Computational Statistics

Generalized Linear Mixed Model

Chenyin Gao

Dept. Statistics, Sun Yat-sen University

Contents

1 Progect Summary

1.1 Model Notation

In this project, we consider a clustering problem. Suppose we have observed n observations, each observation is a binary process, i.e. the response $Y_{ij} = 0$ or 1, $i = 1, \cdots, n$, $j = 1, \cdots, T$. Here n is the number of subjects and T is the length of observation. In general, T might vary across subjects, time points may also be different. In this project, however, we simply assume that all subjects have common time length and time points. We also assume that these subjects belong to two clusters. For each cluster, the conditional expectation of response variable is:

$$\begin{aligned} P_{ij} &= \mathbb{E}(Y_{ij}|U_i = 1, X_{1,ij}, Z_{1,i}) = g^{-1}(\beta_1 X_{1,ij} + Z_{1,i}) \\ P_{ij} &= \mathbb{E}(Y_{ij}|U_i = 2, X_{2,ij}, Z_{2,i}) = g^{-1}(\beta_2 X_{2,ij} + Z_{2,i}) \end{aligned} \tag{1}$$

where U is cluster membership, $X_{c,ij}$ and $Z_c, i(c = 1, 2)$ are fixed and random effects, respectively. The link function $g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$ is given. In a typical clustering problem, U is usually unknown, and hence we treat U as another random effect.

For random effects, we assume that $Z_{c,i} \sim N(0, \sigma_c^2)$ and $\mathbb{P}(U = 1) = \pi_1$ (then $\pi_2 = 1 - \pi_1$). Then the parameter to be estimated is $\Omega = \{\beta_1, \beta_2, \sigma_1, \sigma_2, \pi_1\}$. Treating random effects as missing data, one can write the complete data likelihood function as

$$L(\Omega|Y_{ij}, U_i, Z_{U_i,i}) = \prod_{i=1}^{n} \prod_{c=1}^{2} \{\pi_c f_c(Z_{c,i}) [\prod_{j=1}^{T} f_c(Y_{ij}|Z_{c,i})]\}^{w_{ic}} \tag{2}$$

where $f_c(Z_{c,i})$ is the density function of Normal distribution, $f_c(Y_{ij}|Z_{c,i}) = \mathbb{P}^{Y_{ij}}(1 - \mathbb{P}_{ij})^{1-Y_{ij}}$. w_{ic} is the dummy variable of U_i , i.e.

$$w_{ic} = \begin{cases} 1 & , \quad if\ subject \text{ i } belongs\ to\ cluster\ c \\ 0 & , \quad otherwise \end{cases}$$

1.2 Simulation Setup and Requirement

Generate 100 simulations. In each simulation, set n = 100 and T = 10. The true values of parameter are:$\beta_1 = 1, \beta_2 = 1, \pi_1 = 0.6, \sigma_1 = 2$ and $\sigma_2 = 10$

Use N(0,1) to generate the fixed effect X, and use them for all 100 simulations and use MCEM to evaluate the loglikelihood function. In the E-step, perform K = 500 Gibbs sampling incorporated with a Metropolis-Hastings step, and drop the first 100 as a burn-in procedure.