Generalized Weighted Linear Models Based on Distribution Functions

In-Kwon Yeo *

Abstract

In this paper, a new form of generalized linear models is proposed. The proposed models consist of a distribution function of the mean response and a weighted linear combination of distribution functions of covariates. This form addresses a structural problem of the link function in the generalized linear models. Markov chain Monte Carlo methods are used to estimate the parameters within a Bayesian framework.

Keywords: Bayesian inference, Markov chain Monte Carlo, mixtures of distributions, parametric transformation.

1. Introduction

Suppose that $Y_1, \ldots, Y_n$ are independent random variables each with the probability density function $f(y_i; \theta_i)$. Generalized linear models assume that the Y_i has a density function in the exponential family of the form

$$f(y_i; \theta_i) = \exp\left[a^{-1}(\phi_i)\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi_i)\right],$$

for some specific known functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$. Note that the mean and the variance of the Y_i is derived from the equations $E(Y_i) = \mu_i = b'(\theta_i)$ and $var(Y_i) = a(\phi_i)b''(\theta_i)$, respectively. In the generalized linear models specification, it is assumed that an unknown parameter θ_i depends on a particular regressor $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ and a link function $g(\cdot)$ satisfies a linear models

$$g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown parameters to be estimated. The link function describes the dependence of the μ_i on the linear predictor $\boldsymbol{x}_i^T\boldsymbol{\beta}$ for the regressor $\boldsymbol{x}_i$. For a standard generalized linear model, we have $\boldsymbol{x}_i^T\boldsymbol{\beta} = g(\mu_i) = g(b'(\theta_i))$. The selection of link functions has been one of primary issues in these models. Even though the canonical link function leads to desirable and simple properties of the model, it is not guaranteed that it should always be appropriate and well-behaved. A hardness of generalized linear models for non-statistician is interpreting the regression parameter because of the nonlinearity of link functions. In this paper, we present an alternative model which is also based on the

*Assistant Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University(This work was in part supported by the Korea Research Foundations, Korea, under grant KRF-2002-003-C0028.)

exponential family and provides a flexible model structure and an easy interpretation about the parameters.

2. The Proposed Models

Suppose $Y_1, \ldots, Y_n$ are independent responses and have a distribution in the exponential family with $E(Y_i) = \mu_i \in \Omega$, where Ω is the parameter space. The proposed model assumes that the relationship between the mean response μ_i and the covariate vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ takes the form

$$F_\mu(\mu_i) = \sum_{j=1}^{p} \alpha_j U_j(x_{ij}),$$

where, for $j = 1, \ldots, p$, $\alpha_j \geq 0$ and $\sum_{j=1}^{p} \alpha_j = 1$. The function F_μ is an arbitrary distribution function and U_j's are monotone function and have a value on [0,1]. We will call F_μ the link distribution function and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ the weight parameters, respectively. Since explanatory variable x_{ij} is transformed to $[0, 1]$ through a monotone function U_j and is standardized, the weight parameter α_j measures the relative importance of the j-th explanatory variable in the model and can be also applied to a criterion of the variable selection in which the covariates having a relatively small weight are removed from the model.

In order to parameterize the effect of the covariate x_{ij}, we set

$$U_j(x_{ij}) = F_j(x_{ij})^{\beta_j} \{1 - F_j(x_{ij})\}^{1-\beta_j},$$

where F_j is an arbitrary distribution function and β_j is either 0 or 1. Since F_j is an increasing function, $\beta_j = 1$ implies that the j-th covariate has a positive effect to the mean response and, otherwise, negative. We will also call F_j's the explanatory distribution functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ the effect parameters, respectively. Finally, we have

$$F_\mu(\mu_i) = \sum_{j=1}^{p} \alpha_j F_j(x_{ij})^{\beta_j} \{1 - F_j(x_{ij})\}^{1-\beta_j}. \tag{1}$$

For simple notations, we write $\boldsymbol{F}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i) = \sum_{j=1}^{p} \alpha_j F_j(x_{ij})^{\beta_j} \{1 - F_j(x_{ij})\}^{1-\beta_j}$. Since, generally, the space of mean response is a continuous interval, we assume that the distribution function F_μ is absolutely continuous. Suppose that the F_μ and F_j's are known and some necessary parameters for these distribution functions are predetermined. Then, we have $\mu_i = F_\mu^{-1}\{\boldsymbol{F}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i)\}$ and, from the equation $\mu_i = b'(\theta_i) = h^{-1}(\theta_i)$,

$$\theta_i = h(\mu_i) = h\left[F_\mu^{-1}\{\boldsymbol{F}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i)\}\right] = h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i).$$

The likelihood function is given by

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{y}, \mathbf{X}) = \exp\left[\sum_{i=1}^{n} a^{-1}(\phi_i)\{y_i h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i) - b(h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i))\}\right], \tag{2}$$

where $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

The proposed models can be practically implemented by Bayesian framework rather than Frequentist framework. The book edited by Dey, Ghosh, and Mallick (2000) discusses inferences for generalized linear models and some extended models from a Bayesian point of view.

Many theories and applications discussed in this book can be also applied to the proposed models with a little modification. In this paper, we investigate procedures to estimate the parameters from a Bayesian perspective and also discuss the selection of moderate functions F_μ and F_i's.

3. Bayesian Inferences

In order to perform Bayeisan inference, priors for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are required. The prior considered for the weight parameter $\boldsymbol{\alpha}$ is the Dirichlet distribution with parameter $\boldsymbol{a} = (a_1, \ldots, a_p)^T$. The product of independent Bernoulli distributions with the success probability $\boldsymbol{b} = (b_1, \ldots, b_p)^T$ is a naive choice of the prior for the effect parameter $\boldsymbol{\beta}$. Then, assuming $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are independent, the posterior of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is written as

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}) \propto \exp\left[\sum_{i=1}^{n} a^{-1}(\phi_i)\left\{y_i h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i) - b(h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i))\right\}\right] \prod_{j=1}^{p}\left\{\alpha_j^{a_j/2-1}\left(\frac{b_j}{1-b_j}\right)^{\beta_j}\right\}.$$

This posterior is not analytically tractable. Gibbs sampling is a useful technique which generates samples from the posterior for implementation of the Bayesian model fitting. It requires sampling from the full conditional distributions;

$$\pi(\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{y}, \mathbf{X}) \propto \exp\left[\sum_{i=1}^{n} a^{-1}(\phi_i)\left\{y_i h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i) - b(h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i))\right\} + \sum_{j=1}^{p}(a_j/2 - 1)\log(\alpha_j)\right]$$

$$\pi(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{y}, \mathbf{X}) \propto \exp\left[\sum_{i=1}^{n} a^{-1}(\phi_i)\left\{y_i h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i) - b(h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i))\right\} + \sum_{j=1}^{p}\beta_j \log\left(\frac{b_j}{1-b_j}\right)\right].$$

However, neither $\pi(\boldsymbol{\alpha}|\cdot)$ nor $\pi(\boldsymbol{\beta}|\cdot)$ are standard conditionals, so that an efficient simulation algorithm should be applied to generate samples. We discuss Metropolis-Hastigs algorithm for the proposed models.

It is necessary that a suitable candidate generating density is specified to implement the Metropolis-Hastigs algorithm. Let $q^{\alpha}(\cdot|\boldsymbol{\alpha}^*)$ and $q_j^{\beta}(\cdot|\beta^*)$ be candidate generating densities for $\boldsymbol{\alpha}$ given $\boldsymbol{\alpha}^*$ and for β_j given β^*, respectively. Then, the sampling algorithm is as follows;

1. Initialize $\boldsymbol{\alpha}^{(0)} = (\alpha_1^{(0)}, \ldots, \alpha_p^{(0)})^T$ and $\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, \ldots, \beta_p^{(0)})^T$;
2. Repeat for $t = 1, \ldots, N$.
   (a) Sample a point $\boldsymbol{\alpha}^*$ from $q^{\alpha}(\cdot|\boldsymbol{\alpha}^{(t-1)})$ and a uniform(0,1) random variable U;
   (b) Compute the acceptance probability

$$\gamma^{\alpha}(\boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\alpha}^*) = \min\left(1, \frac{\pi(\boldsymbol{\alpha}^*|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{y}, \mathbf{X}) q^{\alpha}(\boldsymbol{\alpha}^{(t-1)}|\boldsymbol{\alpha}^*)}{\pi(\boldsymbol{\alpha}^{(t-1)}|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{y}, \mathbf{X}) q^{\alpha}(\boldsymbol{\alpha}^*|\boldsymbol{\alpha}^{(t-1)})}\right);$$

   (c) If $U \leq \gamma^{\alpha}(\boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\alpha}^*)$, set $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^*$; Otherwise, set $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^{(t-1)}$;
   (d) Repeat for $j = 1, \ldots, p$,
      i. Sample a point β^* from $q_j^{\beta}(\cdot|\beta_j^{(t-1)})$;

ii. If $\beta^* = \beta_j^{(t-1)}$, set $\beta_j^{(t)} = \beta^*$; Otherwise

A. Sample a uniform(0,1) random variable U and compute the acceptance probability

$$\gamma^{\beta}(\beta_j^{(t-1)}, \beta^*) = \min\left(1, \frac{\pi(\beta^*|\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}_{(-j)}^{(t)}, \boldsymbol{y}, \mathbf{X}) q_j^{\beta}(\beta_j^{(t-1)}|\beta_j^*)}{\pi(\beta_j^{(t-1)}|\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}_{(-j)}^{(t)}, \boldsymbol{y}, \mathbf{X}) q_j^{\beta}(\beta^*|\beta_j^{(t-1)})}\right);$$

B. If $U \leq \gamma^{\beta}(\beta_j^{(t-1)}, \beta^*)$, set $\beta_j^{(t)} = \beta^*$; Otherwise set $\beta_j^{(t)} = \beta_j^{(t-1)}$;

where $\boldsymbol{\beta}_{(-j)}^{(t)} = (\beta_1^{(t)}, \ldots, \beta_{j-1}^{(t)}, \beta_{j+1}^{(t-1)}, \ldots, \beta_p^{(t-1)})^T$ comprises all of $\boldsymbol{\beta}$ except β_j at the t-th repetition and the full conditional distribution of β_j is given as

$$\pi(\beta_j|\boldsymbol{\alpha}, \boldsymbol{\beta}_{(-j)}, \boldsymbol{y}, \mathbf{X}) \propto \exp\left[\sum_{i=1}^{n} a^{-1}(\phi_i)\left\{y_i h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i) - b(h(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i))\right\} + \beta_j \log\left(\frac{b_j}{1-b_j}\right)\right].$$

4. Selection of Link Distribution Function

The link distribution function connects the mean response μ and a weighted linear combination of explanatory distribution functions. If the information is quite limited, the conjugate family can be employed as a proper link distribution function since, according to the Pitman-Koopman Lemma, conjugate priors can only be found in exponential families and the parameter space Ω is consistent with the support of conjugates. Then, the resultant link distribution function is written as $F_{\mu}(u) = F(u; \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is the vector of hyperparameters corresponding a conjugate prior.

Although the conjugate distribution leads to a well-defined link distribution function, we do not guarantee that it is a well-behaved link distribution function. For improvement of fitting data, we consider two classes of link distribution functions indexed by a shape parameter λ. Let $\psi(v, \lambda)$ be an increasing transformation from $[0, 1]$ to $[0, 1]$, for instance, $\psi(v, \lambda) = v^{\lambda}$ for $\lambda > 0$, and

$$\psi(v, \lambda) = \begin{cases} \left\{(v+1)^{\lambda} - 1\right\}/(2^{\lambda} - 1), & \text{if } \lambda \neq 0 \\ \log(v+1)/\log(2), & \text{if } \lambda = 0. \end{cases} \tag{3}$$

Other examples for the ψ are $\mathrm{IB}(v; \lambda, k)$ and $\mathrm{IB}(v; k, \lambda)$, for a fixed k, where $\mathrm{IB}(v; c, d)$ stands for the incomplete beta function associated with the beta density with parameters c and d evaluated at v. Then, the first class of link distribution functions takes the form

$$F_{\mu}(u) = \psi^{-1}\{F_{\gamma}(u), \lambda\}, \tag{4}$$

where $F_{\gamma}(\cdot)$ is an arbitrary distribution function with density function supported on Ω and is sometimes indexed by unknown parameters $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$. The second class of link distribution function has the form $F_{\mu}(u) = F_{\gamma}\{\phi(u, \lambda)\}$, where $\phi(u, \lambda)$ is an increasing transformation mapping of Ω into Ω. When $\Omega = R^+$ or $\Omega = R$, the modulus transformation introduced by John and Draper (1980) and the extended power transformation by Yeo and Johnson (2000) are appropriate for this purpose. Suppose $\pi_{\lambda}(\cdot)$ is the prior density functions

of λ. Specifying a suitable proposal density λ, we can implement the Metropolis-Hastigs algorithm mentioned in Section 3.

In Bayesian literatures, semiparametric link functions using nonparametric specifications have been extensively studied, for instance, Mallick and Gelfand (1994) and Newton, Czado, and Chappell (1996). A mixture model is often employed to specify the link function in generalized linear models. For the proposed model, the nonparametric structure assumes that $F_\mu(\cdot)$ is a random draw from $\mathcal{F}_\mu$ where $\mathcal{F}_\mu$ is a wider class of distribution functions. A mixture model can be also considered in which a dense class of mixtures of standard distributions, $\{F^{(l)}\}$, is modelled for $F_\mu(u) = \sum_{l=1}^{k} \omega_l F^{(l)}(u)$, where $\omega_l \geq 0$ are the mixing weights, $\sum_{l=1}^{k} \omega_l = 1$. If $\omega_l = 1$ for a particular l, the mixture model reduces to a nonparametric specification. Otherwise, this is a semiparametric specification. Usually, the semiparametric specification requires an awkward computation of quantiles and, further, the identifiability is doubted when ω's are unknown and should be estimated with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, simultaneously.

5. Selection of Explanatory Distribution Functions

Suppose that $F_j^*(\cdot)$ is a proper explanatory distribution for explaining the relationship (1). Discrete mixtures of beta densities provide a continuous dense class of models for densities on [0,1]. This implies that an unknown distribution function can be modelled by a mixture of beta distribution functions. Let $F_j^0(\cdot)$ be a centering distribution function for $F_j^*(\cdot)$. We assume that these parameters can be chosen or given so that the resultant $F_j^0(\cdot)$ is well-behaved. Since $\mathrm{IB}(F_j^0(\cdot); c, d)$ provides a rich class of models which describes patterns of dependency of μ on x according to c and d, we can approximate that

$$F_j^*(u) \simeq \sum_{l=1}^{L_j} \omega_l \mathrm{IB}(F_j^0(u); c_l, d_l),$$

where L_j is the number of mixands and ω_l's are the mixing weights, $\omega_l \geq 0$ and $\sum_{l=1}^{L_j} \omega_l = 1$.

Since the specification of L, $\omega = (\omega_1, \ldots, \omega_L)^T$, $\boldsymbol{c} = (c_1, \ldots, c_L)^T$, and $\boldsymbol{d} = (d_1, \ldots, d_L)^T$ is too much to attempt, following Mallick and Gelfand (1994), we fix L, $\boldsymbol{c}$, and $\boldsymbol{d}$ and let only $\boldsymbol{\omega}$ be random given L. The $\boldsymbol{c}$ and $\boldsymbol{d}$ are chosen to provide a set of beta densities which blanket [0,1], for instance, $c_l = \sigma l$ and $d_l = \sigma(L + 1 - l)$, $l = 1, 2, \ldots, L$ where σ is either estimated or given. Two types of priors for $\boldsymbol{\omega}$ are available, the Dirichlet and the multinomial distribution. The Dirichlet distribution leads to a semiparametric setting, while the multinomial distribution leads to a nonparametric specification. In practice, the multinomial distribution is preferable because the Dirichlet distribution deserves a heavy job for the implementation for large p and a complex interpretation for the relationship between μ and covariates.

6. Example of Binary Response Data

Milicer and Szczoka (1966) analyzed data determining the age of menarche of a sample of 3918 Warsaw girls in 1965. Guerrero and Johnson (1982) obtained a remarkable improvement using a probit model where the Box-Cox transformation is taken to the explanatory variable and a normal distribution is assumed for the transformed variable.

During the analyses, we set the uniform (0,1) as the link distribution function $F_\gamma(u)$. Four different $\psi(v,\lambda)$'s, v(Identity), v^λ(Power-1), IB$(v;1,\lambda)$, and equation (3) (Power-2), are considered to improve the fit. An exponential prior was employed for the parameter λ of Power-1, and IB, and a normal for Power-2. The centering distribution was assumed to be uniform (9,18) and the number of mixands for the explanatory distribution $L = 5$, respectively. The prior distribution of mixing weights was a multinomial distribution with 5 categories having the same success probabilities.

Table 1: Summary of MCMC results for ψ's.

	Model selection		Parameter estimation		Pearson's
	β	ω	$\hat{\lambda}$ (s.d.)	$\hat{\sigma}$ (s.d.)	X^2
Identity	1	3	.	2.16(0.14)	161.05
Power-1	1	1	11.44 (0.53)	0.98(0.04)	13.47
IB(1,·)	1	2	0.30 (0.02)	5.69 (0.60)	15.35
Power-2	1	2	5.77 (0.26)	1.76 (0.10)	24.14

Table 1 shows a summary of results based on a run of 200,000 iterations of MCMC with the first 100,000 discarded as burn in. With the maximum likelihood estimates, Pearson's X^2 statistics for logistic, probit, and cloglog regression are 21.31, 21.74, and 190.93, respectively and we see that the proposed models with Power-1 and IB(1,·) much improve the fitting of data. The exploration of some figures shows that the reductions in Pearson's X^2 of Power-1 and IB(1,·) are obtained at low parts of μ comparing with the logistic regression.

References

Dey, D. K., Ghosh, S. K., and Mallick, B. K. (2000), *Generalized linear models: a Bayesian perspective*, Marcel Dekker, New York.

Guerrero, V. M. and Johnson, R. A. (1982), Use of the Box-Cox transformation with binary response models, *Biometrika*, **69**, 309-314.

John, J. A. and Draper, N. R. (1980), An alternative family of transformations, *Applied Statistics*, **29**, 190-197.

Mallick, B. K. and Gelfand, A. E. (1994), Generalized linear models with unknown link functions, *Biometrika*, **81**, 237-245.

Milicer, H. and Szczoka, F. (1966), Age at menarche in Warsaw girls in 1965, *Human Biology*, **38**, 199-203.

Newton, M. A., Czado, C., and Chappell, R. (1996), Bayesian inference for semiparametric binary regression, *Journal of the American Statistical Association*, **91**, 142-153.

Yeo, I. K. and Johnson, R. A. (2000), A new family of power transformations to improve normality or symmetry, *Biometrika*, **87**, 954-959.