Z|X, Y)$. Observe that as in regression settings, we are really only interested in β, the Z's only auxiliary variables to allow us to implement a Gibbs sampler. In other words, this is an example of a *data-augmentation algorithm.*

1. (7 points) Derive and write down the conditional distribution $Z|X, Y, \beta$.
2. (7 points) Derive and write down the conditional distribution $p(\beta|X, Y, U)$?
3. (1 point) Describe the overall Gibbs sampling algorithm, and how you would calculate the posterior mean and covariance of β.

You do not need to implement the algorithm, but it is worth comparing these updates with the corresponding updates for the VB algorithm from the lecture slides (see next).

4. (20 points) Write an R function to implement the VB algorithm for the probit model (given in the slides of Lecture 16-17). This should accept as input a dataset (X, Y) and return the approximations to the posterior distributions over β and the z's. Recall how the algorithm proceeds: start with some arbitrary initial distribution $q(\beta)$ over β, use that to calculate the $q(z_i)$'s, use those to calculate $q(\beta)$, and repeat till these distributions stop changing.

Effect of computer-generated reminders

65 points

An experiment was conducted to study the effects of computer-generated reminders on preventive care. Physicians (specifically, residents and faculty members on the staff of a general medicine clinic) were randomly assigned to either a treatment or control group. The treatment in this experiment is a reminder sent to physicians that encouraged them to inoculate their patients at risk for the flu. The question of interest here is whether the computer-generated reminder increases the likelihood that a patient will be vaccinated.

Your task is to use a probit regression model to learn about the effect of the treatment on patient vaccination, accounting for patient characteristics that may be associated with the response. The observed data is contained in `flu_data.txt`, and consists of the following variables for each of the 2901 patients.

- `treatment`: An indicator that equals 1 if the patient's doctor received treatment, and 0 otherwise.
- `vaccinated`: An indicator that equals 1 if the patient received the vaccine, and 0 otherwise.

- `age`: The patient's age.
- `copd`: equals 1 if the patient has chronic obstructive pulmonary disease, and 0 otherwise.
- `heartd`: An indicator that equals 1 if the patient has heart disease, and 0 otherwise.
- `renal`: An indicator that equals 1 if the patient has kidney disease, and 0 otherwise.
- `liverd`: An indicator that equals 1 if the patient has liver disease, and 0 otherwise.

You can drop patients with missing covariates from your analysis.

2. (25 points) Apply your function from the previous question on the data provided in `flu_data.txt`. Describe how you initialized the algorithm, the number of iterations that you ran it for, and your selected convergence criterion. Plot the approximation to the posterior distribution over β as well as a few of the z_i's. Based on your posterior approximation, what do you conclude about the effect of the treatment on patient vaccination?

3. (20 points) Implement the probit model in Stan, and produce an MCMC approximation to the posterior. Treating these as the truth (recall MCMC becomes arbitrarily accurate with increasing number of samples), comment on your variational approximation deviates from the true posterior distribution.

3. (20 points) Implement the *logistic regression* model in Stan, and produce an MCMC approximation to the posterior. Comment on any difference between this as the results from the probit model.