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

Anestis Touloumis

1 Introduction

The **R** package **SimCorMultRes** is suitable for simulation of correlated ordinal or nominal multinomial responses (with three or more response categories) and of correlated binary responses conditional on a model specification for the marginal probabilities, which is accomplished by extending existing threshold approaches that give rise to regression models for independent binary and multinomial responses. We describe some theoretical details of these approaches and we provide simple examples to illustrate the use of the core functions in **SimCorMultRes**.

Let Y_{it} be the binary or 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. Note that we assume that $Y_{it} \in \{0, 1\}$ for binary responses and $Y_{it} \in \{1, 2, \ldots, J \geq 3\}$ for multinomial responses.

2 Correlated Nominal Multinomial Responses

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

$$\log\left(\frac{\Pr(Y_{it} = j|\mathbf{x}_{it})}{\Pr(Y_{it} = I|\mathbf{x}_{it})}\right) = (\beta_{t0j} - \beta_{t0J}) + (\boldsymbol{\beta}_{tj} - \boldsymbol{\beta}_{tJ})'\mathbf{x}_{it} = \beta^*_{t0j} + \boldsymbol{\beta}^{*\prime}_{tj}\mathbf{x}_{it}, \tag{1}$$

where β_{t0j} and $\boldsymbol{\beta}_{tj}$ is the j-th response category specific intercept and parameter vector at the t-th measurement occasion respectively. The popular identifiability constraints $\beta_{t0J} = 0$ and $\boldsymbol{\beta}_{tJ} = \mathbf{0}$ imply that $\beta^*_{t0j} = \beta_{t0j}$ and $\boldsymbol{\beta}^*_{tj} = \boldsymbol{\beta}_{tj}$ for all $j = 1, \ldots, J-1$.
Define

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

where $\mu_{itj} = \beta_{0j} + \boldsymbol{\beta}'_{tj}\mathbf{x}_{it}$ 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. Random variables associated with different subjects are independent. That is, the random variables $e_{i_1t_1j_1}$ and $e_{i_2t_2j_2}$ are independent provided that $i_1 \neq i_2$.
3. Category specific random variables for each subject at a given measurement occasion are independent, i.e., e_{itj_1} and e_{itj_2} are independent for all i, t and $j_1 \neq j_2$.

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. The above threshold approach extends the principle of maximum random utility (McFadden, 1973) to correlated nominal 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 of model (1) in a matrix form

(`lin.pred`) and the correlation matrix in the multivariate normal distribution of the NORTA method (`cor.matrix`). The `lin.pred` argument should be an $N \times (TJ)$ matrix such that the i-th row 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 $N = 500$, $J = 4$, $T = 3$, $(\beta_{t01}, \beta_{t02}, \beta_{t03}, \beta_{t04}) = (\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}) = (1, 3, 5, 0)$ and $\boldsymbol{\beta}_t = \boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4) = (2, 4, 6, 0)$ for all t, 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 following **R** code is used to simulate nominal multinomial responses under this sampling scheme

```
> library("SimCorMultRes")
> set.seed(1)
> N <- 500
> ncategories <- 4
> clsize <- 3
> Xmat <- matrix(rnorm(N),N,ncategories)
> betas <- c(1,2,3,4,5,6)
> lin.pred <- 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)
> lin.pred <- matrix(lin.pred,N,ncategories*clsize)
> cor.matrix <- diag(1,12)
> Y <- rmult.bcl(clsize,ncategories,lin.pred,cor.matrix)
```

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

$$\Pr(Y_{it} \leq j | \mathbf{x}_{it}) = F(\beta_{t0j} + \boldsymbol{\beta}_t' \mathbf{x}_{it}) \tag{2}$$

where β_{t0j} is the j-th category specific intercept at the t-th measurement occasion and $\boldsymbol{\beta}_t$ is the parameter vector at the t-th measurement occasion, and F is a cumulative distribution function. The response category specific intercepts are assumed to be monotone increasing, that is

$$-\infty = \beta_{t00} < \beta_{t01} < \beta_{t02} < \cdots < \beta_{t0(J-1)} < \beta_{t0J} = \infty$$

for all t. Define

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

where $\mu_{it} = \boldsymbol{\beta}_t' \mathbf{x}_{it}$ 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. Random variables associated with different subjects are independent, i.e., the random variables $e_{i_1t_1}$ and $e_{i_2t_2}$ are independent for all $i_1 \neq i_2$.

It can be shown that using the threshold

$$Y_{it} = j \Leftrightarrow \beta_{t0(j-1)} < U_{it} \leq \beta_{t0j}$$

correlated ordinal multinomial responses under the marginal cumulative link model specification in (2) are generated. This threshold extends the approach of McCullagh (1980) to generating correlated multinomial responses.

The function `rmult.clm()` requires the user to provide the common cluster size T (`clsize`), the linear predictor of model (2) excluding the response category intercepts in a matrix form (`lin.pred`), the correlation matrix in the multivariate normal distribution of the NORTA method (`cor.matrix`), the response category specific intercepts β_{0tj}'s (`intercepts`) and the cumulative distribution function F (`link`). The `lin.pred` argument should be an $N \times T$ matrix such that the i-th row 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 $N = 500$, $J = 5$, $T = 4$, $(\beta_{t01}, \beta_{t02}, \beta_{t03}, \beta_{t04}) = (\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}) = (-1.5, -0.5, 0.5, 1.5)$ and $\boldsymbol{\beta}_t = \beta = 1$ for all t, a single time-stationary covariate for each subject drawn from a standard normal distribution and a latent correlation matrix equal to

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

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

```
> set.seed(12345)
> N <- 500
> clsize <- 4
> intercepts <- c(-1.5,-0.5,0.5,1.5)
> cor.matrix <- toeplitz(c(1,0.85,0.5,0.15))
> lin.pred <- rsmvnorm(N,toeplitz(c(1,rep(0.85,clsize-1))))
> Y <- rmult.clm(clsize,lin.pred,cor.matrix,intercepts,"probit")
```

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

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

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

3.2 Marginal continuation ratio model

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

$$\Pr(Y_{it} = j | Y_{it} \geq j, \mathbf{x}_{it}) = F(\beta_{t0j} + \boldsymbol{\beta}_t^{'} \mathbf{x}_{it}) \tag{3}$$

where β_{t0j} and $\boldsymbol{\beta}_t$ is the j-th category specific intercept and the parameter vector at the t-th measurement occasion respectively, and F is a cumulative distribution function. The response category specific intercepts are assumed to be monotone increasing at each measurement occasion, that is

$$-\infty = \beta_{t00} < \beta_{t01} < \beta_{t02} < \ldots < \beta_{t0(J-1)} < \beta_{t0J} = \infty$$

for all t. Define

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

where $\mu_{it} = \boldsymbol{\beta}_t' \mathbf{x}_{it}$ 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. Random variables associated with different subjects are independent, i.e., the random variables $e_{i_1 t_1 j_1}$ and $e_{i_2 t_2 j_2}$ 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., 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

$$Y_{it} = j, \text{ given } Y_{it} \geq j \Leftrightarrow U_{itj} \leq \beta_{t0j}$$

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 generating correlated multinomial responses.

The function `rmult.crm()` requires the user to provide the common cluster size T (`clsize`), the linear predictor of model (3) excluding the response category intercepts in a matrix form (`lin.pred`), the correlation matrix of the multivariate normal distribution in the NORTA method (`cor.matrix`), the category specific intercepts β_{0j}'s (`intercepts`) and the cumulative distribution function F (`link`). The `lin.pred` argument should be an $N \times T$ matrix such that the i-th row 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 $N = 500$, $J = 5$, $T = 4$, $(\beta_{t01}, \beta_{t02}, \beta_{t03}, \beta_{t04}) = (\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}) = (-1.5, -0.5, 0.5, 1.5)$ and $\boldsymbol{\beta}_t = \beta = 1$ for all t 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
> clsize <- 4
> intercepts <- c(-1.5,-0.5,0.5,1.5)
> cor.matrix <- diag(1,16)
> x <- rnorm(N)
> lin.pred <- matrix(rep(x,clsize),N,clsize,byrow=TRUE)
> Y <- rmult.crm(clsize,lin.pred,cor.matrix,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
```

4 Correlated Binary Responses

The function `rbin()` simulates correlated binary responses under the marginal model

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

where β_{t0} and $\boldsymbol{\beta}_t$ is the intercept and the parameter vector at the t-th measurement occasion respectively, and F is a cumulative distribution function.
For subject i, define $\mu_{it} = \boldsymbol{\beta}_t' \mathbf{x}_{it}$ and define the random variables $\{e_{it}\}$ such that:

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

2. Random variables associated with different subjects are independent, i.e., the random variables $e_{i_1t_1}$ and $e_{i_2t_2}$ are independent for all $i_1 \neq i_2$.

It can be shown that using the threshold

$$Y_{it} = I(e_{it} \leq \beta_{t0} + \mu_{it}),$$

correlated binary responses under the marginal model specification in (4) are generated. Here, $I(A)$ is the indicator function of the event A.

The function `rbin()` requires the user to provide the common cluster size T (`clsize`), the linear predictor of model (4) excluding the intercept in a matrix form (`lin.pred`), the correlation matrix in the multivariate normal distribution of the NORTA method (`cor.matrix`), the intercepts $\beta_{01}, \ldots, \beta_{0T}$ (`intercepts`) and the cumulative distribution function F (`link`). The `lin.pred` argument should be an $N \times T$ matrix such that the i-th row has elements $(\mu_{i1}, \mu_{i2}, \cdots, \mu_{iT})$.

For example, suppose that we want to simulate correlated binary responses from a marginal cumulative probit model with $N = 500$, $J = 5$, $\beta_{t0} = \beta_0 = 1$ and $\boldsymbol{\beta}_t = \beta = 1$ for all t, a single time-stationary covariate for each subject drawn from a standard normal distribution and a latent correlation matrix equal to

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

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

```
> set.seed(1)
> N <- 500
> clsize <- 4
> intercepts <- 1
> cor.matrix <- toeplitz(c(1,0.85,0.5,0.15))
> lin.pred <- matrix(rnorm(N),N,clsize)
> Y <- rbin(clsize,lin.pred,cor.matrix,intercepts,"probit")
```

The simulated clustered binary responses for the first six subjects are

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

```
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    1    1    1    1
[3,]    1    1    1    0
[4,]    1    1    1    1
[5,]    1    1    1    1
[6,]    0    0    1    1
```

References

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

[2] McFadden, D. (1974). *Conditional logit analysis of qualitative choice behavior*. New York: Academic Press, 105–142.

[3] Tutz, G. (1991). Sequential models in categorical regression, *Computational Statistics & Data Analysis*, **11**, 27–295.