simplexreg: an R package for regression analysis of proportional data using the simplex distribution

Article:

Zhang, Peng, Qiu, Zhenguo and Shi, Chengchun (2016) simplexreg: an R package for regression analysis of proportional data using the simplex distribution. Journal of Statistical Software, 71 (11). ISSN 1548-7660

10.18637/jss.v071.i11

Journal of Statistical Software

August 2016, Volume 71, Issue 11. *doi: 10.18637/jss.v071.i11*

simplexreg: An R Package for Regression Analysis of Proportional Data Using the Simplex Distribution

Peng Zhang
Zhejiang University

Zhenguo Qiu
Alberta Health Services

Chengchun Shi
North Carolina State University

Abstract

Outcomes of continuous proportions arise in many applied areas. Such data are typically measured as percentages, rates or proportions confined in the unitary interval. In this paper, the R package **simplexreg** which provides dispersion model fitting of the simplex distribution is introduced to model such proportional outcomes. The maximum likelihood method and generalized estimating equations techniques are available for parameter estimation in cross-sectional and longitudinal studies, respectively. This paper presents methods and algorithms implemented in the package, including parameter estimation, model checking as well as density, cumulative distribution, quantile and random number generating functions of the simplex distribution. The package is applied to real data sets for illustration.

Keywords: dispersion models, proportional data, random variable generation, R, simplex distribution.

1. Introduction

The theory of generalized linear models (GLMs, McCullagh and Nelder 1989) attests that regression analysis requires an appropriate recognition about the type of response variable. While the normal distribution is popular in practice, Jørgensen (1997) pointed out that the normal distribution is an exception, rather than the rule, except for data with small dispersions. Fisher (1953) reminded us of the importance of describing data in their natural habitat. Analysis of non-normal data should therefore take into account the actual type of data if such knowledge is available. Nelder and Wedderburn (1972) were the first to show, by introducing the class of GLMs, that a large variety of non-normal data may be analyzed by a united technique. The GLMs were originally developed for exponential families of distributions, but the main ideas were later extended to a wider class of models, the so called dispersion models (Jørgensen 1997). The major contribution behind dispersion models is that the notions of location and scale may be generalized to position and dispersion, respectively, so that a comprehensive range of non-normal distributions is covered and more data types such as positive data, positive data with zero, count data, binomial data and directional data can be dealt with by dispersion models in their natural habitat. An important subclass of dispersion models are the exponential dispersion models, which are the now familiar class of generalized linear models. In the meanwhile, the residual sum of squares from the analysis of variance is generalized to the notion of deviance, making the analysis of deviance available as a general inference tool in model fitting and model selection.

In applied fields, outcomes of proportions often arise. Few models are suitable for fitting such data. The beta distribution is often used in Bayesian statistics as the conjugate prior distribution for binomial proportions. With suitable parameter transformations of the beta distribution, Ferrari and Cribari-Neto (2004) proposed a beta regression model for rates or proportions and implemented the related statistical estimation and inference methods in the R (R Core Team 2016) package **betareg** (Cribari-Neto and Zeileis 2010). Another R package **gamlss** (Rigby and Stasinopoulos 2005; Stasinopoulos and Rigby 2007) provides semi-parametric regression type models for proportional data. However, with an additional dispersion parameter, the simplex distribution (Barndorff-Nielsen and Jørgensen 1991) based GLM is more robust for analysis of continuous proportional data (Zhang and Qiu 2014). The simplex distribution includes a large class of distributions whose domains are confined in (0, 1). As shown by Jørgensen (1997), such a distribution is actually a dispersion model and shares many common analytic properties with the exponential dispersion models. Therefore, the GLM for continuous proportional data can be developed on the lines of the classical GLMs (Song and Tan 2000). In the literature, the simplex marginal model (Song and Tan 2000; Song, Qiu, and Tan 2004), and the simplex mixed-effects model (Qiu, Song, and Tan 2008) have been extensively studied. In this paper, we will briefly present some properties of the simplex distribution, which provides a foundation for the inference and computation in the package **simplexreg** (Zhang, Qiu, and Shi 2016). The package is implemented in the R system and available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=simplexreg`. There are three major lines of functions:

1. Calculation of the density, cumulative distribution, quantile and random number generating functions for the simplex distribution.
2. Statistical inference in the simplex generalized linear model (SGLM, Zhang and Qiu 2014) via maximum likelihood (ML) for cross-sectional proportional data set.
3. Analysis of longitudinal proportional data set via an extended version of generalized estimating equations (GEE) using the simplex distribution.

The paper is organized as follows. Section 2 presents the methods to calculate density, distribution and quantile functions as well as the function generating random variables of the simplex distribution. Section 3 illustrates the methodology of modeling proportional data using simplex generalized regression. The generalized estimating equations for longitudinal proportional outcomes are given in Section 4. Then we address model diagnostics in Section 5. Section 6 presents the details of the **simplexreg** package. Section 7 further conducts analyses based on the simplex distribution in R with real data sets. Finally, plans for extending the package are described in Section 8.

2. The simplex distribution

Simplex distributions are effectively derived from the generalized inverse Gaussian distribution (Barndorff-Nielsen and Jørgensen 1991). Consider the class of renormalized saddle-point approximations (when $\sigma^2 \to 0$) defined by

$$p(y; \alpha_1, \alpha_2, \mu, \sigma^2) = c(\alpha_1, \alpha_2, \mu, \sigma^2) y_1^{\alpha_1 - 1}(1 - y_1)^{\alpha_2 - 1} \exp\{-\frac{1}{2\sigma^2} d_{\alpha_1, \alpha_2}(y; \mu)\}, \quad y \in (0, 1),$$

where $d_{\alpha_1,\alpha_2}(y; \mu)$ is a unit deviance defined by

$$d_{\alpha_1,\alpha_2}(y; \mu) = \mu^{2\alpha_1 - 1}(1 - \mu)^{2\alpha_2 - 1} \frac{y(1 - y)}{(y - \mu)^2},$$

and $c(\alpha_1, \alpha_2, \mu, \sigma^2)$ is the normalized term. The parameters of the distribution are $(\alpha_1, \alpha_2)^\top \in \mathcal{R}^2$, $\sigma^2 > 0$ and $\mu \in (0, 1)$. The distribution is called the general simplex distribution, and is denoted by $Y \sim S(\alpha_1, \alpha_2, \mu, \sigma^2)$. If $\alpha_1, \alpha_2 > 0$, the limiting case $\sigma^2 \to \infty$ is the beta distribution with parameters α_1 and α_2. The special case with $(\alpha_1, \alpha_2) = (-\frac{1}{2}, -\frac{1}{2})$ gives the standard simplex distribution, which we from now on refer to as the simplex distribution.

With mean $\mu \in (0, 1)$ and dispersion parameter $\sigma^2 > 0$, the simplex distribution has a density function taking a similar expression to a normal density,

$$f(y; \mu, \sigma^2) = \left[2\pi\sigma^2\{y(1 - y)\}^3\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2} d(y; \mu)\right\}, \quad y \in (0, 1), \tag{1}$$

where the unit deviance function is

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

with the unit variance function $V(\mu) = \mu^3(1 - \mu)^3$.

Then a random variable Y which follows a simplex distribution with mean μ and dispersion parameter σ^2 is denoted by $Y \sim S^-(\mu, \sigma^2)$. Jørgensen (1997) gave the variance of Y, τ^2, as

$$\tau^2 = \mu(1 - \mu) - \frac{1}{\sqrt{2\sigma^2}} \exp\left\{\frac{1}{2\sigma^2\mu^2(1 - \mu)^2}\right\} \Gamma\left\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1 - \mu)^2}\right\}, \tag{2}$$

where $\Gamma(a, b)$ is the incomplete Γ-function defined by $\Gamma(a, b) = \int_b^\infty t^{a-1} b^t dt$.

Another important property about the density of the simplex distribution worth to mention is that when the dispersion parameter $\sigma^2 \to 0$, the small-dispersion asymptotic theory (Jørgensen 1997) leads to

$$\frac{Y - \mu}{\sigma\sqrt{V(\mu)}} \xrightarrow{d} N(0, 1). \tag{3}$$

It has been proved that the simplex distribution has a uni-mode if $\sigma \leq 4/\sqrt{3}$; otherwise, it yields multi-modes.

Given values of parameters, μ and σ, the calculation of the simplex density function is straightforward (see (1)). As for the distribution and quantile functions, we calculate the normalized cumulative distribution and quantile functions instead of simplifying the computation if the dispersion parameter is small. When the dispersion parameter is large, however, the distribution function is calculated through a numerical integration while the quantile function is obtained by solving nonlinear equations.

The interest in developing algorithms to generate random variables from the simplex distribution lies in many aspects, one of which is the need in simulation studies. Also, the effective algorithm to generate random variables from the simplex distribution is the base of Markov chain Monte Carlo methods in Bayesian inference. Simplex random variable generation can be developed based on a certain transformation which is motivated by the fact that the simplex distribution is transformed from the inverse Gaussian mixture distribution (Qiu 2001).

The inverse Gaussian mixture distribution (Jørgensen 1991), denoted by $X \sim M\text{-}IG(\xi, \epsilon^2, p)$, is the mixture of the inverse Gaussian distribution (with probability $1-p$) and its complementary reciprocal (with probability p), with probability density:

$$f(x; \xi, \epsilon^2, p) = (2\pi\epsilon^2 x^3)^{-\frac{1}{2}} \left(1 - p + \frac{px}{\xi}\right) \exp\left\{-\frac{1}{2\epsilon^2}\frac{(x-\xi)^2}{\xi^2 x}\right\}, \quad x > 0.$$

Suppose $X \sim M\text{-}IG(\xi, \epsilon^2, p)$, let

$$y = \frac{x}{1+x}, \quad \mu = \frac{\xi}{1+\xi}, \quad \text{and} \quad \sigma^2 = \frac{\epsilon^2}{(1-p)^2} = \epsilon^2(1-\xi)^2.$$

Then $Y = \frac{X}{1+X} \sim S^-(\mu, \sigma^2)$, noting that the Jacobian is $(1-y)^{-2}$ in this transformation. Therefore, to generate random variables from the simplex distribution $S^-(\mu, \sigma^2)$, we can first produce random variables from the inverse Gaussian mixture distribution, $M\text{-}IG(\xi, \epsilon^2, p)$. Jørgensen (1991) presented the method to generate the inverse Gaussian mixture random variables from the inverse Gaussian distribution, denoted by $IG(\xi, \epsilon^2)$, and χ^2_1 distribution. To generate inverse Gaussian random variables, we adopt the method proposed by Michael, Schucany, and Hass (1976), which is based on the property that the kernel of the inverse Gaussian density has a χ^2-distribution. The proposed transformation method for a simplex random number generation is built only upon the χ^2-generator and uniform-generator, with the process listed as follows:

1. Set $p = \mu$, $\xi = \frac{\mu}{1-\mu}$ and $\epsilon^2 = \sigma^2(1-\mu)^2$.

2. Generate random variable $X_1 \sim IG(\xi, \epsilon^2)$:

   - Generate random variable $Z \sim \chi^2(1)$, and U from the uniform $(0,1)$ distribution.
   - Set $Z_1 = \xi + \frac{\xi^2\epsilon^2 Z}{2} - \frac{\xi\epsilon^2}{2}\sqrt{\frac{4\xi Z}{\epsilon^2} + \xi^2 Z^2}$.
   - Choose $\frac{\xi^2}{Z_1}$ to be X_1 if $U > \frac{\xi}{\xi + Z_1}$ or choose Z_1 otherwise.

3. Generate the random variable $X \sim M\text{-}IG(\xi, \epsilon^2, p)$:

   - Generate random variable $X_2 \sim \xi^2\epsilon^2\chi^2(1)$.
   - Let X equals to $X_1 + X_2$ with probability p and X_1 with probability $1-p$.

4. Apply the "simplex" transformation function $Y = \frac{X}{1+X}$.

The following properties of the simplex distribution will be used in the inference and computation of simplex models:

Proposition 1 *Let* $Y \sim S^-(\mu; \sigma^2)$*. Then* $\mathsf{E}(Y) = \mu$*, and*

(i) $\mathsf{E}\left\{d(Y;\mu)\right\} = \sigma^2$

(ii) $\mathsf{E}\left\{(Y-\mu)d'(Y;\mu)\right\} = -2\sigma^2$.

(iii) $\mathsf{E}\left\{(Y-\mu)d(Y;\mu)\right\} = 0$.

(iv) $\mathsf{E}\left\{d'(Y;\mu)\right\} = 0$.

(v) $\frac{1}{2}\mathsf{E}\left\{d''(Y;\mu)\right\} = \frac{3\sigma^2}{\mu(1-\mu)} + \frac{1}{\mu^3(1-\mu)^3}$.

(vi) $\mathsf{VAR}\left\{d(Y;\mu)\right\} = 2\left(\sigma^2\right)^2$.

(vii) $\mathsf{VAR}\left\{u(Y;\mu)\right\} = \frac{3\sigma^4}{\mu(1-\mu)} + \frac{\sigma^2}{\mu^3(1-\mu)^3}$.

The proof can be found in Song and Tan (2000), Qiu (2001), Song *et al.* (2004) and Zhang and Qiu (2014).

3. Simplex generalized linear models

Consider cross-sectional proportional responses y_i, $0 < y_i < 1$, $i = 1, 2, \ldots, m$. Let x_i be a p-element vector of covariates for subject $i = 1, 2, \ldots, m$, z_i be a subset of x_i. The goal is to model the means of the responses and the dispersion as functions of these covariates.

Assuming y_i are realizations of the random variable Y_i, $Y_i|x_i \sim S^-(\mu_i, \sigma_i^2)$, we write a generalized linear model

$$\eta_i = g(\mu_i) = x_i^\top \beta, \tag{4}$$

where the function $g : (0,1) \to (-\infty, \infty)$ is the link function, and β is the vector of regression parameters of interest, and the dispersion σ_i^2 in the simplex heterogeneous model may be modeled with some covariates:

$$h(\sigma_i^2) = z_i^\top \gamma, \tag{5}$$

where the function $h : (0, \infty) \to (-\infty, \infty)$ is the link function, and γ is the vector of regression coefficients associated with the dispersion σ_i^2. The homogeneous simplex model is obtained by removing the varying dispersion effect in (5).

This is an extension of the generalized linear models mentioned in Jørgensen (1997), noting that (i) given $x_i, Y_i, i = 1, \ldots, m$ are independently distributed with mean $\mathsf{E}(Y_i|x_i) = \mu_i$; (ii) the parameters μ_i may vary from subject to subject; (iii) the link function can be any monotonic and differentiable function. However, the simplex density function is not a member of the exponential family with the general form and the variance of Y_i does not vary with the x_i's through μ_i alone, which, in fact, depends on the dispersion parameter.

To estimate the parameters β and γ for the SGLM with varying dispersion, we follow the classical theory of GLMs and use the iteratively re-weighted least squares algorithm for the maximum likelihood estimation of β and γ. Clearly, the log-likelihood takes the form

$$\ell(\beta, \gamma) = -\frac{1}{2}\sum_{i=1}^{m}\left\{\frac{d(y_i;\mu_i)}{\sigma_i^2} - \log \sigma_i^2\right\}, \tag{6}$$

subject to a constant term.

Define the surrogate or “working" vector $s = (s_1, s_2, \dots, s_m)^\top$ as follows: The ith component is given by

$$s_i = g(\mu_i) + \frac{u(y_i, \mu_i)}{\sigma_i^2 w_i g'(\mu_i)}, \tag{7}$$

where $u(y_i, \mu_i)$ is the score function defined as

$$u(y; \mu) = \frac{y - \mu}{\mu(1-\mu)} \left\{ d(y; \mu) + \frac{1}{\mu^2(1-\mu)^2} \right\}, \tag{8}$$

and weights are given by

$$w_i = \frac{\mathsf{E}d''(y_i; \mu_i)}{2\sigma_i^2 \{g'(\mu_i)\}^2} = \frac{1}{\sigma_i^2 \{g'(\mu_i)\}^2} \left(\frac{3\sigma_i^2}{\pi_i} + \frac{1}{\pi_i^3} \right), \tag{9}$$

where $\pi_i = \mu_i(1-\mu_i)$ by Proposition 1(v).

The expression in (7) of the surrogate response is a natural extension to dispersion models. For exponential family models, this surrogate response reduces to $s_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$, a common form seen in the classical GLMs. By the deviations of (6), we may estimate β and γ by the standard iteratively re-weighted least squares method that iteratively solves the score equations of β and γ through updating

$$\begin{aligned} \beta^{(k+1)} &= \beta^{(k)} + \left(\sum_{i=1}^m x_i^\top w_i^{(k)} x_i \right)^{-1} \sum_{i=1}^m x_i^\top w_i^{(k)} \left(s_i^{(k)} - \eta_i^{(k)} \right), \\ \gamma^{(k+1)} &= \gamma^{(k)} + \left(\frac{1}{2} \sum_{i=1}^m z_i^\top z_i \right)^{-1} \sum_{i=1}^m z_i^\top v_i^{(k)}, \end{aligned}$$

where $v_i^{(k)} = \frac{d(y_i; \mu_i^{(k)})}{2\sigma_i^{2\,(k)}} - \frac{1}{2}$, with the logarithmic link function h.

In addition, Proposition 1(i) leads to a moment method, an alternative approach we employed to estimate the constant dispersion parameter σ^2 in the simplex homogeneous model. The moment estimator coincides with its MLE. By bias correction, the estimation of σ^2 has the closed form:

$$\hat{\sigma}^2 = \frac{1}{m-p} \sum_{i=1}^m d(y_i; \mu_i), \tag{10}$$

in the invariant dispersion simplex GLMs. Details of the discussion on estimation and modeling of dispersion parameters can be found in Zhang and Qiu (2014).

4. Simplex marginal models for longitudinal data analysis

Let y_{ij}, $j = 1, \dots, n_i$ be the sequence of observed repeated measurements on the ith of m subjects, and t_{ij}, $j = 1, \dots, n_i$, be the sequence of corresponding times on which the measurements are taken on each subject, and x_{ijk}, $k = 1, \dots, p$, be p explanatory variables where x_{ij1} may be set to 1 corresponding to an intercept. We assume that y_{ij} are realizations of random variables Y_{ij} which follow simplex distributions $Y_{ij} \sim S^-(\mu_{ij}, \sigma_{ij}^2)$, where $\mu_{ij} \in (0, 1)$

are the mean parameters and $\sigma_{ij}^2 > 0$ are the dispersion parameters, and both may be specified as functions of covariates. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top$. We assume that $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ are independent.

A heterogeneous marginal simplex model consists of three components. The first component is a model to describe the population-averaged effects, where the mean parameter μ_{ij} depends on the time-varying covariates $\mathbf{x}_{ij}$ via a generalized linear model of the form

$$\eta_{ij} = g(\mu_{ij}) = \mathbf{x}_{ij}^\top \beta, \tag{11}$$

where g is a known link function and $\beta = (\beta_0, \ldots, \beta_{p-1})^\top$ are the regression coefficients to be estimated. The second component is a model to describe the pattern of dispersion parameters σ_{ij}^2 as a function of covariates $\mathbf{z}_{ij}$ (maybe a subset of $\mathbf{x}_{ij}$, which can be omitted in the homogeneous dispersion model), given by

$$h(\sigma_{ij}^2) = \mathbf{z}_{ij}^\top \gamma, \tag{12}$$

where h is a known link function and $\gamma = (\gamma_0, \ldots, \gamma_{r-1})^\top$ with γ_0 corresponding to the intercept term. The third component is for modeling the correlation structure. The correlation between Y_{ij} and Y_{ik} is a function of the location parameters and perhaps of additional parameters, $\alpha = (\alpha_1, \ldots, \alpha_q)^\top$, namely,

$$\mathsf{COR}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \alpha), \tag{13}$$

where $\rho(\cdot)$ is a known function. Various types of correlation structures may be used for the ρ function. Amongst others, three commonly used in the analysis of longitudinal data are the exchangeable, auto-regressive (AR) and m-dependent correlations.

Denote the mean vector of subject i by $\mu_i = (\mu_{i1}, \ldots, \mu_{in_i})^\top$. Let the score vector for subject i be $\mathbf{u}_i = (u_{i1}, \ldots, u_{in_i})^\top$, with u_{ij} evaluated by (8), $\mathbf{s}_i = \mathrm{diag}\{\mu_{ij}^3(1-\mu_{ij})^3\}\mathbf{u}_i$ be the working vector, and $\mathbf{R}(\alpha)$ be an $n_i \times n_i$ working correlation matrix with a $q \times 1$ vector of correlation parameters α. The working covariance matrix for $\mathbf{s}_i$ is

$$\mathbf{V}_i = \mathrm{diag}^{1/2}\left\{\mathsf{VAR}(s_{ij})\right\} \mathbf{R}(\alpha) \mathrm{diag}^{1/2}\left\{\mathsf{VAR}(s_{ij})\right\},$$

where $\mathsf{VAR}(u_{ij})$ in $\mathsf{VAR}(s_{ij})$ is calculated by Proposition 1(vii). The GEE1 (Liang and Zeger 1986) for the simplex margin corresponding to the estimating equation for β is given by

$$\Psi_1(\beta, \gamma, \alpha) = \sum_{i=1}^{m} \mathbf{D}_i^\top \mathbf{A}_i \mathbf{V}_i^{-1} \mathbf{s}_i = \mathbf{0}, \tag{14}$$

where $\mathbf{A}_i = \mathrm{diag}\left\{\sigma_{ij}^{-2} v(\mu_{ij}) \mathsf{VAR}(u_{ij})\right\}$ and $\mathbf{D}_i^\top = \partial \mu_i^\top / \partial \beta$. The estimating equation for the dispersion component is given as follows,

$$\Psi_2(\beta, \gamma, \alpha) = \sum_{i=1}^{m} \left(\frac{\partial \sigma_i^\top}{\partial \gamma}\right) \Sigma_i^{-1} (\mathbf{d}_i - \sigma_i) = \mathbf{0}, \tag{15}$$

where $\mathbf{d}_i = (d(y_{i1}; \mu_{i1}), \ldots, d(y_{in_i}; \mu_{in_i}))^\top$, Σ_i is a working covariance matrix, and $\sigma_i = \mathsf{E}(\mathbf{d}_i) = (\sigma_{i1}^2, \ldots, \sigma_{in_i}^2)^\top$.

Following Prentice and Zhao (1991), the additional set of estimating equations for the correlation parameters based on the standardized score residuals, is defined by

$$r_{ij} = \frac{u_{ij}}{\sqrt{\mathsf{VAR}(u_{ij})}} = \frac{u_{ij}}{\sigma_{ij}\sqrt{\frac{1}{2}\mathsf{E}d''(y_{ij};\mu_{ij})}}. \tag{16}$$

It is easy to see that such score residuals satisfy moment properties of $\mathsf{E}(r_{ij}) = 0$, $\mathsf{VAR}(r_{ij}) = 1$ and $\mathsf{E}(r_{ij}r_{ij'}) = \mathsf{COR}(u_{ij}, u_{ij'}) = \mathsf{COR}(s_{ij}, s_{ij'})$. The estimating equation for the correlation parameter α then takes the form

$$\Psi_3(\beta,\gamma,\alpha) = \sum_{i=1}^{m}\left(\frac{\partial\xi_i^\top}{\partial\alpha}\right)\mathbf{H}_i^{-1}(\mathbf{r}_i - \xi_i) = \mathbf{0}, \tag{17}$$

where $\mathbf{r}_i = (r_{i1}r_{i2}, r_{i1}r_{i3}, \ldots, r_{in_i-1}r_{in_i})^\top$, $\mathbf{H}_i$ is a working covariance matrix and $\xi_i = \mathsf{E}(\mathbf{r}_i)$. Details of the sensitivity and variability matrices for the GEEs are referred in Song and Tan (2000) and Song *et al.* (2004). Using the Newton-scoring algorithm, the solution of the joint Equations 14, 15 and 17 can be obtained numerically by iteratively updating the values of the parameters.

5. Model diagnostics

Fitting data with a certain model means choosing appropriate forms for the predictor, the link function and the distribution function. In general, Pearson's χ^2 and the deviance perform the important roles as general goodness-of-fit statistics.

The Pearson residual takes the form

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathsf{VA\hat{R}}(y_i)}} = \frac{y_i - \hat{\mu}_i}{\hat{\tau}_i}, \tag{18}$$

where $\hat{\tau}_i$ has no analytical form expression as it involves the incomplete gamma function in (2). For over-dispersed data, the dispersion parameter σ^2 is large and thus the variance of the response approaches to $\mu(1-\mu)$. This leads to an approximate Pearson residual:

$$r_i^a = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-\hat{\mu}_i)}}. \tag{19}$$

Replacing parameters by their corresponding estimates, the simplex distribution assumption can be checked by the plot of $\hat{r_i}^P$ or $\hat{r_i}^a$ against $\hat{\mu_i}$, which aims to examine the mean-variance relation.

An informal check for the link function assumption could be done by McCullagh and Nelder (1989)'s plot of the adjusted dependent variable a_i against the linear predictor $\hat{\eta}_i$. In our setting, define

$$a_i = g(\mu_i) + \left\{\frac{3\sigma^4}{\mu_i(1-\mu_i)} + \frac{\sigma^2}{V(\mu_i)}\right\}^{-1/2} u(y_i;\mu_i), \tag{20}$$

where $u(y_i;\mu_i)$ and $V(\mu_i) = \mu_i^3(1-\mu_i)^3$ are the score function (8) and variance function of the simplex model, respectively. It follows from Proposition 1 that $\mathsf{E}(a_i) = g(\mu_i)$ since $\mathsf{E}(u_i) = 0$, and $\mathsf{VAR}(a_i) = \mathsf{E}\{a_i - g(\mu_i)\}^2 = 1$.

In longitudinal analysis, the scatter plots of the standardized score residuals defined by (16) at different lags can informally be used to check the assumption of the working correlation structure.

As an extension of GLMs, the measure of the discrepancy or the goodness-of-fit of the simplex GLM can be formed by deviance. Noting that for the unit deviance function, $d(y; y) = d(\mu; \mu) = 0$, setting the perfect fit $\hat{\mu}_i = y_i$ gives the log-likelihood of the saturated model. Hence, by (6), the deviance function is

$$D = \sum_{i=1}^{m} D(y_i; \hat{\mu}_i) = 2 \sum_{i=1}^{m} \{\ell_i(y_i; y_i) - \ell_i(\hat{\mu}_i; y_i)\} = \sum_{i=1}^{m} d(y_i; \hat{\mu}_i)/\sigma_i^2, \tag{21}$$

which follows a χ^2_{m-p}.

It becomes difficult in determining the degree $m - p$ of the χ^2 distribution for the goodness-of-fit test in the presence of within-subject dependence for longitudinal data. Qiu (2001) proposed the partial deviance $D_j^p = \sum_{i=1}^{m_j} d(y_{ij}; \hat{\mu}_{ij})/\sigma^2_{ij}, \;\; j \in T$, where T denotes a collection of all distinct times on which observations are made. Cross-sectionally, y_{ij}'s are independent and hence D_j^p follows approximately $\chi^2_{m_j-p}$, with m_j being the total number of y_{ij}'s observed cross-sectionally at time t_j. A series of the goodness-of-fit χ^2 tests can be performed along these time occasions. Both observed partial deviance D_j^p statistics and the corresponding critical values determined by $\chi^2_{m_j-p}$ can be depicted and compared at each time point. The plot displays a detailed scenario of testing for the goodness-of-fit over the spectrum of time, and hence is more informative than an overall test based on a single summary statistics.

6. The simplexreg package

The **simplexreg** package carries out generalized linear model regression as well as generalized estimation equations based on the simplex distribution and provides related functions of the simplex distribution. Some routines, including updating parameters for both the homogeneous and heterogeneous simplex marginal models via the Newton-Raphson method, are written in C with the GNU Scientific Library (GSL; Galassi *et al.* 2009) to get support for vector and matrix operation tasks and facilitate the computation. All the C programs are written in double precision.

6.1. Functions of the simplex distribution

In the **simplexreg** package, the function `dsimplex` gives the density function, `psimplex` provides the distribution function, `qsimplex` calculates the quantile function and `rsimplex` gives random numbers generated from the simplex distribution. They thus possess the same forms as other distribution functions in R:

```
dsimplex(x, mu, sig)
psimplex(q, mu, sig)
qsimplex(p, mu, sig)
rsimplex(n, mu, sig)
```

where `x` and `q` are vectors of quantiles, `p` is the vector of probability, `n` is the user-specified number of samples, arguments `mu` and `sig` are the mean parameter μ and the square root

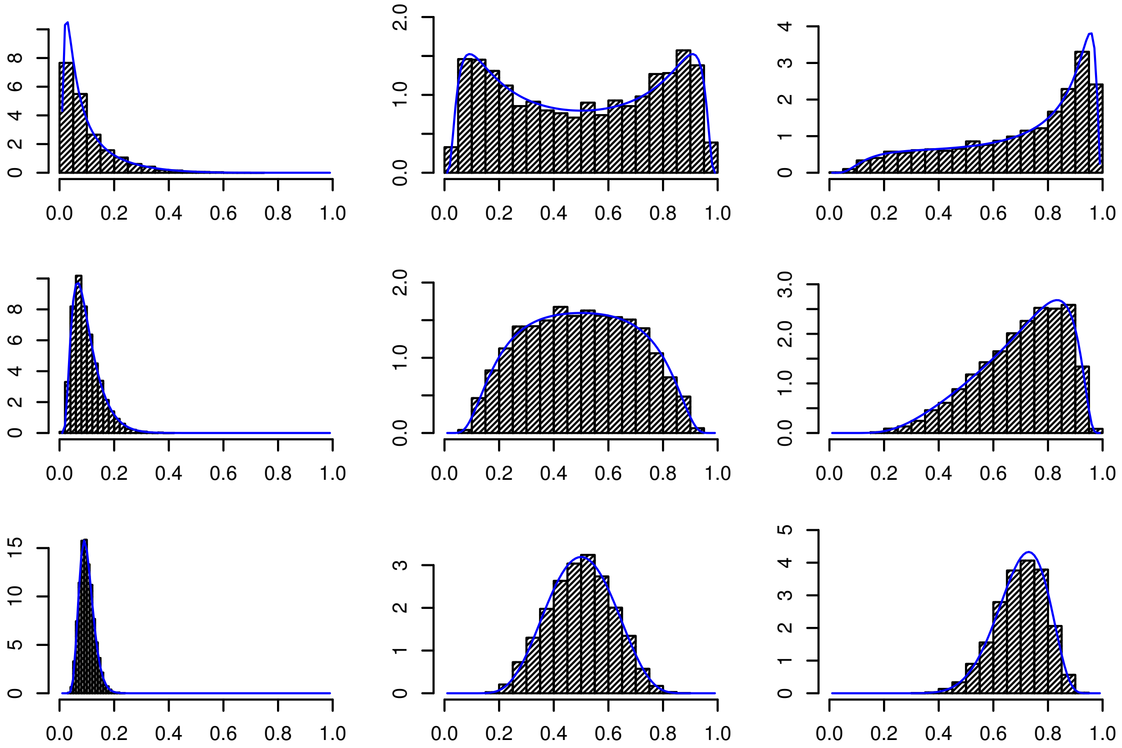

Figure 1: Histograms of random numbers of the simplex distributions. $\mu = (0.1, 0.5, 0.7)$ from left plots to right plots; $\sigma^2 = (4^2, 2^2, 1)$ from top plots to bottom plots. Solid lines are corresponding to simplex densities.

of the dispersion parameter σ^2 of the simplex distribution, respectively. Numerical integration is used in calculating distribution and quantile functions. However, to speed up, the corresponding functions of the normal distributions are computed when the approximation according to the small dispersion asymptotic theory is close enough, namely, $|\text{error}| < 10^{-6}$.

To illustrate `rsimplex` and `dsimplex`, histograms of random numbers obtained from the simplex random number generator are compared to the densities of the corresponding simplex distributions (Figure 1).

6.2. Regression analysis of the simplex model

The function `simplexreg` fits proportional data with simplex regression models. The arguments are similar to packages implementing regression models in R:

```
simplexreg(formula, data, subset, na.action,
  link = c("logit", "probit", "cloglog", "neglog"), corr = "Ind", id = NULL,
  control = simplexreg.control(...), model = TRUE, y = TRUE, x = FALSE, ...)
```

The regression model and data can be specified via `formula` and `data`. Methods, such as for the generic functions `print`, `summary` and `coef`, are available for the returned S3 class object '`simplexreg`'. The function `simplexreg.fit` gives an alternative approach to specify the model:

```
simplexreg.fit(y, x, z = NULL, t = NULL, link = "logit", corr = "Ind",
  id = NULL, control = simplexreg.control())
```

The `corr` argument specifies the correlation structure in the marginal model, providing three options: independent, exchangeable and auto-regressive of order 1, given by `"Ind"`, `"Exc"` and `"AR1"` respectively. The default value is `"Ind"`, reducing the marginal model to a simplex GLM. To specify both mean and dispersion equations (see (4) and (5)) in the simplex model, we employ the form `y ~ x1 + x2 | z1 + z2` (Zeileis and Croissant 2010), where `y ~ x1 + x2` specifies the mean model and covariates `z1` and `z2` are associated with the dispersion parameter σ^2. Without the latter part, a homogeneous dispersion model will be fitted.

To obtain initial values for regression coefficients β, the package fits the data with a linear model for logit transformed responses. The initial values for γ are from a log-linear model treating $d(y_i;\mu_i)$ as the response which has a Gamma distribution.

For longitudinal proportional data, the marginal simplex models consist of three components, the population-average effects, the pattern of dispersion and the correlation structure. A `formula` of homogeneous dispersion takes the form `y ~ x1 + x2 | 1 | t` and a `formula` in the form `y ~ x1 + x2 | z1 + z2 | t` is used for heterogeneous simplex marginal models (`corr = "Exc" or "AR1"`). The parameter `t` in the `formula` as well as in the function `simplexreg.fit` corresponds to the time covariate. A factor identifying clusters of the observations should also be specified by the argument `id`. And variables `y`, `x`, `z`, `t` are required to be sorted in accordance with the clusters.

The function provides four options for the link function of the mean, i.e., the function g in (4) and (11): `link = "logit", "probit", "cloglog", "neglog"`, corresponding to the logit, probit, complementary log-log and negative-log functions, respectively. However, when it comes to the link of dispersion, only the logarithmic function is supported. And function `simplexreg.control` controls the fitting process of simplex models.

For the returned object of class '`simplexreg`', the `summary` method lists a standard output, including Wald statistics as well as the p values for the regression coefficients. And based on the fitted values, a χ^2 test is performed for the simplex GLM model and the result is also reported. Argument `type` in the `summary` function specifies the type of residuals included in the output. The `coef` and `vcov` functions extract coefficients and their covariance matrix, respectively. Akaike's information criterion (AIC) defined as $AIC = 2\ell(\beta,\gamma) - 2p$ and Bayesian information criterion (BIC) $BIC = 2\ell(\beta,\gamma) - p\log n$ where p is the number of parameters, are calculated via the `AIC` and `BIC` methods. For simplex marginal models, these functions are not supported are not supported since the model is non-likelihood-based. Function `predict` provides predicted values of the mean or the dispersion for the responses or new observations. The `plot` method draws graphs for visually examining the correlation structure and the model assumption. Its arguments include:

```
plot(x, type = c("residuals", "corr", "GOF"), res = "adjvar", lag = 1, ...)
```

where `x` is the S3 class object returned from `simplexreg` fitting, `type` specifies the types of graphs. Residuals analysis is given by `type = "residuals"` with one of the four types chosen for `res`: `stdPerr` the exact standard Pearson residual r_i^P given in (18), `appstdPerr` the approximated Pearson residual r_i^a given in (19), `stdscor` the standardized score residuals r_{ij} detailed in (16), and the `adjvar` adjusted dependent variable a_i in (20). The first three can be plotted against the mean μ to examine the mean-variance relation as well as detect model assumption violation. The plot of adjusted dependent variable against the linear predictor η (see (4) and (11)) can be used to check the link function. All these residuals could be obtained the `residuals` method. Model diagnosis and residual analysis using this function is further demonstrated with examples in Section 7.

When `type = "corr"`, a graph is drawn to explore the correlation structure. Standardized score residuals are used to examine the auto-correlation at `lag`. Partial deviances against the time covariates are plotted when `type = "GOF"` (leveraging the **plotrix** package, Lemon 2006).

7. Data examples

Two examples are used to illustrate the capacities of **simplexreg**. The first models the recovery rate of CD34+ cells after peripheral blood stem cell (PBSC) transplants and the second is the longitudinal study of decay of intraocular gas (C_3F_8) presented in Song and Tan (2000), Song *et al.* (2004) and Qiu *et al.* (2008). These two proportional data sets are modeled via the simplex GLM technique and the GEE method, respectively. These analyses are done using R version 2.15.3.

7.1. PBSC study

Autologous peripheral blood stem cell (PBSC) transplants have been widely used for rapid hematologic recovery following myeloablative therapy for various malignant hematological disorders. Hematopoietic reconstitution largely depends on the reinfusion of sufficient numbers of stem cells to engraft in the bone marrow micro-environment, as indicated in Allan *et al.* (2002). The dose of viable CD34+ cells is considered an important marker of adequacy of PBSC harvest, as well as a predictor of hematopoietic engraftment. Studies have shown that the process of freezing, cryopreservation and thawing prior to reinfusion could inevitably damage PBSCs and remarkably decrease the number of viable CD34+ cells, as demonstrated by Yang, Acker, Cabuhat, Letcher, Larratt, and McGann (2005). The loss of viable CD34+ cells is usually assessed by post-cryopreservation recovery rates of the number of post-thaw viable CD34+ cells and that of pre-freeze viable CD34+ cells. It is of scientific interest to investigate the mechanism by which the post-cryopreservation recovery rates are infected.

A study enrolled 242 patients who consented to autologous PBSC transplant after myeloablative doses of chemotherapy between the years 2003 and 2008 at the Edmonton Hematopoietic Stem Cell Lab in the Cross Cancer Institute – Alberta Health Services. Age, gender and clinical characteristics, such as cancer type and chemotherapy type, were recorded. The patients are 18 to 71 years old and most of them are male (170), diagnosed with multiple myeloma, non-Hodgkin's lymphoma, acute leukaemia, solid tumors, amyloidosis or others. Patients received primary chemotherapy, with 1 day protocol, 3 day protocol, G-CSF only or other types, for mobilizing CD34+ cells. The PBSC collection was initiated when the circulating CD34+ count in the peripheral blood reached or exceeded 15 cells/μL. PBSC products were cryopreserved and stored in a liquid nitrogen vapor until reinfusion. PBSC samples were assessed on the day of collection (pre-freeze) and post cryopreservation (post-thaw) for absolute viable CD34+ cells. Post-cryopreservation viability was calculated as the percentage of the absolute number of viable cells or colonies over the number of pre-freeze cells.

The data set object `sdac` is included in the **simplexreg** package with the first five rows of the data frame shown as follows,

```
R> library("simplexreg")
R> data("sdac", package = "simplexreg")
R> head(sdac, n = 5)

age gender rcd   ageadj chemo
 62       M 0.75  22     0
 39       M 0.83  0      1
 43       M 0.94  3      1
 58       M 0.86  18     0
 43       M 0.54  3      0
```

where `rcd` denotes the recovery rate of CD34+ cells. We factorize the treatment with a dummy variable `chemo` indicating if a patient receives a chemotherapy on a one-day protocol (0) or on a 3-day protocol (1). For simplicity, two `chemo` categories, G-CSF and other, are combined into either 1 day protocol or 3 day protocol, by number of days they took. To reflect the age structure in the patient population, we adjusted the age by setting `age`< 40 as the baseline age and subtracting other ages by 40, ending up in a new covariate `ageadj`. The range of CD34+ cells recovery rates is 40%–100%. The two extreme values, 100%, are replaced by 99% in the regression analysis to avoid the boundary issue for proportional data.

The data are fitted with simplex regression models using both the homogeneous and heterogeneous structures of dispersion. The mean response model is given by

$$\mathrm{logit}(\mu_i) = \beta_0 + \beta_1 \texttt{ageadj} + \beta_2 \texttt{chemo}.$$

The dispersion parameter in the heterogeneous simplex model is regressed on age, with

$$\log(\sigma_i^2) = \gamma_0 + \gamma_1 \texttt{age}.$$

The `summary` output of both models is given below:

```
R> sim.glm1 <- simplexreg(rcd ~ ageadj + chemo, data = sdac)
R> sim.glm2 <- simplexreg(rcd ~ ageadj + chemo | age, data = sdac)
R> summary(sim.glm1)

Call:
simplexreg(formula = rcd ~ ageadj + chemo, data = sdac)

standard Pearson residuals:
    Min      1Q  Median      3Q     Max
-2.8257 -0.5853 -0.0083  0.4974  1.3964

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.100226   0.140683   7.821 5.26e-15 ***
ageadj      0.013575   0.006519   2.082   0.0373 *
```

```
chemo       0.266092   0.124991   2.129   0.0333 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: 105.3,  p-value: 0.4877576
Deviance: 236
Number of Fisher Scoring iterations:  6

R> summary(sim.glm2)

Call:
simplexreg(formula = rcd ~ ageadj + chemo | age, data = sdac)

standard Pearson residuals:
    Min      1Q  Median      3Q     Max
-3.0821 -0.5386 -0.0032  0.4956  1.4484

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.115550   0.141396   7.890 3.03e-15 ***
ageadj      0.013013   0.006452   2.017   0.0437 *
chemo       0.251921   0.121807   2.068   0.0386 *

Coefficients (dispersion model with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.60750    0.36687   7.107 1.18e-12 ***
age         -0.01500    0.00688  -2.181   0.0292 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: 99.51,  p-value: 0.4332267
Deviance: 239
Number of Fisher Scoring iterations:  6
```

The p values of the χ^2 tests for both models are 0.488 and 0.433, respectively, implying no lack-of-fit. The chemotherapy type, adjusted by `ageadj`, is shown to be significant in the model. In addition, the result also indicates that the dispersions in the post-cryopreservation recovery rates are associated with patients' ages. Comparing the mean coefficients of the two models, it is clear that the dispersion assumption does not have a significant impact on the coefficients in the mean model.

For the purpose of model selection, information criteria are returned by the function `AIC`.

```
R> AIC(sim.glm1, sim.glm2)

         df       AIC
sim.glm1  4 -202.5467
sim.glm2  5 -189.0300
```

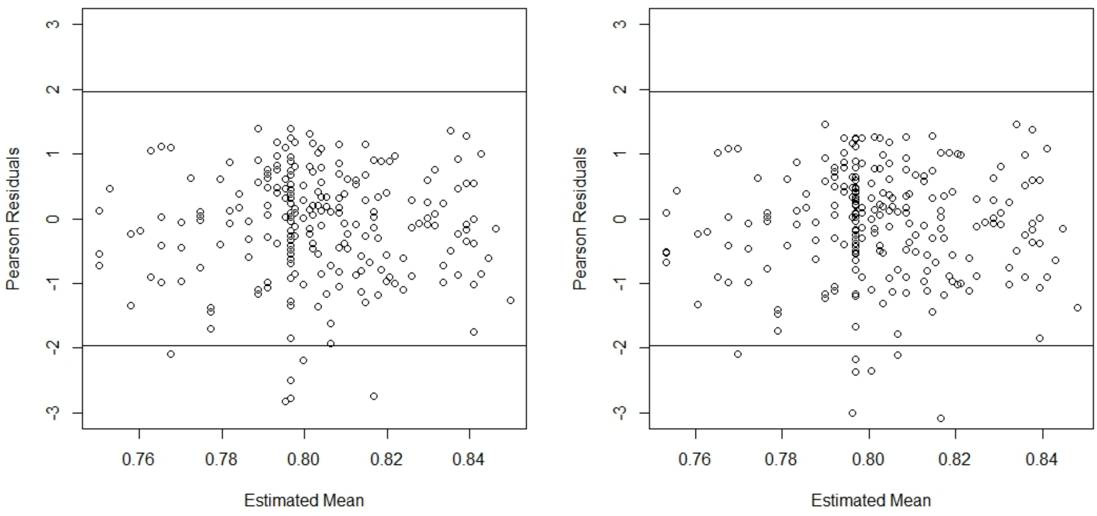

Figure 2: Checking the distribution by exact Pearson residuals, PBSC study. Homogeneous and heterogeneous models from left plot to right plot.

The `AIC` values show that the heterogeneous model has a better fit than the homogeneous one. Apart from the `AIC` criterion, the significance of the `age` coefficient in the dispersion model also indicates that the homogeneous dispersion assumption may be violated.

The model assumption can be checked based on values of defined residuals. For example, plots of the exact Pearson residuals against the estimated mean $\hat{\mu}_i$'s for the homogeneous and heterogeneous models are shown in Figure 2.

```
R> plot(sim.glm1, type = "residuals", res = "stdPerr", ylim = c(-3, 3))
R> plot(sim.glm2, type = "residuals", res = "stdPerr", ylim = c(-3, 3))
```

We could see that there is no clear pattern in the Pearson residuals plot and about 97% points lie in the horizontal band between -1.96 and 1.96.

7.2. Eye surgery study

Song *et al.* (2004) re-analyze the ophthalmological data C_3F_8 on the use of intraocular gas in retinal repair surgeries reported in Meyers, Ambler, Tan, Werner, and Huang (1992). The corresponding data frame, `retinal`, is included in the package. The outcome variable was the percent of gas (`Gas`) left in the eye. The gas, with three different concentration levels, 15%, 20% and 25% (`Level`), was injected into the eye before surgery for 31 patients. They were then followed three to eight (average of 5) times over a three-month period, and the volume of gas in the eye at the follow-up times were recorded as a percentage of the initial gas volume. The primary interest was to investigate whether concentration levels of the gas injected in patients' eyes affect the decay rate of the gas.

Before setting up the model, we first explore the correlation structure via `plot` under the naive assumption (independent observations). Let argument `lag = k`, the function plots r_{ij} against r_{ik} for all i and $j < k$, $|t_{ij} - t_{ik}| = k$.

```
R> data("retinal", package = "simplexreg")
R> sim.glm3 <- simplexreg(Gas ~ LogT + LogT2 + Level | LogT + Level | Time,
```

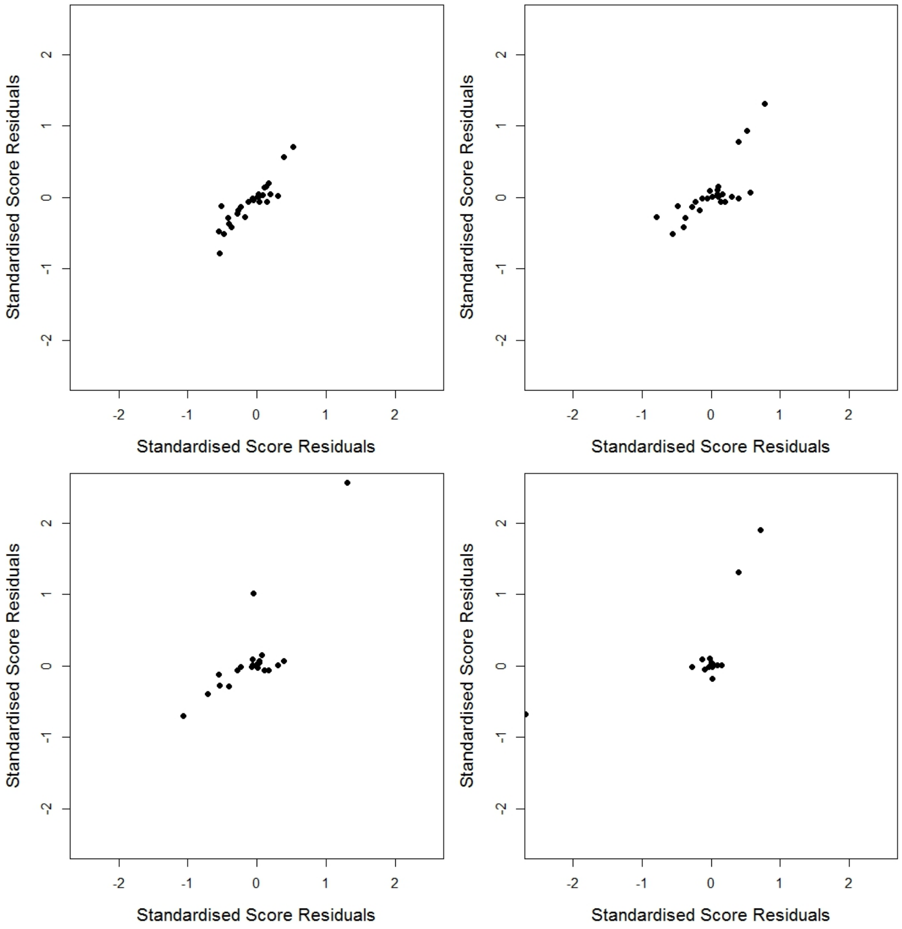

Figure 3: Examine auto-correlation at four different lags, Eye Surgery study. `lag = 1, 2, 3, 4` from left to right and from top to bottom.

```
+    data = retinal, id = ID)
R> plot(sim.glm3, type = "corr", xlim = c(-2.5, 2.5), ylim = c(-2.5, 2.5),
+    pch = 16)
R> plot(sim.glm3, type = "corr", lag = 2, xlim = c(-2.5, 2.5),
+    ylim = c(-2.5, 2.5), pch = 16)
R> plot(sim.glm3, type = "corr", lag = 3, xlim = c(-2.5, 2.5),
+    ylim = c(-2.5, 2.5), pch = 16)
R> plot(sim.glm3, type = "corr", lag = 4, xlim = c(-2.5, 2.5),
+    ylim = c(-2.5, 2.5), pch = 16)
```

From Figure 3 we can tell that the auto-correlation at lag `1` seems to be strongest while it decreases at lag `2` and `3`, and finally becomes insignificant at lag `4`. Consequently, it is clear that the AR(1) structure may fit the data. We fit a simplex marginal model with heterogeneous dispersion and AR(1) correlation structure for the data, following Song *et al.* (2004),

```
R> sim.gee2 <- simplexreg(Gas ~ LogT + LogT2 + Level | LogT + Level | Time,
+    corr = "AR1", id = ID, data = retinal)
```

```
R> summary(sim.gee2)

Call:
simplexreg(formula = Gas ~ LogT + LogT2 + Level | LogT + Level | Time,
    data = retinal, corr = "AR1", id = ID)

standard Pearson residuals:
    Min      1Q  Median      3Q     Max
-4.5801 -0.3452  0.0591  0.3910  4.6374

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.72142    0.20272  13.425  < 2e-16 ***
LogT         0.03394    0.31195   0.109 0.913359
LogT2       -0.32946    0.08515  -3.869 0.000109 ***
Level        0.40924    0.21689   1.887 0.059180 .

Coefficients (dispersion model with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.1532     0.3514  17.511  < 2e-16 ***
LogT         -0.4574     0.1694  -2.699  0.00695 **
Level        -0.4919     0.3563  -1.381  0.16735

Coefficients (correlation):
      Estimate Std. Error z value Pr(>|z|)
alpha  -0.3491     0.1865  -1.872   0.0612 .
rho     0.7054     0.1315   5.363 8.17e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Overall Deviance: 181
Number of Fisher Scoring iterations:  25
```

Apart from the regression coefficients in the mean and dispersion equations, information about the auto-correlation coefficient, `alpha` and `rho`, is also involved. The lag-1 auto-correlation, ρ, shown significant in the model, indicates that correlation for observations of the same patients is strong.

The graph of the adjusted dependent variable $\hat{a}_{ij}$ against the linear predictor $\hat{\eta}_{ij}$ is shown in the left panel of Figure 4. Overall 97% points fall into the 95% confidence band and those points show an increasing linear trend. This implies that the `logit` link function is a reasonable choice for the data.

```
R> plot(sim.gee2, type = "residuals", ylim = c(-6, 6), pch = 16)
R> plot(sim.gee2, type = "GOF", xlab = "Days after Gas Injection",
+    ylim = c(0, 40))
```

Consider those time points at which the cross-sectional clusters size $m_j > p = 4$ in order to compute the needed critical values. For the unequally spaced time points, we found $m_j =$

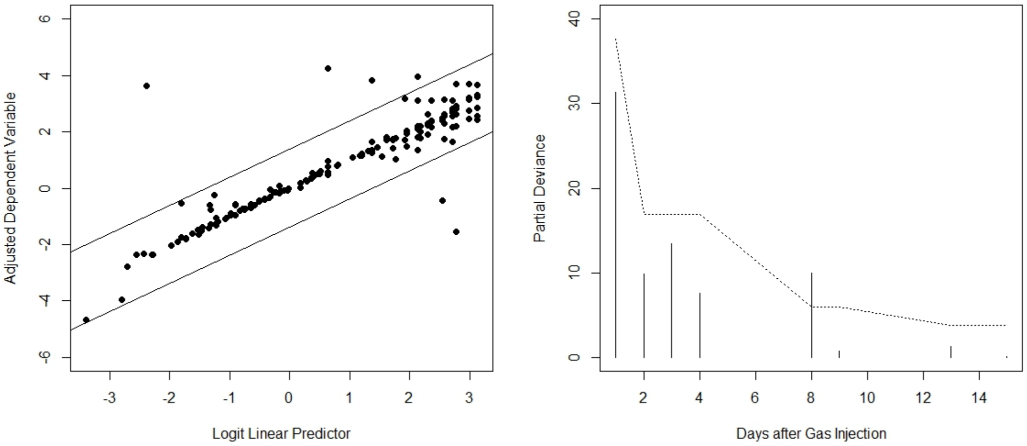

Figure 4: Link function checking (left) and goodness-of-fit test (right), Eye Surgery study.

$29, 13, 13, 13, 6, 6, 5, 5$ corresponding to day $j = 1, 2, 3, 4, 8, 9, 13, 15$. The partial deviances are depicted in the right panel of Figure 4, indicating an overall good fitting. Among the 8 time points, only the partial deviance on the 8th day gives the evidence of lack of fit, with a small margin.

8. Conclusion

In this paper, we describe the capabilities of the **simplexreg** package for conducting the simplex regression analysis in R. Statistical inferences and residual analyses of the simplex regression models via maximum likelihood and generalized estimation equations methods in R were presented, together with properties of the simplex distributions.

The density, cumulative distribution, quantile and the random generating functions for the simplex distribution were implemented in the package. The simplex random number generator is shown efficient and accurate, providing a powerful tool for a simulation based inference approach, such as the Markov chain Monte Carlo method, to fit hierarchical simplex generalized linear models for multilevel proportional data using a Bayesia approach.

The multi-dimensional simplex model is an important extension to the simplex GLM discussed in this paper. Jørgensen and Lauritzen (2000) proposed a class of multivariate dispersion models for multivariate non-normal responses. When over-dispersion appears in the compositional data, the multivariate simplex distribution (Jørgensen and Lauritzen 2000) can be considered as the underlying distribution for multivariate regression modeling. Developing regression analysis of the multivariate simplex distribution will be our future work through further investigation on this study.

9. Acknowledgments

The authors gratefully acknowledge NSFC 107104-N11405 for support.

References

Allan DS, Keeney M, Howson-Jan K, Popma J, Weir K, Bhatia M, Sutherland DR, Chin-Yee IH (2002). "Number of Viable CD34+ Cells Reinfused Predicts Engraftment in Autologous Hematopoietic Stem Cell Transplantation." *Bone Marrow Transplantation*, **29**, 967–972. doi:10.1038/sj.bmt.1703575.

Barndorff-Nielsen OE, Jørgensen B (1991). "Some Parametric Models on the Simplex." *Journal of Multivariate Analysis*, **39**, 106–116. doi:10.1016/0047-259x(91)90008-p.

Cribari-Neto F, Zeileis A (2010). "Beta Regression in R." *Journal of Statistical Software*, **34**(2), 1–24. doi:10.18637/jss.v034.i02.

Ferrari SLP, Cribari-Neto F (2004). "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics*, **31**, 799–815. doi:10.1080/0266476042000214501.

Fisher RA (1953). "Dispersion on a Sphere." *Proceedings of the Royal Society of London A*, **217**, 295–305. doi:10.1098/rspa.1953.0064.

Galassi M, Davies J, Theiler J, Gough B, Jungman G, Alken P, Booth M, Rossi F (2009). *GNU Scientific Library Reference Manual, Version 1.3*. URL http://www.gnu.org/software/gsl.

Jørgensen B (1991). "On the Mixture of the Inverse Gaussian Distribution with its Complementary Reciprocal." *Scandinavian Journal of Statistics*, **18**, 77–89.

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

Jørgensen B, Lauritzen SL (2000). "Multivariate Dispersion Models." *Journal of Multivariate Analysis*, **74**, 267–281. doi:10.1006/jmva.1999.1885.

Lemon J (2006). "**plotrix**: A Package in the Red Light District of R." *R News*, **6**(4), 8–12. URL https://CRAN.R-project.org/doc/Rnews/.

Liang KY, Zeger SL (1986). "Longitudinal Data Analysis Using Genrealized Linear Models." *Biometrika*, **73**, 13–22. doi:10.2307/2336267.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. Chapman and Hall, London.

Meyers S, Ambler JS, Tan M, Werner J, Huang S (1992). "Variation of Perfluorpropane Disappearance after Vitrectomy." *Retina*, **12**, 359–363. doi:10.1097/00006982-199212040-00012.

Michael JR, Schucany W, Hass RW (1976). "Generating Random Variates Using Transformations with Multiple Roots." *The American Statistician*, **30**, 88–90. doi:10.2307/2683801.

Nelder JA, Wedderburn RWM (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society A*, **135**, 370–384. doi:10.2307/2344614.

Prentice RL, Zhao LP (1991). "Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses." *Biometrics*, **47**, 825–839. doi:10.2307/2532642.

Qiu Z (2001). *Simplex Mixed Models for Longitudinal Proportional Data*. Ph.D. thesis, Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada.

Qiu Z, Song PXK, Tan M (2008). "Simplex Mixed-Effects Models for Longitudinal Proportional Data." *Scandinavian Journal of Statistics*, **35**, 577–596. doi:10.1111/j.1467-9469.2008.00603.x.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rigby R, Stasinopoulos DM (2005). "Generalized Additive Models for Location, Scale and Shape." *Applied Statistics*, **54**, 507–554. doi:10.1111/j.1467-9876.2005.00510.x.

Song PXK, Qiu Z, Tan M (2004). "Modeling Heterogeneous Dispersion in Marginal Simplex Models for Continuous Longitudinal Proportional Data." *Biometrical Journal*, **46**, 540–553. doi:10.1002/bimj.200110052.

Song PXK, Tan M (2000). "Marginal Models for Longitudinal Continuous Proportional Data." *Biometrics*, **56**, 496–502. doi:10.1111/j.0006-341x.2000.00496.x.

Stasinopoulos DM, Rigby RA (2007). "Generalized Additive Models for Location Scale and Shape (GAMLSS) in R." *Journal of Statistical Software*, **23**(7), 1–46. doi:10.18637/jss.v023.i07.

Yang H, Acker JP, Cabuhat M, Letcher B, Larratt L, McGann LE (2005). "Association of Post-Thaw Viable CD34+ Cells and CFU-GM with Time to Hematopoietic Engraftment." *Bone Marrow Transplantation*, **35**, 881–887. doi:10.1038/sj.bmt.1704926.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. doi:10.18637/jss.v034.i01.

Zhang P, Qiu Z (2014). "Regression Analysis of Proportional Data Using Simplex Distribution." *Science China Mathematics (Chinese Version)*, **44**, 89–104. doi:10.1360/012013-200.

Zhang P, Qiu Z, Shi C (2016). ***simplexreg**: Regression Analysis of Proportional Data Using Simplex Distribution*. R package version 1.3, URL https://CRAN.R-project.org/package=simplexreg.

Affiliation:

Peng Zhang
Department of Mathematics
Zhejiang University
310027 Hangzhou, China
E-mail: pengz@zju.edu.cn, 3100102177@zju.edu.cn

Zhenguo Qiu
Surveillance, CancerControl
Alberta Health Services
Edmonton, Alberta T5J 3H1, Canada
Email: zhenguo.qiu@albertahealthservices.ca

Chengchun Shi
Department of Statistics
North Carolina State University
27606 Raleigh, North Carolina, United States of America
Email: cshi4@ncsu.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
August 2016, Volume 71, Issue 11 *Submitted:* 2014-02-02
doi:10.18637/jss.v071.i11 *Accepted:* 2015-06-21