A Bayesian binary regression model with exponential power link

Martín, Jacinto
E-mail: jrmartin@unex.es

Naranjo, Lizbeth
E-mail: lizbeth@unex.es

Pérez, Carlos Javier
E-mail: carper@unex.es

Universidad de Extremadura, Departamento de Matemáticas
Avenida de Elvas s/n
06006 Badajoz, Spain

Suppose that n independent binary random variables $Y_1, \ldots, Y_n$ are observed, where Y_i is distributed as Bernoulli with success probability $\mathrm{p}(Y_i = 1|\boldsymbol{\beta}) = \Psi(\mathbf{x}_i^T\boldsymbol{\beta})$. $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters, $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ik})$ is a vector of known covariates, and Ψ is a known nonnegative function ranging between 0 and 1. The standard approach to model the dependence of binary data on explanatory variables under the generalized linear models setting is performed through a cumulative density function (cdf) Ψ. For instance, the probit model is obtained when Ψ is the standard normal cdf and the logit model when Ψ is the logistic cdf (see, for example, McCullagh and Nelder, 1989). With the upsurge of MCMC methods in the nineties, the new simulation tools allowed to efficiently apply Bayesian regression methods.

A Bayesian approach of a binary regression model is proposed. The inverse of the exponential power cdf is used as the link function. The exponential power (EP) family (Box and Tiao, 1973) includes the normal distribution and incorporates additional shapes, including platykurtic and leptokurtic ones. This means that distributions with both lighter and heavier tails compared to the normal one can be achieved. These distributions allow the modeling of kurtosis, providing, in general, more flexible fits to experimental data than the normal distribution. The proposed approach contains, among others, the probit model and an approximation to the logistic model as special cases.

The approach's implementation is based on two main ideas. The first one is to develop a data augmentation framework by introducing latent variables in a similar way as in Albert and Chib (1993). The second one is using the mixture representation for the EP distribution suggested by Walker and Gutiérrez-Peña (1999). These two ideas are exploited to derive efficient Gibbs sampling algorithms for both informative and non informative settings. In contrast to the work of Haro-López *et al.* (2000), the full conditional distributions can be directly generated from by using Gibbs sampling, so no Metropolis-Hasting method needs to be applied.

We introduce n independent latent variables $Z_1, \ldots, Z_n$, where Z_i is distributed as exponential power $\mathrm{EP}(\mathbf{x}_i^T\boldsymbol{\beta}, 1, \theta)$ and define $Y_i = 1$ if $Z_i > 0$, and $Y_i = 0$ if $Z_i \leq 0$. By using random variables U_i, the distribution of Z_i is written as a mixture:

$$\begin{aligned} Z_i|\mathbf{y}, \boldsymbol{\beta}, \sigma, \theta, u_i &\sim \begin{cases} \mathrm{U}\left(x_i^T\boldsymbol{\beta} - \sigma u_i^{\theta/2}, x_i^T\boldsymbol{\beta} + \sigma u_i^{\theta/2}\right) I\left[z_i > 0\right] & \text{if } y_i = 1 \\ \mathrm{U}\left(x_i^T\boldsymbol{\beta} - \sigma u_i^{\theta/2}, x_i^T\boldsymbol{\beta} + \sigma u_i^{\theta/2}\right) I\left[z_i \leq 0\right] & \text{if } y_i = 0 \end{cases} \\ U_i|\theta &\sim \mathrm{Ga}\left(1 + \theta/2, 1/2\right), \end{aligned}$$

where $I[\cdot]$ denotes the indicator function.

The following step is defining the prior distributions. We consider $\boldsymbol{\beta} \sim \mathcal{F}$ and $\theta \sim \mathrm{U}(0, 2]$, where $\mathcal{F}$ is the multivariate normal distribution family $\mathrm{N}_k(\mathbf{b}, \mathbf{B})$. The parameter θ is allowed to vary over its

range, providing flexibility to fit different shapes. Also, an improper prior distribution like $\pi(\boldsymbol{\beta}) \propto 1$ is considered.

The full conditional distributions of $\mathbf{Z}$, $\mathbf{U}$, $\boldsymbol{\beta}$ and θ are derived:

- The full conditional distributions of $Z_1, \ldots, Z_n$ are independent, with

$$Z_i|\mathbf{y}, \mathbf{u}, \boldsymbol{\beta}, \theta \sim \begin{cases} \mathrm{U}\left(\max\left\{0, x_i^T\boldsymbol{\beta} - \sigma u_i^{\theta/2}\right\}, \max\left\{0, x_i^T\boldsymbol{\beta} + \sigma u_i^{\theta/2}\right\}\right) & \text{if } y_i = 1 \\ \mathrm{U}\left(\min\left\{0, x_i^T\boldsymbol{\beta} - \sigma u_i^{\theta/2}\right\}, \min\left\{0, x_i^T\boldsymbol{\beta} + \sigma u_i^{\theta/2}\right\}\right) & \text{if } y_i = 0 \end{cases}.$$

- The full conditional distributions of $U_1, \ldots, U_n$ are independent, with

$$U_i|\mathbf{y}, \mathbf{z}, \boldsymbol{\beta}, \theta \sim \mathrm{Exp}\left(\tfrac{1}{2}\right) I\left[u_i > \left|\frac{z_i - x_i^T\boldsymbol{\beta}}{\sigma}\right|^{2/\theta}\right].$$

- We denote $\boldsymbol{\beta}^T = (\beta_j, \boldsymbol{\beta}_{(-j)}^T)$, $\boldsymbol{\beta}_{(-j)}^T = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k)$. Since the distribution of $\boldsymbol{\beta}$ is a multivariate normal distribution, then the conditional distribution of β_j given $\boldsymbol{\beta}_{(-j)}$ is a normal distribution. Then, the posterior distribution of β_j conditioned on $\mathbf{Z}$, $\mathbf{U}$, $\boldsymbol{\beta}_{(-j)}$ and θ is given by

$$\beta_j|\mathbf{y}, \mathbf{z}, \mathbf{u}, \boldsymbol{\beta}_{(-j)}, \theta \sim \mathrm{N}(\mathrm{b}_j^*, \mathrm{B}_j^*) I\left[\beta_j \in \left(\underline{\beta}_j, \overline{\beta}_j\right)\right]$$

for $j = 1, \ldots, k$, where

$$\begin{aligned}
\underline{\beta}_j &= \max_i \left\{ \frac{z_i - \mathbf{x}_{i(-j)}^T\boldsymbol{\beta}_{(-j)} - \sigma u_i^{\theta/2}}{x_{ij}} \right\}, \\
\overline{\beta}_j &= \min_i \left\{ \frac{z_i - \mathbf{x}_{i(-j)}^T\boldsymbol{\beta}_{(-j)} + \sigma u_i^{\theta/2}}{x_{ij}} \right\}, \\
\mathrm{b}_j^* &= \mathrm{b}_j - \mathrm{B}_{j(-j)}\mathrm{B}_{(-j)(-j)}^{-1}\mathrm{b}_{(-j)}, \\
\mathrm{B}_j^* &= \mathrm{B}_{jj} - \mathrm{B}_{j(-j)}\mathrm{B}_{(-j)(-j)}^{-1}\mathrm{B}_{(-j)j}, \\
\mathbf{x}_{i(-j)}^T &= (x_{i1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{ik}), \\
\mathbf{b}^T &= (\mathrm{b}_j, \mathrm{b}_{(-j)}), \\
\mathbf{B}^T &= \begin{pmatrix} \mathrm{B}_{jj} & \mathrm{B}_{j(-j)} \\ \mathrm{B}_{(-j)j} & \mathrm{B}_{(-j)(-j)} \end{pmatrix}.
\end{aligned}$$

If $\pi(\boldsymbol{\beta}) \propto 1$ is considered, then

$$\beta_j|\mathbf{y}, \mathbf{z}, \mathbf{u}, \boldsymbol{\beta}_{(-j)}, \theta \sim \mathrm{U}\left(\underline{\beta}_j, \overline{\beta}_j\right).$$

- The posterior density of θ given $\mathbf{Z}$, $\mathbf{U}$ and $\boldsymbol{\beta}$ is given by

$$\mathrm{p}(\theta|\mathbf{y}, \mathbf{z}, \mathbf{u}, \boldsymbol{\beta}) \propto \frac{1}{\Gamma(1 + \theta/2)^n 2^{n\theta/2}} I\left[\theta \in (\underline{\theta}, \overline{\theta}]\right]$$

where

$$\underline{\theta} = \max\left\{0, \max_{i \in \Theta^+} \frac{2\log(|z_i - x_i^T\boldsymbol{\beta}|/\sigma)}{\log(u_i)}\right\}, \Theta^+ = \{i : \log(u_i) > 0\},$$

$$\overline{\theta} = \min\left\{2, \min_{i \in \Theta^-} \frac{2\log(|z_i - x_i^T\boldsymbol{\beta}|/\sigma)}{\log(u_i)}\right\}, \Theta^- = \{i : \log(u_i) < 0\}.$$

The following example shows the applicability of the proposed approach and its high performance when comparing with other competing models.

Adult respiratory distress syndrome

Adult respiratory distress syndrome (ARDS) is a complication in many critically ill patients. The usual diagnosis of ARDS is based on clinical findings of refractory respiratory failure and x-rays of the lungs showing fluid accumulation. Rocker *et al.* (1988) used lung images obtained after labeling the plasma protein transferrin in patients meeting the clinical criteria for ARDS and in patients who did not, to calculate a lung protein accumulation index (P) that is larger as there is more protein in the lungs. They also recorded other characteristics of these patients, including their sex (S), age (A), x-ray lung fluid score (R), and amount of oxygen in their blood (O). Figure 1 shows the data ($n = 44$ patients), the points drawn with light gray color correspond to the patients with absence of ARDS ($y_i = 0$), and the points drawn with dark gray color correspond to the patients with presence of ARDS ($y_i = 1$).

ARDS data

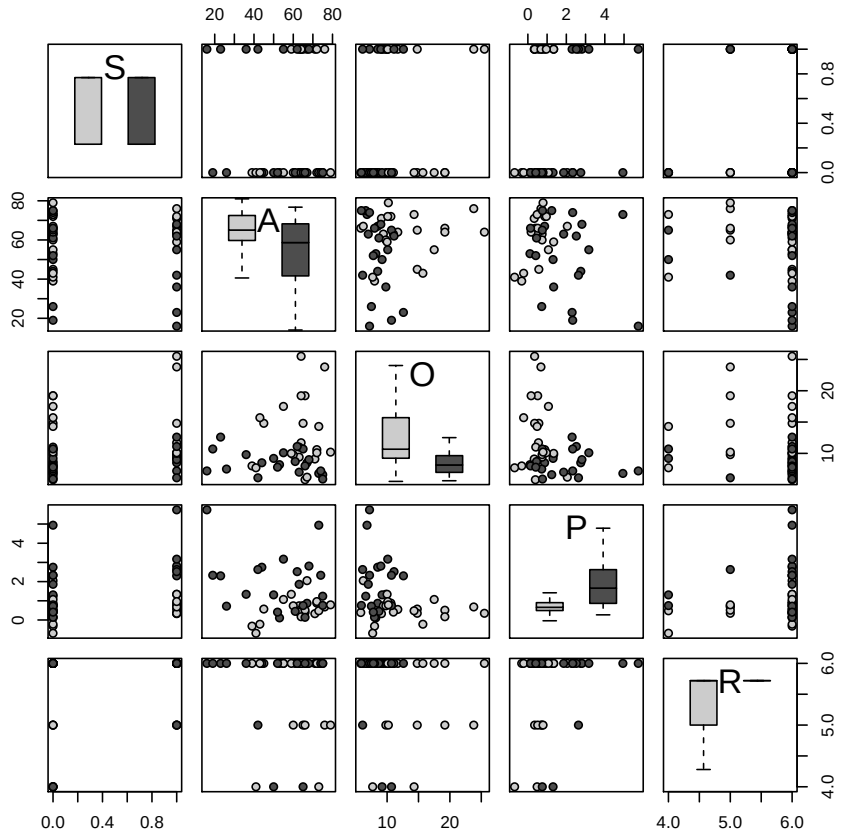

Figure 1: ARDS data: the points drawn with light gray color correspond to the patients with absence of ARDS, and the points drawn with dark gray color correspond to the patients with presence of ARDS.

After analyzing the most influential covariates, simplified models are considered:

$$\Psi^{-1}(p_i) = \beta_0 + \beta_1 A_i + \beta_2 O_i + \beta_3 P_i,$$

with $i = 1, \ldots, 44$.

In a non informative setting, we generated a total of 110,000 iterations of MCMC for each model. Then, it is considered a burn-in of 10,000 and saving one out of 10 observations. DIC is used to compare the models (see Table 1).

Estimated DICs

Model	DIC	$\overline{D}$	$\widehat{\rho_D}$
Normal	39.250	35.850	3.400
EP-link ($\theta = 1$)	39.746	36.156	3.590
t-link ($\nu = 8$)	40.210	36.255	3.955
EP-link ($\theta = 1.35162$)	38.991	35.703	3.288
EP-link (θ r. v.)	38.239	35.323	2.916

Table 1: Estimated DICs for models fitted to ARDS data.

The results obtained with normal-link and EP-link ($\theta = 1$) are very close. The differences are due to simulations, since they represent the same probit model. An interesting result is that the logit model with the EP-link approximation is better than the one obtained with the t(8)-link. Finally, note that the EP-link that considers θ as a random variable is the best model with the best fit ($\overline{D} = 35.323$) and lowest number of effective parameters ($\widehat{\rho_D} = 2.916$).

Now, we illustrate the application of the informative approach for the EP-link (θ r.v) model. A multivariate normal prior distribution for $\boldsymbol{\beta}$ is used, $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{b}, \mathbf{B})$, which is also the most commonly used prior for regression parameters. In order to specify $\mathbf{b}$ and $\mathbf{B}$, we utilize the empirical Bayesian approach (see Carlin and Louis, 1996), leading to

$$\mathbf{b} = \begin{pmatrix} 6.99 \\ -0.06 \\ -0.56 \\ 2.02 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 10.813 & -0.085 & -0.520 & 0.356 \\ -0.085 & 1.001 & 0.001 & -0.008 \\ -0.520 & 0.001 & 1.053 & -0.048 \\ 0.357 & -0.008 & -0.048 & 1.736 \end{pmatrix}.$$

Based on the prior obtained, we derive the posterior distributions. With the same MCMC specification as in the noninformative setting, the chain seems to have converged.

We used this generated sample to evaluate the posterior distribution. The simulated parameters are summarized in Table 2, whereas Figure 2 displays the posterior distributions with the 90% and 95% HDP intervals.

Summary of the posterior estimates

Variable	Mean	Median	Standard deviation	95% HDP interval
β_0	7.6003	7.6142	2.4501	$(2.6912, 12.2177)$
β_1 (A)	-0.0572	-0.0555	0.0334	$(-0.1224, 0.0072)$
β_2 (O)	-0.6358	-0.6164	0.2441	$(-1.1235, -0.1814)$
β_3 (P)	2.0583	2.0161	0.7441	$(0.6192, 3.4684)$
θ	1.7120	1.7720	0.2494	$(1.2500, 2.000)$

Table 2: Summary of the posterior estimates for the parameters of the EP-link model.

The coefficient of protein (P) is positive, indicating that high lung protein is associated with a high probability of having ARDS. Both age (A) and oxygen (O) have negative effects. Thus, higher

Posterior distributions and HDP intervals

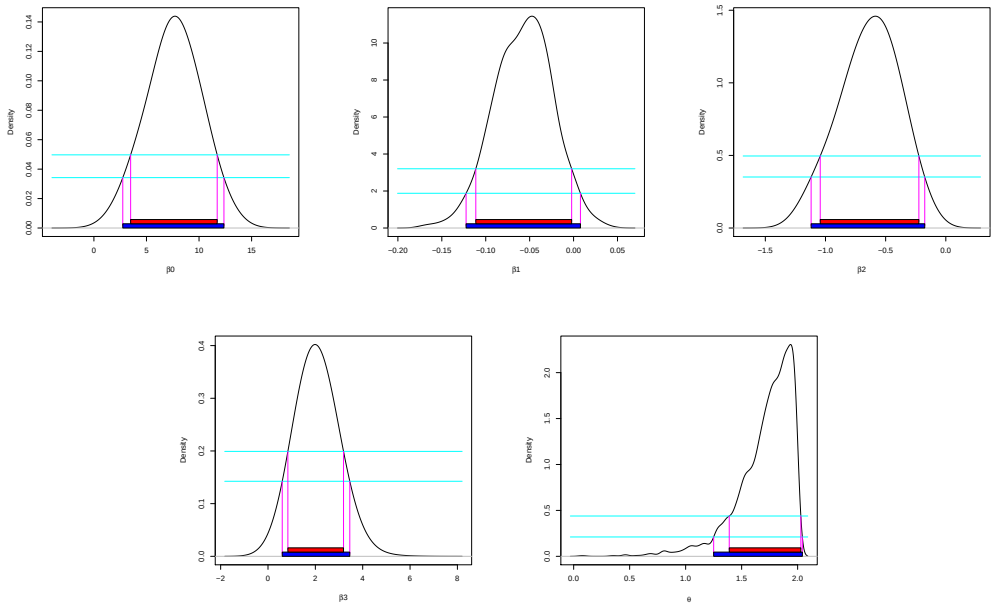

Figure 2: Estimated posterior distributions and 90% and 95% HDP intervals for the parameters of the EP-link model.

oxygen is associated with lower probability of ARDS, which makes sense because ARDS is clinically defined in terms of low oxygenation. Similarly, higher age is associated with a lower probability of ARDS. This is justified because in this sample the ARDS patients tend to be younger.

A very interesting result has been obtained for the shape parameter θ, since its 95% HDP interval is $(1.25, 2)$. The flexibility of the EP-link model with θ as a random variable has allowed that this model chooses the values of θ for the best fit. Note that this range of values (that give leptokurtic cdfs) is far from $\theta = 1$, that corresponds to the probit model.

REFERENCES

Albert, J., Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88, 669-679.

Box, G. E. P., Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, Massachusetts.

Carlin, B. P., Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall, London.

Haro-López, R., Mallick, B. K., Smith, A. F. M. (2000). Binary regression using adaptative robust link functions. In Dey, D. K., Ghosh, S. K., Mallick, B. K. (Eds.), *Generalized Linear Models: A Bayesian Perspective.* Marcel Dekker.

McCullagh, P., Nelder, J. A. (1989). *Generalized Linear Models.* Chapman & Hall, Monographs on Statistics and Applied Probability 37.

Rocker, G. M., Pearson, D., Stephens, M. Shale, D. J. (1988) An assessment of a double-isotope method for the detection of transferrin accumulation in the lungs of patients with widespread pulmonary infiltrates. *Clinical Science* 75, 45-52.

Walker, S. G., Gutiérrez-Peña, E. (1999). Robustifying Bayesian procedures. In Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. (Eds.), *Bayesian Statistics 6.* Oxford University Press, 685-710.

ABSTRACT

A flexible Bayesian approach of a generalized linear model is proposed to describe the dependence of binary data on explanatory variables. The inverse exponential power cumulative distribution function is used as the link of the binary regression model. The power exponential family provides distributions with both lighter and heavier tails compared to the normal one, and includes the normal and an approximation to the logistic distribution as particular cases. The idea of using a data augmentation framework and a mixture representation of the power exponential distribution is exploited to derive efficient Gibbs sampling algorithms for both informative and non informative settings. An example illustrate the performance of the proposed approach when comparing with other competing models.