A general skew-probit link for binary response

Jorge Luis Bazán, IME-USP, jbazan@ime.usp.br

Heleno Bolfarine, IME-USP, hbolfar@ime.usp.br

Marcia D´. Branco, IME-USP, mbranco@ime.usp.br

ABSTRACT: We introduce a general skew-probit link for the modelling of binary response which can be appropriate when the probability of a given binary response approaches 0 at a different rate than it approaches 1. As particular cases we obtain the skew-probit link due a Chen et al. (1999), the skew-probit link due a Bazán et al. (2004) and the usual probit link. Properties of the this asymmetric link are investigated. A Bayesian MCMC inference approach is developed and a comparisons of the fit for several models of regression with link different are studied.

Palavras-chave: link skew-probit, binary response, bayesian estimation, comparisons of models.

1. INTRODUCTION

Traditionality in the modelling of the binary data symmetrical link such as probit and logit are used. However, as Chen et al. (1999) mentions, when the probability of a given binary response approaches 0 at a different rate than it approaches 1, symmetric links are inappropriate. This paper is devoted to the development of a parametric family of asymmetric links. Let $y = (y_1, y_2, \ldots, y_n)'$ denote an $n \times 1$ vector de n independent dichotomous random variables. Also let $x_i = (x_{i1}, \ldots, x_{in})'$ be a $k \times 1$ vector of covariates, where x_{i1} may be 1, which corresponds to an intercept, X denotes the $n \times k$ design matrix with rows x_i', and $\beta = (\beta_1, \ldots, \beta_k)'$ is a $k \times 1$ vector of regression coefficients. Assume that $y_i = 1$ with probability p_i and $y_i = 0$ with probability $1 - p_i$. In a traditional modelling for binary data, it is usually assumed that

$$p_i = F(\eta_i(\beta)) = F(x_i'\beta), \quad i = 1, \ldots, n \tag{1}$$

where $F(.)$ denotes a cumulative distribution function (cdf), F^{-1} is typically called a link function and $\eta_i(\beta) = x_i'\beta$ is ith linear predictor.

When F is a cdf of a symmetric distribution, the resulting link is symmetric and the response curve for p_i has a symmetric form about $p_i = 0.5$. Examples are obtained when F is a cdf elliptical distributions such as Normal, Logistic, Student-T, Double Exponential and Cauchy distributions (see Albert & Chib, 1993). Consequently, when a cdf of an asymmetric distribution is considered, the response curve for p_i has not a symmetric form about $p_i = 0.5$. Thus, asymmetric links are obtained, for example, when asymmetric distribution such as the Gumbel (that define the loglog link) and Weibull distributions are considered. However, when this ap-

proach is used, it is not possible to model asymmetry by using a parameter and no relationship is established between symmetric and asymmetric links. Also, families of asymmetrical links can be obtained by considering a modification of the linear predictor η_i for $h(\eta_i, \lambda)$ where $h(.)$ is a continue function, λ is an asymmetry parameter and considering F the cdf of a symmetric distribution. It is the case of the Burr, Logistic and Normal distributions (see for example Czado and Munk, 2000). Another more general family of links are obtained when F is chosen on the class of elliptical scale mixtures cdf (see Basu and Mukhopadhyay, 2000).

In this communication a new family of parametric skew probit links is proposed. In this link: a) p_i is obtained by considering the cdf of a distribution function evaluated at the linear predictor, b) the parameter of asymmetry is associated with the distribution of $F(.)$ and is independent of the linear predictor, and c) the augmented likelihood function is not needed for model formulation. The skew link proposed is more general than the probit link, and introduces a parameter that controls the rate of increasing (or decreasing) of the probability of success (failure) of he binary response. For the building of the new link, the cumulative distribution function of the skew normal distribution (Dalla Valle, 2004) is considered. As particular case we obtain the skew probit link proposed by Chen et al. (1999), denoted by CDS Skew probit, and the skew probit link due Bazan et al. (2004), denoted by BBB Skew probit.

2. METHODOLOGY

To introduce the class of skew probit model, we consider

$$p_i = \Phi_{SN}(x_i'\beta \; ; \; \mu, \sigma^2, \lambda), \quad i = 1, \ldots, n \tag{2}$$

where $\Phi_{SN}(\; . \; ; \mu, \sigma^2, \lambda)$ denote the cdf of the skew-normal distribution with asymmetrical parameter $\lambda \in \mathcal{R}$, location μ and scale σ^2. Note that if $\mu = 0$, $\sigma^2 = 1$ and $\lambda = 0$ the probit model follows. Also, if $\mu = 0$, $\sigma^2 = 1 + \lambda^2$ and $-\lambda$ is the asymmetrical parameter, the CDS skew probit follows. If $\mu = 0$, $\sigma^2 = 1$ and λ is the asymmetrical parameter, the BBB skew probit follows. Let $D_{obs} = (n, y, X)$ denote the observed data. Then, the likelihood function for the class of skew-probit models is given by

$$L(\beta, \lambda | D_{obs}) = \prod_{i=1}^{n} [\Phi_{SN}(\, x_i'\beta \, ; \, \mu, \sigma^2, \lambda)]^{y_i} [1 - \Phi_{SN}(x_i'\beta \, ; \, \mu, \sigma^2, \lambda)]^{1-y_i}. \tag{3}$$

To implement a Bayesian estimation procedure we incorporate general class of independent prior distributions for $\boldsymbol{\beta}$ and λ.

$$\pi(\boldsymbol{\beta}, \lambda) = \pi_1(\boldsymbol{\beta})\pi_2(\lambda). \tag{4}$$

For $\pi_1(.)$ can be considered common priors in the probit model including improper priors. For $\pi_2(.)$, as in Bazán et al (2004), we use a reparametrization of the skew-probit model in terms of $\delta = \frac{\lambda}{(1+\lambda^2)^{1/2}}$, which take values in the interval $(-1, 1)$, so that we can consider that $\delta \sim U(-1, 1)$.

3. RESULTS AND DISCUSSION

Using the likelihood in (3) and the prior distribution in (4) we obtain posterior distributions. However, such an approach is complicated. An approach based on data augmentation as considered in Albert and Chib (1993) can be used. It is easily shown that the skew probit link is equivalent to considering that

$$y_i = I(Z_i > 0) = \begin{cases} 1, & Z_i > 0; \\ 0, & Z_i \leq 0, \end{cases} \tag{5}$$

where $Z_i \sim SN(\mu + x_i'\beta, \sigma^2, -\lambda)$, $i = 1, \ldots, n$, and $I(.)$ is indicator function. Then, the *complete-data likelihood function* for the skew probit model with $D = (\boldsymbol{Z}^T, y^T)^T$ is given by

$$L(Z, \beta, \lambda | D) = \prod_{i=1}^{n} \phi_{SN}(Z_i; \mu + x_i'\beta, \sigma^2, -\lambda) I(Z_i, y_i), \tag{6}$$

where $I(Z_i, y_i) = I(Z_i > 0)I(y_i = 1) + I(Z_i \leq 0)I(y_i = 0)$, $i = 1, \ldots, n$. An alternative data augmentation approach to the skew probit model follows by considering that

$$Z_i = \mu + x_i'\beta + e_i, \qquad e_i \sim SN(\mu, \sigma^2, -\lambda), \quad i = 1, \ldots, n \tag{7}$$

Note that the error e_i in the structure introduced are latent residuals i.i.d and can be used for model checking. In addition, using the stochastical representation for the skew normal distribution (Henze, 1986) we can write

$$e_i = \sigma(-\delta V_i - (1 - \delta^2) W_i) + \mu, \quad i = 1, \ldots, n \tag{8}$$

where $V_i \sim HN(0, 1)$, the half normal distribution, and $W_i \sim N(0, 1)$, the standard normal distribution. It follows that the conditional distribution $e_i | V_i = v_i$ is a normal distribution with mean $\mu - \sigma\delta v_i$ and variance $(1 - \delta^2)\sigma^2$. According to this result, simulation of Z_i in the linear structure (6) can be performed in two steps. By considering this result and latent structure in (8) a hierarchical formulation of the model is given as follow:

$$Z_i | v_i, y_i, \beta, \delta \sim N(x_i'\beta + \mu - \sigma\delta v_i, (1 - \delta^2)\sigma^2) I(z_i, y_i)$$

$$V_i \sim HN(0, 1)$$

$$\beta \sim \pi_1(.)$$

$$\delta \sim U(-1, 1)$$

This hierarchical structures can be easily introduced in WinBugs software. Note that all of the full conditional distributions for Gibbs sampling are straightforward to derive and sample from as in Albert and Chib (1993). In general for the CDS and BBB skew probit $\mu = 0$ and σ^2 is known.

We illustrate the Bayesian approaches developed in this paper for the application of the proposed skew-probit model using a Beetle Mortality data. These well-known data were also analyzed for Czado (1994) concluding that a asymmetric link is more convenient and improves significantly the fit for logit and probit regression. In order for illustrated the use of the model proposed we used CDS and BBB skew-probit for fit a binary regression for the data. Also are considered cloglog and probit links. To compare the models, we computed the deviance information criterion (DIC), described in Spiegelhalter et al. (2002), the Expected Akaike Information Criterion (EAIC) and the Expected Bayesian Information (Schwarz) Criterion (EBIC) proposed in Brooks (2002) and the sum-of-square of latent residuals (SSLR) (in the case of the cloglog link is not applied). For a detailed description these criterions see Bazán et al. (2004).

Table 1: Comparison of the models for Beetle Mortality data

models	*DIC*	*EBIC*	*EAIC*	*SSLR*
probit	375.4	391.9	379.4	482.2
cloglog	368.7	385.2	372.7	
skew probit CDS	272.2	184.1	167.4	481.4
skew probit BBB	258.0	317.4	300.7	249.6

Note, in the Table 1, that clearly the two skew-probit models are more convenient for the data set analyzed and are more convenient than the asymmetric cloglog link. Note also that BBB skew-probit is better when considering DIC and SSLR and that CDS skew-probit is better when EBIC and EAIC are used. We consider that the BBB skew probit is more adequate since that EBIC and EAIC are penalized with a penalization depending on the data and DIC is penalized by the effective number of parameters (Spiegelhalter et al., 2002) and SSLR shows the discrepancy between the observations and the posterior estimations.

Applications of the skew-probit proposed in item response model can revised in Bazán et al (2004) and in regression can be seen in Chen et al (1999). Binary data are source, for example to binomial models, epidemiological studies, multilevel modelling, longitudinal data analysis, meta-analysis and item response theory. When a symmetric link is not adequate in this cases, the skew-probit proposed can be easily implemented for this models.

4. CONCLUSIONS

1. The general skew-probit link introduced has a particular cases the skew-probit link to due a Chen et al (1999), the skew-probit link to due a Bazán et al (2004) and the probit link. 2. The skew- probit link introduces a parameter for asymmetry of response curves that is

easily interpreted and define a class of asymmetric links. 3. Implementations of the approach can be easily obtained by considering the versions of the augmented likelihood proposed. 4. Applications in several areas where symmetric links are not justified can be obtained with the proposed model.

5. REFERENCES

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

Basu, S. and Mukhopadhyay, S. (2000). Binary Response Regression With Normal Scale Mixtures Links, in *Generalized Linear Models: A Bayesian Perspective*,eds. D.K. Dey, S.K. Ghosh, and B.K: Mallick, New York: Marcel Dekker.

Bazán, J., Bolfarine, H., and Branco, D´. M. (2004). A new family of asymmetric models for item response theory: A SKEW-NORMAL IRT FAMILY. 49p. *Technical report* (RT-MAE-2004-17). Department of Statistics. University of São Paulo.

Brooks, S. P. (2002). Discussion on the paper by Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). *Journal Royal Statistical Society, Series B*, **64**,3, 616-618.

Chen, M-H, Dey, D. K., and Shao, Q-M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, **94**, 448, 1172-1186.

Czado, C. (1994). Bayesian Inference of Binary Regression Models With Parametric Link. *Journal of Statistical Planning and Inference*, **41**, 121-140.

Czado, C. and Munk, A. (2000). Noncanonical links in generalized linear models -when is the effort justified?. *Journal of Statistical Planning and Inference*, **87**, 317-345.

Dalla Valle, A. (2004). The skew-normal distribution, in *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Genton, M. G., Ed., Chapman & Hall / CRC, Boca Raton, FL, pp. 3-24.

Henze, N. (1986). A probabilistic representation of the "skew-normal" distribution. *Scandinavian Journal Statistical*. **13**, 271-275.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal Royal Statistical Society, Series B*, **64**,3, 583-639.