Chilean Journal of Statistics
Vol. 9, No. 2, September 2018, 3-17

Likelihood analysis for a class of simplex mixed models

Wagner Hugo Bonat*, José Evandeilton Lopes, Silvia Emiko Shimakura and Paulo Justiniano Ribeiro Jr

Department of Statistics, Paraná Federal University , Curitiba, Brazil.

(Received: August, 2017 · Accepted in final form: August, 2018)

Abstract

This paper describes the specification, estimation and comparison of simplex mixed models based on the likelihood paradigm. This class of models is suitable to deal with restricted response variables, such as rates, percentages, indexes and proportions. The estimation of simplex mixed models is challenged by the intractable integral in the likelihood function. We compare results obtained with three numerical integration methods Laplace, Gauss-Hermite and Quasi-Monte Carlo to solve such integral. The specification of simplex mixed models includes the choice of a link function for which we compare models fitted with logit, probit, complement log-log and Cauchy link functions. Furthermore, results from the simplex mixed models fitted to two datasets are compared with fits of beta, linear and non-linear mixed effects models. The first is a study concerning life quality of industry workers with data collected according to a hierarchical sampling scheme. The second corresponds to water quality measurements taken at 16 operating hydroelectric power plants in Paraná State, Brazil. Our results showed that the simplex mixed models provide the best fit between the approaches considered for the two datasets analyzed. None of the choices of the link function outperformed the others. Simulation studies were designed to check the properties of the maximum likelihood estimators and the computational implementation. The Laplace method provides the best balance between computational complexity and accuracy. The data sets and `R` code are available in the supplementary material.

Keywords: Life quality · Likelihood · Numerical integration · Simplex distribution · Water quality.

1. Introduction

Statistical regression models are used to establish relations between explanatory and response variables in many fields of science. In particular, the linear regression model is probably the most used statistical method in the literature. Limitations of this model to deal with variance heterogeneity and discrete response variables motivated developments resulting in the class of generalized linear models (GLM) (Nelder and Wedderburn, 1972). GLMs extend linear regression models to deal with distributions of response variables belonging to the exponential family. Based on the ideas laid by Nelder and Wedderburn (1972), statistical modelling literature has grown quickly, mainly to deal with binary, binomial and count data.

In spite of its flexibility the standard GLM family has no suitable distributions to model restricted response variables such as rates, percentages, indexes and proportions. In this

*Corresponding author. Email: wbonat@ufpr.br

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

situation a frequently adopted approach is to use the beta regression model as proposed by Paolino (2001), Cepeda (2001), Kieschnick and McCullough (2003) and Ferrari and Cribari-Neto (2004) in the context of independent observations. We are interested in situations where the observations cannot be assumed independent such as for repeated measurements and longitudinal data analysis. We follow an approach similar to the analysis of dependent data in the class of generalized linear mixed models (GLMM), a natural extension of GLMs obtained by adding Gaussian random effects to the linear predictor.

To deal with the combination of restricted response variables and dependent observations Bonat et al. (2015) discuss the inference for the class of beta mixed models. Through a couple of data sets, the authors showed that beta mixed models provide a better fit than orthodox linear mixed models and non-linear mixed models. Similar models were proposed by Figueroa-Zúñiga et al. (2013) and Bonat et al. (2015) but with inference based upon a Bayesian paradigm.

In this paper we present an alternative model to deal with a combination of restricted response variables and dependent observations based on the simplex distribution (Barndorff-Nielsen and Jørgensen, 1991; Jørgensen, 1997). Such distribution is a natural choice to model restricted response variable, since its domain is the unit interval. The simplex distribution is a flexible two parameters distribution with a diversity of shapes for the distribution function and, as a member of the exponential dispersion family, can be parametrised with orthogonal parameters.

The literature about simplex regression models is sparse, probably Kieschnick and McCullough (2003) was the first to use the simplex distribution in the context of regression models. The simplex regression model is implemented by the `simplexreg` package (Zhang et al., 2014) for the `R` environment for statistical computing (R Core Team, 2015). Bonat et al. (2012) proposed a comprehensive approach to specify regression models for response variables in the unit interval, where the simplex regression model is a special case. López (2013) presents Bayesian inference for simplex regression models and a comparison with beta regression models.

Song (2007) presents simplex regression models in the context of correlated data adopting generalized estimating equations and quadratic estimating equations for estimation. Qiu et al. (2008) uses *penalized quasi-likelihood* and *restricted maximum likelihood* for estimation of simplex mixed models in the context of longitudinal data analysis. Zhang and Wei (2008) propose simplex mixed models with likelihood based inference using an stochastic approximation algorithm. The likelihood function for simplex mixed models requires the solution of an analytically intractable integral. Here, we adopt an approach based on the marginal likelihood function and use two datasets to assess three numerical methods to solve the integral: Laplace, Gauss-Hermite and Quasi-Monte Carlo. Four link functions are considered: `logit`, `probit`, `complement log-log` and `cauchy`. Additionally, we compare the results obtained with the simplex mixed model, with the ones obtained with beta, linear and non-linear mixed models. The first dataset is a study concerned with the life quality index of industry workers with data collected according to a hierarchical sampling scheme. The second corresponds to water quality indicators measured quarterly at 16 operating hydroelectric power plants in Paraná State, Brazil. Both were previous analysed by Bonat et al. (2015) using beta mixed models. Furthermore, simulation studies were designed to check the properties of the maximum likelihood estimators and the computational implementation.

Section 2 presents the simplex mixed model. Section 3 presents the estimation procedure, including a review on the numerical integration methods. Section 4 presents the results of the simulation study. Section 5 summarizes the results of the data analyses. Finally, Section 6 provides some discussions and recommendations for future work. `R` code and

data sets are provided in the supplementary material.[1]

2. Simplex mixed models

We specify the simplex mixed model with a structure similar to a generalized linear mixed model. The random variable $Y \in (0,1)$ is said to follow a simplex distribution, $Y \sim S^-(\mu, \sigma^2)$, with the probability density function given by

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2[y(1-y)]^3}} \exp\left\{-\frac{1}{2\sigma^2} d(y,\mu)\right\},$$

where $\mu \in (0,1)$ and $\sigma^2 > 0$ are the mean and the dispersion parameters, respectively; and

$$d(y,\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}$$

is the unit deviance function. For specification of simplex mixed models, let y_{ij} be the observation $j = 1,\ldots,n_i$ within the unit sample $i = 1,\ldots,N$ of the random variable Y_{ij}. A hierarchical description for the simplex mixed model is the following:

$$Y_{ij}|\boldsymbol{b}_i \sim S^-(\mu_{ij}, \sigma^2)$$
$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top\boldsymbol{b}_i.$$

The model assumes that the observations from the response variable Y_{ij} are conditionally independent given a q-dimensional vector of Gaussian random effects, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. The linear predictor is linked to the mean by a link function g and consists of the sum of fixed effects $\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}$ and random effects $\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i$. The vectors $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ contain values of p and q covariates, respectively. Finally, $\boldsymbol{\beta}$ is a vector of p regression parameters.

In the context of simplex mixed models the link function $g : (0,1) \to \Re$ plays an important role, since it links the linear predictor to the mean of the response variable. The `logit` link function is a frequent choice. In this paper along with the `logit` link function we investigate and compare the fit of some alternative functions, such as `probit`, complement log-log (`clog-log`) and `Cauchy`. Table 1 presents the expressions for each link function, its inverse and first derivative. Here η denotes the linear predictor, Φ is the cumulative distribution function of the standard Gaussian distribution, tan and csc are the tangent and cosecant functions, respectively.

Table 1. Expressions related to the link functions `logit`, `probit`,`clog-log` and `Cauchy`.

Link function	$g(\mu)$	$g^{-1}(\eta)$	$g'(\mu)$
`Logit`	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{\exp^\eta}{1+\exp^\eta}$	$\mu(1-\mu)^{-1}$
`Probit`	$\Phi^{-1}(\mu)$	$\Phi(\eta)$	$\sqrt{2\pi}\exp(\mu^2/2)$
`Clog-log`	$\log(-\log(1-\mu))$	$1-\exp(-\exp\eta)$	$(\mu-1)\log(1-\mu)^{-1}$
`Cauchy`	$\tan(\pi(\mu-\frac{1}{2}))$	$\pi\csc^2(\pi\eta)$	$\pi\csc^2(\pi\cdot\mu)$

[1] Available at www.leg.ufpr.br/doku.php/publications:simplexmix

3. Estimation of simplex mixed models

The marginalised likelihood is used to estimate model parameters. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma^2)^\top$ be the parameter vector and $\boldsymbol{y}_i$ be an n_i-dimensional vector of measurements from the i^{th} sample unit. The contribution to the likelihood of each independent sample unit is

$$L_i(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma^2; \boldsymbol{y}_i) \equiv \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma^2) f(\boldsymbol{b}_i|\boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{b}_i.$$

Therefore, the marginal likelihood function is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma^2; \boldsymbol{y}_i) = \prod_{i=1}^{N} L_i(\boldsymbol{y}_i; \boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma^2). \tag{1}$$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is obtained by maximization of the log-likelihood function (1). Note that for each evaluation of this function we need to numerically solve N q-dimensional integrals.

The two numerical methods are applied, one within each step of the other: the integration of the random effects within each step of the maximization for parameter estimation. We use the `BFGS` algorithm implemented in the `R` (R Core Team, 2015) function `optim()` for the maximization procedure. The numerical integration plays an important role in the estimation of simplex mixed models, since it will be computed many times within the numerical maximization algorithm. When the dimension of the random effects is low, say $q \leq 5$, a frequent choice is the Gauss-Hermite method. In this paper, along with Gauss-Hermite method we also use the Laplace and Monte Carlo methods. These methods were chosen because each one uses a different approach to solve the integral. The Gauss-Hermite method is based on a quadrature procedure, basically it means that the integral will be approximated by a finite sum. The Monte Carlo method uses samples from the integral to approximate it as an expectation. Finally, the Laplace approximation uses a Taylor series expansion to approximate the integrand by a function analytically tractable. In what follows, we provide a short description of these methods.

3.1 Gauss-Hermite

The Gauss-Hermite method has been designed to approximate integral as follows

$$\int_{\Re} \exp(-x^2) f(x) \mathrm{d}x \approx \sum_{i=1}^{n} w_i f(g_i),$$

where n is the number of points used for the approximation, g_i's are roots of the Hermite polynomial $H_n(g) (i = 1 < 2, \ldots, n)$ and w_i are weights given by,

$$w_i = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 [H_n(g_i)]^2}.$$

The Gauss-Hermite method approximates the integral by a weighted sum of the function evaluated at the Gauss-Hermite points and integration weights. It is easily implemented in `R`, as the function `gauss.quad()` from package `statmod` (Smyth et al., 2013) provides the weights and the Gauss-Hermite points.

3.2 Laplace approximation

The Laplace approximation method Tierney and Kadane (1986) has been designed to approximate integrals as follows

$$\int_{\Re} \exp\{Q(x)\}\mathrm{d}x \approx (2\pi)^{\frac{q}{2}} |Q''(\hat{x})|^{-\frac{1}{2}} \exp\{Q(\hat{x})\},$$

where $Q(x)$ is a known, unimodal and bounded function of a q-dimensional variable x. Let $\hat{x}$ be the value for which $Q(x)$ is maximized. The method requires obtaining $\hat{x}$ the maximum of the integrand and the Hessian $Q''(\hat{x})$, the matrix of second derivatives, either analytically or numerically. The latter is used here.

3.3 Monte Carlo and Quasi-Monte Carlo Integration

The Monte Carlo method Pan and Thompson (2007) has been proposed to estimate the value of integrals written as an expectation. Suppose we want to estimate the integral of a given function $f(x)$ whose domain is the real line $\Re$. Let $p(x)$ be a probability density function in the same domain. The following equation suggests an estimator for the value of the integral

$$\int_{\Re} f(x)dx \approx \int_{\Re} \frac{f(x)}{p(x)} p(x)\mathrm{d}x,$$

since it is equivalent to $\mathrm{E}\left(\frac{f(x)}{p(x)}\right)$ with respect to the density $p(x)$. The expectation is estimated by generating random numbers according to $p(x)$, computing $f(x)/p(x)$ for each sample and averaging the values. The number of samples determines the accuracy of the estimator. A natural choice for $p(x)$ is the standard Gaussian distribution, since the domain of this distribution is the real line.

The Monte Carlo method is quite easy to use, but in the context of simplex mixed models it is computed within a numerical maximization process. In that case the Monte Carlo method presents an inconvenient problem, since it is based on simulations, it can return different values for the integral evaluated at the same points. This fact can slow down or even prevent convergence of the maximization process.

To overcome this problem the Quasi-Monte Carlo method suggests to change the simulated values by a low-discrepancy sequence (Pan and Thompson, 2007). The package `fOptions` (Wuertz, 2012) for `R` (R Core Team, 2015) has routines to obtain such sequences by two methods, *Halton* and *Sobol*. In this paper we choose to report the results based on the *Halton* method, since for most fitted models the results were identical. Generic functions to use these methods in the context of simplex mixed models are provided in the supplementary material.

4. Simulation study

In this section we present a simulation study to verify the properties of the maximum likelihood estimators and the computational implementation. We generated 500 data sets considering 25 measures taken at an increasing number of subjects (10, 20 and 40) resulting in samples of size 250, 500 and 1000, respectively. We fitted the models using the three integration methods, namely, Laplace (LA), Gauss-Hermite (GH) and Quasi-Monte Carlo (QMH). The number of integration points used by the Gauss-Hermite method was fixed

at 100. Similarly, the number of samples used by the Quasi-Monte Carlo method was fixed at 200. These numbers were fixed based on preliminaries fits. We also ran the simulation study using different numbers of integration points or sample sizes. For the Gauss-Hermite method, we also considered 150 and 200 integration points. For the Quasi-Monte Carlo, we also considered 500 and 1000 sample sizes. For both methods, the results obtained by using the different numbers of integration points or sample sizes were really similar. Thus, we opted to report the results for the values aforementioned, i.e. 100 integration points for the GH method and 200 samples for the QMH method.

We use the `logit` link function and consider models with an intercept ($\beta_0 = 0.5$) and slope ($\beta_1 = 0.3$). The covariate is a sequence from -1 to 1. In order to explore the proportion of the variance coming from each of the two random components in the model, we designed three simulation scenarios. The scenarios 1, 2 and 3 assume that $\sigma^2 = 0.75$ and $\sigma_I^2 = 0.25$, $\sigma^2 = 0.5$ and $\sigma_I^2 = 0.5$ and $\sigma^2 = 0.25$ and $\sigma_I^2 = 0.75$, respectively. In this notation, σ^2 represents the dispersion parameter associated with the simplex distribution and σ_I^2 denotes the variance associated with the Gaussian random effect. In this way, we have a often case of simplex models with random intercepts. Figure 1 shows the expected bias plus and minus expected standard error for the parameters in each scenario.

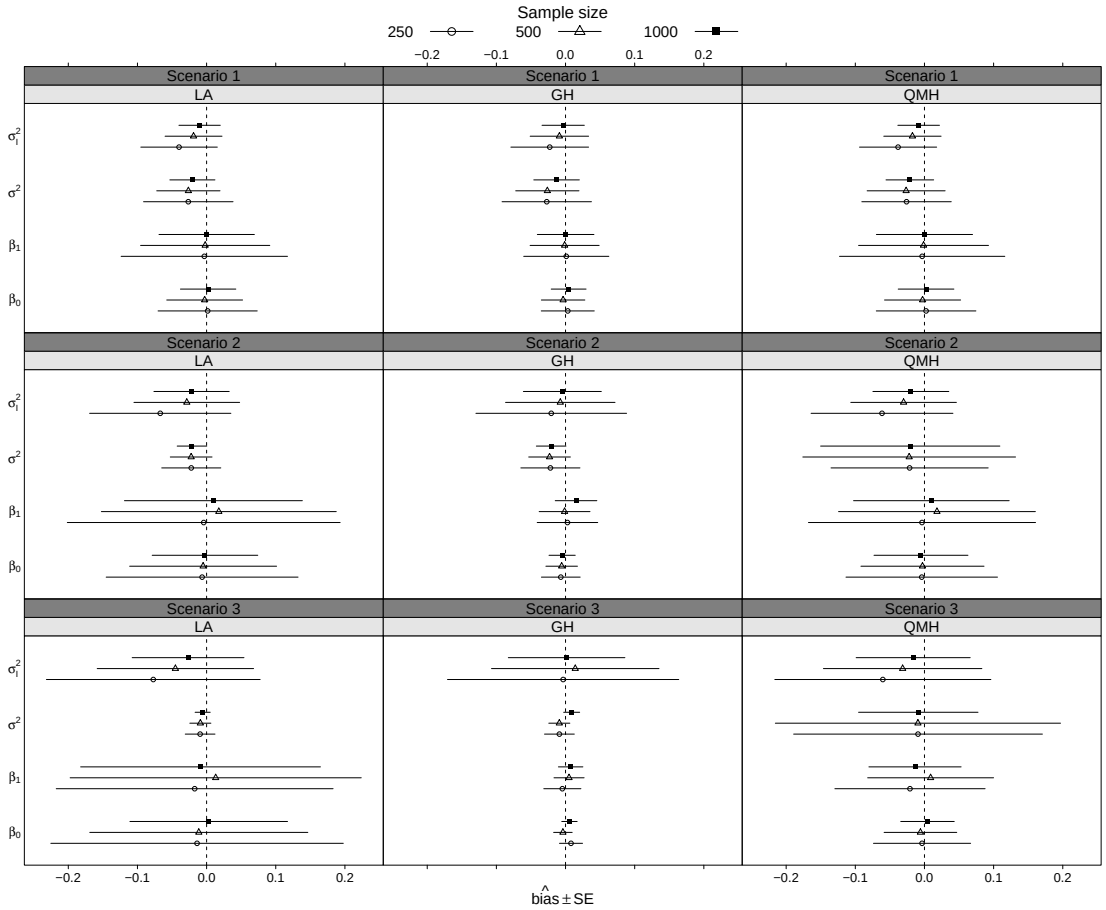


Figure 1. Expected bias plus and minus expected standard error for each parameter and simulation scenario.

The results in Figure 1 show that for all scenarios both, the expected bias and expected standard error, tend to zero as the sample size is increased. In general all methods tend to underestimate the parameters in the covariance structure for small sample sizes whilst improving for increasing sample sizes. The biases are close to zero for large samples, as expected. It is clear the decreasing of the values of expected standard errors for increasing

sample sizes, indicating the consistency of our estimators. In general the methods agree in terms of estimates and disagree in terms of standard errors mainly for the ones associated with the regression parameters.

The scenario 1 presents the easiest case for the estimation algorithm, since the proportion of the variance from the Gaussian random effect is small. In that scenario the three integration methods present estimates and standard errors similar mainly for the parameters that describe the covariance structure (σ^2 and σ_I^2). For the regression parameters the three integration methods provide similar estimates. However, the `GH` method presents expected standard errors smaller than the ones obtained by the `Laplace` and `QMH` methods.

In the scenario 2 the proportion of variance from the Gaussian random effect increases, thus the differences between the integration methods also appear clearer. The expected standard errors for the regression parameters obtained by the `GH` method are smaller than the ones obtained by the `Laplace` and `QMH` methods. Finally, in the scenario 3 the most proportion of the variance comes from the Gaussian random effect, thus the expected standard errors computed by the `Laplace` and `QMH` methods for the regression parameters increase as expected. On the other hand, the standard errors for the regression parameters computed by the `GH` method are really small. It gives us reason to believe that the `GH` method underestimate the standard errors for the regression parameters. To verify it, Table 2 presents the coverage rate for each simulation scenario, sample size and integration method. The nominal level of confidence was fixed at 95%.

Table 2. Coverage rate by simulation scenario, sample size and integration method.

	Methods								
	`Laplace`			`GH`			`QMH`		
	250	500	1000	250	500	1000	250	500	1000
	Scenario 1								
β_0	0.8654	0.9375	0.9393	0.6600	0.6882	0.6960	0.8694	0.9334	0.9432
β_1	0.9056	0.9233	0.9313	0.7000	0.6902	0.7140	0.9116	0.9193	0.9249
σ^2	0.8895	0.8991	0.8404	0.8880	0.8967	0.8380	0.8915	0.8971	0.8377
σ_I^2	0.8121	0.8770	0.8868	0.8540	0.8987	0.9220	0.8232	0.8850	0.9026
	Scenario 2								
β_0	0.9117	0.9245	0.9144	0.4596	0.3870	0.4460	0.8016	0.8526	0.8520
β_1	0.8932	0.9116	0.9239	0.5320	0.5570	0.5000	0.8243	0.8571	0.8830
σ^2	0.8644	0.8620	0.8242	0.8633	0.8545	0.8235	0.8698	0.8861	0.8568
σ_I^2	0.7905	0.8620	0.8741	0.8716	0.9082	0.9117	0.7954	0.8616	0.8758
	Scenario 3								
β_0	0.9083	0.9184	0.9315	0.2926	0.2500	0.2222	0.5041	0.5093	0.5598
β_1	0.8891	0.9090	0.9070	0.3684	0.3490	0.2729	0.6193	0.6261	0.6339
σ^2	0.8891	0.8624	0.8117	0.8989	0.8773	0.8019	0.9074	0.8785	0.8588
σ_I^2	0.8123	0.8694	0.8973	0.8989	0.9386	0.9299	0.8106	0.8785	0.9138

The results presented in Table 2 show that the coverage rate for the regression parameters are sensitive to the simulation scenarios. In the scenario 1 both `Laplace` and `QMH` presented coverage rate for the regression parameters close to the nominal level (95%). However, the results were getting worst from scenario 1 to 3 for the `QMH` method. The `Laplace` method is the unique where the coverage rate is similar in all simulation scenarios. The `GH` method presents the worst results in terms of coverage rate for the regression parameters in all simulation scenarios. The coverage rate for the covariance parameters were slightly below the nominal level for all integration methods and simulation scenarios.

5. Data analyses

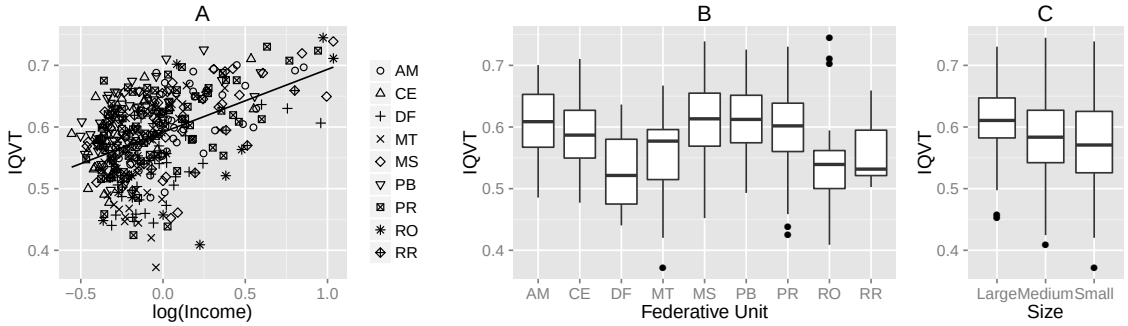
5.1 Income and life quality of Brazilian industry workers

The first dataset is from a poll realized by the Industry Social Service (Serviço Social da Indústria - SESI) to assess important factors associated with the workers' life quality. The data were collected in 2010 following a sampling plan developed by SESI, using an specific questionnaire and included eight Brazilian States and the Federal District. The dataset corresponds to observations from 365 companies. The response variable (IQVT) is an index that measures the worker's life quality in the industry. This index is computed following the same criterion adopted by the United Nations (UN) to compute Human Development Index (HDI). The resulting values are in the unit interval and the closer to one the higher the workers life quality.

The data analysis considers two covariates related to the companies for which the impact on IQVT is of particular interest, namely, company's size and average income. The first can be related to the capability of managing and providing life quality to the workers. The second is given by the total of salaries divided by the number of workers expressing the capacity to fulfill individual basic needs such as food, health, housing and education. The income is expressed in logarithmic scale centred around the average.

The main goal is to specify a suitable regression model to evaluate the influence of these two covariates on the IQVT. The federative unit where the company is based is expected to influence the index considering varying local legislation, taxing and further economic and political conditions. This is accounted for by including a random effect, regarding the eight States as a sample of the country's federative units.

Figure 2. IQVT related to (centred log) average income, company size and federative unit.



Plots in Figure 2 suggest that IQVT is associated with income (A), location (B) and size (C) and the following simplex mixed model (SMM) is specified for the IQVT data:

$$
\begin{aligned}
Y_{ij}|\boldsymbol{b}_i &\sim S^{-}(\mu_{ij}, \sigma^2) \\
g(\mu_{ij}) &= (\beta_0 + b_{i1}) + \beta_1 Medium_{ij} + \beta_2 Small_{ij} + (\beta_3 + b_{i2}) Income_{ij} \\
\boldsymbol{b}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}) \text{ with } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_I^2 & \rho(\sigma_I \sigma_S) \\ \rho(\sigma_I \sigma_S) & \sigma_S^2 \end{bmatrix}.
\end{aligned}
$$

The model is parametrised such that β_0 is associated with large size companies and β_1 and β_2 are differential effects for medium and small size companies, respectively. A random intercept b_{i1} and slope b_{i2} associated with *income* account for the effect of the federative units. Model parameters to be estimated are the regression coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$, the random effects covariance parameters $(\sigma_I^2, \sigma_S^2, \rho)$ and the dispersion parameter σ^2.

The final model is chosen after fitting and comparing a sequence of nested models.

Model 1 has only the intercept. The covariates *size*, *income*, the random intercept and the random slope associated to *income* are sequentially added defining Models 2 to 5. The models with random effects are initially fitted using Laplace approximation and a `logit` link function. We also fit similar models using the beta mixed model (`BMM`) proposed by Bonat et al. (2015), the orthodox linear mixed model (`LMM`) and the non-linear mixed model (`NLMM`) using a `logit` function. Table 3 shows parameter estimates for the `SMM` and the maximised log-likelihood values for `BMM`, `LMM` and `NLMM`.

Table 3. Parameter estimates and standard errors for the simplex models (top) and maximised log-likelihood for the alternative models (bottom) for the IQVT data.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
β_0	0.3474	0.4468	0.4551	0.4213	0.4230
β_1		−0.1071	−0.0879	−0.0704	−0.0710
β_2		−0.1632	−0.1446	−0.1335	−0.1355
β_3			0.4213	0.4726	0.4680
σ^2	0.3262	0.3095	0.2404	0.1814	0.2313
σ^2_I				0.0161	0.0164
σ^2_S					0.0004
ρ					0.9832
Model	Maximised log-likelihood				
`SMM`	473.0003	482.5772	528.6662	570.3376	570.4720
`BMM`	472.1979	481.5058	526.9421	561.7935	561.7954
`LMM`	470.4207	479.9673	523.8509	558.8973	558.9004
`NLMM`	470.4207	479.9673	523.7791	558.9608	558.9608

The results in Table 3 show that Model 4 provides the best fit. The gain in likelihood values from model 4 to model 5 do not justify the addition of the random slope. This result is reassured by the estimates of σ^2_S and ρ. These estimates indicate that the random intercept alone captures the extra variability induced by the repeated measures structure. Model 4 shows that the covariates *size* and *income* have a significant effect on the response variable IQVT. In general the IQVT increases when the income increases and decreases 2.89% and 5.66% for medium and small companies, respectively, in comparison with large companies.

The maximised log-likelihood values also show that the SMM provides a better fit than the alternative BMM, LMM and NLMM. The difference is more pronounced for models with random effects. All modelling strategies identify Model 4 as the best fit.

In order to check the effect of different numerical integration methods and link functions, we refit Model 4 using Laplace, Gauss-Hermite (`GH`) and Quasi-Monte Carlo (`QMC`) methods combined with four link functions, `logit`, `probit`, `clog-log` and `Cauchy`. To reach the required accuracy for `GH` and `QMC` methods we used $n = 380$ integration points. The maximised log-likelihood values are given in Table 4 and show that all numerical integration methods provide similar results. The `GH` method seems to be more sensible to the choice of the link function, while the `Laplace` and `QMC` present similar values for all link functions. The choice of the link function has no effect for this data set.

Finally, Table 5 presents estimates and standard errors obtained by using the `LAPLACE`, `GH` and `QMC` methods for the Model 4.

The results in Table 5 show that the `Laplace` and `QMC` methods provide virtually the same estimates and standard errors. The `GH` method present smaller estimates and standard errors for the regression parameters and larger estimate and standard error for the variance of the random effect, although the differences are small in their magnitude. The standard error computed by the `GH` method for the intercept (β_0) is 43.15% smaller than the ones computed by the `Laplace` method. These results agree with our simulation study

Table 4. Maximised log-likelihood values by different numerical integration methods and link functions fitting the Model 4 for the IQVT data.

Link functions	Methods		
	`Laplace`	`GH`	`QMC`
`logit`	570.3376	571.9386	570.5104
`probit`	570.3349	570.7279	570.4370
`clog-log`	570.2071	572.5328	570.2071
`cauchy`	570.2286	565.1114	570.3773

Table 5. Estimates and standard errors obtained by different integration methods for model 4 - IQVT data.

Parameter	Methods		
	`Laplace`	`GH`	`QMC`
β_0	0.4213(0.0482)	0.4026(0.0208)	0.4232(0.0476)
β_1	−0.0704(0.0265)	−0.0655(0.0262)	−0.0706(0.0265)
β_2	−0.1335(0.0290)	−0.1304(0.0285)	−0.1338(0.0289)
β_3	0.4726(0.0377)	0.4651(0.0350)	0.4718(0.0377)
σ	0.1814(0.0134)	0.1813(0.0134)	0.1814(0.0134)
σ_I	0.0161(0.0082)	0.0215(0.0102)	0.0162(0.0083)

showing that in general the `GH` method underestimate the standard error for the regression parameters. We highlight that for this data analysis such underestimation is weak, since the proportion of the variance coming from the Gaussian random effect is small.

5.2 Water quality on power plant reservoirs

This example is concerned with water quality indicators measured quarterly at 16 operating hydroelectric power plants during 2004 in Paraná State, Brazil. The water quality indicators are: dissolved oxygen, temperature, faecal coliform, water pH, biochemical oxygen demand (DBO), total nitrogen, total phosphorus, turbidity and total solids. The indicators are combined to produce a single water quality index (IQA, acronym in Portuguese) based upon a study conducted in the 70's by the US National Sanitation Foundation and adapted by the Brazilian company CETESB - Companhia de Tecnologia de Saneamento Ambiental.

Monitoring aims to detect changes in water quality, possibly attributable to the presence of the dams. Water quality measurements taken at locations considered directly affected and unaffected by the reservoir are compared. More specifically, measurements taken upstream the main river are considered unaffected reference values whereas measurements taken at the reservoir and downstream are considered potentially affected by the water contention and passage throughout the power plant. The main interest is the covariate `LOCAL`, with levels `upstream`, `reservoir` and `downstream` controlled for the effects of the `power plant` and the `QUARTER` of data collection. The dataset has 190 observations with 12 measurements (4 quarters × 3 locations) for each of the 16 power plants with only two missing data.

Plots in Figure 3 summarises and relate the IQA data to the potential covariates. Figure 3(A) shows a left asymmetry typical for this kind of data. Figure 3(B) suggests a significant variation between power plants. In a similar way Figure 3(C) suggests that observations `upstream` present smaller values than at the `reservoir` and `downstream`. Finally, Figure 3(D) shows smaller values on the first and fourth quarters (warmer periods), a pattern expected to be repeated over the years. Based on this exploratory analysis we propose that IQA at the i^{th} relative location, j^{th} power plant and t^{th} quarter be modelled

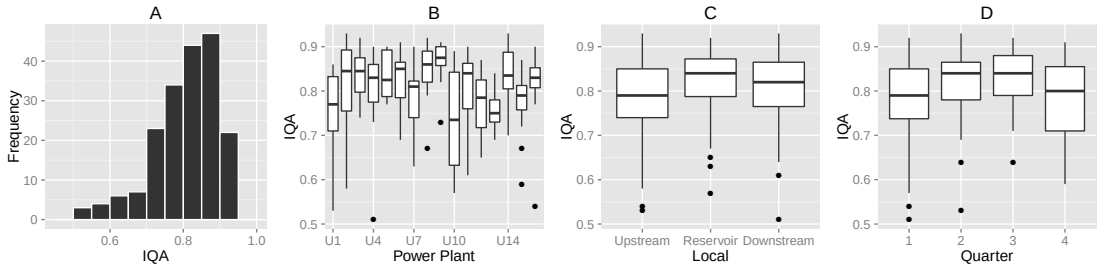


Figure 3. Summaries for the IQA data.

by the following simplex mixed model:

$$Y_{ijt}|b_j \sim S^{-}(\mu_{ijt}, \phi)$$
$$g(\mu_{ijt}) = \beta_0 + \beta_{1i} + \beta_{2t} + b_j$$
$$b_j \sim N(0, \sigma_U^2).$$

Under the adopted parametrization, β_{1i}, $i = 2, 3$ quantifies the changes from `upstream` to `reservoir` and `downstream`, respectively. Likewise β_{2t}, $t = 2, 3, 4$ are differences between the first quarter and the others. The random intercept b_j captures the deviations of each power plant from the overall mean. Following the previous example, we fitted a sequence of nested models. Model 1 is the null model with just the intercept. Model 2 includes the covariate `LOCAL` and Model 3 the covariate `QUARTER`. Model 4 adds a random intercept associated with the power plants. We fitted the Model 4 using the Laplace approximation and `logit` link function. Furthermore, we fitted the same set of models using the `BMM`, `LMM` and `NLMM`, using the `logit` link function where applicable. Table 6 shows the parameter estimates for the `SMM` along with the maximised log-likelihood values for the `BMM`, `LMM` and `NLMM`.

Table 6. Parameter estimates and standard errors for the simplex models (top) and maximised log-likelihood by alternative models (bottom) for the IQA data.

Parameter	Model 1	Model 2	Model 3	Model 4
β_0	1.3934	1.2599	1.1028	1.1497
β_{12}		0.2517	0.2518	0.2128
β_{13}		0.1614	0.1622	0.1712)
β_{22}			0.2393	0.1518
β_{23}			0.3519	0.3146
β_{24}			0.0710	−0.0127
σ^2	1.9496	1.8654	1.7183	1.4130
σ_U^2				0.0269
Model	Maximised log-likelihood			
`SMM`	220.9769	224.9744	232.4383	246.3474
`BMM`	215.3817	218.9064	224.6237	231.0469
`LMM`	198.2386	202.1282	208.6868	213.6839
`NLMM`	198.2386	202.1282	208.7399	214.8852

The results in Table 6 show that the two covariates `LOCAL` and `QUARTER` have a significant effect on IQA levels. It is also clear that the random effect associated with the power plant improves the model fit. Based on Model 4 we conclude that the IQA increases from `upstream` to `reservoirs` and `downstream` by 4.83% and 3.5%, respectively. The IQA increases from the first to second and third quarters and decreases from the first to

the fourth quarter, although the last difference is not significant. These results are also confirmed by the alternative models.

The maximised log-likelihood values indicate that the `SMM` provides the best fit among the fitted models. It is interesting to highlight that the simplex regression model without random effects already provides a better fit than `BMM`, `LMM` and `NLMM`. Table 7 presents the maximised log-likelihood values obtained by fitting Model 4, using different link functions and numerical integration methods. For this dataset we used $n = 160$ integration points. As in the previous example, all numerical integration methods provide similar values for the maximised log-likelihood function. The `GH` method seems to be more sensible to the choice of the link function while `Laplace` and`QMC` methods always return similar values. Yet again, the link function, has no effect.

Table 7. Maximised log-likelihood values by numerical integration methods and link functions fitting the Model 4 for the IQA data.

Link	Methods		
functions	`Laplace`	`GH`	`QMC`
`logit`	246.3474	246.9509	246.7217
`probit`	246.2955	247.7303	246.6783
`clog-log`	246.2094	248.4117	246.5769
`cauchy`	246.4287	246.7981	246.8354

Table 8 presents estimates and standard errors obtained by the different integration methods for the Model 4.

Table 8. Estimates and standard error for Model 4 by different integration methods for the IQA data.

Parameter	Methods		
	`Laplace`	`GH`	`QMC`
β_0	1.1497(0.1046)	1.1564(0.0990)	1.1521(0.1059)
β_1	0.2128(0.0884)	0.2086(0.0882)	0.2126(0.0886)
β_2	0.1713(0.0827)	0.1688(0.0821)	0.1704(0.0828)
β_3	0.1518(0.1073)	0.1462(0.1069)	0.1509(0.1075)
β_4	0.3146(0.0960)	0.3122(0.0953)	0.3142(0.0963)
β_5	−0.0127(0.1027)	−0.0163(0.1026)	−0.0120(0.1035)
σ	1.4130(0.1449)	1.4078(0.1444)	1.4059(0.1444)
σ_I	0.0269(0.0147)	0.0295(0.0158)	0.0297(0.0161)

Similarly we have seen for the IQVT data, the results in Table 8 show that the three integration methods provide similar estimates, but differ slightly in standard errors. The `GH` method provides smaller estimates (except for the intercept) and standard errors for the regression parameters than the `Laplace` method, but such differences are really small in their magnitude.

6. Discussion

This paper reports results of analysis using simplex mixed models under likelihood based inference. We have described how to specify, fit and compare simplex mixed models by analysing two datasets. Model specification includes the choice of a link function for which we consider the logit, probit, complement-log log and Cauchy. The choice of the link function has no effect on the model fitting measures and related inferences.

The estimation of simplex mixed models involves solving an intractable integral when evaluating the likelihood function. Three numerical approaches to solve such integral were

considered, the Laplace, Gauss-Hermite and Quasi-Monte Carlo methods. Such choices are justified by the fact that each of these methods use different ways to solve the integral. The first approximates the integrand, the second uses a finite sum and the third is based on the concept of the expectation of a function. In spite of all numerical integration methods providing similar results in our data examples, the convenience of using each one is not the same. In our data analyses the Gauss-Hermite method proved hard to be tuned in terms of the number of integration points, resulting in substantially different values for the maximised log-likelihood according to the number of integration points. It is important to highlight that in the context of simplex mixed models is not possible to obtain a closed-form expression for the Fisher information matrix. Thus, we replace it by the observed information matrix obtained numerically using the Richardson method (Lindfield and Penny, 1989; Gilbert and Varadhan, 2012). However, when computing such approximation by using the Gauss-Hermite and Monte Carlo methods to approximate the log-likelihood function, we detected that for the number of integration points and samples used, the numerical approximations were not accurate, mainly for the components associated with the regression coefficients. It implies that the standard errors associated with the regression coefficients are underestimated. Furthermore, the simulation study also confirmed that the Gauss-Hermite and Monte Carlo methods can strongly underestimate the standard errors associated with the regression parameters.

For the IQVT analysis, we needed $n = 380$ integration points to reach a value comparable to the Laplace method. In general this method seems to underestimate the maximised log-likelihood value. Similar issues appear in the Quasi-Monte Carlo method with the additional problem that in our data analyses this method showed to be really sensitive to the initial values. Furthermore, we combined different integration methods with different link functions. The issues above appear more frequently when combining probit link function with Gauss-Hermite and Quasi-Monte Carlo methods. Based on our experience fitting the models shown here, we recommend to use the combination Laplace approximation and logit link function when fitting simplex mixed models. An additional advantage of the Laplace approximation is that it can be used when the dimension of the random effects is high, for example in the case of times series or spatial data (Bonat and Ribeiro Jr, 2016).

We compared the simplex mixed models with the recently proposed beta mixed models and also with linear and non-linear mixed models using the logit link function. Maximised log-likelihood values are substantially higher for the simplex mixed models. For maximization of the approximated log-likelihood function we used the `BFGS` algorithm implemented in the `R` function `optim`. We have also used alternative numerical maximization methods such as Nelder-Mead and Conjugate Gradient with similar results.

The estimation of simplex mixed models is a complex numerical problem, since many numerical algorithms are involved in the procedure. For all models presented in this paper we used the strategy of obtaining the profile likelihood. Although such technique is not a convergence check, in the present case it provides a more detailed exploration of the log-likelihood function preventing against local maximum or non-convergence. Details about how to implement the profile likelihood in `R` can be found in (Bolker and R Core Team, 2014). We provide the `R` code and the data sets in the supplementary material.

Possible topics for further investigation and extensions include designing simulation studies comparing the simplex and beta mixed models. The simulation studies presented in Section 3 showed that the maximum likelihood estimators are slightly biased for the variance parameters. Thus, a topic for future investigation is to extend the restricted maximum likelihood method (Noh and Lee, 2007) to the class of simplex mixed models presented in this paper. Another interesting point for future reasearch is to propose tools for model checking as the quantile-quantile plots with simulated envelopes (Moral et al., 2017).

Acknowledgements

We thanks Milton Matos de Souza and Sonia Beraldi de Magalhães from *Serviço Social da Indústria (SESI)* for the IQVT data. We also thanks the Paraná Energy Company (COPEL) and Nicole M. Brassac de Arruda from the *Instituto de Tecnologia para o Desenvolvimento - LACTEC* for the IQA data. The second author is supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) - Brazil. Anonymous referee reports provided very helpful suggestions to improve the manuscript.

References

Barndorff-Nielsen. O.E., and Jørgensen, B. (1991). Some parametric models on the simplex. Journal of Multivariate Analysis 39, 106-116.

Bolker, B., and R Core Team (2014). bbmle: Tools for general maximum likelihood estimation. R package version 1.0.17. URL: CRAN.R-project.org/package=bbmle

Bonat, W.H., Ribeiro Jr, P.J., and Zeviani, W.M. (2012). Regression models with response on the unit interval: Specification, estimation and comparison. Biometric Brazilian Journal 30, 415-431.

Bonat, W.H., Ribeiro Jr, P.J., and Zeviani, W.M. (2015). Likelihood analysis for a class of beta mixed models. Journal of Applied Statistics 42, 252-266.

Bonat, W.H., Ribeiro Jr, P.J., and Shimakura, S.E. (2015). Bayesian analysis for a class of beta mixed models. Chilean Journal of Statistics 6, 3-13.

Bonat, W.H., and Ribeiro Jr, P.J. (2016). Practical likelihood analysis for spatial generalized linear mixed models. Environmetrics 27, 83-89.

Cepeda, E. (2001). Variability Modeling in Generalized Linear Models. Unpublished Ph.D. Thesis. Mathematics Institute, Universidade Federal do Rio de Janeiro, Brazil.

Ferrari, S.L.P, and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. Journal of Applied Statistics 31, 799-815.

Figueroa-Zúñiga, J.I. and Arellano-Valle, R.B., and Ferrari, S.L.P. Mixed beta regression: A Bayesian perspective. Computational Statistics and Data Analysis 61, 137-147.

Gilbert, P., and Varadhan, R. (2012). numDeriv: Accurate numerical derivatives. R package version 2012.9-1. URL: CRAN.R-project.org/package=numDeriv

Jørgensen, B. (1997). The Theory of Dispersion Models. Chapman & Hall, London.

Kieschnick, R., and McCullough, B.D. (2003). Regression analysis of variates observed on $(0, 1)$: percentages, proportions and fractions. Statistical Modelling 3, 193-213.

Lindfield, G.R., and Penny, J.E.T. (1989). Microcomputers in Numerical Analysis. Halsted Press, New York.

López, F.O. (2013). A Bayesian approach to parameter estimation in simplex regression model: A comparison with beta regression. Revista Colombiana de Estadística 36, 1-21.

Moral, R.A., Hinde, J., and Demétrio, C.G.B. (2017). Half-normal plots and overdispersed models in R: The hnp Package. Journal of Statistical Software 81, 123. doi: 10.18637/jss.v081.i10.

Nelder, J.A., and Wedderburn, R.W. (1972). Generalized linear models. Journal of the Royal Statistical Society, Series A 135, 370-384.

Noh, M. and Lee, Y. (2007). REML estimation for binary data in GLMMs. Journal of Multivariate Analysis 98, 896-915.

Pan, J., and Thompson, R. (2007). Quasi-Monte Carlo estimation in generalized linear mixed models. Computational Statistics & Data Analysis 51, 5765-5775.

Paolino, P. (2001). Maximum likelihood estimation of models with Beta-distributed dependent variables. Political Analysis 9, 325-346.

Qiu, Z., Song, P.X.K., and Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data. Scandinavian Journal of Statistics 35, 577-596.

R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: www.R-project.org

Smyth, G., Hu, Y., Dunn, P., and Phipson, B. (2013). statmod: Statistical Modeling. R package version 1.4.17. URL: CRAN.R-project.org/package=statmod

Song, P.X.K. (2007). Correlated Data Analysis: Modeling, Analytics, and Applications. Springer, New York.

Tierney, L., and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81, 82-86.

Wuertz, D. (2012). fOptions: Basics of Option Valuation. R package version 2160.82. URL: CRAN.R-project.org/package=fOptions

Zhang, P., Qiu, Z., and Shi, C. (2014). simplexreg: Regression analysis of proportional data using simplex distribution. R package version 1.0. URL: CRAN.R-project.org/package=simplexreg

Zhang, W., and Hongjie Wei, H. (2008). Maximum likelihood estimation for simplex distribution nonlinear mixed models via the stochastic approximation algorithm. Rocky Mountain: Journal of Mathematics 38, 1863-1875.