Simulating Correlated Binary and Multinomial Responses under Marginal Model Specification: The SimCorMultRes Package

by Anestis Touloumis

Abstract We developed the R package **SimCorMultRes** to facilitate simulation of correlated categorical (binary and multinomial) responses under a desired marginal model specification. The simulated correlated categorical responses are obtained by applying threshold approaches to correlated continuous responses of underlying regression models and the dependence structure is parametrized in terms of the correlation matrix of the latent continuous responses. This article provides an elaborate introduction to the **SimCorMultRes** package demonstrating its design and usage via three examples. The package can be obtained via CRAN.

Introduction

Fitting marginal models with correlated binary or multinomial responses is required in many applications in which the responses are assumed to be correlated. The obvious instance of such studies is longitudinal studies (Diggle et al., 2002) where the categorical responses for each subject are collected across time points and form a cluster. For each cluster, the associated covariates are also recorded as they might influence the true marginal probabilities. Ordinary regression models designed for independent responses might not lead to consistent estimators of the marginal regression parameters or of their standard errors. For this reason, many authors have developed and proposed procedures for estimating the regression parameters of a marginal model with categorical responses that are robust to misspecification of the dependence structure, including maximum likelihood methods (Fitzmaurice and Laird, 1993; Glonek and McCullagh, 1995), copula approaches (Masarotto and Varin, 2012), quasi-least squares approaches (Shults and Chaganty, 1998), generalized quasi-likelihood methods (Sutradhar and Das, 1999; Sutradhar, 2003) and generalized estimating equations (GEE) approaches (Lipsitz et al., 1991; Chaganty and Joe, 2004; Touloumis et al., 2013). Although the asymptotic properties of these methods are well-established, the evaluation of their performance in finite samples under misspecification of the correlation structure relies on simulations. The crucial step of these empirical studies is to simulate correlated categorical responses that satisfy a desired marginal model and dependence structure specification.

Motivated by this, we present the R package **SimCorMultRes** (Touloumis, 2016) which makes it easy to simulate correlated categorical responses under a given marginal model and dependence structure configuration. The package implements marginal models for correlated binary responses (two response categories) as well as for correlated multinomial responses (three or more response categories) while taking into account the nature of the response categories (ordinal or nominal). In summary, the correlated binary/multinomial responses are obtained as realizations of an underlying continuum. This means that latent regression models with correlated continuous responses are utilized so as to generate the correlated categorical responses that satisfy the desired marginal model specification. The categorical responses are obtained by applying threshold approaches to the correlated continuous responses. In order to avoid theoretical pitfalls outlined in the next paragraph, the desired dependence structure is expressed in terms of the correlation matrix of the latent responses. To the best of our knowledge, **SimCorMultRes** is the first package in R that allows direct simulation of correlated categorical responses under a marginal model specification with categorical and/or continuous covariates.

To fully appreciate the features of **SimCorMultRes**, we briefly compare it with two R packages: i) **GenOrd** (Barbiero and Ferrari, 2015), that implements the methods presented by Ferrari and Barbiero (2012) and its features being discussed in greater detail in Barbiero and Ferrari (in press), and ii) **MultiOrd** (Amatya and Demirtas, 2016), that is described in Amatya and Demirtas (2015) and relies on the simulation techniques proposed by Demirtas (2006). These packages are designed to simulate random vectors of correlated binary or ordinal responses subject to fixed but common marginal probabilities across all subjects and a predefined correlation matrix for the correlated categorical responses. Therefore, unlike **SimCorMultRes**, it is not straightforward to utilize **GenOrd** or **MultiOrd** for simulating categorical responses conditional on a regression model specification for the marginal probabilities, especially when the marginal probabilities vary across subjects. In addition, **SimCorMultRes** has the

unique feature (to the best of our knowledge) to simulate correlated nominal responses. Another difference between **SimCorMultRes** and the R packages **GenOrd** and **MultiOrd** is that the former requires the association among the categorical responses to be directly expressed via their correlation matrix and that the joint specification of the marginal probabilities and of the correlation matrix leads to a valid joint distribution for the correlated categorical responses. A necessary condition for this is that the so-called Frétchet-Hoeffding bounds are satisfied, which can be verified by employing the method of Demirtas and Hedeker (2011). As noted by one of the reviewers, both **GenOrd** and **MultiOrd** have built-in mechanisms to check the restrictions imposed by the Frétchet-Hoeffding bounds. Unfortunately, even if these restrictions are met, it is still theoretically possible that a legitimate joint distribution does not exist for the correlated categorical responses (Bergsma and Rudas, 2002). To circumvent this difficulty, the methodology implemented in **SimCorMultRes** always defines the joint distribution of the correlated categorical responses in terms of the joint distribution of correlated latent random variables and thus, it allows the user to generate correlated categorical responses under any configuration of the marginal probabilities provided that the user-defined correlation matrix of the latent continuous responses is positive definite, a condition that can be more easily verified.

The remainder of this paper is organized as follows. First we present the theoretical background of the threshold approaches implemented in **SimCorMultRes**. In particular, we introduce the general two-stage algorithm for simulating correlated categorical responses, focusing on the threshold approaches that give rise to the marginal models with correlated categorical responses and on the modified version of the NORmal To Anything (NORTA) method (Cario and Nelson, 1997), the default simulation method of correlated latent random variables in **SimCorMultRes**. Next, we describe the core and utility functions of the package. Then, we demonstrate the use of **SimCorMultRes** by considering the problems of evaluating two estimation methods for marginal models with correlated nominal multinomial responses, of assessing the quality of an approximation that links the uniform local odds ratios structure with the correlation parameter of an underlying bivariate normal distribution, and of simulating correlated categorical random variables under no marginal model specification. Finally, we summarize the features of **SimCorMultRes** and discuss future extensions.

Theoretical background

In this section, we introduce the threshold approaches that give rise to marginal models with correlated binary, ordinal or nominal responses. Since the thresholds are applied to correlated continuous responses, simulation of correlated continuous responses is required. This step can be performed in various ways, eġ, directly from an appropriate multivariate distribution, by utilizing distributional properties about the sum or the difference of random vectors or by employing copula approaches. Herein we discuss a simple and straightforward simulation method that is based on the NORTA method, and we present a general algorithm that combines the threshold approaches with the modified NORTA method, enabling us to generate correlated categorical responses subject to a marginal model specification in a unified manner. However, we underline that the use of the NORTA method is optional in the general algorithm and that it can be replaced with another simulation method/technique as long as the distributional restrictions regarding the correlated continuous variables that are imposed by the thresholds are met.

For notational ease, adopt a longitudinal set-up for generating the correlated binary or multinomial variables. Let Y_{it} be the random variable of subject i $(i = 1, \ldots, N)$ at time t $(t = 1, \ldots, T)$ and let $\mathbf{x}_{it}$ denote the associated covariates vector. To be consistent with the notation in the majority of the literature, let $Y_{it} \in \{0, 1\}$ when there are two response categories and let $Y_{it} \in \{1, 2, \ldots, J \geq 3\}$ for at least three categories.

Binary responses

Suppose the aim is to simulate correlated binary variables such that the marginal probabilities satisfy the model

$$\Pr\left(Y_{it} = 1 | \mathbf{x}_{it}\right) = F\left(\beta_{t0} + \boldsymbol{\beta}_t' \mathbf{x}_{it}\right) \tag{1}$$

where β_{t0} is the intercept and $\boldsymbol{\beta}_t$ is the covariates parameter vector at time t, respectively, and where F is a cumulative distribution function (c.d.f.).

Now, consider the multivariate latent regression model

$$\mathbf{U}_i^B = \begin{pmatrix} U_{i1}^B \\ \vdots \\ U_{iT}^B \end{pmatrix} = \begin{pmatrix} \mu_{i1}^B \\ \vdots \\ \mu_{iT}^B \end{pmatrix} + \begin{pmatrix} e_{i1}^B \\ \vdots \\ e_{iT}^B \end{pmatrix} = \boldsymbol{\mu}_i^B + \mathbf{e}_i^B$$

where $\mu_{it}^B = \boldsymbol{\beta}_t' \mathbf{x}_{it}$ and $\{\mathbf{e}_i^B : i = 1, \ldots, N\}$ are independent random vectors such that $e_{it}^B \sim F$ for all i and t. Under these assumptions, generation of binary responses under the threshold

$$Y_{it} = I\left(e_{it}^B \leq \beta_{t0} + \mu_{it}^B\right) = I\left(U_{it}^B \leq \beta_{t0} + 2\mu_{it}^B\right)$$

gives rise to the marginal model (1), where $I(A)$ denotes the indicator function of the event A. This approach is a straightforward extension of the gold-standard simulation method proposed by Emrich and Piedmonte (1991), in the sense that it also permits marginal modeling of the univariate probabilities through covariates. Implementation of the method of Emrich and Piedmonte (1991) can be found in the orphaned R package **mvtBinaryEP** (By and Qaqish, 2011).

Ordinal responses

Options for marginal modelling of correlated ordinal responses include the marginal cumulative link model

$$\Pr\left(Y_{it} \leq j | \mathbf{x}_{it}\right) = F\left(\beta_{tj0} + \boldsymbol{\beta}_t' \mathbf{x}_{it}\right) \tag{2}$$

and the marginal continuation-ratio model

$$\Pr\left(Y_{it} = j | Y_{it} \geq j, \mathbf{x}_{it}\right) = F\left(\beta_{tj0} + \boldsymbol{\beta}_t' \mathbf{x}_{it}\right). \tag{3}$$

In both models, F is a c.d.f. and $\boldsymbol{\beta}_t$ is the parameter vector at time t when the corresponding $(J-1)$ category-specific intercepts $\left(\beta_{t10}, \beta_{t20}, \ldots, \beta_{t(J-1)0}\right)$ are excluded.

First, consider the marginal cumulative link model (2) and suppose the multivariate latent regression model

$$\mathbf{U}_i^{O1} = \begin{pmatrix} U_{i1}^{O1} \\ \vdots \\ U_{iT}^{O1} \end{pmatrix} = \begin{pmatrix} \mu_{i1}^{O1} \\ \vdots \\ \mu_{iT}^{O1} \end{pmatrix} + \begin{pmatrix} e_{i1}^{O1} \\ \vdots \\ e_{iT}^{O1} \end{pmatrix} = \boldsymbol{\mu}_i^{O1} + \mathbf{e}_i^{O1}$$

holds, where $\mu_{it}^{O1} = -\boldsymbol{\beta}_t' \mathbf{x}_{it}$, and $\{\mathbf{e}_i^{O1} : i = 1, \ldots, N\}$ are independent random vectors such that $e_{it}^{O1} \sim F$ for all i and t. To generate an ordinal response Y_{it} that satisfies model (2), one can categorize U_{it}^B by using the corresponding category-specific intercepts according to the threshold

$$Y_{it} = j \Leftrightarrow \beta_{t(j-1)0} < U_{it}^{O1} \leq \beta_{tj0}$$

where

$$-\infty = \beta_{t00} < \beta_{t10} < \beta_{t20} < \cdots < \beta_{t(J-1)0} < \beta_{tJ0} = \infty.$$

This threshold approach extends the approach discussed in McCullagh (1980) from cumulative link models with independent ordinal responses to marginal cumulative link models with correlated ordinal responses.

Next, consider the marginal continuation-ratio model (3) and suppose the following multivariate latent regression model holds

$$\mathbf{U}_i^{O2} = \begin{pmatrix} \mathbf{U}_{i1}^{O2} \\ \vdots \\ \mathbf{U}_{iT}^{O2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{i1}^{O2} \\ \vdots \\ \boldsymbol{\mu}_{iT}^{O2} \end{pmatrix} + \begin{pmatrix} \mathbf{e}_{i1}^{O2} \\ \vdots \\ \mathbf{e}_{iT}^{O2} \end{pmatrix} = \boldsymbol{\mu}_i^{O2} + \mathbf{e}_i^{O2}$$

where $\mathbf{U}_{it}^{O2} = \left(U_{it1}^{O2}, \ldots, U_{itJ}^{O2}\right)'$, $\boldsymbol{\mu}_{it}^{O2} = -\left(\boldsymbol{\beta}_t' \mathbf{x}_{it}, \ldots, \boldsymbol{\beta}_t' \mathbf{x}_{it}\right)'$ and $\mathbf{e}_{it}^{O2} = \left(e_{it1}^{O2}, \ldots, e_{itJ}^{O2}\right)'$ for all i and t, and $\{\mathbf{e}_i^{O2} : i = 1, \ldots, N\}$ are independent random vectors such that:

1. $e_{itj}^{O2} \sim F$ for all i, t and j,
2. e_{itj}^{O2} and $e_{itj'}^{O2}$ are independent for all $j \neq j'$ (local independence assumption).

The marginal continuation-ratio model (3) arises by applying the threshold

$$Y_{it} = j, \text{ given } Y_{it} \geq j \Leftrightarrow U_{itj}^{O2} \leq \beta_{tj0}$$

to the components of $\mathbf{U}_{it}$'s in a sequential order. This approach extends the latent variable representation described in Tutz (1991) which gives rise to the continuation-ratio model for independent ordinal responses (see Agresti, 2013).

Nominal responses

Consider the marginal baseline-category logit model

$$\log\left[\frac{\Pr\left(Y_{it}=j|\mathbf{x}_{it}\right)}{\Pr\left(Y_{it}=J|\mathbf{x}_{it}\right)}\right]=\left(\beta_{tj0}-\beta_{tJ0}\right)+\left(\boldsymbol{\beta}_{tj}-\boldsymbol{\beta}_{tJ}\right)'\mathbf{x}_{it}=\beta^{*}_{tj0}+\boldsymbol{\beta}^{*\prime}_{tj}\mathbf{x}_{it} \tag{4}$$

where β_{tj0} and $\boldsymbol{\beta}_{tj}$ is the j-th category-specific intercept and parameter vector at time t, respectively. For identifiability reasons, restrictions such as $\beta_{tJ0}=0$ and $\boldsymbol{\beta}_{tJ}=\mathbf{0}$ for all t are required, which imply that $\beta^{*}_{tj0}=\beta_{tj0}$ and $\boldsymbol{\beta}^{*}_{tj}=\boldsymbol{\beta}_{tj}$ for all t and for all $j=1,\ldots,J-1$. Note that model (4) relates with the baseline-category logit model (see Agresti, 2013) and hence it is appropriate for marginal modelling of correlated nominal responses.

To connect the marginal baseline-category logit model (4) with underlying regression models, consider the multivariate latent regression model

$$\mathbf{U}^{NO}_{i}=\begin{pmatrix}\mathbf{U}^{NO}_{i1}\\ \vdots\\ \mathbf{U}^{NO}_{iT}\end{pmatrix}=\begin{pmatrix}\boldsymbol{\mu}^{NO}_{i1}\\ \vdots\\ \boldsymbol{\mu}^{NO}_{iT}\end{pmatrix}+\begin{pmatrix}\mathbf{e}^{NO}_{i1}\\ \vdots\\ \mathbf{e}^{NO}_{iT}\end{pmatrix}=\boldsymbol{\mu}^{NO}_{i}+\mathbf{e}^{NO}_{i}$$

where $\mathbf{U}^{NO}_{it}=\left(U^{NO}_{it1},\ldots,U^{NO}_{itJ}\right)'$, $\boldsymbol{\mu}^{NO}_{it}=\left(\beta_{t10}+\boldsymbol{\beta}'_{t1}\mathbf{x}_{it},\ldots,\beta_{t(J-1)0}+\boldsymbol{\beta}'_{tJ}\mathbf{x}_{it}\right)'$ and $\mathbf{e}^{NO}_{it}=\left(e^{NO}_{it1},\ldots,e^{NO}_{itJ}\right)'$ for all i and t, and $\{\mathbf{e}^{NO}_{i}:i=1,\ldots,N\}$ are independent random vectors such that:

1. e^{NO}_{itj} follow a standard extreme distribution for all i, t and j,
2. the assumption of choice independence is met at each measurement occasion, that is e^{NO}_{itj} and $e^{NO}_{itj'}$ are independent for all $j\neq j'$.

The threshold

$$Y_{it}=j\Leftrightarrow U_{itj}=\max\{U_{it1},\ldots,U_{itJ}\}$$

extends the principle of maximum random utility (McFadden, 1974) and it generates correlated nominal responses that give rise to the marginal baseline-category logit model (4).

Simple version of the NORTA method

Li and Hammond (1975) proposed a simple method for generating continuous random vectors with given marginal distributions and a prescribed correlation matrix. Cario and Nelson (1997) introduced the NORTA method which essentially modifies the approach of Li and Hammond (1975) to account for any type of marginal distributions (discrete, continuous or mixed). Here, we describe a simple version of the NORTA method in which the desired marginal distributions are continuous and identical which is required by all the threshold approaches implemented in **SimCorMultRes**.

Let F be the c.d.f. of the target marginal distribution. To generate a p-variate random vector $\mathbf{W}=\left(W_1,\ldots,W_p\right)'$ with correlation matrix $\text{cor}\left(\mathbf{W}\right)=\mathbf{R}_W$ such that $W_k\sim F$ for all $k=1,\ldots,p$, the following NORTA transformation can be utilized:

1. Generate a random vector $\mathbf{Z}=\left(Z_1,\ldots,Z_p\right)'$ from a standard multivariate normal distribution with correlation matrix $\text{cor}\left(\mathbf{Z}\right)=\mathbf{R}_Z$. The elements of $\mathbf{R}_Z$ are calculated by solving numerically $p\left(p-1\right)/2$ equations, such that each equation relates $\text{cor}\left(Z_k,Z_{k'}\right)$ with $\text{cor}\left(W_k,W_{k'}\right)$ for all $k<k'$. The exact formulae are given by Li and Hammond (1975).
2. Apply the transformation $W_k=F^{-1}\left[\Phi\left(Z_k\right)\right]$ for all k, where Φ is the cumulative distribution of the standard normal distribution.

If $F=\Phi$, then the second step of the above modified NORTA algorithm is not needed. Otherwise, the correlation matrices $\mathbf{R}_Z$ and $\mathbf{R}_W$ are expected to differ. In fact, Cario and Nelson (1997) showed that under mild conditions it is possible to have $\mathbf{R}_Z\approx\mathbf{R}_W$. For example, if F is the cumulative distribution function of the standard logistic distribution (which might be the case in the marginal models for correlated binary and ordinal responses), then $\mathbf{R}_Z\approx\mathbf{R}_W$ due to the well-known approximation $\Phi\left(x\right)=F\left(x\pi/3\right)$ for all $x\in\Re$. This simplifies the computational task as the $p\left(p-1\right)/2$ equations are not needed to be solved and issues regarding non-existence of a valid correlation matrix $\mathbf{R}_Z$ for a given choice of the correlation matrix $\mathbf{R}_W$ (Li and Hammond, 1975) are avoided provided that $\mathbf{R}_Z$ is positive-definite under mild conditions (Cario and Nelson, 1997).

General generative process

We propose a simple and efficient two-staged general algorithm for generating correlated categorical responses:

Stage 1. *Marginal model specification:* Provide the covariates, the regression parameters and the link function (if required) of the desired marginal model (1), (2), (3), or (4).

Stage 2. *Simulation of continuous random vectors via the NORTA method and threshold approach:* Define the desired dependence structure by fixing $\mathbf{R}_Z$ to generate the continuous random vectors $\mathbf{U}_i$'s from the multivariate latent regression model implied by the marginal model specification selected in Stage 1. Apply the corresponding threshold approach to obtain the correlated binary, ordinal or nominal responses.

In the marginal models described above, we usually choose F to be the c.d.f. of a standard normal, logistic or extreme value distribution. In either of these cases, it can be shown that the simulated categorical responses are independent if and only if $\mathbf{R}_Z$ is the identity matrix, which is true if and only if the random variables in the latent random vectors $\mathbf{e}_i$'s are independent (Cario and Nelson, 1997). For all other forms of $\mathbf{R}_Z$, correlated categorical responses will be generated.

Expressing the association structure in terms of $\mathbf{R}_Z$ ensures the existence of a joint distribution for the correlated categorical responses regardless of the marginal model specification which is not the case when the association is expressed directly via the correlation matrix of the correlated categorical responses. This well-known fact has been mentioned by Bergsma and Rudas (2002) among others, and it has been exemplified in the case of correlated binary and multinomial responses by Chaganty and Joe (2004), Chaganty and Joe (2006) and Touloumis et al. (2013), respectively. The simplest scenario where adopting a common correlation matrix for the correlated categories responses across subjects is problematic is when the linear predictor in the marginal model is allowed to vary freely on the real line. In this case, only the identity matrix is a feasible value for the correlation matrix.

As mentioned before, the proposed version of the NORTA method is not the only option to simulate continuous random vectors in Stage 2 and instead, alternative simulation techniques can be easily employed. However, the user must be cautious in order to respect the corresponding marginal distributional assumptions and the assumption of local independence or choice independence whenever the marginal models (3) or (4) are used, respectively.

We emphasize that the proposed algorithm can also handle the situation in which no marginal model specification is provided. For more details, please refer to the third example below.

Description of SimCorMultRes

SimCorMultRes contains four core functions (`rbin`, `rmult.bcl`, `rmult.clm` and `rmult.crm`) that enable the user to generate correlated categorical responses and two utility functions (`rnorta` and `rsmvnorm`) initially designed for internal use in the core functions. We describe in detail the arguments and the output of the core and utility functions.

Core functions

Each core function in **SimCorMultRes** simulates correlated categorical responses under a marginal model specification. In particular, `rbin` simulates correlated binary responses that satisfy the marginal model (1), `rmult.clm` simulates correlated ordinal responses that satisfy the marginal cumulative link model (2), `rmult.crm` simulates correlated ordinal responses that satisfy the marginal continuation-ratio model (3) and `rmult.bcl` simulates correlated nominal responses that satisfy the marginal baseline-category logit model (4).

The common cluster size (`clsize`) of the subjects is required in all core functions.

The `ncategories` argument in `rmult.bcl` indicates the number of nominal response categories. The number of ordinal response categories in `rmult.clm` and `rmult.crm` is indirectly defined by the `intercepts` argument. It contains the values of the threshold parameters which can be provided either as a $T \times (J-1)$ matrix or as a vector of length $J-1$. In the first case, the (t,j)-th element of `intercepts` corresponds to β_{tj0} and in the second case, it is assumed that $\beta_{tj0} = \beta_{j0}$ for all t in the marginal models (2) or (3). The `intercepts` argument is also employed in `rbin` to specify whether the intercepts in the marginal model (1) are time-dependent. If $\beta_{t0} = \beta_0$ for all t, then `intercepts` should be a single number that reflects the value of β_0. Otherwise, it should be a vector of size T with the t-th element equal to the value of β_{t0}.

The values for the marginal regression parameters (`betas`) should be provided as a numeric vector whenever $\boldsymbol{\beta}_t = \boldsymbol{\beta}$ for all t in models (1), (2) or (3), and whenever $\beta_{tj} = \beta_j$ and $\boldsymbol{\beta}_{tj} = \boldsymbol{\beta}_j$ for all t in

model (4). In all other cases, `betas` should be provided as a matrix with T rows such that the t-th row contains the value of the marginal parameter vector at time t. It is important to emphasize that (category-specific) intercept values should not be included in `betas` unless the function `rmult.bcl` is used.

The functional relationship of the covariates in the marginal model (`xformula`) is specified similarly as in other regression models with the single difference that no response variable should be provided. The covariates defined in `xformula` can be imported via the `xdata` argument in "long" format, meaning that each row contains all the subject-specific covariates information at a given time. When `xdata` is missing, then the covariates are extracted from the environment that the core function is called.

The `link` argument in `rbin`, `rmult.clm` or `rmult.crm` determines the c.d.f. F in the marginal models (1), (2) or (3) respectively, i. e., the link function. Options for the link function include the probit (`"probit"`), the logit (`"logit"`), the complimentary log-log (`"cloglog"`) and the cauchit (`"cauchit"`). It is worth mentioning that there is no `link` argument in the function `rmult.bcl` because the marginal distribution of the latent continuous random variables e_{itj}^{NO}'s is always the standard extreme value distribution.

In all core functions, the latent random vectors $\mathbf{e}_i$'s can be either simulated using the proposed NORTA approach or provided by the user via the `rlatent` argument. In the first case, the correlation matrix $\mathbf{R}_Z$ of the multivariate normal distribution (`cor.matrix`) in the modified NORTA method and the `link` argument, wherever present, are required. Checks are carried out to ensure that `cor.matrix` is a positive-definite correlation matrix and whenever `rmult.crm` or `rmult.bcl` is employed, `cor.matrix` is forced to satisfy the restrictions of the latent dependence structure that are implied by the threshold approach associated with models (3) or (4), respectively. In the case where the preferred simulation method is not the NORTA method, `rlatent` should contain the values of the latent random vectors while `cor.matrix` and `link` are ignored. Examples of using the `rlatent` argument can be found in the help files and the vignette of **SimCorMultRes**.

The output of any core function is displayed as a list with three items: (i) a matrix with the simulated responses such that the (i, t)-th element corresponds to the realization of Y_{it} (`Ysim`), (ii) a data frame (`simdata`) that contains the simulated responses (y), the covariates specified by `xformula`, subjects' identities (id) and the measurement occasions (time), and (iii) the NORTA generated or user-defined latent random vectors (`rlatent`).

Utility functions

The utility function `rnorta` offers a more general implementation of the NORTA method described earlier. The user needs to specify the number of random vectors (`R`), the correlation matrix $\mathbf{R}_Z$ of the multivariate normal distribution (`cor.matrix`) and the names of the quantile functions of the desired marginal distributions (`distr`). The optional `qparameters` argument permits users to consider parameter values for the marginal distributions other than the default (obtained when `qparameters = NULL`). The function returns `R` random vectors with marginal distributions specified by `distr` (and `qparameters`) when `cor.matrix` is the correlation matrix of the multivariate normal distribution in the NORTA method. We highlight that `rnorta` has been extended to handle situations that are beyond the scope of simulation of correlated categorical responses subject to a marginal model specification. Unlike the simple version of the NORTA method needed for our purposes, `rnorta` does not require marginal distributions to be identical. In fact, any univariate discrete or continuous distribution whose quantile function is available in R can be employed in `distr` provided that the required R package is available.

The function `rsmvnorm` generates `R` random vectors from a multivariate normal distribution with mean vector the zero vector and covariance matrix `cor.matrix`.

Note that an error message is returned whenever `cor.matrix` in functions `rnorta` or `rsmvnorm` is not a positive-definite correlation matrix.

Empirical illustration

We now illustrate the use of **SimCorMultRes** to: i) evaluate the performance of GEE approaches for estimating the regression parameters of a marginal baseline-category logit model, ii) to verify approximations that relate a uniform local odds ratios structure to the correlation coefficient of a bivariate normal distribution (Goodman, 1979) and, iii) to simulate correlated categorical random variables with fixed arbitrary univariate probabilities that are not subject to a marginal model specification.

Parameter estimation of marginal models

The motivation behind the creation of **SimCorMultRes** lies on evaluating statistical methods that estimate the regression coefficients of marginal models with correlated binary or multinomial responses. To exemplify this, we employ two GEE models for estimating a marginal model with correlated nominal responses: i) the local odds ratios GEE approach (Touloumis et al., 2013) and ii) the independence "working" model, which treats all observations as independent when solving the estimating equations. Although the two competing GEE models are asymptotically equally efficient, in the sense that they both produce consistent estimators for the marginal regression parameters and of their standard errors, the regression coefficient estimators of the independence "working" model are expected to be slightly less precise than those of the local odds ratios GEE approach in small and moderate sample sizes due to the fact that the independence "working" model does not account for the dependence among the correlated responses (Touloumis et al., 2013).

To investigate this assertion for the case of correlated nominal responses, we employed the marginal baseline-category logit model

$$\log\left[\frac{\Pr\left(Y_{it}=j|\mathbf{x}_{it}\right)}{\Pr\left(Y_{it}=5|\mathbf{x}_{it}\right)}\right]=\beta_{j0}+\beta_{j1}x_{it} \tag{5}$$

where $(\beta_{10},\beta_{11},\beta_{20},\beta_{21},\beta_{30},\beta_{31},\beta_{40},\beta_{41})=(2,1,1,2,1.5,1.5,2.5,0.5)$ and $x_{it}\overset{i.i.d.}{\sim}\mathrm{N}(0,1)$ for all $i=1,\ldots,100$ and $t=1,2,3,4$. Further, the correlation matrix $\mathbf{R}_Z$ among the normally distributed variables Z_{itj}'s in the NORTA method was given by

$$\mathrm{cor}\left(Z_{itj},Z_{it'j'}\right)=\begin{cases}1 & \text{if } t=t' \text{ and } j=j'\\ 0 & \text{if } t=t' \text{ and } j\neq j'\\ 0.56^{tj-t'j'} & \text{if otherwise.}\end{cases}$$

```
> library("SimCorMultRes")
> library("multgee")
Loading required package: gnm
Loading required package: VGAM
Loading required package: stats4
Loading required package: splines
> set.seed(1)
> N <- 100
> clsize <- 4
> ncategories <- 5
> betas <- c(2, 1, 1, 2, 1.5, 1.5, 2.5, 0.5, 0, 0)
> x <- rnorm(N * clsize)
> cor.matrix <- toeplitz(0.56^seq(0, clsize * ncategories - 1))
> for (i in 1:clsize) {
+   diag.index <- 1:ncategories + (i - 1) * ncategories
+   cor.matrix[diag.index, diag.index] <- diag(1, ncategories)
+ }
```

Conditional on the above marginal model specification and dependence structure, we simulated correlated nominal responses and we fitted the local odds ratios GEE approach with an RC-type dependence structure and the independence "working" model using the R package **multgee** (Touloumis, 2015). We replicated this procedure 1000 times and at each iteration we recorded the estimates of the marginal regression parameter vector of the two competing models:

```
> B <- 1000
> indeGEEcoefs <- matrix(NA_real_, B, 8)
> RCGEEcoefs <- matrix(NA_real_, B, 8)
> for (b in 1:B) {
+   SimNomRes <- rmult.bcl(clsize = clsize, ncategories = ncategories,
+                          betas = betas, xformula = ~x, cor.matrix = cor.matrix)
+   fitRC <- try(nomLORgee(y ~ x, id = id, repeated = time, data = SimNomRes$simdata,
+                          LORstr = "RC", add = 0.05), silent = TRUE)
+   if (!inherits(fitRC, "try-error")) {
+     if (fitRC$convergence$conv)
+       RCGEEcoefs[b, ] <- coef(fitRC)
+   }
+   fitinde <- try(nomLORgee(y ~ x, id = id, repeated = time, data = SimNomRes$simdata,
+                            LORstr = "independence"), silent = TRUE)
```

Table 1: Simulation results for the local odds GEE approach (LOR) and the independence "working" model (IEE) for estimating the regression parameter vector and their standard errors (second row) in the marginal model (5).

Model	$\beta_{10} = 2$	$\beta_{11} = 1$	$\beta_{20} = 1$	$\beta_{21} = 2$	$\beta_{30} = 1.5$	$\beta_{31} = 1.5$	$\beta_{40} = 2.5$	$\beta_{41} = 0.5$
IEE	2.0701	1.0308	1.0494	2.0530	1.5682	1.5441	2.5702	0.5209
	0.3567	0.3232	0.4011	0.3601	0.3740	0.3412	0.3627	0.2995
LOR	2.0471	0.9951	1.0378	1.9832	1.5487	1.4943	2.5473	0.4946
	0.3512	0.3105	0.3944	0.3490	0.3673	0.3319	0.3562	0.2873
SRE	1.0522	1.0927	1.0405	1.0852	1.0526	1.0738	1.0573	1.0921

```
+   if (!inherits(fitinde, "try-error")) {
+     if (fitinde$convergence$conv)
+       indeGEEcoefs[b, ] <- coef(fitinde)
+   }
+ }
```

Although the local odds GEE approach did not always converge, the convergence rate for the local odds ratios GEE model was high

```
> convergence <- c(mean(!is.na(indeGEEcoefs)), mean(!is.na(RCGEEcoefs))) * 100
> convergence
[1] 100.0  99.7
```

and therefore, we can conduct a fair comparison by excluding the results from those 3 iterations in which the local odds ratios GEE approach failed to converge.

Table 1 summarizes the simulation results by displaying the simulated mean and standard error of the regression estimates from the two competing GEE models and the simulated relative efficiency (SRE) for each regression parameter of model (5). For a given coefficient of model (5), the SRE criterion was defined as the ratio of the simulated mean square error of the corresponding Monte Carlo estimate based on the local odds ratios GEE approach to that based on the independence "working" model. Values of the SRE criterion greater (less) than 1.0 imply that the local odds ratios GEE approach is more (less) efficient than the independence "working" model in estimating this specific regression parameter. As expected, the two GEE models seem to estimate consistently the marginal model (5), with the local odds ratios GEE approach being 4.05%–9.27% more efficient in estimating each regression coefficient.

The results of Table 1 were calculated using the following R commands:

```
> simindemean <- colMeans(indeGEEcoefs, na.rm = TRUE)
> simindesd <- apply(indeGEEcoefs, 2, function(x) sd(x, na.rm = TRUE))
> simRCmean <- colMeans(RCGEEcoefs, na.rm = TRUE)
> simRCsd <- apply(RCGEEcoefs, 2, function(x) sd(x, na.rm = TRUE))
> simindesmse <- (betas[-c(9:10)] - simindemean)^2 + simindesd^2
> simRCsmse <- (betas[-c(9:10)] - simRCmean)^2 + simRCsd^2
> SRE <- simindesmse/simRCsmse
> rbind(simindemean, simindesd, simRCmean, simRCsd, SRE)
```

Uniform association model and bivariate normal distribution

Let f_{ab} denote the observed frequency of the cell (a, b) in a two-way contingency table and let F_{ab} be the corresponding expected frequency under some model, for $a = 1, \ldots, A$ and $b = 1, \ldots, B$. Goodman (1979) proposed the uniform association model

$$\log(F_{ab}) = \nu + \kappa_a + \lambda_b + \phi \tag{6}$$

where the parameters ν, $\{\kappa_a : a = 1, \ldots, A\}$, $\{\lambda_b : b = 1, \ldots, B\}$ and ϕ are identifiable once restrictions, such as sum to zero constraints (Agresti, 2013), are applied to $\{\kappa_a : a = 1, \ldots, A\}$ and $\{\lambda_b : b = 1, \ldots, B\}$. The association between the row and column variables is modelled parsimoniously by assuming a common value ϕ for the $(A-1) \times (B-1)$ log local odds ratios. The key property of the uniform association model is that ϕ relates to the correlation parameter ρ of an underlying bivariate normal distribution (Goodman, 1979) via the approximations

$$\phi \approx \frac{\rho}{1-\rho^2} \frac{11}{12} \tag{7}$$

or

$$\rho \approx \left(\sqrt{1+\eta^2} - \eta\right) \times 13/12 \tag{8}$$

where $\eta = (2\phi)^{-1}$. The validity of these approximations has been explored only for $\rho = 0.5$ by Goodman (1979). Here, we perform a more detailed empirical investigation by considering a grid of values for ρ, namely $\rho = 0.05, 0.10, \ldots, 0.95$.

For each value of ρ, we simulated 1000 random vectors from a bivariate normal distribution with mean vector the zero vector and covariance matrix the correlation matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In a similar fashion as in Goodman (1979), correlated ordinal responses were generated by applying the threshold approach linked to model (2), with $F = \Phi$ and equi-distanced category-specific intercepts $(\beta_{10}, \beta_{20}, \beta_{30}, \beta_{40}, \beta_{50}, \beta_{60}, \beta_{70}) = (-3, -2, -1, 0, 1, 2, 3)$. The sampling scheme does not involve any covariates, that is $\boldsymbol{\beta}_t = \mathbf{0}$ and $\mathbf{x}_{it} = \mathbf{0}$ for all t. Next, we cross-classified the correlated simulated responses to obtain a 8×8 contingency table and we estimated ϕ by fitting the uniform association model (6). We repeated this procedure 10000 times:

```
> library("SimCorMultRes")
> set.seed(123)
> commonlogoddsratio <- function(N, rho, intercepts, B) {
+   cor.matrix <- toeplitz(c(1, rho))
+   x <- rep(0, 2 * N)
+   ans <- rep(0, B)
+   for (b in 1:B) {
+     CorOrdRes <- rmult.clm(clsize = 2, intercepts = intercepts, betas = 0,
+                            xformula = ~x, link = "probit", cor.matrix = cor.matrix)
+     simdata <- data.frame(table(CorOrdRes$Ysim[, 1], CorOrdRes$Ysim[, 2]))
+     if (any(simdata[, 3] == 0))
+       simdata[, 3] <- simdata[, 3] + 0.001
+     colnames(simdata) <- c("x", "y", "Freq")
+     fit <- glm(Freq ~ x + y + as.numeric(x):as.numeric(y), family = poisson(),
+                data = simdata)
+     ans[b] <- as.numeric(coef(fit)[length(coef(fit))])
+   }
+   ans
+ }
> N <- 1000
> intercepts <- c(-3, -2, -1, 0, 1, 2, 3)
> B <- 10000
> rho <- seq(0.05, 0.95, 0.05)
> logoddsratio <- rep(0, length(rho))
> for (i in seq_along(rho)) {
+   simdata <- commonlogoddsratio(N, rho[i], intercepts, B)
+   logoddsratio[i] <- mean(simdata)
+ }
There were 50 or more warnings (use warnings() to see the first 50)
> eta <- 1/(2 * logoddsratio)
> rhophi <- (sqrt(1 + eta^2) - eta) * 13/12
```

The produced `warnings()` reflect the fact that we have added 0.001 to each cell of the two-way contingency table whenever an observed zero count occurred to ensure the existence of the maximum likelihood estimator of ϕ (Birch, 1963). We estimated the underlying correlation parameter ρ with $\rho_{\hat{\phi}}$ obtained by replacing ϕ in (8) with its Monte Carlo counterpart $\hat{\phi}$. The following R commands were run to obtain Figure 1.

```
> absdif <- abs(rhophi - rho)
> plot(rho, absdif, xlab = expression(rho), ylab = expression(abs(rho[hat(phi)] -
+                                                                rho)), xaxt = "n")
> axis(1, at = seq(0.05, 0.95, 0.1), labels = seq(0.05, 0.95, 0.1))
```

Figure 1 displays the absolute difference between the true correlation parameter ρ and $\rho_{\hat{\phi}}$. In general, approximation (8) seems to work well for weak correlation patterns, that is when $\rho \leq 0.20$. The

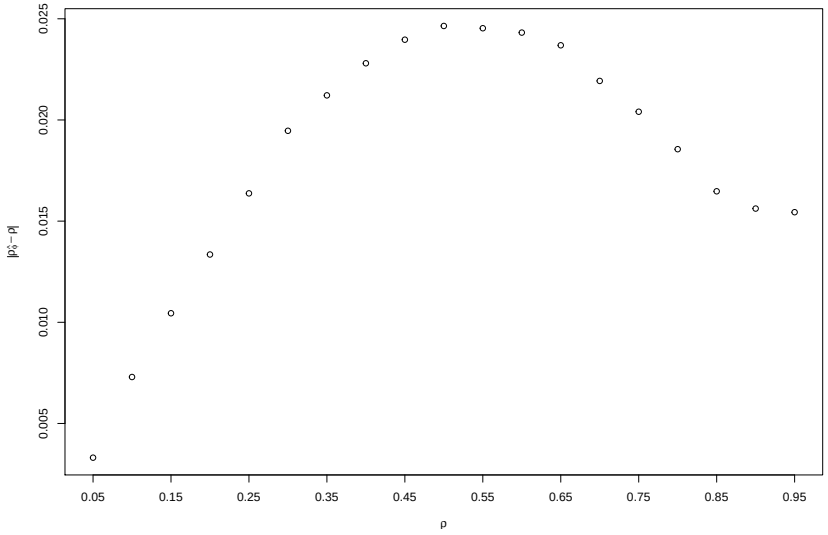

Figure 1: The absolute difference between the true correlation parameter ρ and $\rho_{\hat{\phi}}$, the correlation implied by the association model (6).

simulated absolute difference increases slightly for $0.25 \leq \rho \leq 0.55$ and then it decreases as $0.6 \leq \rho \leq 0.95$. In addition, the Monte Carlo estimates of ϕ increase as the true value of ρ increases, which suggests that ϕ does capture the strength of the underlying correlation structure. Therefore, we may conclude that approximations (7) and (8) can adequately describe the relationship between the uniform local odds ratios parameter ϕ in the uniform association model (6) and the correlation parameter ρ of an underlying bivariate normal distribution.

Simulating correlated categorical responses under no marginal model specification

For completeness' sake, we illustrate how to utilize **SimCorMultRes** in order to generate correlated categorical random variables conditional on a desired dependence structure and known marginal probabilities that are not determined by a regression model.

Suppose the goal is to simulate 5000 trivariate vectors $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})'$ of multinomial responses such that $Y_{it} \in \{1, 2, 3, 4\}$,

$$\begin{array}{llll} \Pr(Y_{i1} = 1) = 0.1 & \Pr(Y_{i1} = 2) = 0.3 & \Pr(Y_{i1} = 3) = 0.4 & \Pr(Y_{i1} = 4) = 0.2 \\ \Pr(Y_{i2} = 1) = 0.2 & \Pr(Y_{i2} = 2) = 0.2 & \Pr(Y_{i2} = 3) = 0.2 & \Pr(Y_{i2} = 4) = 0.4 \\ \Pr(Y_{i3} = 1) = 0.2 & \Pr(Y_{i3} = 2) = 0.4 & \Pr(Y_{i3} = 3) = 0.3 & \Pr(Y_{i3} = 4) = 0.1 \end{array}$$

and a common uniform local odds ratio structure

$$\phi_{tt'} = \frac{\Pr(Y_{it} = j, Y_{it'} = j') \Pr(Y_{it} = j+1, Y_{it'} = j'+1)}{\Pr(Y_{it} = j, Y_{it'} = j'+1) \Pr(Y_{it} = j+1, Y_{it'} = j')} = 2$$

holds for all $i = 1, \ldots, 5000$, $t < t'$ and $j, j' = 1, 2, 3$. The above sampling scheme can be reparametrized in terms of the threshold approach related to the marginal cumulative link model (2) while utilizing the conclusions of the previous example to obtain the desired dependence structure.

To this direction, first define $\beta_{tj0} = \Phi^{-1}[\Pr(Y_{it} \leq j)]$ for all t ($t = 1, 2, 3$) and j ($j = 1, 2, 3$) as the category-specific intercepts of a marginal cumulative probit model with no covariates:

```
> library(SimCorMultRes)
> set.seed(123)
> N <- 5000
> clsize <- 3
> mprobs_1 <- c(0.1, 0.3, 0.4, 0.2)
> mprobs_2 <- c(0.2, 0.2, 0.2, 0.4)
> mprobs_3 <- c(0.2, 0.4, 0.3, 0.1)
> cprobs_1 <- cumsum(mprobs_1[-4])
> cprobs_2 <- cumsum(mprobs_2[-4])
> cprobs_3 <- cumsum(mprobs_3[-4])
> intercepts <- qnorm(rbind(cprobs_1, cprobs_2, cprobs_3))
```

```
> x <- rep(0, clsize * N)
```

Next, for the pairwise dependence structure, approximate the desired pairwise uniform local odds ratios $\phi_{12}, \phi_{13}, \phi_{23}$ via the correlation parameters of an underlying trivariate normal distribution with mean vector the zero vector. Since the desired pairwise local odds ratios are all equal ($\phi_{12} = \phi_{13} = \phi_{23} = 2$), we may assume that the corresponding correlation parameters are all equal. This common correlation parameter ρ can be sufficiently approximated by equation (8), which suggests that $\rho \approx 0.5543$:

```
> CommomLOR <- log(2)
> eta <- 1/(2 * CommomLOR)
> rhophi <- (sqrt(1 + eta^2) - eta) * 13/12
> rhophi
[1] 0.5543136
```

To sum up, the desired correlated multinomial responses can be simulated under a cumulative probit model with no covariates and an exchangeable correlation matrix for the underlying trivariate normal distribution with correlation parameter equal to 0.5543:

```
> cor.matrix <- toeplitz(c(1, rhophi, rhophi))
> simdata <- rmult.clm(clsize = clsize, intercepts = intercepts, betas = 0,
+                      xformula = ~x, link = "probit", cor.matrix = cor.matrix)
```

The simulated category-specific probabilities satisfy the desired marginal configuration

```
> t(apply(simdata$Ysim, 2, function(x) table(x)/N))
          1      2      3      4
[1,] 0.0970 0.2986 0.4068 0.1976
[2,] 0.1996 0.2046 0.1978 0.3980
[3,] 0.2068 0.3868 0.3068 0.0996
```

and a simulated correlation matrix for the latent random variables

```
> cor(simdata$rlatent)
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.5476759 0.5527712
[2,] 0.5476759 1.0000000 0.5534464
[3,] 0.5527712 0.5534464 1.0000000
```

This approach can also be employed to generate correlated binary random variables with known marginal probabilities provided that the desired correlation structure of the binary responses can be expressed in terms of a correlation matrix in the NORTA method. In this case **SimCorMultRes** is essentially implementing the simulation method of Emrich and Piedmonte (1991) without performing the first step of their algorithm.

Summary

We have presented the R package **SimCorMultRes** that simulates correlated binary or multinomial random variables conditional on a marginal model specification while expressing the dependence structure via the correlation structure of latent random variables. We outlined the underlying theory that **SimCorMultRes** is based on and illustrated the use of the package with three examples. To the best of our knowledge, **SimCorMultRes** is the first R package that targets specifically on the generation of correlated binary, nominal or ordinal responses under marginal model specification. In some instances, it could also be used to simulate correlated categorical responses even when no model specification is provided for the marginal probabilities by exploiting the relationship of association measures for discrete variables and the bivariate normal distribution. This can be achieved by following a similar approach as the one adopted in the third example herein. The results in this paper were obtained using **SimCorMultRes** version 1.4.1 and R 3.3.1.

Although the NORTA method is the default tool for simulating the latent random vectors denoted by $\mathbf{e}_i$'s, it is extremely important to emphasize that these can be provided by the user via the `rlatent` argument in the core functions. For example, generating correlated binary responses under a marginal logit model specification and with an exchangeable correlation matrix, can be accomplished by taking the difference of two independent random vectors from the multivariate Gumbel distribution each with correlation matrix the desired correlation matrix. This approach can be found in standard textbooks, such as Balakrishnan (1992). A working example, can be found in the vignette of this package.

A future direction is to increase the scope of marginal regression models for nominal and ordinal responses, e.g., by including threshold approaches that give rise to a marginal adjacent-categories logit model and allowing category-specific regression parameters in the marginal models with ordinal responses.

Acknowledgments

The author wishes to thank the associate editor and two anonymous referees for their valuable comments and suggestions which significantly improved this manuscript and led to a more flexible implementation of the related methodologies in **SimCorMultRes**.

Bibliography

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, 3rd edition, 2013. [p81, 82, 86]

A. Amatya and H. Demirtas. Multiord: An R package for generating correlated ordinal data. *Communications in Statistics-Simulation and Computation*, 44:1683–1691, 2015. [p79]

A. Amatya and H. Demirtas. *MultiOrd: Generation of multivariate ordinal variates*, 2016. URL http://CRAN.R-project.org/package=MultiOrd. R package version 2.2. [p79]

N. Balakrishnan. *Handbook of the Logistic Distribution*. CRC Press, 1992. [p89]

A. Barbiero and P. A. Ferrari. *GenOrd: Simulation of ordinal and discrete variables with given correlation matrix and marginal distributions*, 2015. URL http://CRAN.R-project.org/package=GenOrd. R package version 1.4.0. [p79]

A. Barbiero and P. A. Ferrari. An R package for the simulation of correlated discrete variables. *Communications in Statistics-Simulation and Computation*, in press. doi: 10.1080/03610918.2016.1146758. [p79]

W. Bergsma and T. Rudas. Marginal models for categorical data. *The Annals of Statistics*, 30:140–159, 2002. [p80, 83]

M. W. Birch. Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society B*, 25:220–233, 1963. [p87]

K. By and B. Qaqish. *mvtBinaryEP: Generates correlated binary data*, 2011. URL https://CRAN.R-project.org/package=mvtBinaryEP. R package version 1.0.1. [p81]

M. C. Cario and B. L. Nelson. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Northwestern University, IEMS Technical Report, 1997. [p80, 82, 83]

N. R. Chaganty and H. Joe. Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society B*, 66:851–860, 2004. [p79, 83]

N. R. Chaganty and H. Joe. Range of correlation matrices for dependent Bernoulli random variables. *Biometrika*, 93:197–206, 2006. [p83]

H. Demirtas. A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation*, 76:1017–1025, 2006. [p79]

H. Demirtas and D. Hedeker. A practical way for computing approximate lower and upper correlation bounds. *The American Statistician*, 65:104–109, 2011. [p80]

P. Diggle, K. Y. Liang, and S. L. Zeger. *Longitudinal Data Analysis*. Oxford Statistical Science Series, London, 2002. [p79]

L. J. Emrich and M. R. Piedmonte. A method for generating high-dimensional multivariate binary variables. *American Statistician*, 49:302–304, 1991. [p81, 89]

P. A. Ferrari and A. Barbiero. Simulating ordinal data. *Multivariate Behavioral Research*, 47:566–589, 2012. [p79]

G. M. Fitzmaurice and N. M. Laird. A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80:141–151, 1993. [p79]

G. F. V. Glonek and P. McCullagh. Multivariate logistic models. *Journal of the Royal Statistical Society B*, 57:533–546, 1995. [p79]

L. A. Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74:537–552, 1979. [p84, 86, 87]

S. T. Li and J. L. Hammond. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Transactions on Systems, Man and Cybernetics*, 5:557–561, 1975. [p82]

S. R. Lipsitz, N. M. Laird, and D. P. Harrington. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika*, 78:153–160, 1991. [p79]

G. Masarotto and C. Varin. Gaussian copula marginal regression. *Electronic Journal of Statistics*, 6: 1517–1549, 2012. [p79]

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society B*, 42:109–142, 1980. [p81]

D. McFadden. Conditional logit analysis of qualitative choice behaviour. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, 1974. [p82]

J. Shults and N. R. Chaganty. Analysis of serially correlated data using quasi-least squares. *Biometrics*, 54:1622–1630, 1998. [p79]

B. C. Sutradhar. An overview on regression models for discrete longitudinal responses. *Statistical Science*, 18:377–393, 2003. [p79]

B. C. Sutradhar and K. Das. On the accuracy of efficiency of estimating equation approach. *Biometrika*, 86:459–465, 1999. [p79]

A. Touloumis. R package multgee: A generalized estimating equations solver for multinomial responses. *Journal of Statistical Software*, 64:1–14, 2015. [p85]

A. Touloumis. *SimCorMultRes: Simulates correlated multinomial responses*, 2016. URL `http://CRAN.R-project.org/package=SimCorMultRes`. R package version 1.4.1. [p79]

A. Touloumis, A. Agresti, and M. Kateri. Generalized estimating equations for multinomial responses using a local odds ratios parameterization. *Biometrics*, 69:633–640, 2013. [p79, 83, 85]

G. Tutz. Sequential models in categorical regression. *Computational Statistics & Data Analysis*, 11: 275–295, 1991. [p81]

Anestis Touloumis
Computing, Engineering and Mathematics
University of Brighton
Brighton, United Kingdom
`A.Touloumis@brighton.ac.uk`