Simulating Correlated Binary and Multinomial Responses with **SimCorMultRes**

Anestis Touloumis

2017-07-11

Contents

1 Introduction

The **R** package **SimCorMultRes** is suitable for simulation of correlated binary responses (exactly two response categories) and of correlated nominal or ordinal multinomial responses (three or more response categories) conditional on a regression model specification for the marginal probabilities of the response categories. This vignette briefly describes the simulation methods proposed by Touloumis (2016) and illustrates the use of the core functions of **SimCorMultRes**. A more detailed description of **SimCorMultRes** can be found in Touloumis (2016).

2 Areas of Applications

This package was created to facilitate the task of carrying out simulation studies and evaluating the performance of statistical method(s) for estimating the regression parameters in a marginal model with clustered binary and multinomial responses. Examples of such statistical methods include maximum likelihood methods, copula approaches, quasi-least squares approaches, generalized quasi-likelihood methods and generalized estimating equations (GEE) approaches among others (see references in Touloumis 2016).

In addition, **SimCorMultRes** can be used to generate correlated binary and multinomial random variables conditional on a desired dependence structure and known marginal probabilities even if these are not determined by a regression model (see third example in Touloumis 2016) or to explore approximations of association measures for discrete variables that arise as realizations of an underlying continuum (see second example in Touloumis 2016).

3 Simulation Methods

Let Y_{it} be the binary or multinomial response for subject i ($i = 1, \ldots, N$) at measurement occasion t ($t = 1, \ldots, T$), and let $\mathbf{x}_{it}$ be the associated covariates vector. We assume that $Y_{it} \in \{0, 1\}$ for binary responses and $Y_{it} \in \{1, 2, \ldots, J \geq 3\}$ for multinomial responses.

3.1 Correlated nominal responses

The function `rmult.bcl` simulates nominal responses under the marginal baseline-category logit model

$$\log\left[\frac{\Pr(Y_{it} = j|\mathbf{x}_{it})}{\Pr(Y_{it} = J|\mathbf{x}_{it})}\right] = (\beta_{tj0} - \beta_{tJ0}) + (\boldsymbol{\beta}_{tj} - \boldsymbol{\beta}_{tJ})'\mathbf{x}_{it} = \beta^*_{tj0} + \boldsymbol{\beta}^{*\prime}_{tj}\mathbf{x}_{it}, \tag{1}$$

where β_{tj0} is the j-th category-specific intercept at measurement occasion t and $\boldsymbol{\beta}_{tj}$ is the j-th category-specific parameter vector associated with the covariates at measurement occasion t. The popular identifiability constraints $\beta_{tJ0} = 0$ and $\boldsymbol{\beta}_{tJ} = \mathbf{0}$ for all t, imply that $\beta^*_{tj0} = \beta_{tj0}$ and $\boldsymbol{\beta}^*_{tj} = \boldsymbol{\beta}_{tj}$ for all $t = 1, \ldots, T$ and $j = 1, \ldots, J-1$. The threshold

$$Y_{it} = j \Leftrightarrow U^{NO}_{itj} = \max\{U^{NO}_{it1}, \ldots, U^{NO}_{itJ}\}$$

generates clustered nominal responses that satisfy the marginal baseline-category logit model (1), where

$$U^{NO}_{itj} = \beta_{tj0} + \boldsymbol{\beta}'_{tj}\mathbf{x}_{it} + e^{NO}_{itj},$$

and where the random variables $\{e^{NO}_{itj} : i = 1, \ldots, N, t = 1, \ldots, T \text{ and } j = 1, \ldots, J\}$ satisfy the following conditions:

1. e^{NO}_{itj} follows the standard extreme value distribution for all i, t and j (mean $= \gamma \approx 0.5772$, where γ is Euler's constant, and variance $= \pi^2/6$).
2. $e^{NO}_{i_1t_1j_1}$ and $e^{NO}_{i_2t_2j_2}$ are independent random variables provided that $i_1 \neq i_2$.
3. $e^{NO}_{itj_1}$ and $e^{NO}_{itj_2}$ are independent random variables provided that $j_1 \neq j_2$.

For each subject i, the association structure among the clustered nominal responses $\{Y_{it} : t = 1, \ldots, T\}$ depends on the joint distribution and correlation matrix of $\{e^{NO}_{itj} : t = 1, \ldots, T \text{ and } j = 1, \ldots, J\}$. If the random variables $\{e^{N}_{itj} : t = 1, \ldots, T \text{ and } j = 1, \ldots, J\}$ are independent then so are $\{Y_{it} : t = 1, \ldots, T\}$.

Example 3.1 (Simulation of clustered nominal responses using the NORTA method)**.** Suppose the aim is to simulate nominal responses from the marginal baseline-category logit model

$$\log\left[\frac{\Pr(Y_{it} = j|\mathbf{x}_{it})}{\Pr(Y_{it} = 4|\mathbf{x}_{it})}\right] = \beta_{j0} + \beta_{j1}x_{i1} + \beta_{j2}x_{it2}$$

where $N = 500$, $T = 3$, $(\beta_{10}, \beta_{11}, \beta_{12}, \beta_{20}, \beta_{21}, \beta_{22}, \beta_{30}, \beta_{31}, \beta_{32}) = (1, 3, 2, 1.25, 3.25, 1.75, 0.75, 2.75, 2.25)$ and $\mathbf{x}_{it} = (x_{i1}, x_{it2})'$ for all i and t, with $x_{i1} \overset{iid}{\sim} N(0, 1)$ and $x_{it2} \overset{iid}{\sim} N(0, 1)$. For the dependence structure, suppose that the correlation matrix $\mathbf{R}$ in the NORTA method has elements

$$\mathbf{R}_{t_1j_1,t_2j_2} = \begin{cases} 1 & \text{if } t_1 = t_2 \text{ and } j_1 = j_2 \\ 0.95 & \text{if } t_1 \neq t_2 \text{ and } j_1 = j_2 \\ 0 & \text{otherwise} \end{cases}$$

for all $i = 1, \ldots, N$.

```r
# parameter vector
betas <- c(1, 3, 2, 1.25, 3.25, 1.75, 0.75, 2.75, 2.25, 0, 0, 0)
# sample size
N <- 500
# number of nominal response categories
ncategories <- 4
# cluster size
clsize <- 3
set.seed(1)
# time-stationary covariate x_{i1}
x1 <- rep(rnorm(N), each = clsize)
# time-varying covariate x_{it2}
x2 <- rnorm(N * clsize)
# create covariates dataframe
xdata <- data.frame(x1, x2)
set.seed(321)
library(SimCorMultRes)
# correlation matrix for the NORTA method
cor.matrix <- kronecker(toeplitz(c(1, rep(0.95, clsize - 1))), diag(ncategories))
# simulation of clustered nominal responses
CorNorRes <- rmult.bcl(clsize = clsize, ncategories = ncategories, betas = betas,
    xformula = ~x1 + x2, xdata = xdata, cor.matrix = cor.matrix)
suppressPackageStartupMessages(library("multgee"))
# fitting a GEE model
fit <- nomLORgee(y ~ x1 + x2, data = CorNorRes$simdata, id = id, repeated = time,
    LORstr = "time.exch")
# checking regression coefficients
round(coef(fit), 2)
#> beta10   x1:1   x2:1 beta20   x1:2   x2:2 beta30   x1:3   x2:3
#>   1.07   3.18   1.99   1.35   3.40   1.70   0.89   3.06   2.22
```

3.2 Correlated ordinal responses

Simulation of clustered ordinal responses is feasible under either a marginal cumulative link model or a marginal continuation-ratio model.

3.2.1 Marginal cumulative link model

The function `rmult.clm` simulates ordinal responses under the marginal cumulative link model

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

where F is a cumulative distribution function (cdf), β_{tj0} is the j-th category-specific intercept at measurement occasion t and $\boldsymbol{\beta}_t$ is the regression parameter vector associated with the covariates at measurement occasion t. The category-specific intercepts at each measurement occasion t are assumed to be monotone increasing, that is

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

for all t. Using the threshold

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

clustered ordinal responses that satisfy the marginal cumulative link model (2) are generated, where

$$U_{it}^{O1} = -\boldsymbol{\beta}_t' \mathbf{x}_{it} + e_{it}^{O1},$$

and where $\{e_{it}^{O1} : i = 1, \dots, N \text{ and } t = 1, \dots, T\}$ are random variables such that:

1. $e_{it}^{O1} \sim F$ for all i and t.
2. $e_{i_1t_1}^{O1}$ and $e_{i_2t_2}^{O1}$ are independent random variables provided that $i_1 \neq i_2$.

For each subject i, the association structure among the clustered ordinal responses $\{Y_{it} : t = 1, \dots, T\}$ depends on the pairwise bivariate distributions and correlation matrix of $\{e_{it}^{O1} : t = 1, \dots, T\}$. If the random variables $\{e_{it}^{O1} : t = 1, \dots, T\}$ are independent then so are $\{Y_{it} : t = 1, \dots, T\}$.

Example 3.2 (Simulation of clustered ordinal responses conditional on a marginal cumulative probit model with time-varying regression parameters)**.** Suppose the goal is to simulate correlated ordinal responses from the marginal cumulative probit model

$$\Pr(Y_{it} \leq j | \mathbf{x}_{it}) = \Phi(\beta_{j0} + \beta_{t1} x_i)$$

where Φ denotes the cdf of the standard normal distribution (mean $= 0$ and variance $= 1$), $N = 500$, $T = 4$, $(\beta_{10}, \beta_{20}, \beta_{30}, \beta_{40}) = (-1.5, -0.5, 0.5, 1.5)$, $(\beta_{11}, \beta_{21}, \beta_{31}, \beta_{41}) = (1, 2, 3, 4)$ and $\mathbf{x}_{it} = x_i \overset{iid}{\sim} N(0, 1)$ for all i and t. For the dependence structure, assume that $\mathbf{e}_i^{O1} = (e_{i1}^{O1}, e_{i2}^{O1}, e_{i3}^{O1}, e_{i4}^{O1})'$ are iid random vectors from a tetra-variate normal distribution with mean vector the zero vector and covariance matrix the correlation matrix

$$\begin{pmatrix} 1.00 & 0.85 & 0.50 & 0.15 \\ 0.85 & 1.00 & 0.85 & 0.50 \\ 0.50 & 0.15 & 1.00 & 0.85 \\ 0.15 & 0.85 & 0.50 & 1.00 \end{pmatrix}.$$

```r
set.seed(12345)
# sample size
N <- 500
# cluster size
clsize <- 4
# category-specific intercepts
intercepts <- c(-1.5, -0.5, 0.5, 1.5)
# time-varying regression parameters associated with covariates
betas <- matrix(c(1, 2, 3, 4), 4, 1)
# time-stationary covariate
x <- rep(rnorm(N), each = clsize)
# correlation matrix for the NORTA method
cor.matrix <- toeplitz(c(1, 0.85, 0.5, 0.15))
# simulation of ordinal responses
CorOrdRes <- rmult.clm(clsize = clsize, intercepts = intercepts, betas = betas,
    xformula = ~x, cor.matrix = cor.matrix, link = "probit")
# first eight rows of the simulated dataframe
head(CorOrdRes$simdata, n = 8)
#>   y         x id time
#> 1 1 0.5855288  1    1
#> 2 2 0.5855288  1    2
#> 3 2 0.5855288  1    3
#> 4 1 0.5855288  1    4
#> 5 1 0.7094660  2    1
#> 6 1 0.7094660  2    2
#> 7 1 0.7094660  2    3
#> 8 1 0.7094660  2    4
```

3.2.2 Marginal continuation-ratio model

The function `rmult.crm` simulates clustered ordinal responses under the marginal continuation-ratio model

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

where β_{tj0} is the j-th category-specific intercept at measurement occasion t, $\boldsymbol{\beta}_t$ is the regression parameter vector associated with the covariates at measurement occasion t and F is a cdf. This is accomplished by utilizing the threshold

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

where

$$U_{itj}^{O2} = -\boldsymbol{\beta}_t' \mathbf{x}_{it} + e_{itj}^{O2},$$

and where $\{e_{itj}^{O2} : i = 1, \ldots, N \ , t = 1, \ldots, T \text{ and } j = 1, \ldots, J-1\}$ satisfy the following three conditions:

1. $e_{itj}^{O2} \sim F$ for all i, t and j.
2. $e_{i_1 t_1 j_1}^{O2}$ and $e_{i_2 t_2 j_2}^{O2}$ are independent random variables provided that $i_1 \neq i_2$.
3. $e_{itj_1}^{O2}$ and $e_{itj_2}^{O2}$ are independent random variables provided that $j_1 \neq j_2$.

For each subject i, the association structure among the clustered ordinal responses $\{Y_{it} : t = 1, \ldots, T\}$ depends on the joint distribution and correlation matrix of $\{e_{itj}^{O2} : j = 1, \ldots, J \text{ and } t = 1, \ldots, T\}$. If the random variables $\{e_{itj}^{O2} : j = 1, \ldots, J \text{ and } t = 1, \ldots, T\}$ are independent then so are $\{Y_{it} : t = 1, \ldots, T\}$.

Example 3.3 (Simulation of clustered ordinal responses conditional on a marginal continuation-ratio probit model)**.** Suppose simulation of clustered ordinal responses under the marginal continuation-ratio probit model

$$\Pr(Y_{it} = j | Y_{it} \geq j, \mathbf{x}_{it}) = \Phi(\beta_{j0} + \beta x_{it})$$

with $N = 500$, $T = 4$, $(\beta_{10}, \beta_{20}, \beta_{30}, \beta_{40}, \beta) = (-1.5, -0.5, 0.5, 1.5, 1)$ and $\mathbf{x}_{it} = x_{it} \overset{iid}{\sim} N(0,1)$ for all i and t is desired. For the dependence structure, assume that $\left\{\mathbf{e}_i^{O2} = \left(e_{i11}^{O2}, \ldots, e_{i44}^{O1}\right)' : i = 1, \ldots, N\right\}$ are iid random vectors from a multivariate normal distribution with mean vector the zero vector and covariance matrix the 16×16 correlation matrix with elements

$$\text{corr}(e_{it_1 j_1}^{O2}, e_{it_2 j_2}^{O2}) = \begin{cases} 1 & \text{for } j_1 = j_2 \text{ and } t_1 = t_2 \\ 0.24 & \text{for } t_1 \neq t_2 \\ 0 & \text{otherwise.} \end{cases}$$

```r
set.seed(1)
# sample size
N <- 500
# cluster size
clsize <- 4
# category-specific intercepts
intercepts <- c(-1.5, -0.5, 0.5, 1.5)
# regression parameters associated with covariates
betas <- 1
# time-varying covariate
x <- rnorm(N * clsize)
# number of ordinal response categories
ncategories <- 5
# correlation matrix for the NORTA method
cor.matrix <- diag(1, (ncategories - 1) * clsize) + kronecker(toeplitz(c(0,
```

```r
    rep(0.24, ncategories - 2))), matrix(1, clsize, clsize))
# simulation of ordinal responses
CorOrdRes <- rmult.crm(clsize = clsize, intercepts = intercepts, betas = betas,
    xformula = ~x, cor.matrix = cor.matrix, link = "probit")
# first six clusters with ordinal responses
head(CorOrdRes$Ysim)
#>     t=1 t=2 t=3 t=4
#> i=1   2   1   3   1
#> i=2   1   4   1   1
#> i=3   2   2   1   3
#> i=4   3   5   2   2
#> i=5   2   1   1   1
#> i=6   3   3   4   5
```

3.3 Correlated binary responses

The function `rbin` simulates binary responses under the marginal model specification

$$\Pr(Y_{it} = 1|\mathbf{x}_{it}) = F(\beta_{t0} + \boldsymbol{\beta}_t'\mathbf{x}_{it}) \tag{4}$$

where β_{t0} is the intercept at measurement occasion t, $\boldsymbol{\beta}_t$ is the regression parameter vector associated with the covariates at measurement occasion t and F is a cdf. The threshold

$$Y_{it} = 1 \Leftrightarrow U_{it}^B \leq \beta_{t0} + 2\boldsymbol{\beta}_t'\mathbf{x}_{it},$$

generates clustered binary responses that satisfy the marginal model (4), where

$$U_{it}^B = \boldsymbol{\beta}_t'\mathbf{x}_{it} + e_{it}^B, \tag{5}$$

and where $\{e_{it}^B : i = 1, \dots, N \text{ and } t = 1, \dots, T\}$ are random variables such that:

1. $e_{it}^B \sim F$ for all i and t.
2. $e_{i_1t_1}^B$ and $e_{i_2t_2}^B$ are independent random variables provided that $i_1 \neq i_2$.

For each subject i, the association structure among the clustered binary responses $\{Y_{it} : t = 1, \dots, T\}$ depends on the pairwise bivariate distributions and correlation matrix of $\{e_{it}^B : t = 1, \dots, T\}$. If the random variables $\{e_{it}^B : t = 1, \dots, T\}$ are independent then so are $\{Y_{it} : t = 1, \dots, T\}$.

Example 3.4 (Simulation of clustered binary responses conditional on a marginal probit model using NORTA method)**.** Suppose the goal is to simulate clustered binary responses from the marginal probit model

$$\Pr(Y_{it} = 1|\mathbf{x}_{it}) = \Phi(0.2x_i)$$

where $N = 100$, $T = 4$ and $\mathbf{x}_{it} = x_i \overset{iid}{\sim} N(0, 1)$ for all i and t. For the association structure, assume that the random variables $\mathbf{e}_i^B = (e_{i1}^B, e_{i2}^B, e_{i3}^B, e_{i4}^B)'$ in (5) are iid random vectors from the tetra-variate normal distribution with mean vector the zero vector and covariance matrix the correlation matrix $\mathbf{R}$ given by

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.90 & 0.90 & 0.90 \\ 0.90 & 1.00 & 0.90 & 0.90 \\ 0.90 & 0.90 & 1.00 & 0.90 \\ 0.90 & 0.90 & 0.90 & 1.00 \end{pmatrix}. \tag{6}$$

This association configuration defines an exchangeable correlation matrix for the clustered binary responses, i.e. $\text{corr}(Y_{it_1}, Y_{it_2}) = \rho_i$ for all i and t. The strength of the correlation (ρ_i) is decreasing as the absolute value of the time-stationary covariate x_i increases. For example, $\rho_i = 0.7128$ when $x_i = 0$ and $\rho_i = 0.7$ when $x_i = 3$ or $x_i = -3$. Therefore, a strong exchangeable correlation pattern for each subject that does not differ much across subjects is implied with this configuration.

```
set.seed(123)
# sample size
N <- 100
# cluster size
clsize <- 4
# intercept
intercepts <- 0
# regression parameter associated with the covariate
betas <- 0.2
# correlation matrix for the NORTA method
cor.matrix <- toeplitz(c(1, 0.9, 0.9, 0.9))
# time-stationary covariate
x <- rep(rnorm(N), each = clsize)
# simulation of clustered binary responses
CorBinRes <- rbin(clsize = clsize, intercepts = intercepts, betas = betas, xformula = ~x,
    cor.matrix = cor.matrix, link = "probit")
library(gee)
# fitting a GEE model
binGEEmod <- gee(y ~ x, family = binomial("probit"), id = id, data = CorBinRes$simdata)
#> Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
#> running glm to get initial regression estimate
#> (Intercept)           x
#>   0.1315121   0.2826005
# checking the estimated coefficients
summary(binGEEmod)$coefficients
#>              Estimate Naive S.E.   Naive z Robust S.E. Robust z
#> (Intercept) 0.1315121 0.06399465 2.055048   0.1106696 1.188331
#> x           0.2826006 0.07191931 3.929412   0.1270285 2.224703
```

Example 3.5 (Simulation of clustered binary responses under a conditional marginal logit model without utilizing the NORTA method)**.** Consider now simulation of correlated binary responses from the marginal logit model

$$\Pr(Y_{it} = 1|\mathbf{x}_{it}) = F(0.2x_i)$$

where F is the cdf of the standard logistic distribution (mean $= 0$ and variance $= \pi^2/3$), $N = 100$, $T = 4$ and $\mathbf{x}_{it} = x_i \overset{iid}{\sim} N(0,1)$ for all i and t. This is similar to the marginal model configuration in Example 3.4 except from the link function. For the dependence structure, assume that the correlation matrix of $\mathbf{e}_i^B = (e_{i1}^B, e_{i2}^B, e_{i3}^B, e_{i4}^B)'$ in (5) is equal to the correlation matrix $\mathbf{R}$ defined in (6). To simulate $\mathbf{e}_i^B$ without utilizing the NORTA method, one can employ the tetra-variate extreme value distribution (Gumbel 1958). In particular, this is accomplished by setting $\mathbf{e}_i^B = \mathbf{U}_i - \mathbf{V}_i$ for all i, where $\mathbf{U}_i$ and $\mathbf{V}_i$ are independent random vectors from the tetra-variate extreme value distribution with dependence parameter equal to 0.9, that is

$$\Pr\left(U_{i1} \leq u_{i1}, U_{i2} \leq u_{i2}, U_{i3} \leq u_{i3}, U_{i4} \leq u_{i4}\right) = \exp\left\{-\left[\sum_{t=1}^{4} \exp\left(-\frac{u_{it}}{0.9}\right)\right]^{0.9}\right\}$$

and

$$\Pr\left(V_{i1} \leq v_{i1}, V_{i2} \leq v_{i2}, V_{i3} \leq v_{i3}, V_{i4} \leq v_{i4}\right) = \exp\left\{-\left[\sum_{t=1}^{4} \exp\left(-\frac{v_{it}}{0.9}\right)\right]^{0.9}\right\}.$$

It follows that $e_{it}^B \sim F$ for all i and t and $\text{corr}(\mathbf{e}_i^B) = \mathbf{R}$ for all i.

```
set.seed(8)
# simulation of epsilon variables
library(evd)
#>
#> Attaching package: 'evd'
#> The following objects are masked from 'package:VGAM':
#>
#>     dfrechet, dgev, dgpd, dgumbel, pfrechet, pgev, pgpd, pgumbel,
#>     qfrechet, qgev, qgpd, qgumbel, rfrechet, rgev, rgpd, rgumbel,
#>     venice
rlatent1 <- rmvevd(N, dep = sqrt(1 - 0.9), model = "log", d = clsize)
rlatent2 <- rmvevd(N, dep = sqrt(1 - 0.9), model = "log", d = clsize)
rlatent <- rlatent1 - rlatent2
# simulation of clustered binary responses
CorBinRes <- rbin(clsize = clsize, intercepts = intercepts, betas = betas, xformula = ~x,
    rlatent = rlatent)
# fitting a GEE model
binGEEmod <- gee(y ~ x, family = binomial("logit"), id = id, data = CorBinRes$simdata)
#> Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
#> running glm to get initial regression estimate
#> (Intercept)           x
#>  0.04146261  0.09562709
# checking the estimated coefficients
summary(binGEEmod)$coefficients
#>               Estimate Naive S.E.   Naive z Robust S.E.  Robust z
#> (Intercept) 0.04146261  0.1008516 0.4111249   0.1790511 0.2315686
#> x           0.09562709  0.1107159 0.8637160   0.1949327 0.4905647
```

3.4 No marginal model specification

To achieve simulation of clustered binary, ordinal and nominal responses under no marginal model specification, perform the following intercepts:

1. Based on the marginal probabilities calculate the intercept of a marginal probit model for binary responses (see Example 3.6) or the category-specific intercepts of a cumulative probit model (see third example in Touloumis 2016) or of a baseline-category logit model for multinomial responses (see Example 3.7).
2. Create a pseudo-covariate say `x` of length equal to the number of cluster size (`clsize`) times the desired number of clusters of simulated responses (say `R`), that is `x = clsize * R`. This step is required in order to identify the desired number of clustered responses.
3. Set `betas = 0` in the core functions `rbin` (see Example 3.6) or `rmult.clm`, or set 0 all values of the `beta` argument that correspond to the category-specific parameters in the core function `rmult.bcl` (see Example 3.7).
4. set `xformula = ~ x`.
5. Run the core function to obtain realizations of the simulated clustered responses.

Example 3.6 (Simulation of clustered binary responses without covariates)**.** Suppose the goal is to simulate 5000 clustered binary responses with $\Pr(Y_t = 1) = 0.8$ for all $t = 1, \ldots, 4$. For simplicity, assume that the clustered binary responses are independent.

```r
set.seed(123)
# sample size
N <- 5000
# cluster size
clsize <- 4
# intercept
intercepts <- qnorm(0.8)
# pseudo-covariate
x <- rep(0, each = clsize * N)
# regression parameter associated with the covariate
betas <- 0
# correlation matrix for the NORTA method
cor.matrix <- diag(clsize)
# simulation of clustered binary responses
CorBinRes <- rbin(clsize = clsize, intercepts = intercepts, betas = betas, xformula = ~x,
    cor.matrix = cor.matrix, link = "probit")
library(gee)
# simulated marginal probabilities
colMeans(CorBinRes$Ysim)
#>    t=1    t=2    t=3    t=4
#> 0.8024 0.7972 0.7948 0.8088
```

Example 3.7 (Simulation of clustered nominal responses without covariates)**.** Suppose the aim is to simulate $N = 5000$ clustered nominal responses with $\Pr(Y_t = 1) = 0.1$, $\Pr(Y_t = 2) = 0.2$, $\Pr(Y_t = 3) = 0.3$ and $\Pr(Y_t = 4) = 0.4$, for all i and $t = 1, \ldots, 3$. For the sake of simplicity, we assume that the clustered responses are independent.

```r
# sample size
N <- 5000
# cluster size
clsize <- 3
# pseudo-covariate
x <- rep(0, each = clsize * N)
# parameter vector
betas <- c(log(0.1/0.4), 0, log(0.2/0.4), 0, log(0.3/0.4), 0, 0, 0)
# number of nominal response categories
ncategories <- 4
set.seed(1)
# correlation matrix for the NORTA method
cor.matrix <- kronecker(toeplitz(c(1, rep(0, clsize - 1))), diag(ncategories))
# simulation of clustered nominal responses
CorNorRes <- rmult.bcl(clsize = clsize, ncategories = ncategories, betas = betas,
    xformula = ~x, cor.matrix = cor.matrix)
# simulated marginal probabilities
apply(CorNorRes$Ysim, 2, table)/N
#>      t=1    t=2    t=3
#> 1 0.1000 0.0996 0.1036
#> 2 0.2034 0.2000 0.2000
#> 3 0.2874 0.3130 0.2894
#> 4 0.4092 0.3874 0.4070
```

4 How to Cite

```
citation("SimCorMultRes")

To cite R package SimCorMultRes in publications, please use:

  Touloumis, A. (2016). Simulating Correlated Binary and
  Multinomial Responses under Marginal Model Specification: The
  SimCorMultRes Package. The R Journal 8:2, 79-91.

A BibTeX entry for LaTeX users is

  @Article{,
    title = {Simulating Correlated Binary and Multinomial Responses under
         Marginal Model Specification: The SimCorMultRes Package},
    author = {Anestis Touloumis},
    year = {2016},
    journal = {The R Journal},
    volume = {8},
    number = {2},
    pages = {79-91},
    url = {https://journal.r-project.org/archive/2016/RJ-2016-034/index.html},
  }
```

References

Gumbel, E. J. 1958. *Statistics of Extremes*. Columbia University Press, New York.

Touloumis, A. 2016. "Simulating Correlated Binary and Multinomial Responses under Marginal Model Specification: The SimCorMultRes Package." *The R Journal* 8 (2): 79–91.