STAT 5060: Advanced Modeling and Data Analysis
Assignment 1

Academic year 23/24, first term

Due date: Oct 24, 2023

1. Consider a GLM with count data as follows: for $i = 1, \cdots, n$,

$$y_i \sim Poisson(\mu_i), \quad \log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\beta} = (1, -1, 0.5, 1)^T$, $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3})^T$, $x_{i1} \sim U(0, 1)$, and $(x_{it2}, x_{it3}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (0, 0)^T$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$.

   (a) Generate data using the above setting with sample size $n = 400$.

   (b) Estimate $\boldsymbol{\beta}$ based on Model (1) and the generated data.

   (c) Repeat steps (a) and (b) for 10 times and calculate the Bias and RMS of the parameter estimates.

[Hint: (i) Bias of $\hat{\beta}$ is given by $(\frac{1}{S}\sum_{j=1}^{S} \hat{\beta}_j) - \beta_0$, where β_0 is the true value of β, $\hat{\beta}_j$ is the estimate of β at the jth replication, S is the number of replications; RMSE of $\hat{\beta}$ is given by $\left[\frac{1}{S}\sum_{j=1}^{S}(\hat{\beta}_j - \beta_0)^2\right]^{\frac{1}{2}}$. (ii) The R packages and the corresponding functions are marked in red. (iii) In this problem, use the stats package, via the glm.fit function]

2. Consider an extended GLM with nominal data as follows: for $i = 1, \cdots, n$,

$$\begin{aligned} & y_i \sim Categorical(\pi_{i1}, \cdots, \pi_{i4}), \quad \pi_{ij} = P(y_i = j), \\ & \log \frac{\pi_{ij}}{\pi_{i4}} = \mathbf{x}_i^T \boldsymbol{\beta}_j, \quad j = 1, 2, 3, \end{aligned} \tag{2}$$

where $\boldsymbol{\beta}_1 = (-1, 1, -1)^T$, $\boldsymbol{\beta}_2 = (-1, -1, 1)^T$, $\boldsymbol{\beta}_3 = (1, -1, 1)^T$, and $\mathbf{x}_i = (1, x_{i1}, x_{i2})^T$ with $x_{i1} \sim U(0, 1)$ and $x_{i2} \sim N(0, 1)$.

   (a) Generate data using the above setting with sample size $n = 800$.

   (b) Estimate $\boldsymbol{\beta}$ based on Model (2) and the generated data.

   (c) Repeat steps (a) and (b) for 10 times and calculate the Bias and RMS of the parameter estimates.

[Hint: consider the nnet package, via the multinom function]

3. Consider a GLM with longitudinal binary data as follows: for $i = 1, \cdots, n$, $t = 1, \cdots, T$,

$$y_{it} \sim Bernoulli(\pi_{it}), \quad \text{logit}(\pi_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta} + u_i, \tag{3}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients, $\mathbf{x}_{it} = (1, x_{it1}, x_{it2})^T$, $x_{it1} \sim Bernoulli(0.7)$, $x_{it2} \sim N(0, 1)$, u_i is a subject-specific random effect, and $u_i \sim N(0, \sigma^2)$. The true values of the parameters are $\boldsymbol{\beta} = (-0.7, 0.4, -0.5)^T$ and $\sigma^2 = 1$.

(a) Generate data using the above setting with $n = 800$ and $T = 4$.

(b) Estimate $\boldsymbol{\beta}$ and σ^2 based on Model (3) and the generated data.

(c) Repeat steps (a) and (b) for 10 times and calculate the Bias and RMSE of the parameter estimates.

4. Reanalyze Example 3.3 using cumulative logit models with and without the random intercept u_i and compare the results obtained from these two competing models.

[Hint: consider the ordinal package, via the clmm and clmm2 functions; the mixor package, via the mixor function; the MCMCglmm package, via family="ordinal"; the brms package, e.g. via family="cumulative"]