Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2569

Regression Models for Correlated Binary Data with Random Effects Assuming a Mixture of Normal Distributions

Jorge Alberto Achcar

Vanderly Janeiro

Nº 96

RELATÓRIOS TÉCNICOS DO ICMC

São Carlos

Outubro/1999

.

RESUMO

Neste relatório nós apresentamos uma análise Bayesiana de modelos de regressão logística para dados binários correlacionados com efeitos aleatórios assumindo uma mistura de distribuições normais. Este modelo da uma grande flexibilidade para ser ajustado a dados binários correlacionados em muitas aplicações. Considerando os algoritmos Gibbs sampling e Metropolis-Hastings, nós obtemos estimativas de Monte Carlo para as quantidades á posteriori de interesse.

Regression Models for Correlated Binary Data with Random Effects Assuming a Mixture of Normal Distributions

Jorge Alberto Achcar

Vanderly Janeiro

Universidade de São Paulo

ICMC - Caixa Postal 668

13560-970 - São Carlos - SP - Brazil

Abstract

In this paper we present a Bayesian analysis of logistic regression models for correlated binary data with random effects assuming a mixture of normal distributions. This model gives a great flexibility to be fitted by correlated binary data in many applications. Considering Gibbs sampling with Metropolis-Hastings algorithms, we obtain Monte Carlo estimates for the posterior quantities of interest.

1 Introduction

Consider two or more measurements taken at one time for the same subjects or when repeated measurements are taken over time, where we observe a binary (0-1) response y_{ij} on the ith observation and jth variable, $i = 1, \ldots, n$ and $j = 1, \ldots, J$. Associated to each response y_{ij} let $\underline{x}_{ij} = (x_{ij1}, \ldots, x_{ijp_j})'$ be the corresponding p_j-dimensional row regression vector. Let $\underline{y}_i = (y_{i1}, \ldots, y_{iJ})'$ and $\underline{y} = (\underline{y}_1, \ldots, \underline{y}_n)'$ be the observed data, where $y_{i1}, \ldots, y_{iJ}$ are dependent and $\underline{y}_1, \ldots, \underline{y}_n$ are independent.

Different models are proposed in the literature for modelling correlated binary data in the presence of covariates (see for example, Dey and Chen, 1996; Prentice, 1988; or Ochi and Prentice, 1984).

Prentice (1988), consider the random effect binary logistic regression model for y_{ij} given $\underline{x}_{ij}$,

$$P\{Y_{ij} = y_{ij}|\alpha_i, \underline{\beta}_j, \underline{x}_{ij}\} = p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} \tag{1}$$

where $p_{ij} = \frac{\exp\{\alpha_i + \underline{\beta}_j' \underline{x}_{ij}\}}{1+\exp\{\alpha_i + \underline{\beta}_j' \underline{x}_{ij}\}}$, $\underline{\beta}_j' = (\beta_{1j}, \ldots, \beta_{p_j j})$, $j = 1, \ldots, J$ are p_j-dimensional vector of regression coefficients. Observe that α_i denotes a random effect on the ith observation, which captures the correlation among $y_{i1}, \ldots, y_{iJ}$.

Dey and Chen (1996) assume,

$$\alpha_i \sim N(0, \sigma_\alpha^2) \tag{2}$$

where the random effects α_i are independent.

Other models are considered in the literature to analyse correlated binary data. Chib and Greenberg (1998) consider a multivariate probit model; a generalization of multivariate probit models is given by multivariate t-link models (see for example, Dey and Chen, 1996).

The use of classical methods based on the usual asymptotical approximations could involve very intensive computation and the accuracy of the obtained inferences could be not appropriate.

Albert and Jais (1998), introduce a Bayesian analysis for the logistic regression model (1) considering the use of the Gibbs sampler (see for example, Gelfand and Smith, 1990) to obtain the posterior quantities of interest. Day and Chen (1995) consider a hierarchical Bayesian analysis of model (1). They also introduce some model diagnostics using simulation based approach for model adequacy.

In this paper, we consider a random effect logistic regression model for correlated binary data assuming a mixture of normal distributions for the random effects α_i, given by,

$$\pi(\alpha_i) = \sum_{k=1}^{K} p_k \Phi_k(\alpha_i|\mu_k, \sigma_k^2) \tag{3}$$

where $\sum_{k=1}^{K} p_k = 1$ and Φ_k denotes a normal density $N(\mu_k, \sigma_k^2)$.

For a Bayesian analysis of this model, we consider the use of Markov Chain Monte Carlo (MCMC) methods to simulate samples of the joint posterior distribution for the parameters.

2 A Bayesian Analysis Assuming a Mixture of Normal Distributions for α_i

Let us assume that the random effects α_i, $i = 1, \ldots, n$ are independent with a mixture of $K = 2$ normal distributions (3), with $p_1 + p_2 = 1$.

From the logistic regression model (1), the likelihood function for $\underline{\alpha}, \underline{\beta}_1, \ldots, \underline{\beta}_J$ where $\underline{\alpha} = (\alpha_1, \ldots, \alpha_n)'$ is given by,

$$L(\underline{\alpha}, \underline{\beta}_1, \ldots, \underline{\beta}_J) = \prod_{i=1}^{n} \prod_{j=1}^{J} \frac{\exp\{(\alpha_i + \underline{\beta}_j' \underline{x}_{ij}) y_{ij}\}}{1 + \exp\{\alpha_i + \underline{\beta}_j' \underline{x}_{ij}\}} \tag{4}$$

Assuming prior independence among the parameters, consider the following prior densities for $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p_1, \beta_{lj}$, $l = 1, \ldots, p_j$; $j = 1, \ldots, J$:

$$\begin{aligned}
&\mu_k \sim N(d_k, e_k^2);\ d_k,\ e_k \ \text{ known},\ k = 1, 2; \\
&\sigma_k^2 \sim \mathcal{IG}(a_k, b_k);\ a_k,\ b_k;\ \text{known},\ k = 1, 2; \\
&p_1 \sim B(f, g);\ f,\ g;\ \text{known}; \\
&\beta_{lj} \sim N(b_{lj}, c_{lj}^2);\ b_{lj},\ c_{lj};\ \text{known},\ l = 1, \ldots, p_j; \quad j = 1, \ldots, J,
\end{aligned} \tag{5}$$

where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2; $\mathcal{IG}(a, b)$ denotes an inverse gamma distribution with mean $b/(a-1)$ and variance $b^2/[(a-1)^2(a-2)]$ and $B(f, g)$ denotes a beta distribution with mean $f/(f+g)$ and variance $fg/[(f+g)^2(f+g+1)]$.

The joint posterior for $\underline{\theta} = (\underline{\alpha}, \mu_1, \mu_2, p_1, \sigma_1^2, \sigma_2^2, \underline{\beta}_1, \ldots, \underline{\beta}_J)$ is given by,

$$\pi(\underline{\theta} | \underline{y}, \underline{x}) \propto \Psi(\underline{\theta}) \left\{ \prod_{i=1}^{n} \sum_{k=1}^{2} p_k \Phi_k(\alpha_i | \mu_k, \sigma_k^2) \right\} \tag{6}$$

where

$$\Psi(\underline{\theta}) = \left\{ \prod_{k=1}^{2} (\sigma_k^2)^{-(a_k+1)} \exp(-\frac{b_k}{\sigma_k^2}) \right\} \cdot \left\{ \prod_{k=1}^{2} \exp\left[-\frac{1}{2e_k^2} (\mu_k - d_k)^2 \right] \right\} \cdot$$

$$p_1^{f-1} (1-p_1)^{g-1} \cdot \left\{ \prod_{l=1}^{p_j} \prod_{j=1}^{J} \exp\left[-\frac{1}{2c_{lj}^2} (\beta_{lj} - b_{lj})^2 \right] \right\} \cdot$$

$$\frac{\exp\left\{ \sum_{i=1}^{n} \sum_{j=1}^{J} (\alpha_i + \underline{\beta}_j' \underline{x}_{ij}) y_{ij} \right\}}{\prod_{i=1}^{n} \prod_{j=1}^{J} \left\{ 1 + \exp(\alpha_i + \underline{\beta}_j' \underline{x}_{ij}) \right\}}.$$

To obtain better performance for the Gibbs sampling algorithm, we consider the introduction of latent variables (see for example, Tanner and Wong, 1987) given by $\underline{Z}_i = (Z_{i1}, Z_{i2}), \quad i = 1, \ldots, n$, where $Z_{i1} | \underline{\theta}, \underline{y}, \underline{x} \sim b(1, h_{i1})$ (a Bernoulli distribution) with h_{i1} given by

$$h_{i1} = \frac{p_1 \Phi_1(\alpha_i | \mu_1, \sigma_1^2)}{p_1 \Phi_1(\alpha_i | \mu_1, \sigma_1^2) + (1-p_1) \Phi_2(\alpha_i | \mu_2, \sigma_2^2)} \tag{7}$$

That is,

$$\pi(\underline{Z}_i) \propto h_{i1}^{Z_{i1}} (1 - h_{i1})^{Z_{i2}} \tag{8}$$

where $Z_{i1} = 1$ with probability h_{i1} ($Z_{i1} = 0$ with probability $1 - h_{i1}$). Observe that $Z_{i1} + Z_{i2} = 1$.

Thus, we have,

$$\pi(\underline{Z}_1, \ldots, \underline{Z}_n) \propto \frac{\prod_{i=1}^{n} \prod_{k=1}^{2} [p_k \Phi_k(\alpha_i | \mu_k, \sigma_k^2)]^{Z_{ik}}}{\prod_{i=1}^{n} \{ \sum_{k=1}^{2} p_k \Phi_k(\alpha_i | \mu_k, \sigma_k^2) \}} \tag{9}$$

Combining equation (9) with equation (6), we obtain,

$$\pi(\underline{Z}_1, \ldots, \underline{Z}_n, \underline{\theta} | \underline{y}, \underline{x}) \propto \Psi(\underline{\theta}) \left\{ \prod_{i=1}^{n} \prod_{k=1}^{2} [p_k \Phi_k(\alpha_i | \mu_k, \sigma_k^2)]^{Z_{ik}} \right\} \tag{10}$$

where $\Psi(\underline{\theta})$ is given in (6).

To generate samples of the joint posterior distribution (10) we use the Gibbs sampling algorithm. Starting with initial values $\underline{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$, we follow the following steps:

(i) Generate a sample $\underline{Z}^{(1)} = (\underline{Z}_1^{(1)}, \ldots, \underline{Z}_n^{(1)})$ from equation (8).

(ii) Generate a sample of $\underline{\theta}$ from the conditional distributions

$$\begin{aligned} &\pi(\theta_1|\theta_2^{(0)}, \ldots, \theta_p^{(0)}, \underline{Z}^{(1)}, \underline{x}, \underline{y}),\ \pi(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)}, \underline{Z}^{(1)}, \underline{x}, \underline{y}), \\ &\ldots, \pi(\theta_p|\theta_1^{(1)}, \ldots, \theta_{p-1}^{(1)}, \underline{Z}^{(1)}, \underline{x}, \underline{y}). \end{aligned} \tag{11}$$

Then, continue iterations by repeating steps (i) and (ii).

The conditional distributions for the Gibbs algorithm are given by,

(i)

$$\pi(\alpha_i|\underline{\beta}, \underline{\mu}, \underline{\sigma}^2, p_1, \underline{y}, \underline{x}, \underline{Z}) \sim N\left\{\frac{Z_{i1}\mu_1\sigma_2^2 + Z_{i2}\mu_2\sigma_1^2}{Z_{i1}\sigma_2^2 + Z_{i2}\sigma_1^2}, \frac{\sigma_1^2\sigma_2^2}{Z_{i1}\sigma_2^2 + Z_{i2}\sigma_1^2}\right\}\Psi_1(\underline{\alpha}, \underline{\beta}),$$

where

$$\Psi_1(\underline{\alpha}, \underline{\beta}) = \exp\left\{\alpha_i y_{1\cdot} - \sum_{j=1}^{J} \log[1 + \exp(\alpha_i + \underline{\beta}_j' \underline{x}_{ij})]\right\},$$

$$y_{i\cdot} = \sum_{j=1}^{J} y_{ij}\ , \quad i = 1, \ldots, n, \tag{12}$$

$$\underline{\mu} = (\mu_1, \mu_2); \quad \underline{\sigma} = (\sigma_1^2, \sigma_2^2)$$

(ii)

$$\pi(\sigma_k^2|\underline{\alpha}, \underline{\beta}, \underline{\mu}, \sigma_{(k)}^2, p_1, \underline{y}, \underline{x}, \underline{Z}) \sim \mathcal{IG}\left(a_k + \frac{v_k}{2}; b_k + \frac{\sum_{i=1}^{n} Z_{ik}(\alpha_i - \mu_k)^2}{2}\right)$$

where $v_k = \sum_{i=1}^{n} Z_{ik} \quad k = 1, 2; \quad \sigma_{(k)}^2 = \sigma_j^2, \quad j \neq k; \quad j, k = 1, 2,$

(iii)

$$\pi(p_1|\underline{\alpha}, \underline{\beta}, \underline{\mu}, \underline{\sigma}^2, \underline{y}, \underline{x}, \underline{Z}) \sim B(f + v_1, g + v_2),$$

(iv)

$$\pi(\beta_{lj}|\underline{\alpha},\underline{\beta}_{(lj)},\underline{\mu},\underline{\sigma}^2,p_1,\underline{y},\underline{x},\underline{Z}) \sim N(b_{lj},c_{lj}^2)\Psi_2(\underline{\alpha},\underline{\beta}),$$

where $\Psi_2(\underline{\alpha},\underline{\beta}) = \exp\left\{\beta_{lj}a_{jl} - \sum_{i=1}^n \log[1+\exp(\alpha_i + \underline{\beta}_j'\underline{x}_{ij})]\right\}$,

$a_{jl} = \sum_{i=1}^n x_{ijl}y_{ij}, \quad \underline{\beta}_{(lj)} = (\beta_{1j},\ldots,\beta_{l-1,j},\beta_{l+1,j},\ldots,\beta_{p_jj})$,

$l = 1,\ldots,p_j; \quad j = 1,\ldots,J,$

(v)

$$\pi(\mu_k|\underline{\alpha},\underline{\beta},\mu_{(k)},\underline{\sigma}^2,p_1,\underline{y},\underline{x},\underline{Z}) \sim N\left\{\frac{d_k\sigma_k^2 + e_k^2\sum_{i=1}^n \alpha_i Z_{ik}}{\sigma_k^2 + e_k^2\sum_{i=1}^n Z_{ik}}, \frac{e_k^2\sigma_k^2}{\sigma_k^2 + e_k^2\sum_{i=1}^n Z_{ik}}\right\}, \quad k = 1,2$$

where $\mu_{(k)} = \mu_j, \quad j \neq k, \quad j,k = 1,2.$

Observe that the variables α_i, and β_{lj}; $i = 1,\ldots,n$; $l = 1,\ldots,p_j$; $j = 1,\ldots,J$ should be generated using the Metropolis-Hastings algorithm (see for example, Smith and Roberts, 1993).

3 A Bayesian Analysis Assuming a Normal Distribution for α_i

If we assume that the random effects $\alpha_i, \quad i = 1,\ldots,n$ have a normal distribution $N(0,\sigma_\alpha^2)$ with prior distributions,

$$\begin{aligned} &\sigma_\alpha^2 \sim \mathcal{IG}(a,b);\ a,\ b \quad \text{known;} \\ &\beta_{lj} \sim N(b_{lj},c_{lj}^2);\ b_{lj},c_{lj}^2 \quad \text{known;} \quad l = 1,\ldots,p_j;\ j = 1,\ldots,J, \end{aligned} \tag{13}$$

and prior independence, the conditional distributions for the Gibbs sampling algorithm are given by,

(i)

$$\pi(\alpha_i|\sigma_\alpha^2,\underline{\beta},\underline{y},\underline{x}) \sim N(0,\sigma_\alpha^2)\Psi_3(\underline{\alpha},\underline{\beta}) \tag{14}$$

where

$\Psi_3(\underline{\alpha},\underline{\beta}) = \exp\left\{\alpha_i\sum_{j=1}^J y_{ij} - \sum_{j=1}^J \log[1+\exp(\alpha_i + \underline{\beta}_j'\underline{x}_{ij})]\right\}$,

(ii)

$$\pi(\sigma_\alpha^2|\underline{\alpha},\underline{\beta},\underline{y},\underline{x}) \sim \mathcal{IG}\left(\frac{n}{2}+a, b+\frac{\sum_{i=1}^n \alpha_i^2}{2}\right)$$

(iii)

$$\pi(\beta_{lj}|\underline{\alpha},\sigma_\alpha^2,\underline{\beta}_{(lj)},\underline{y},\underline{x}) \sim N(b_{lj},c_{lj}^2)\Psi_4(\underline{\alpha},\underline{\beta})$$

where

$\Psi_4(\underline{\alpha},\underline{\beta}) = \exp\left\{\beta_{lj}a_{jl} - \sum_{i=1}^n \log[1+\exp(\alpha_i+\underline{\beta}_j'\underline{x}_{ij})]\right\}$,

$a_{jl} = \sum_{i=1}^n x_{ijl}y_{ij}$,

$l = 1,\ldots,p_j; \quad j = 1,\ldots,J.$

Observe that we need to use the Metropolis-Hastings algorithm to generate the variables α_i and β_{lj}.

It is interesting to observe that if we consider α fixed with prior distributions,

$$\begin{aligned} \alpha &\sim N(\mu_0,\sigma_0^2);\ \mu_0,\ \sigma_0^2 \quad \text{known;} \\ \beta_{lj} &\sim N(b_{lj},c_{lj}^2);\ b_{lj},\ c_{lj}^2 \quad \text{known;} \end{aligned} \tag{15}$$

the conditional distributions for the Gibbs sampling algorithm are given by,

(i)

$$\pi(\alpha|\underline{\beta},\underline{y},\underline{x}) \sim N(\mu_0,\sigma_0^2)\Psi_5(\alpha,\underline{\beta}), \tag{16}$$

where

$\Psi_5(\alpha,\underline{\beta}) = \exp\left\{\alpha y_{..} - \sum_{i=1}^n\sum_{j=1}^J \log[1+\exp(\alpha+\sum_{l=1}^{p_j}\beta_{lj}x_{ijl})]\right\}$,

$y_{..} = \sum_{i=1}^n\sum_{j=1}^J y_{ij}$.

(ii)

$$\pi(\beta_{lj}|\underline{\beta}_{(lj)},\alpha,\underline{y},\underline{x}) \sim N(b_{lj},c_{lj}^2)\Psi_6(\alpha,\underline{\beta}_j),$$

where

$\Psi_6(\alpha,\underline{\beta}_j) = \exp\left\{a_{lj}\beta_{lj} - \sum_{i=1}^n \log[1+\exp(\alpha+\sum_{l=1}^{p_j}\beta_{lj}x_{ijl})]\right\}$

where a_{lj} and $\underline{\beta}_{(lj)}$ are defined in (12).

4 An Example

In table 1, we have the captures ($y_{ij} = 1$ for captured; $y_{ij} = 0$ for not captured) of peromyscus maniculatus collected by V. Reid at East Stuart Gulch, Colorado (data set introduced by Huggins, 1991). The columns represent the sex (m or f), the ages (y: young, sa: semi-adult, a: adult), the weights in grams, and the capture histories of 36 individuals over 6 trapping occasions.

For the data set of table 1, we assume the logistic regression model (1), that is,

$$P\{Y_{ij} = y_{ij} | \alpha_i, \underline{\beta}_1, \underline{\beta}_2, \underline{\beta}_3, x_{ij}\} = p_{ij}^{y_{ij}} (1 - p_{ij})^{1 - y_{ij}} \tag{17}$$

where

$$p_{ij} = \frac{\exp\{\alpha_i + x_{ij1}\beta_{1j} + x_{ij2}\beta_{2j} + x_{ij3}\beta_{3j}\}}{1 + \exp\{\alpha_i + x_{ij1}\beta_{1j} + x_{ij2}\beta_{2j} + x_{ij3}\beta_{3j}\}}$$

$i = 1, \ldots, 36$; $j = 1, \ldots, 6$; $\underline{\beta}_1 = (\beta_{11}, \ldots, \beta_{16})$; $\underline{\beta}_2 = (\beta_{21}, \ldots, \beta_{26})$; $\underline{\beta}_3 = (\beta_{31}, \ldots, \beta_{36})$; $x_{ij1} = 1(0)$ for m(f); $x_{ij2} = 1$(y), 2(sa) or 3(a); $x_{ij3} =$ weight $- 14$.

Table 1: Huggins Data

Captures of *peromyscus maniculatus*																	
m	y	12	1	1	1	1	1	1	f	y	10	1	0	0	1	0	0
f	y	15	1	0	0	1	1	1	f	a	23	0	1	0	0	1	0
m	y	15	1	1	0	0	1	1	f	y	7	0	1	1	0	0	1
m	y	15	1	1	0	1	1	1	m	y	8	0	1	0	0	0	1
m	y	13	1	1	1	1	1	1	m	a	19	0	1	0	1	0	1
m	y	13	0	1	1	0	1	0	f	a	22	0	0	1	1	1	1
f	y	5	0	1	0	1	0	1	f	y	10	0	0	1	0	1	1
m	a	20	0	1	0	0	0	1	f	y	14	0	0	1	1	1	1
m	y	12	0	1	0	0	1	1	f	a	19	0	0	1	0	0	0
f	y	6	0	0	1	0	0	0	f	a	20	0	0	0	1	0	0
m	a	21	1	1	0	1	1	1	m	sa	16	0	0	0	1	1	1
m	y	11	1	1	1	1	1	0	f	y	11	0	0	0	1	1	0
m	sa	15	1	1	1	0	0	1	m	y	14	0	0	0	0	1	0
m	y	14	1	1	1	1	1	1	f	y	11	0	0	0	0	1	0
m	y	13	1	1	0	1	1	1	m	a	24	0	0	0	0	1	0
f	a	22	1	1	1	0	1	1	m	y	9	0	0	0	0	0	1
m	y	14	1	1	1	1	1	1	m	sa	16	0	0	0	0	0	1
m	y	11	1	0	1	1	1	0	f	a	19	0	0	0	0	0	1

Assuming α fixed, maximum likelihood estimators for $\alpha, \underline{\beta}_1, \underline{\beta}_2$ and $\underline{\beta}_3$ are given by

$\widehat{\alpha} = 1.629,$

$\widehat{\underline{\beta}_1} = (1.3060, 0.9334, -0.6831, 0.0606, 1.3575, 0.4252),$

$\widehat{\underline{\beta}_2} = (-1.9722, -1.2288, -1.0135, -1.0931, -1.3414, -0.7094),$

$\widehat{\underline{\beta}_3} = (0.3256, 0.1863, 0.0876, 0.1560, 0.3142, 0.1114).$

Also assuming α fixed for a Bayesian analysis of model (17) with prior densities (15) and $\mu_0 = 1.6$, $\sigma_0 = 0.2$, $b_{11} = 1.3$, $b_{12} = 0.9$, $b_{13} = -0.6$, $b_{14} = 0.06$, $b_{15} = 1.3$, $b_{16} = 0.4$, $b_{21} = -1.9$, $b_{22} = -1.2$, $b_{23} = -1.0$, $b_{24} = -1.0$, $b_{25} = -1.3$, $b_{26} = -0.7$, $b_{31} = 0.3$, $b_{32} = 0.1$, $b_{33} = 0.08$, $b_{34} = 0.1$, $b_{35} = 0.3$, $b_{36} = 0.1$ and $c_{lj} = 0.2$, $l = 1, 2, 3$; $j = 1, 2, \ldots, 6$, we generate 5 separate Gibbs chains, each of which ran for 2000 iterations. We monitored the convergence of the Gibbs samples using the Gelman and Rubin (1992) method, which utilizes the analysis of variance technique to determine if further iterations are needed. For each parameter, we discarded the 500 first iterations ("burn-in-samples") and we considered the $10th, 20th, \ldots$ iterations.

Monte Carlo estimates for the posterior means for $\alpha, \underline{\beta}_1, \underline{\beta}_2$, and $\underline{\beta}_3$ aproximated from the generated Gibbs samples (see (16)) are given by,

$\widetilde{\alpha} = 1.6255,$

$\widetilde{\underline{\beta}_1} = (1.3177, 0.9018, -0.5956, 0.0629, 1.3072, 0.3999),$

$\widetilde{\underline{\beta}_2} = (-1.9060, -1.2081, -0.9963, -1.0933, -1.3101, -0.7078),$ and

$\widetilde{\underline{\beta}_3} = (0.3057, 0.1807, 0.0766, 0.1471, 0.3076, 0.1057).$

For a Bayesian analysis of the logistic regression model (17) assuming random effects $\alpha_i,\ i = 1, 2, \ldots, 36$ with a normal distribution $N(0, \sigma_\alpha^2)$ and prior distributions (13) with $a = 4,\ b = 1/3$ and the same values for the prior distributions (15) with α fixed, we also generated 5 separate Gibbs chains, each of which ran for 6000 iterations. Monte Carlo estimates for the posterior means of $\sigma_\alpha^2, \underline{\beta}_1, \underline{\beta}_2$ and $\underline{\beta}_3$ approximated from the generated Gibbs Samples (see (14)) are given by,

$\widetilde{\sigma_\alpha^2} = 0.3267,$

$\widetilde{\underline{\beta}_1} = (1.3715, 0.9660, -0.5516, 0.1321, 1.3635, 0.4574),$

$\widetilde{\underline{\beta}_2} = (-1.2067, -0.4186, -0.2174, -0.3053, -0.4614, 0.1390$, and

$\widetilde{\underline{\beta}_3} = (0.2272, 0.0785, -0.0480, 0.0550, 0.2052, 0.0056)$.

Assuming a mixture of $K = 2$ normal distributions (3) for the random effects α_i, $i = 1, 2, \ldots, 36$ and prior distribution (5) with $d_1 = -0.4$, $d_2 = 0.4$, $e_1^2 = e_2^2 = 0.2$, $a_1 = a_2 = 4$, $b_1 = b_2 = 1/3$, $f = g = 1$ and the same values for b_{lj} and c_{lj}^2, $l = 1, 2, 3$; $j = 1, 2, \ldots, 6$ considered for the prior distributions (13) and (15), we also generated 5 separated Gibbs chains, each of which ran for 6000 iterations.

In table 2, we have the posterior summaries obtained for the parameters.

Table 2: Posterior Summaries (mixture of two normal distributions for α_i)

Parameter	Mean	s.d.	95 % credible interval	$\widehat{\mathrm{R}}$
μ_1	-0.6478	0.3881	(-1.4215 ; 0.1078)	1.0287
μ_2	1.0367	0.2925	(0.5044 ; 1.6452)	1.0366
σ_1^2	0.1317	0.1172	(0.0380 ; 0.4470)	1.0147
σ_2^2	0.1317	0.1172	(0.0380 ; 0.4470)	1.0147
p_1	0.3586	0.1477	(0.0890 ; 0.6551)	1.0289
β_{11}	1.2975	0.1948	(0.9284 ; 1.6791)	1.0014
β_{12}	0.8999	0.1990	(0.5140 ; 1.2726)	1.0010
β_{13}	-0.6039	0.2042	(-1.0033 ; -0.2117)	1.0011
β_{14}	0.0577	0.1934	(-0.3100 ; 0.4333)	1.0007
β_{15}	1.2930	0.1964	(0.9037 ; 1.6887)	0.9999
β_{16}	0.3962	0.2001	(-0.0247 ; 0.7842)	1.0011
β_{21}	-1.2995	0.2061	(-1.7026 ; -0.8977)	1.0000
β_{22}	-0.5030	0.2001	(-0.8949 ; -0.1181)	1.0016
β_{23}	-0.3006	0.2015	(-0.7041 ; 0.0911)	1.0052
β_{24}	-0.3952	0.2032	(-0.7832 ; -0.0161)	1.0017
β_{25}	-0.5020	0.1972	(-0.8984 ; -0.0960)	1.0045
β_{26}	0.0599	0.1961	(-0.3277 ; 0.4368)	1.0045
β_{31}	0.3047	0.2019	(-0.0871 ; 0.7049)	1.0019
β_{32}	0.1725	0.2024	(-0.2293 ; 0.5590)	1.0022
β_{33}	-0.0824	0.2047	(-0.4873 ; 0.3255)	1.0033
β_{34}	0.1416	0.1938	(-0.2369 ; 0.5188)	1.0018
β_{35}	0.3078	0.1909	(-0.0688 ; 0.7014)	1.0024
β_{36}	0.1159	0.2024	(-0.2681 ; 0.5214)	1.0031

We also have in table 2, the estimated potential scale reductions $\widehat{\mathrm{R}}$ (see Gelman and Rubin, 1992) for all the parameters. In this case, the number of iterations considered was sufficient for approximate convergence ($\sqrt{\widehat{\mathrm{R}}} < 1.1$ for all the parameters).

In table 2, we observe from 95% credible intervals for the regression parameters β_{lj}, $l = 1, 2, 3$; $j = 1, 2, \ldots, 6$ that the covariates x_1 (sex) and x_2 (ages) present significative

effects on the probabilities of captures for most times $J = 1, 2, \ldots, 6$, but the covariate x_3 (weight) does not present a significative effect (the 95% credible intervals for β_{3j}, $j = 1, 2, \ldots, 6$ include zero).

Table 3: Pearson Residuals (mixture of two normal distributions for α_i)

J					
1	2	3	4	5	6
0.7292	0.5234	0.7775	0.7326	0.4920	0.4802
1.0291	-1.3546	-1.3195	0.7104	0.6896	0.5731
0.6363	0.5569	-0.8247	-1.2248	0.4274	0.5561
0.5220	0.4568	-1.0054	0.6698	0.3506	0.4562
0.6319	0.4845	0.8176	0.6887	0.4257	0.4572
-1.0433	0.7350	1.2402	-0.9571	0.6457	-1.4417
-0.2121	1.7467	-1.9948	1.4402	-0.3116	1.0218
-0.3549	1.2684	-0.4972	-0.8447	-1.1849	0.6461
-0.9246	0.7763	-0.8671	-0.9203	0.7298	0.7122
-0.1507	-0.3809	0.8559	-0.4549	-0.2218	-0.6329
0.7499	0.4399	-0.5949	0.6354	0.2248	0.2966
0.9131	0.6135	0.8023	0.8455	0.6171	-1.8279
1.0799	0.6346	1.2488	-1.1343	-2.0543	0.4783
0.5442	0.4458	0.8546	0.6436	0.3661	0.4328
0.6951	0.5330	-1.1118	0.7577	0.4683	0.5030
1.2183	0.6259	1.2808	-1.6596	0.3638	0.3374
0.5427	0.4446	0.8522	0.6419	0.3651	0.4316
1.0104	-1.4732	0.8877	0.9356	0.6828	-1.6519
2.8697	-0.6760	-1.2454	1.3177	-0.5160	-1.0031
-0.4908	1.1184	-0.3847	-0.9146	0.6075	-1.6124
-0.2847	1.4847	0.5498	-0.7920	-0.4197	0.9191
-0.4560	1.2084	-0.9275	-0.6289	-0.6716	0.9900
-0.6897	0.7453	-0.4531	1.0436	-2.2939	0.4748
-0.7434	-1.4469	1.4143	0.6653	0.4017	0.3726
-0.4361	-0.8460	0.6416	-0.9498	1.5486	0.7966
-0.8355	-1.2442	0.7264	0.7615	0.8033	0.6066
-0.2595	-0.6159	2.2669	-0.6701	-0.8650	-1.2437
-0.2682	-0.5959	-0.3757	1.5664	-0.8955	-1.1698
-0.8469	-1.3490	-0.6035	1.0459	0.5314	0.5748
-0.4004	-0.7270	-1.1790	1.2443	1.6842	-1.0486
-0.7195	-0.8784	-0.4582	-0.6084	0.9349	-0.9047
-0.2882	-0.5233	-0.8487	-0.5785	2.3399	-0.7548
-0.8582	-1.1996	-0.2142	-0.7930	0.3477	-1.6348
-0.4268	-0.7251	-0.7154	-0.5426	-0.6296	1.1622
-0.5324	-0.8479	-0.3793	-0.6010	-1.1828	0.9144
-0.2380	-0.5648	-0.4045	-0.6145	-0.7933	0.8768

In table 3, we have the Pearson residuals d_{ij} for the random effect logistic regression model with a mixture of two normal distributions for α_i to measure the discrepancy

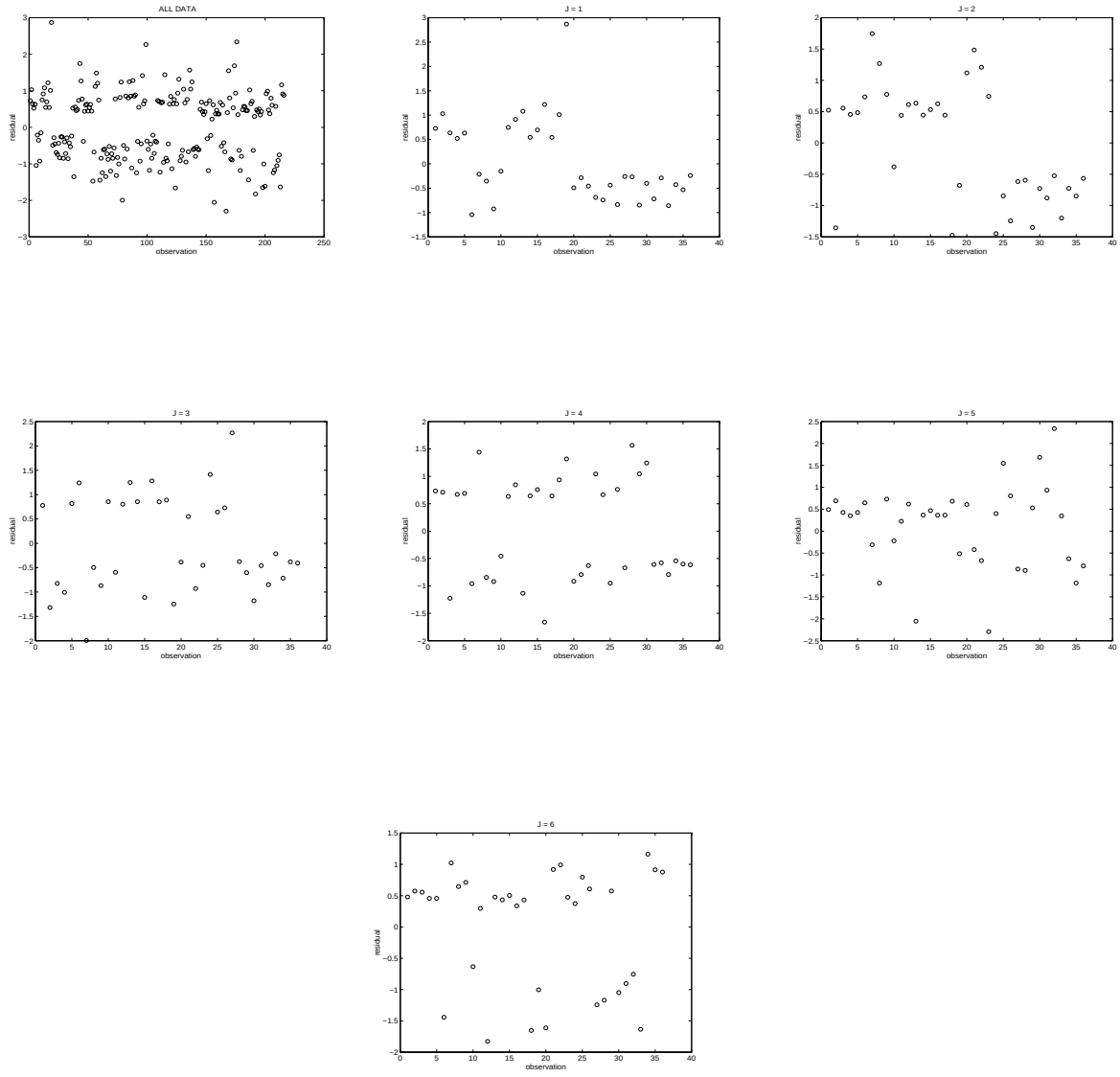
between the data and the model, which are given by

$$d_{ij} = \frac{y_{ij} - \widetilde{p}_{ij}}{\sqrt{\widetilde{p}_{ij}\widetilde{q}_{ij}}}, \tag{18}$$

where $\widetilde{p}_{ij}$ is given in (17) with the Bayes estimates for the parameters and $\widetilde{q}_{ij} = 1 - \widetilde{p}_{ij}$, $i = 1, \ldots, 36$; $j = 1, 2, \ldots, 6$.

In figure 1, we have the plots of the Pearson residuals d_{ij} against i for each value of $j = 1, 2, \ldots, 6$. We observe good fit of the random effect logistic regression model with a mixture of two normal distributions for α_i considering the data set of table 1.

Figure 1. Pearson Residuals (mixture of two normal distributions for α_i)



To check the overall performance for a model, we can consider the total Pearson residual discrepancy measure which is given by

$$D = \sum_{i=1}^{n} \sum_{j=1}^{J} \frac{(y_{ij} - \widetilde{p}_{ij})^2}{\widetilde{p}_{ij}\widetilde{q}_{ij}} \tag{19}$$

For the random effect logistic regression model with a mixture of two normal distributions for α_i, we have $D = 183.3444$. Considering a normal distribution $N(0, \sigma_\alpha^2)$ for the random effects α_i, we have $D = 185.7211$ and considering a fixed α, we have $D = 207.5838$. That is, we have better fit of the data set of table 1 for the random effect model with a mixture of two normal distribution for α_i, $i = 1, 2, \ldots, n$.

In table 3, we observe that the $19th$ observation have a large value for the Pearson residuals d_{ij} (an outlier), especially for $J = 1$. Taking out the $19th$ observation, we have $D = 175.0432$ if we assume random effects α_i with a mixture of two normal distributions. If we assume random effects α_i with a normal distribution, we have $D = 177.0421$ and considering a fixed effect α, we have $D = 198.2011$.

In table 4, we have a summary of the Monte Carlo estimates for the posterior means of β_{lj}, $l = 1, 2, 3$; $j = 1, 2, \ldots, 6$ based on the Gibbs samples and considering the different models.

Table 4: Posterior Means for β_{lj} ($l = 1, 2, 3$; $j = 1, 2, \ldots, 6$)

Parameter	MLE	α Fixed	$\alpha_i \sim N(0, \sigma_\alpha^2)$	$\alpha_i \sim$ Mixture of Normals
β_{11}	1.3060	1.3177	1.3715	1.2975
β_{12}	0.9334	0.9018	0.9660	0.8999
β_{13}	-0.6831	-0.5956	-0.5516	-0.6039
β_{14}	0.0606	0.0629	0.1321	0.0577
β_{15}	1.3575	1.3072	1.3635	1.2930
β_{16}	0.4252	0.3999	0.4574	0.3962
β_{21}	-1.9722	-1.9060	-1.2067	-1.2995
β_{22}	-1.2288	-1.2081	-0.4186	-0.5030
β_{23}	-1.0135	-0.9963	-0.2174	-0.3006
β_{24}	-1.0931	-1.0933	-0.3053	-0.3952
β_{25}	-1.3414	-1.3101	-0.4614	-0.5020
β_{26}	-0.7094	-0.7078	0.1390	0.0599
β_{31}	0.3256	0.3057	0.2272	0.3047
β_{32}	0.1863	0.1807	0.0785	0.1725
β_{33}	0.0876	0.0766	-0.0480	-0.0824
β_{34}	0.1560	0.1471	0.0550	0.1416
β_{35}	0.3142	0.3076	0.2052	0.3078
β_{36}	0.1114	0.1057	0.0056	0.1159

5 Some Conclusions

In many applications of correlated binary data considering logistic regression models, the usual assumption of normality for the random effects could not be appropriate. The use of a mixture of normal distributions for the random effects gives a great flexibility to analyse correlated binary data, as we observed with the data set table 1.

The use of MCMC methods for a Bayesian analysis of this model is a suitable way to get the posterior summaries of interest.

References

[1] ALBERT, I.; JAIS, J.P. (1998). Gibbs Sampler For The Logistic Model in the Analysis of Longitudinal Binary Data. *Statistics in Medicine*, n.17, p.2905-2921.

[2] CHIB, S.; GREENBERG, E. (1998). Analysis of Multivariate Probit Models. *Biometrika*, n.85, p.347-361.

[3] DEY, D.K.; CHEN, M.-H (1996). Bayesian Analysis of Correlated Binary Data Models. *Technical Report, Dep. of Statistics, University of Connecticut, U.S.A.*

[4] GELFAND, A.E.; SMITH, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, n.85, p.398-409.

[5] GELMAN, A.; RUBIN, B.D. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, n.4, p.457-511.

[6] HUGGINS, R.M. (1991). Some Practical Aspects of a Conditional Likelihood Approach to Capture Experiments. *Biometrics*, n.47, p.725-732.

[7] OCHI, Y. e PRENTICE, R.L. (1984). Likelihood Inference in a Correlated Probit Regression Model.*Biometrics*, n.71, p.531-543.

[8] PRENTICE, R.L. (1988). Correlated Binary Regression with Covariate Specific to each Binary Observation. *Biometrics*, n.44, p.1033-1048.

[9] SMITH, A.F.M.; ROBERT, G. O. (1993). Bayesian Computation via the Gibbs Sampler and Related MCMC Methods. *Journal of the Royal Statistical Society*, series B, n.55, p.3-24.

[10] TANNER, M.; WONG, W. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, n.82, p.528-550.