# Vignette for the **R** package **SimCorMultRes**

Anestis Touloumis

## 1 Introduction

The **R** package **SimCorMultRes** is suitable for simulation of correlated ordinal or nominal multinomial responses conditional on a given marginal model specification for the univariate probabilities. The key idea is to modify existing threshold approaches that give rise to models for independent ordinal or nominal multinomial responses so as to introduce dependencies between the multinomial responses. This vignette describes the latent threshold approaches employed in **SimCorMultRes** and it illustrates via simple examples the use of the core functions.

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

## 2 Correlated Nominal Multinomial Responses

The function `rmult.bcl()` simulates correlated nominal multinomial responses under the marginal baseline category logit model specification

$$\log\left(\frac{\Pr(Y_{it} = j|\mathbf{x}_{it})}{\Pr(Y_{it} = I|\mathbf{x}_{it})}\right) = (\beta_{0j} - \beta_{0J}) + (\boldsymbol{\beta}_j - \boldsymbol{\beta}_J)'\mathbf{x}_{it} = \beta^*_{0j} + \boldsymbol{\beta}^{*\prime}_j\mathbf{x}_{it}, \tag{1}$$

where $\beta_{0j}$ is the $j$-th response category specific intercept and $\boldsymbol{\beta}_j$ is the $j$-th response category specific parameter vector. The popular identifiability constraints $\beta_{0J} = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$ imply that $\beta^*_{0j} = \beta_{0j}$ and $\boldsymbol{\beta}^*_j = \boldsymbol{\beta}_j$ for all $j = 1, \ldots, J-1$.
Define

$$U_{itj} = \mu_{itj} + e_{itj},$$

where $\mu_{itj} = \beta_{0j} + \boldsymbol{\beta}'_j\mathbf{x}_{it}$ is the $j$-th response category specific linear predictor for subject $i$ at the $t$-th measurement occasion, and where the random variables $\{e_{itj}\}$ satisfy the following conditions:

1. Marginally, $e_{itj}$ follows a standard extreme value distribution for all $i$, $t$ and $j$.
2. The random variables associated with different subjects are independent, i.e., conditional on $t_1$, $t_2$, $j_1$ and $j_2$ the random variables $e_{i_1t_1j_1}$ and $e_{i_2t_2j_2}$ are independent for all $i_1 \neq i_2$.
3. The category specific random variables for each subject at a given measurement occasion are independent, i.e., conditional on $i$ and $t$ the random variables $e_{itj_1}$ and $e_{itj_2}$ are independent for all $j_1 \neq j_2$.

It can be shown that using the threshold

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

correlated nominal multinomial responses that satisfy the marginal baseline category logit model specification in (1) are generated. This approach extends the principle of maximum random utility (McFadden, 1973) to correlated multinomial responses.

The function `rmult.bcl()` requires the user to provide the common cluster size $T$ (`clsize`), the number of nominal response categories $J$ (`ncategories`), the linear predictor in a matrix form (`lin.pred`) and the correlation matrix used in the NORTA method (`cor.matrix`) for the dependence structure of $\{\mathbf{e}_{itj}\}$. The `lin.pred` argument should be an $N \times (TJ)$ matrix such that the $i$-th row corresponds to subject $i$ and has

elements $(\mu_{i11}, \ldots, \mu_{i1J}, \mu_{i21}, \ldots, \mu_{i2J}, \cdots, \mu_{iT1}, \ldots, \mu_{iTJ})$.

For example, suppose that we want to simulate nominal multinomial responses under a marginal baseline category logit model with $N = 500$, $J = 4$, $T = 3$, $(\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}) = (1, 3, 5, 0)$, $(\beta_1, \beta_2, \beta_3, \beta_4) = (2, 4, 6, 0)$ and a time-stationary covariate for each subject drawn from a standard normal distribution. For the sake of simplicity, suppose that $\{e_{itj}\}$ are independent. The **R** code to simulate clustered nominal multinomial responses under this configuration is

```
> library("SimCorMultRes")
> set.seed(1)
> N <- 500
> ncategories <- 4
> clustersize <- 3
> Xmat <- matrix(rnorm(N),N,ncategories)
> betas <- c(1,2,3,4,5,6)
> linpred <- matrix(c(betas[c(2,4,6)],0),N,4,byrow=TRUE)*Xmat+
+            matrix(c(betas[c(1,3,5)],0),N,4,byrow=TRUE)
> linpred <- matrix(linpred,N,ncategories*clustersize)
> cormat <- diag(1,12)
> Y <- rmult.bcl(clsize=3,ncategories=4,lin.pred=linpred,cor.matrix=cormat)
```

The simulated clustered nominal multinomial responses for the first six subjects are

```
> head(Y$Ysim)
```

```
     [,1] [,2] [,3]
[1,]    3    3    1
[2,]    2    3    3
[3,]    4    2    3
[4,]    3    3    3
[5,]    3    3    3
[6,]    1    1    3
```

# 3 Correlated Ordinal Multinomial Responses

Generation of correlated ordinal multinomial responses is feasible under either a marginal cumulative link model or a marginal continuation ratio model specification.

## 3.1 Marginal cumulative link model

The function `rmult.clm()` simulates correlated ordinal multinomial responses under the marginal cumulative link model specification

$$\Pr(Y_{it} \leq j | \mathbf{x}_{it}) = F(\beta_{0j} + \boldsymbol{\beta}' \mathbf{x}_{it}) \tag{2}$$

where $\beta_{0j}$ is the $j$-th response category specific intercept, $\boldsymbol{\beta}$ is the parameter vector associated with the covariates and $F$ is a cumulative distribution function. Note that the response category specific intercepts are assumed to be monotone increasing

$$-\infty = \beta_{00} < \beta_{01} < \beta_{02} < \cdots < \beta_{0(J-1)} < \beta_{0J} = \infty.$$

Define

$$U_{it} = \mu_{it} + e_{it},$$

where $\mu_{it} = \boldsymbol{\beta}' \mathbf{x}_{it}$ is the linear predictor minus the corresponding response category specific intercept for subject $i$ at the $t$-th measurement occasion, and where the random variables $\{e_{it}\}$ satisfy the following conditions:

1. Marginally, $e_{it}$ follows the distribution specified by $F$ for all $i$ and $t$.

2. The random variables associated with different subjects are independent, i.e., conditional on $t_1$ and $t_2$ the random variables $e_{i_1 t_1}$ and $e_{i_2 t_2}$ are independent for all $i_1 \neq i_2$.

It can be shown that using the threshold

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

correlated ordinal multinomial responses under the marginal cumulative link model specification in (2) are generated. This approach is based on the threshold approach mentioned in McCullagh (1980).

The function `rmult.clm()` requires the user to provide the common cluster size $T$ (`clsize`), the linear predictor excluding the response category intercepts in a matrix form (`lin.pred`), the correlation structure of the latent random variables $\{e_{it}\}$ (`corr`), the response category specific intercepts $\beta_{0j}$'s (`cuts`) and the cumulative distribution function $F$ (`link`). The `lin.pred` argument should be an $N \times T$ matrix such that the $i$-th row corresponds to subject $i$ and has elements $(\mu_{i1}, \mu_{i2}, \cdots, \mu_{iT})$.

For example, suppose that we want to simulate correlated ordinal multinomial responses from a marginal cumulative probit model with $N = 500$, $J = 5$, $T = 4$, $(\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}) = (-1.5, -0.5, 0.5, 1.5)$, $\boldsymbol{\beta} = \beta = 1$ and a single time-stationary covariate for each subject drawn from a standard normal distribution. Further, suppose that the latent dependence structure is given by the correlation matrix

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

The following **R** code generates the clustered ordinal multinomial responses under this configuration

```
> set.seed(1)
> N <- 500
> clustersize <- 4
> intercepts <- c(-Inf,-1.5,-0.5,0.5,1.5,Inf)
> cormat <- toeplitz(c(1,0.85,0.5,0.15))
> x <- rnorm(N)
> linpred <- matrix(rep(x,clustersize),N,clustersize,byrow=TRUE)
> Y <- rmult.clm(clsize=clustersize,lin.pred=linpred,corr=cormat,
+                cuts=intercepts,link="probit")
```

The simulated clustered ordinal multinomial responses for the first six subjects are

```
> head(Y$Ysim)
```

```
     [,1] [,2] [,3] [,4]
[1,]    4    3    5    3
[2,]    2    4    3    2
[3,]    1    2    1    3
[4,]    4    5    2    3
[5,]    4    3    3    3
[6,]    4    3    2    4
```

## 3.2 Marginal continuation ratio model

The function `rmult.crm()` simulates correlated ordinal multinomial responses under the marginal continuation ratio model specification

$$\Pr(Y_{it} = j | Y_{it} \geq j, \mathbf{x}_{it}) = F(\beta_{0j} + \boldsymbol{\beta}^{'} \mathbf{x}_{it}) \tag{3}$$

where $\beta_{0j}$ is the $j$-th response category specific intercept, $\boldsymbol{\beta}$ is the parameter vector associated with the covariates and $F$ is a cumulative distribution function. Note that the response category specific intercepts are assumed to be monotone increasing

$$-\infty = \beta_{00} < \beta_{01} < \beta_{02} < \ldots < \beta_{0(J-1)} < \beta_{0J} = \infty.$$

Define

$$U_{itj} = \mu_{it} + e_{itj},$$

where $\mu_{it} = \boldsymbol{\beta}'\mathbf{x}_{it}$ is the linear predictor minus the corresponding response category specific intercept for subject $i$ at the $t$-th measurement occasion, and where the random variables $\{e_{itj}\}$ satisfy the following conditions:

1. Marginally, $e_{itj}$ follows the distribution specified by $F$ for all $i$, $t$ and $j$.
2. The random variables associated with different subjects are independent, i.e., conditional on $t_1$, $t_2$, $j_1$ and $j_2$ the random variables $e_{i_1t_1j_1}$ and $e_{i_2t_2j_2}$ are independent for all $i_1 \neq i_2$.
3. The category specific random variables for each subject at a given measurement occasion are independent, i.e., conditional on $i$ and $t$ the random variables $e_{itj_1}$ and $e_{itj_2}$ are independent for all $j_1 \neq j_2$.

It can be shown that using the threshold

$$Y_{it} = j, \text{ given } Y_{it} \geq j \Leftrightarrow U_{itj} \leq \beta_{0j}$$

correlated ordinal multinomial responses that satisfy the marginal continuation ratio model specification in (3) are generated. This approach extends the latent variable representation described in Tutz (1991) to correlated multinomial responses.

The function `rmult.crm()` requires the user to provide the common cluster size $T$ (`clsize`), the linear predictor excluding the response category intercepts in a matrix form (`lin.pred`), the correlation matrix used in the NORTA method (`cor.matrix`) for the generation of $\{e_{itj}\}$, the category specific intercepts $\beta_{0j}$'s (`cuts`) and the cumulative distribution function $F$ (`link`). The `lin.pred` argument should be an $N \times T$ matrix such that the $i$-th row corresponds to subject $i$ and has elements $(\mu_{i1}, \mu_{i2}, \cdots, \mu_{iT})$.

For example, suppose that we want to simulate ordinal multinomial responses under a marginal continuation ratio probit model with $N = 500$, $J = 5$, $T = 4$, $(\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}) = (-1.5, -0.5, 0.5, 1.5)$, $\boldsymbol{\beta} = \beta = 1$ and a single time-stationary covariate for each subject drawn from a standard normal distribution. To simplify matters further, suppose that $\{e_{itj}\}$ are independent. The following **R** code generates the clustered ordinal multinomial responses under this configuration

```
> set.seed(1)
> N <- 500
> clustersize <- 4
> intercepts <- c(-Inf,-1.5,-0.5,0.5,1.5,Inf)
> cormat <- diag(1,16)
> x <- rnorm(N)
> linpred <- matrix(rep(x,clustersize),N,clustersize,byrow=TRUE)
> Y <- rmult.crm(clsize=clustersize,lin.pred=linpred,cor.matrix=cormat,
+                cuts=intercepts,link="probit")
```

The simulated clustered ordinal multinomial responses for the first six subjects are

```
> head(Y$Ysim)
```

```
     [,1] [,2] [,3] [,4]
[1,]    2    5    2    3
[2,]    5    5    2    4
[3,]    2    5    5    4
[4,]    2    4    3    1
[5,]    5    4    2    5
[6,]    1    5    5    5
```

# References


[1] McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society B* (**42**), 109-142.

[2] McFadden, D. (1973). *Conditional logit analysis of qualitative choice behavior*. Institute of Urban and Regional Development, University of California.

[3] Tutz, G. (1991). Sequential models in categorical regression, *Computational Statistics & Data Analysis* (**11**), 275-295.