PLISKA
STUDIA MATHEMATICA BULGARICA

ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **17** (2005), 97–107

PLISKA
STUDIA MATHEMATICA
BULGARICA

PARAMETERIZED LINK FUNCTIONS IN GENERALIZED LINEAR RANDOM EFFECT MODELS: A CASE STUDY ON BREAST CANCER TREATMENT

Dario Gregori, Rosalba Rosato, Giovannino Ciccone, Lara Lusa

In non-linear random effects some attention has been very recently devoted to the analysis of suitable transformation of the response variables separately (Taylor 1996) or not (Oberg & Davidian 2000) from the transformations of the covariates and, as far as we know, no investigation has been carried out on the choice of link function in such models. In our study we consider the use of a random effect model when a parameterized family of links (Aranda-Ordaz 1981, Prentice 1996, Pregibon 1980, Stukel 1988 and Czado 1997) is introduced. We point out the advantages and the drawbacks associated with the choice of this data-driven kind of modeling. Difficulties in the interpretation of regression parameters, and therefore in understanding the influence of covariates, as well as problems related to loss of efficiency of estimates and overfitting, are discussed. A case study on radiotherapy usage in breast cancer treatment is discussed.

1. Introduction

In the framework of hierarchical generalized linear models (HGLM) link function relates the linear predictor to the expected value and its form is usually assumed to be known and fixed. Formulating a random effect regression problem through a threshold model, it can be easily shown that the choice of the link function results from the hypothesis made on the distribution of the first level residuals of a latent continuous variable. The most common choices for the residuals' distribution are the standard normal and the logistic distribution, which respectively lead to the probit and logit link.

In some cases it might be useful to improve the modeling flexibility, allowing the first level residuals to have a distribution that depends on one or more parameters, and therefore letting the link function to be a member of a class indexed by parameters to be estimated.

While the performance of some families of link functions has been investigated for the generalized linear models (GLM) (Pregibon 1980, Stukel 1988 and Czado 1997), not much work has been done for hierarchical generalized linear models. In non-linear random effects models some attention has been very recently devoted to analyzing suitable transformation of the response variables separately (Taylor 1996) or not (Oberg & Davidian 2000) from the transformations of the covariates and, as far as we know, no investigation has been carried out on the choice of link function in such models. Nevertheless, the effects of a mis-specified link function are, at least in principle, not trivial and can affect, besides the interpretation of the model, the point estimation and the statistical significance of fixed effects and variance components, and the overall fit of the model. In the next sections, we will investigate the bias introduced in a random effect analysis from mis-specifying the link function. Thus, the impact of such bias will be discussed with reference to a real data set on breast cancer treatment.

2. A Simulation Study

Several simulations were performed to assess the effect of link mis-specification for random intercept models with binary dependent variable and to evaluate the performance of estimation procedures that implied a data-driven choice of the link function. The Aranda-Ordaz Asymmetric family (AOA family) of link functions (Aranda-Ordaz 1981) was used to generate data. For the AOA family, the linear predictor η and $P(Y=1)=\pi$ are related by

$$\eta = \log\left[\frac{(1-\pi)^{-\lambda}-1}{\lambda}\right]. \tag{1}$$

Some well known link functions are members of the AOA family, which reduces to the logit link for $\lambda = 1$ and to the complementary log log link for $\lambda \to 0$. The probit link is approximated when $\lambda = 0.38$. The parameter λ must satisfy the constrain $\lambda > -\exp(-\eta)$.

The Monte Carlo data sets used in the simulations were produced according to the following procedure

1. ***true* values** are fixed for second-level variance component τ^2 and fixed effects α and β;

2. **covariate's values** x_j are simulated independently from $N(0.5, 1)$ at a cluster level;

3. **random intercepts** b_j are generated independently from $N(0, 1)$;

4. **linear predictor** is calculated for each observation; for the ith observation within the jth cluster it is assumed to be

$$\eta_{ij} = \alpha + \beta x_{ij} + \tau\, b_j$$

   with $j = 1, \ldots, J$, $i = 1, \ldots, n_j$; the number of observations within each cluster is assumed to be constant ($n_j = n$); the total number of observations is $N = nJ$;

5. **link function** is assumed to belong to AOA family, $P(Y_{ij} = 1) = \pi_{ij}$ is derived from Eq. 1;

6. **response variable** y_{ij} is assigned a 0 or 1 value according to the result obtained comparing the value simulated from a uniform random variable with π_{ij}.

Different values of λ, τ^2, number of clusters and number of observations were used to generate simulated data; each simulation used 1000 Monte Carlo data sets. The model fitting was performed using the SAS NLMIXED Procedure (SAS Institute 1999), maximizing an approximated likelihood integrated over the random effects, using a dual quasi-Newton algorithm.

2.1. Effect of link mis-specification

In order to assess the effect of link mis-specification, random intercept logistic models were fitted on the simulated Monte Carlo data sets, using AOA family at various levels of λ.

Notice that direct comparison of the estimated fixed effects and variance components with their *true* values is not appropriate in this framework, because of their heavy dependence on the link function. Only comparisons in terms of estimated probabilities can quantify the effects of the link choice. Estimated parameters have to be adjusted for the scale, if their crude comparison is of interest, namely for the differences in the variance of the first level residuals (σ_e^2), which typically depends on the parameters that index the link family. Only the remaining differences can be attributed to the choice of the model. Therefore when evaluating the relative bias of the estimates, a correction for the scale is applied to reduce the *true* values to the logit scale ($\beta_c^\star = \beta^\star \sqrt{\mathrm{var}_{logit}/\mathrm{var}_{AOAF}(\lambda)}$, Relative

Model	$\alpha^\star$	$\beta^\star$	$\tau^{\star 2}$	λ	n	J	N
S1	0.5	0.5	0.5	(0.05, 0.1, 0.4, 0.8, 0.9, 1.1, 1.2, 2, 4)	15	50	750
S2	0.5	0.5	(0.5, 0.8, 1.2)	(0.4, 0.9, 1.2)	(5, 15, 50)	(150, 50, 15)	750
S3	0.5	0.5	(0.5, 0.8, 1.2)	(0.4, 0.9, 1.2)	500	15	7500
S4	0.5	0.5	(0.5, 0.8, 1.2)	(0.4, 0.9, 1.2)	1000	15	15000

Table 1: Simulation plan

λ	0.05	0.10	0.40	0.80	0.90	1.10	1.20	2	4
α	1.571	1.486	1.010	0.630	0.562	0.487	0.378	0.042	-0.457
	(0.226)	(0.228)	(0.181)	(0.145)	(0.140)	(0.153)	(0.133)	(0.111)	(0.103)
β	0.910	0.868	0.688	0.546	0.520	0.498	0.464	0.372	0.276
	(0.189)	(0.193)	(0.156)	(0.128)	(0.120)	(0.122)	(0.116)	(0.101)	(0.092)
τ^2	1.647	1.508	0.885	0.555	0.509	0.470	0.398	0.254	0.140
	(0.548)	(0.501)	(0.302)	(0.204)	(0.191)	(0.178)	(0.165)	(0.116)	(0.083)
Est. ρ	0.334	0.314	0.212	0.144	0.134	0.125	0.108	0.072	0.041
True ρ	0.200	0.198	0.181	0.149	0.140	0.124	0.116	0.066	0.020
Rel. Bias α	1.451	1.334	0.677	0.174	0.085	0.011	0.185	0.877	3.469
Rel. Bias β	0.419	0.363	0.142	0.017	0.004	0.034	0.001	0.091	0.490
Rel. Bias τ^2	1.004	0.859	0.220	0.036	0.053	0.012	0.075	0.089	1.046
Rel. Bias ρ	0.669	0.589	0.173	0.030	0.046	0.011	0.067	0.083	1.003
ΔP_1	0.195	0.180	0.088	0.020	0.009	0.015	0.019	0.057	0.079
ΔP_2	0.314	0.209	0.220	0.007	0.002	0.004	0.006	0.015	0.464

Table 2: Estimation of a logit model from data generated according to the S1 plan

Bias (β)=$\mid \beta - \beta_c^\star \mid / \beta_c^\star$). The intraclass correlation coefficient $\rho = \frac{\tau^2}{\sigma_e^2+\tau^2}$ does not depend on the scale and its comparison, when using different link functions, is therefore appropriate.

To assess the effect of link mis-specification in terms of estimated probabilities, two quantities were computed: the first (ΔP_1) given by the area between two curves of partial derivatives of probabilities with respect to the covariate, one referred to the estimated probabilities ($\partial\pi_{est} \setminus \partial x$), the other to the simulated probabilities ($\partial\pi_{sim}\setminus\partial x$). The second chosen quantity (ΔP_2) is the relative difference between estimated and simulated probabilities, normalized by the amount of error imputable to the estimation procedure. ΔP_1 and ΔP_2 respectively measure the differences in fitted values attributable to the estimation of the slope (β) and to overall estimation of the model.

The summary of the S1 Model (Table 1) is given in Table 2.1..

Estimated parameters remain statistically significant for all estimated models (with the exception of α for $\lambda = 2$ model) and the relative precision (β/SE(β)) of the parameters slightly increases with λ. As expected, the relative bias of parameters substantially increases as λ takes values that differ from the unity. The same happens for the intraclass correlation coefficient, which tends to be overestimated for data simulated from models that differ substantially from the logistic. Estimated and simulated probabilities show negligible differences when simulated model is close to the logistic, while these differences become substantial for more extreme values. This should be kept in mind when selecting a link function; indeed, values for which estimated probabilities significantly differ form simulated ones include probit and complementary log log link. Point estimates remain stable varying second level variance (S2, S3, S4), number of clusters (S2) or overall number of observations (S3, S4). Their precision increases as τ^2 decreases, as well as when the number of clusters or observations increases.

2.2. Estimation of the model

The second set of simulations was related to the estimation of random intercept models with a parametric link function. a Monte Carlo data sets where generated as in the previous set of simulations. The *correct* model was fitted, allowing the link function to be a member of the AOA family, and estimating λ from data instead of keeping it fixed. Data with different number of clusters and corresponding to various values of λ were generated. Fixed effects α and β and τ^2 were kept fixed at 0.5, each cluster had 15 observations. The summary of the estimated models is given in Table 3.

In a separate simulation exercise we assessed that, using these settings, simulating data with 50 clusters from a logistic model ($\lambda = 1$) led to consistent estimates when the logistic model was fit, except for a small downward bias in the variance component. This is no longer true when λ needs to be estimated. Indeed, models generated using 50 clusters are heavily biased in their fixed effects and variance components estimates. Moreover their standard errors are respectively about 20, 8 and 68 times wider for α, β and τ^2. Some convergence problems were encountered, especially for $\lambda = 0.1$, for which 12.5% of the models did non converge.

Increasing the number of clusters and consequently the number of observations, estimates presented a substantial smaller bias and a higher precision; for $\lambda = 1$ the width of confidence intervals was about 4 times wider for α and about twice wider for β and τ^2 when compared to estimates obtained estimating a logistic model; there is therefore an important loss of efficiency of the estimates. Con-

		λ							
		0.1		0.5		0.8		1.2	
J=50	α	0.881	(2.019)	1.120	(3.124)	0.986	(2.595)	1.299	(3.714)
	β	0.658	(0.798)	0.733	(1.132)	0.705	(1.072)	0.779	(1.305)
	τ^2	1.733	(16.031)	2.979	(19.727)	2.010	(10.810)	3.129	(15.784)
	λ	0.403	(1.266)	1.054	(2.639)	1.301	(2.839)	2.100	(4.458)
J=500	α	0.538	(0.149)	0.530	(0.235)	0.457	(0.218)	0.417	(0.282)
	β	0.516	(0.075)	0.513	(0.102)	0.485	(0.091)	0.472	(0.108)
	τ^2	0.531	(0.142)	0.537	(0.205)	0.474	(0.162)	0.453	(0.203)
	λ	0.139	(0.148)	0.529	(0.295)	0.729	(0.322)	1.042	(0.478)
	CI coverage	0.718		0.959		0.933		0.927	
J=1000	α	0.517	(0.106)	0.504	(0.155)	0.441	(0.146)	0.395	(0.180)
	β	0.507	(0.052)	0.501	(0.061)	0.476	(0.091)	0.466	(0.067)
	τ^2	0.513	(0.097)	0.506	(0.123)	0.458	(0.106)	0.438	(0.119)
	λ	0.117	(0.108)	0.501	(0.198)	0.706	(0.217)	1.017	(0.311)
	CI coverage	0.819		0.947		0.926		0.896	

Table 3: Estimation of AOA family models from data generated according to the S2 plan ($\alpha = 0.5$, $\beta = 0.5$, $\tau^2 = 0.5$; $n = 15$); standard errors are in brackets.

vergence problems were solved increasing the number of observations. Results obtained would suggest that the method should not be used for small data sets. The coverage of the true value of λ by the 95% confidence intervals was calculated. While this is close to the nominal for $\lambda = 0.5$ and slightly under it for $\lambda = 0.8$, substantial under-coverage can be observed for $\lambda = 0.1$ and $\lambda = 1.2$ (Table 3). This suggests that if a hypothesis testing procedure can be performed to choose between different link functions for medium and big data sets, *p*-values upon which basing them should be adjusted.

3. A case study: the breast cancer and radiotherapy treatment in Piemonte

There's an increasing interest in the health service research related to the appropriateness of care in case on long-term, poor prognosis diseases. In particular, the huge investments and the high costs of cancer treatment are forcing research in the direction of studying the pattern of access to specific technologies, to detect possible inequalities in accessing health care resources.

In this sense, a study was performed in Piemonte (Italy) in 2003, with the aim at identifying the factors influencing likelihood of receiving radiotherapy, which is one of the target treatments of breast cancer in women.

Based on discharge records, 3250 women living in Piemonte from January 2001 to June 2002, who underwent conservative surgery for breast cancer, were studied. For each of them, the treatment with radiotherapy (RT) in the first 6 months after conservative surgery (no mastectomy) was assessed, and it represent the y variable, being 1 if the i-th woman received RT or 0 otherwise . Seventy-two hospitals were involved in the study, organized in 22 districts (ASL). Radiotherapy is a care choice not available in every hospital and/or district. Therefore, a covariate was constructed indicating the presence of radiotherapy in the given district/hospital. Other covariates included disease staging and age.

The model used commonly in the analysis of such data is the random intercept logistic model (Locallio 2001), with the random effect representing the case-mix adjusted estimate of the effect of the j-th hospital on the probability of receiving an RT. Without going into much detail (see Leyland for a complete explanation), the idea is that hospital effect is modeled as a latent random variable representing the unobservable characteristics explaining differences among hospitals not due to difference in patient's characteristics or gravity (the so called case-mix). More precisely, we assumed (Stiratelli) that the conditional expectation μ is expressed as a function of fixed covariates and a random component b

$$\text{logit}(\mu_{ij}) = \mathbf{x}_{ij}\beta + \mathbf{b}_i \tag{2}$$

with variance

$$\text{var}(y_{ij}|\mathbf{b}_i) = \mu_{ij}(1 - \mu_{ij}) \tag{3}$$

and $\mathbf{b}_i$ is an independent Gaussian random vector with mean 0,$(E(\mathbf{b}_i) = 0)$ and covariance D $(\text{cov}(\mathbf{b}_i) = D)$, i.e. $\mathbf{b}_i \sim G(0, D)$. Clearly, the logit transformation of the conditional mean is usually assumed without further investigation.

Table 4 shows the estimates for the canonical logit model and for the alternative AOA link. Significance of the results are changing dramatically with respect to the chosen model. In general, it seems that the logit model tends to be less conservative in identifying relevant covariates. The λ parameter is estimated at 0.625, intermediately between the probit and the logit link. After stratifying for ASL with or without RT, the λ parameters are indicating a log-log link for the ASL without RT and almost a probit for the ALS with RT.

Hospital effects are given in Figure 1, showing a similar beahvior, with a greater variability for AOA model than for logit.

4. Discussion

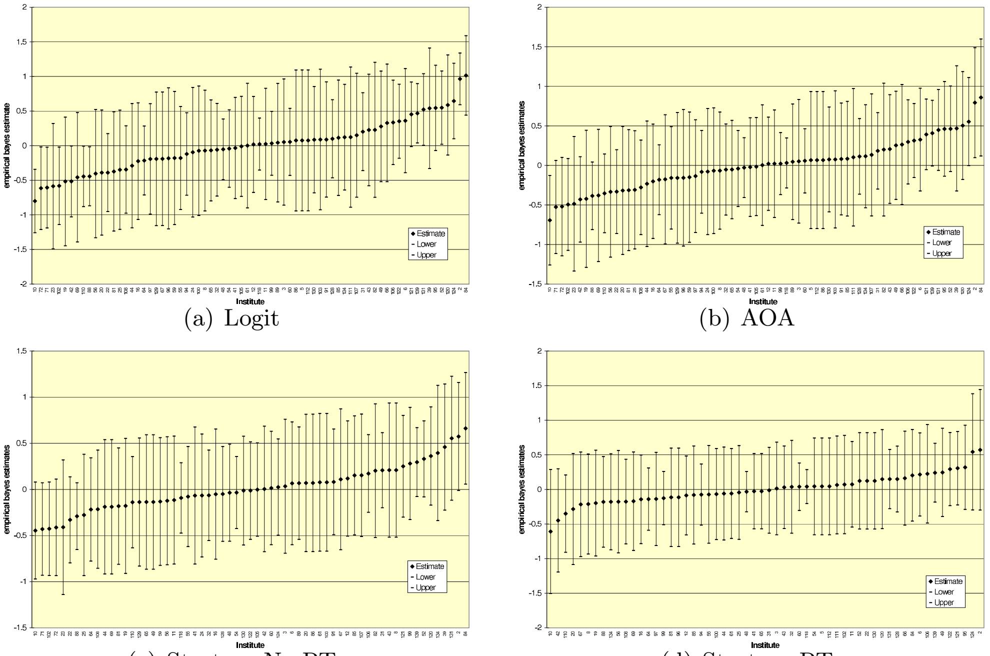
While estimating the form of the link function usually improves the fit of the model when compared to canonical links, there are some drawbacks associated with it. Data might be overfitted, leading to flat likelihoods and numerical problems in the estimating procedure. Letting the estimate of the link function be

Variables	Estimates	Standard Error	p–value
Logit model with random effects			
Intercept	0.654	0.113	¡0.0001
Age >70 years	−1.030	0.094	<0.0001
Disease Staging $g > 1$	0.240	0.140	0.091
ASL with RT	0.229	0.099	0.024
Hospital with RT	0.142	0.185	0.445
var(b)	0.274	0.084	0.002
AOA model with random effects			
Intercept	0.423	0.438	0.338
Age > 70 years	−0.878	0.294	0.003
Disease Staging > 1	0.183	0.153	0.235
ASL with RT	0.188	0.112	0.097
Hospital with RT	0.097	0.173	0.576
var(b)	0.198	0.143	0.172
λ	0.625	0.716	0.386

Table 4: Estimated RE model

	ASL without RT			ASL with RT		
Variables	Estimates	SE	p–value	Estimate	SE	p–value
intercept	0.072	0.370	0.847	0.294	0.434	0.500
Age > 70 years	−0.787	0.265	0.004	−0.747	0.449	0.107
Disease Staging > 1	0.136	0.154	0.383	0.167	0.203	0.412
Hospital with RT	0.140	0.195	0.474	0.294	0.434	0.500
var(b)	0.145	0.116	0.214	0.125	0.174	0.476
λ	−0.001	0.657	0.998	0.471	1.182	0.692

Table 5: Stratified AOA RE model



(a) Logit

(b) AOA

(c) Stratum No RT

(d) Stratum RT

Figure 1: Bayesian empirical estimates of hospital residual effects along with the corresponding 95% confidence intervals.

data-driven within a parametric family of link functions, implies difficulties in the interpretation of parameters and therefore in understanding the influence of covariates. Moreover, estimation of the link has been shown to increase the variance of the estimated parameters and predicted probabilities. Results obtained through simulations showed that important differences in fitted values were obtained using link functions that for GLM do not lead to significant changes when probability values are not extreme. Wrong assumptions about link function have also consequences on the estimation of the within-cluster association, generally overestimating it. Therefore the appropriate choice of the link function for HGLM seems to be even more important than in GLM framework.

The analysis of a real data set on breast cancer care show that the issue is of concrete relevance. Indeed, the use of the canonical link without further investigation have the potential of leading to erroneous estimates, both point and interval, when conducted without the necessary criticism. The estimated λ using the AOA can be eventually used for model checking purposes, in particular when the sample size is limited and the risk of wrong inference is higher, as shown in section 2.2.. However the choice of fitting a model using a parametric link function should be made cautiously, taking into account the fact that the method does not seem to be appropriate for small data sets, that its use leads to loss of efficiency of estimates and that p-values associated with commonly used goodness of fit tests do not quantify appropriately changes in the fit of the model and in some specific quantities that might be of interest for the experimenter.

REFERENCES

[1] Aranda-Ordaz, F. J. On two families of transformations to additivity for binary response data. *Biometrika* **68** (1981), 357–364.

[2] Czado, C. On selecting parametric link transformation families in generalized linear models. *Journal of Statistical Planning and Inference* **61** (1997), 125–139.

[3] Oberg, A., M. Davidian Estimating data transformations in nonlinear mixed effects models. *Biometrics* **56** (2000), 65–72.

[4] Pregibon, D. Goodness of link tests for generalized linear models. *Applied Statistics* **29** (1980), 15–24.

[5] Leyland, A. H., H. Goldstein (eds.) Multilevel Modelling of Health Statistics Wiley Series in Probability and Statistics, John Wiley & Sons, New York, 2001.

[6] Locallio, A. R., J. A. Berlin, T. R. Ten Have, S. E. Kimmel Adjustments fo center in multicenter studies: an overview. *Annals of Internal Medicine* **135** (2001), 112–123.

[7] SAS Institute 1999: SAS/STAT User's Guide, Version 8. SAS Institute Inc.

[8] Stiratelli, R., N. M. Laird, J. H. Ware Random effect models for serial observations with binary response. *Biometrics* **40** (1984), 961–71.

[9] Stukel, T. A. Generalized Logistic Models. *Journal of American Statistical Association* **83** (1988), 426–431.

[10] Taylor, J. M. G., W. G. Cumberland, X. Meng Components of variance models with transformations. *Australian Journal of Statistics* **38** (1996), 183–191.

Dario Gregori
University of Torino,
Department of Public Health
and Microbiology,
Torino, Italy

Rosalba Rosato
San Giovanni Battista Hospital,
Unit of Tumor Epidemiology,
Torino, Italy

Giovannino Ciccone
San Giovanni Battista Hospital,
Unit of Tumor Epidemiology,
Torino, Italy

Lara Lusa
National Institute of Cancer,
Milano,Italy