STAT 5060: Advanced Modeling and Data Analysis
Assignment 2

Academic year 23/24, first term

Due date: Nov 28, 2023

1. "score.csv" is the simulated data from Example 4.3. Please analyse the two-level model in Example 4.3 with the following two models and compare the results.

   (a). The model in page 23 (slides of chapter 4).

   (b). The model without random intercept and random slope.

   The dataset is also available in R as *MASS::nlschools* and *mlmRev::bdf* (for the variable "gender").

2. Conduct a simulation study for the following mixture model:

$$f(\mathbf{y}_i) = \pi_1 f_N(\mathbf{y}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 f_N(\mathbf{y}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \tag{1}$$

$$[\mathbf{y}_i | S_i = k] = \boldsymbol{\alpha}_k + \boldsymbol{\beta}_k \mathbf{x}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_k), \tag{2}$$

   where $f_N(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the multivariate normal distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, $k = 1, 2$; $i = 1, \cdots, 800$; $\mathbf{y}_i$ is a 3×1 observed vector, $\mathbf{x}_i = (x_{i1}, x_{i2})^T$ with $x_{i1} \sim N(0, 1)$, $x_{i2} \sim U(0, 1)$; $\boldsymbol{\alpha}_k$ is a 3×1 vector of intercept and $\boldsymbol{\beta}_k$ is a 3×2 matrix of coeffecients; $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\alpha}_1 = (1, -2, 1)^T$, $\boldsymbol{\alpha}_2 = (-2, 1, 2)^T$, $\boldsymbol{\beta}_1 = \begin{bmatrix} -1 & -1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\boldsymbol{\beta}_2 = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{bmatrix}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \text{diag}\,(0.2, 0.2, 0.2)$.

3. Consider a missing data model as follow:

$$y_{ij} = \mu + \boldsymbol{\beta}^T \mathbf{x}_{ij} + \mathbf{z}_{ij}^T \mathbf{u}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \tag{3}$$

$$\text{logit}(\pi_{ij}) = c + \alpha y_{ij} + \boldsymbol{\gamma}^T \mathbf{x}_{ij}, \quad \pi_{ij} = p(R_{ij} = 1), \tag{4}$$

   where $R_{ij} = 1$ denotes y_{ij} is missing, $i = 1, \cdots, n$, $j = 1, \cdots, m_i$.

   (a). Conduct a simulation study for model (3)-(4) with missing proportion being approximately 30%.

   (b). Fit the dataset generated in (a) with a MAR mechanism. Compare the results based on these two missing mechanisms.