A Flexible Skewed Link Function for Binary Response Data

Xia Wang and Dipak K. Dey

Technical Report #2008-5
June 18, 2008

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

A Flexible Skewed Link Function for Binary Response Data

By Xia Wang and Dipak K. Dey

Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, Connecticut 06269, U.S.A.

xia.wang@uconn.edu dipak.dey@uconn.edu

Summary

The objective of this paper is to introduce a flexible skewed link function for modeling binary data with covariates. The commonly used complementary log-log link is prone to link misspecification because of its positive and fixed skewness parameter. We propose a new link function based on the generalized extreme value (GEV) distribution. The generalized extreme value link is flexible in fitting the skewness in the data with the unknown shape parameter. Using Bayesian methodology, it automatically detects the skewness in the data along with the model fitting. The propriety of posterior distributions under proper and improper priors is explored in details.

Key Words: Complementary log-log link; Generalized extreme value distribution; Latent variable; Posterior distribution

1. Introduction

For binary response data, the commonly used link functions include logit, probit, and complementary log-log. However, these popular links do not always provide the best fit for a given data set. Chen et al. (1999) used the rates at which the probability of a given binary response approaches 0 and 1 to describe a link. By their notation, the logit and the probit links are symmetric link functions since the response function $p(x)$ approaches 0 at the same rate as it approaches 1. The complementary log-log link is positively skewed with the response curve approaching 0 fairly slowly but approaching 1 quite sharply. Logit and probit models are inappropriate when the symmetry is badly violated in a given data. Usually, the complementary log-log link function is used for modeling such asymmetric data. However, the skewness of the complementary log-log link is a positive constant. Then, it may not be an appropriate model if the data is asymmetric the other way. Even when the direction of the skewness is correctly explored beforehand, the constant skewness of the complementary log-log link has still lack of flexibility in model fitting for varying skewness in the data.

Stukel (1988) proposed a class of generalized logistic models for modeling binary data. Stukel's models are very general, and several important and commonly used symmetric and asymmetric link models can be approximated by members of this family. However, in the presence of covariates, Stukel's models yield improper posterior distributions for many types of noninformative improper priors, including the improper uniform prior for the regression coefficients. Using a latent variable approach of Albert & Chib (1993), Chen et al.(1999) proposed another class of skewed links, which can lead to proper posterior distributions for the regression parameters using standard improper priors. However, the model has the limitation that the intercept term is confounded with the skewness parameter. This problem was overcome in Kim et al. (2008) by a class of generalized skewed t-link models.

The constraint on the shape parameter δ as $0 < \delta < 1$, which determines skewness, requires a positively or negatively skewed distribution to be correctly specified before model fitting. This brings the possibility of model misspecification again.

To build an appropriate and extremely flexible model for the binary data, we propose the generalized extreme value distribution as a link function. The advantage of the generalized extreme value link model is that the shape parameter is decided by the model fitting process without any presumption of the direction of skewness in the data. In fact, the complementary log-log link, based on the Gumbel distribution as discussed in §2, is a special case of the generalized extreme value link.

2. Generalized Extreme Value Link Model

Let $y = (y_1, y_2, \cdots, y_n)'$ denote an $n \times 1$ vector of n independent binary random variables. Also, let $x_i = (x_{i1}, x_{i2}, \cdots, x_{ik})'$ be a $k \times 1$ vector of covariates. Suppose X denotes the $n \times k$ design matrix with rows x_i', and $\beta = (\beta_1, \beta_2, \cdots, \beta_k)'$ is a $k \times 1$ vector of regression coefficients. Assume that $y_i = 1$ with probability p_i and $y_i = 0$ with probability $1 - p_i$. Let $w = (w_1, w_2, \cdots, w_n)'$ be a vector of independent latent variables. Then by the latent variable approach of Albert & Chib (1993), the binary response model can be set up as

$$Y_i = \begin{cases} 1 & w_i > 0 \\ 0 & w_i \leq 0, \end{cases} \tag{1}$$

where $w_i = x_i\beta + u_i$, for $i = 1, \cdots, n$, $u_i \sim F$, and F is a cumulatative distribuiton function which decides the link function. Here we propose to use the generalized extreme value distribution for F.

The generalized extreme value distribution is a family of continuous probability distribution developed within the extreme value theory to combine the Gumbel, Fréchet and Weibull families. It has a cumulative distribution function as follows:

$$G(x) = \exp\left[-\left\{1 + \xi\frac{(x-\mu)}{\sigma}\right\}_+^{-\frac{1}{\xi}}\right], \tag{2}$$

where $\mu \in R$ is the location parameter, $\sigma > 0$ is the scale parameter, $\xi \in R$ is the shape parameter and $x_+ = \max(x, 0)$. A more detailed discussion on the extreme value distributions can be found in Coles (2001). Its importance as a link function arises from the fact that the shape parameter ξ in model (2) purely controls the tail behavior of the distribution. Based on the definition of the skewness using the approaching rates (Chen et al., 1999), we find that the generalized extreme value distribution is negatively skewed for $\xi < \log 2 - 1$, and positively skewed for $\xi > \log 2 - 1$. When $\xi = 0$, it gives the Gumbel distribution with $G(x) = \exp\{-\exp(-x)\}$, which is the least positively skewed distribution when ξ is non-negative.

In model (1), assume $u_i \sim GEV(\mu = 0,\ \sigma = 1,\ \xi)$. Then

$$p_i = p(y_i = 1) = 1 - \exp\left\{(1 - \xi x_i \beta)_+^{-\frac{1}{\xi}}\right\} = 1 - GEV(-x_i'\beta;\ \xi), \tag{3}$$

where $GEV(x; \xi)$ represents the cumulative probability at x for the generalized extreme value distribution with $\mu = 0$, $\sigma = 1$, and an unknown shape parameter ξ.

As discussed above, the generalized extreme value link model specified in (1) and (3) is negatively skewed for $\xi < \log 2 - 1$, and positively skewed for $\xi > \log 2 - 1$. The class of the generalized extreme value links also includes the symmetric link as a special case.

For example, by matching the first 3 moments, the probit link can be approximated by the generalized extreme value link with $\mu \approx -0.35579, \sigma \approx 0.99903$, and $\xi \approx -0.27760$. Fig.1 shows the quantile plots between the generalized extreme value model and the probit model. The plot is approximately a straight line between 0.02 and 0.98 quantiles. The discrepancy lies mainly in the tail area.

[Figure 1 here]

Theoretically, ξ in model (1) and (3) can be of any value between $-\infty$ and ∞. However, it is reasonable to assume $-1 \leq \xi < 1$ for the Bayesian computation. Positively large ξ's result in improper posteriors under many improper priors for β, including the uniform priors. Negatively large ξ is rarely seen in practice (Coles, 2001). Our empirical study suggests that the constraint $-1 \leq \xi < 1$ is important not only for the propriety of the posterior estimation, but also for the model identifiability.

3. Prior and Posterior Distributions for the Generalized Extreme Value Link Model

Let $D_{obs} = (n, y, X)$ denote the observed data. We assume $\pi(\xi)$ is proper, and $\pi(\xi) = 0.5$, for $-1 \leq \xi < 1$, which corresponds to the uniform distribution $U[-1, 1)$. Then the joint posterior distribution of (β, ξ) based on D_{obs} is given by

$$\pi(\beta, \xi | D_{obs}) \propto p(y|X, \beta, \xi)\pi(\beta|\, \xi), \tag{4}$$

where $p(y|X, \beta, \xi) = \prod_{i=1}^{n} \{1 - GEV(-x_i\beta; \xi)\}^{y_i} \{GEV(-x_i\beta; \xi)\}^{1-y_i}$.

3.1. Jeffreys' Prior

The Jeffreys' prior for this model has the form $\pi(\beta|\xi) \propto |I(\beta|\xi)|^{\frac{1}{2}}$, where the Fisher information matrix $I(\beta|\xi)$ is $X'\Omega X$, with $\Omega = \text{diag}(\omega_1, \cdots, \omega_n)$, $\omega_i = \{(1-\xi\eta_i)^{-\frac{2}{\xi}-2}\}[\exp\{(1-\xi\eta_i)^{-\frac{1}{\xi}}\} - 1]^{-1}$, and $\eta_i = x_i\beta$, for $i = 1, \cdots, n$. The joint posterior is then given by

$$\pi(\beta, \xi|\, D_{obs}) \propto \prod_{i=1}^{n} \{1 - GEV(-x_i\beta; \xi)\}^{y_i} \{GEV(-x_i\beta; \xi)\}^{1-y_i} |I(\beta|\, \xi)|^{\frac{1}{2}}.$$

3.2. Information Matrix Ridge Prior

We also consider a very general class of proper priors known as Information Matrix Ridge prior, which is given as follows:

$$\pi_{IMR}(\beta|\, \xi) \propto |X'\Omega(\beta|\, \xi)X + \lambda I|^{\frac{1}{2}} \exp\left[-\frac{1}{2c_0}(\beta - \mu_0)'\{X'\Omega(\beta|\, \xi)X + \lambda I\}(\beta - \mu_0)\right], \quad (5)$$

where μ_0 is the location hyperparameter, which can be taken as 0, and $c_0 > 0$ are specified dispersion hyperparameters. The Fisher information matrix $X'\Omega(\beta|\, \xi)X$ is defined as in §3.1 for the Jeffreys' prior. This prior naturally captures the covariance of β via the Fisher information matrix. Also, with the introduction of the "ridge" parameter $\lambda \geq 0$, the matrix $X'\Omega(\beta|\, \xi)X + \lambda I$ can always be positive definite for any design matrix X and for any form of Ω. It can thus easily handle the high-dimensional data with $k > n$ or even the $k < n$ case with collinearity or weak identifiability. Discussions on the logit and the probit links under the Information Matrix Ridge prior for the binary data are given in Gupta and Ibrahim (2008). Next we develop and explore the prior and posterior properties of Information Matrix Ridge prior for the generalized extreme value model. The proofs are postponed in the Appendix.

Theorem 1 *The prior moment generating function of* β *based on the Information Matrix Ridge prior exists for the binomial model with the generalized extreme value link.*

Theorem 2 *The posterior moment generating function of* β *based on the Information Matrix Ridge prior exists for the binomial model with the generalized extreme value link.*

Note that when $k < n$, Jeffreys' prior can be obtained as a special case when $c_0 \to \infty$. The sufficient condition for the existence of the moment generating function under Jeffreys' prior are identical as those for the Information Matrix Ridge prior. Thus we have the following corollary.

Corollary 1 *The prior and posterior moment generating functions of* β *based on the Jeffreys' prior exists for the binomial model with the generalized extreme value link for* $k < n$ *and* $\lambda = 0$.

3.3. Uniform Prior

Let $\tau_i = -1$ if $y_i = 0$ and $\tau_i = 1$ if $y_i = 1$. Define $X^*_{l,m} = (\tau_i x'_i, l < i < m)$ as the $(m-l) \times k$ matrix with rows $\tau_i x'_i$, $l < i \leq m$, where $0 \leq l < m \leq n$. We are led to the following theorem concerning the propriety of the posterior distribution in (4) when $\pi(\beta) \propto 1$, which is an improper uniform prior. The result established in this theorem implies that the proposed generalized extreme value link model is identifiable. The proof of Theorem 3 is also given in Appendix.

Theorem 3 *Suppose that there exist* $p > k$, $0 = m_0 < \cdots < m_p \leq n$, *and positive vectors* $a_1, \cdots, a_p$ *such that* $X^*_{m_{l-1},m_l}$ *is of full rank and* $a'_l X^*_{m_{l-1},m_l} = 0$ *for* $l = 1, \cdots, p$. *Under the improper uniform prior* $\pi(\beta) \propto 1$, *the posterior* (4) *is proper.*

Acknowledgement

The research of the second author was funded in part by SAMSI, Research Triangle Park, N.C., during his visit as part of his sabbatical.

Appendix

Proofs

Proof of Theorem 1. The model specified in (1) and (3) is a binomial model using generalized extreme value link. The conditional density for $y_i|\, x_i, i = 1, \cdots, n$, is given by $p(y_i|x_i, \beta) = \exp\left\{y_i\theta_i - b(\theta_i)\right\}$, where $\theta_i = \log\{p_i/(1-p_i)\}, b(\theta_i) = -\log(1-p_i)$. Using a generalized extreme value link, we have a joint prior for β and ξ as $\pi(\beta,\, \xi) \propto \pi(\beta|\,\xi)\pi(\xi)$. The existence of the prior moment generating function is equivalent to the finiteness of the integral below.

$$\int\int \pi_{IMR}(\beta,\xi)e^{t'(\beta,\xi)}d\beta d\xi \leq \left[\int\{\pi(\beta|\xi)\exp(t'\beta)\}^2 d\beta \int_{-1}^{1} 0.5\exp(2t\xi)d\xi\right]^{\frac{1}{2}}.$$

Since $\int_{-1}^{1}\exp(2t\xi)d\xi$ is finite, we need to prove the finiteness of $\int\{\pi(\beta|\xi)\exp(t'\beta)\}^2 d\beta$. Following Gupta and Ibrahim (2008), it is obvious that the sufficient condition for the finiteness of the above integral is that each integral in the following equation is finite for some τ_j in an open interval containing zero,

$$\prod_{j=1}^{k}\int \exp\left[-\frac{M_1\lambda}{c_0}\{\theta^{-1}(r_j)\}^2 + 2\tau_j\theta^{-1}(r_j)\right]\frac{d^2b(r_j)}{dr_j^2}dr_j, \qquad \text{(A1)}$$

where $\text{rank}(X) = n$, M_1 is such that $|X^{**}| \leq M_1^{-\frac{k}{2}}$, $X^{**\prime} = [X' \vdots x_0']$, with x_0 as a $(k-n)\times k$ matrix selected such that X^{**} is positive definite.

The generalized extreme value link function is given as:

$$\theta(\eta) = \log\left[\{1 - GEV(-\eta;\, \xi)\}\{GEV(-\eta;\, \xi)\}^{-1}\right].$$

It can be shown that $\theta^{-1}(r) = \eta = -GEV^{-1}\left\{(1+e^r)^{-1}\right\}$, where $GEV^{-1}(\cdot)$ is the inverse cumulative distribution function of $GEV(\cdot)$, and $d^2b(r)/dr^2 = \exp(r)\{1+\exp(r)\}^{-2}$. Thus, the sufficient condition in (A1) is equivalent to

$$\int_{-\infty}^{\infty} \exp\left[-\frac{\lambda M}{2c_0}\left\{-GEV^{-1}\left(\frac{1}{1+e^r}\right)\right\}^2 - \tau GEV^{-1}\left(\frac{1}{1+e^r}\right)\right]\frac{e^{\frac{r}{2}}}{1+e^r}dr. \tag{A2}$$

Let $z = -GEV^{-1}\left\{(1+e^r)^{-1}\right\}$ and $gev(-z;\, \xi) = \partial GEV(-z;\, \xi)/\partial z$, which is in fact the probability density function, then (A2) becomes

$$\int_{-\infty}^{\infty} \exp\left(-\frac{\lambda M}{2c_0}z^2 + \tau z\right) gev(-z;\, \xi)[GEV(-z;\, \xi)\{1 - GEV(-z;\, \xi)\}]^{-\frac{1}{2}}dz,$$

which is finite, since $gev(-z;\, \xi)[GEV(-z;\, \xi)\{1 - GEV(-z;\, \xi)\}]^{-\frac{1}{2}}$ is bounded.

Proof of Theorem 2. Similar to the proof of Theorem 1 and using Gupta and Ibrahim (2008), a sufficient condition for the existence of the posterior moment generating function for the Information Matrix Ridge prior is that the one dimensional integral

$$\int \exp\left[-\frac{\lambda M_1}{2c_0}\{\theta^{-1}(r_j)\}^2 + \tau_j\theta^{-1}(r_j) + yr_j\right]\left\{\frac{e^{\frac{r_j}{2}}}{(1+e^{r_j})^2}\right\}dr_j \tag{A3}$$

is finite for some τ_j in an open neighborhood about zero, for $j = 1, \cdots, n$. for $j = n+1, \cdots, k$, the conditions are the same as for the prior moment generating function.

Substitute $z = -GEV^{-1}\{(1+e^r)^{-1}\}$ into the above integral, we get

$$\int \exp\left(-\frac{\lambda M_1}{2c_0}z^2 + \tau z\right)\left\{\frac{1 - GEV(-z;\, \xi)}{GEV(-z;\, \xi)}\right\}^{y-\frac{1}{2}} gev(-z;\, \xi)dz,$$

which is finite since $\{1 - GEV(-z;\, \xi)\}^{y-\frac{1}{2}}\{GEV(-z;\, \xi)\}^{\frac{1}{2}-y} gev(-z;\, \xi)$ is bounded. Thus, the posterior moment generating function exists for the binary data with generalized extreme value link.

Proof of Theorem 3. Let $u, u_1, \cdots, u_n$ be independent random variables with common distribution function F, which is a generalized extreme value distribution with $\mu = 0$, $\sigma = 1$, and a shape parameter ξ. For $0 < a < 1$, it can be shown that $E|u|^a < \infty$ for $-1 \le \xi < 1$. Observing that $1 - F(-x) = EI(u > -x)$ and $F(-x) = EI\{-u \le -(-x)\}$, here I is an indicator function. Now, we have $\{1 - F(-x_i'\beta)\}^{y_i}\{F(-x_i'\beta)\}^{1-y_i} \le EI\{\tau_i u_i \ge \tau_i(-x_i'\beta)\}$ and $\{1 - F(-x_i'\beta)\}^{y_i}\{F(-x_i'\beta)\}^{1-y_i} \ge EI\{\tau_i u_i > \tau_i(-x_i'\beta)\}$. Let $u^* = (\tau_1 u_1, \cdots, \tau_n u_n)$. Using Fubini's theorem, we obtain

$$
\begin{aligned}
&\int_{-1}^{1}\int_{R_k} p(y|X,\beta,\xi)\, d\beta d\xi \\
=&\int_{-1}^{1}\int_{R_n} E\left[\int_{R_k} I\{-\tau_i x_i'\beta < \tau_i u_i, 1 \le i \le n\} d\beta\right] dF(u) d\xi \\
=&\int_{-1}^{1}\int_{R_n} E\left\{\int_{R_k} I(X^*\beta < u^*) d\beta\right\} dF(u) d\xi. \qquad \text{(A4)}
\end{aligned}
$$

Under the condition of Theorem 3, it follows directly from Lemma 4.1 of Chen & Shao (2000) that there exists a constant K such that $||\beta|| \le K \min_{1\le l\le p}(\max_{m_{l-1}<i\le m_l} |w_i|)$ whenever $X^*\beta \le w$, where $w = (w_1, \cdots, w_n)$. Hence, from (A4), we have

$$
\begin{aligned}
&\int_{-1}^{1}\int_{R_k} p(y|X,\beta,\xi)\, d\beta d\xi \\
\le&\int_{-1}^{1}\int_{R_n} \prod_{l=1}^{k} E_{\max_{m_{l-1}<i\le m_l}} |u^*|^{k/p} dF(u) d\xi \\
<&\infty.
\end{aligned}
$$

References

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc.* **88**, 669-79.

Chen, M.-H. & Shao, Q.-M. (2000) Propriety of posterior distribution for dichotomous quantal response models with general link functions. *Proc. Am. Math. Soc.* **129,** 293-302.

Chen, M.-H., Dey, D.K., & Shao, Q.-M. (1999). A New Skewed Link Model for Dichotomous Quantal Response Data. *J. Am. Statist. Assoc.* **94**, 1172-86.

Coles, S.G. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer Verlag, New York.

Gupta, M. & Ibrahim, J.G. (2008). An Information Matrix Prior for Bayesian Analysis in Generalized Linear Models with High Dimensional Data. Presented at ENAR 2008 Spring Meeting.

Kim, S., Chen, M.-H., & Dey, D.K. (2008). Flexible generalized t-link models for binary response data. *Biometrika.* **95**(1), 93-106.

Stukel, T. (1988). Generalized Logistic Models. *J. Am. Statist. Assoc.* **83**, 426-31.

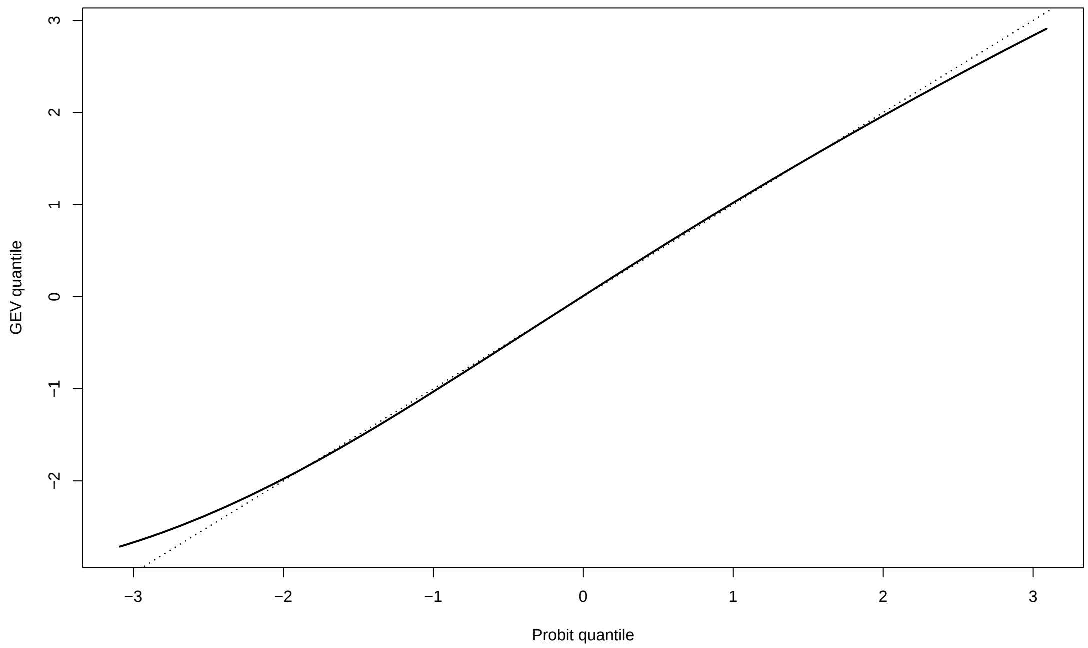

Fig. 1. Plot of generalized extreme value quantiles with $\mu \approx -0.35579, \sigma \approx 0.99903$, and $\xi \approx -0.27760$ against probit quantiles for probabilities between 0.001 and 0.999. The solid line is the quantile plot, and the dotted line is the 45° reference line.