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

Anestis Touloumis

1 Introduction

The **R** package **SimCorMultRes** is suitable for simulation of correlated multinomial responses (with three or more nominal or ordinal response categories) and of correlated binary responses conditional on a model specification for the marginal probabilities. The simulation methods employed herein are extending existing threshold approaches that give rise to regression models for independent binary, nominal or ordinal responses. This vignette describes briefly the threshold approaches that give rise to the implemented marginal models and offers simple examples that illustrate the use of **SimCorMultRes**.

2 Notation

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

3 Correlated Nominal Responses

The function `rmult.bcl()` simulates correlated 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 the t-th measurement occasion and $\boldsymbol{\beta}_{tj}$ is the j-th category-specific parameter vector at the t-th measurement occasion. 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 $j = 1, \dots, J-1$.

Define

$$U_{itj} = \beta_{tj0} + \boldsymbol{\beta}'_{tj}\mathbf{x}_{it} + e_{itj},$$

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, i.e., $e_{i_1t_1j_1}$ 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 (assumption of choice independence), i.e., e_{itj_1} and e_{itj_2} are independent provided that $j_1 \neq j_2$.
4. Subject-specific random variables measured at different occasions may be correlated, i.e., $e_{it_1j_1}$ and $e_{it_2j_2}$ may be correlated provided that $t_1 \neq t_2$.

It can be shown that using the threshold

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

correlated nominal responses that satisfy the marginal baseline-category logit model (1) are generated.

For example, suppose that we want 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 $N=500$, $T=3$, $(\beta_{10},\beta_{11},\beta_{12},\beta_{20},\beta_{21},\beta_{22},\beta_{30},\beta_{31},\beta_{32})=(1,2,1.5,3,4,3.5,5,6,5.5)$ 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 simplicity, assume that $\{e_{itj}\}$ are independent random variables. The following **R** code is used to simulate nominal responses under this sampling scheme

```
> library(SimCorMultRes)
> set.seed(1)
> ncategories <- 4
> N <- 500
> clsize <- 3
> betas <- c(1, 2, 1.5, 3, 4, 3.5, 5, 6, 5.5, 0, 0, 0)
> x1 <- rep(rnorm(N), each = clsize)
> x2 <- rnorm(N * clsize)
> xdata <- data.frame(x1, x2)
> cor.matrix <- diag(1, 12)
> CorNorRes <- rmult.bcl(clsize = clsize, ncategories = ncategories, betas = betas,
+     xformula = ~x1 + x2, xdata = xdata, cor.matrix = cor.matrix)
```

The simulated clustered nominal responses for the first six subjects are

```
> head(CorNorRes$Ysim)
```

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

The same task without utilizing the NORTA method:

```
> library(evd)
> rlatent <- rmvevd(n = N, dep = 1, model = "log", d = clsize * ncategories)
> CorNorRes <- rmult.bcl(clsize = clsize, ncategories = ncategories, betas = betas,
+     xformula = ~x1 + x2, xdata = xdata, rlatent = rlatent)
> head(CorNorRes$Ysim)
```

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

4 Correlated Ordinal Responses

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

4.1 Marginal cumulative link model

The function `rmult.clm()` simulates correlated 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 the t-th measurement occasion and $\boldsymbol{\beta}_t$ is the regression parameter vector at the t-th measurement occasion. The category-specific intercepts at each measurement occasion 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.

Define

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

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., $e_{i_1t_1}$ and $e_{i_2t_2}$ are independent provided that $i_1 \neq i_2$.
3. Subject-specific random variables measured at different occasions may be correlated, i.e., e_{it_1} and e_{it_2} may be correlated provided that $t_1 \neq t_2$.

It can be shown that using the threshold

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

correlated ordinal responses that satisfy the marginal cumulative link model (2) are generated.

For example, suppose that we want to simulate correlated ordinal responses from the marginal cumulative probit model

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

with $N = 500$, $T = 4$, $(\beta_{10}, \beta_{20}, \beta_{30}, \beta_{40}) = (-1.5, -0.5, 0.5, 1.5)$, $(\beta_{11}, \beta_{21}, \beta_{31}, \beta_{41}) = (1, 2, 3, 4)$, $\mathbf{x}_{it} = x_i \overset{iid}{\sim} N(0, 1)$ for all i and t, and a latent correlation matrix equal to

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

Here Φ denotes the cumulative distribution function of the standard normal distribution. The following **R** code generates the clustered ordinal responses under this configuration

```
> set.seed(12345)
> N <- 500
> clsize <- 4
> intercepts <- c(-1.5, -0.5, 0.5, 1.5)
> betas <- matrix(c(1, 2, 3, 4), 4, 1)
> x <- rep(rnorm(N), each = clsize)
> cor.matrix <- toeplitz(c(1, 0.85, 0.5, 0.15))
> CorOrdRes <- rmult.clm(clsize = clsize, intercepts = intercepts, betas = betas,
+     xformula = ~x, cor.matrix = cor.matrix, link = "probit")
```

The simulated clustered ordinal responses for the first six subjects are

```
> head(CorOrdRes$Ysim)
```

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

4.2 Marginal continuation-ratio model

The function `rmult.crm()` simulates correlated 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 the t-th measurement occasion, $\boldsymbol{\beta}_t$ is the regression parameter vector at the t-th measurement occasion and F is a cdf.

Define

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

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., $e_{i_1 t_1 j_1}$ and $e_{i_2 t_2 j_2}$ 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 provided that $j_1 \neq j_2$ (local independence assumption).

4. Subject-specific random variables measured at different occasions may be correlated, i.e., $e_{it_1 j_1}$ and $e_{it_2 j_2}$ may be correlated provided that $t_1 \neq t_2$.

It can be shown that using the threshold

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

correlated ordinal responses that satisfy the marginal continuation-ratio model (3) are generated.

Suppose we want to simulate ordinal multinomial 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. To simplify matters further, suppose that $\{e_{itj}\}$ are independent. The following **R** code generates the clustered ordinal 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 * clsize)
> CorOrdRes <- rmult.crm(clsize = clsize, intercepts = intercepts, betas = 1,
+     xformula = ~x, cor.matrix = cor.matrix, link = "probit")
>
```

The simulated clustered ordinal responses for the first six subjects are

```
> head(CorOrdRes$Ysim)
```

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

5 Correlated Binary Responses

The function `rbin()` simulates correlated 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 parameter vector at measurement occasion t and F is a cdf.

Define

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

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., $e_{i_1 t_1}$ and $e_{i_2 t_2}$ are independent provided that $i_1 \neq i_2$.
3. Subject-specific random variables may be correlated, i.e., e_{it_1} and e_{it_2} may be correlated provided that $t_1 \neq t_2$.

It can be shown that using the threshold

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

correlated binary responses that satisfy the marginal model (4) are generated.

Suppose that the goal is to simulate correlated binary responses from the marginal probit model

$$\Pr(Y_{it} = 1|\mathbf{x}_{it}) = \Phi(0.2x_i) \tag{5}$$

where $N = 5000$, $T = 4$, $\mathbf{x}_{it} = x_i \overset{iid}{\sim} N(0,1)$ for all i and t, and latent correlation matrix

$$\begin{pmatrix} 1.00 & 0.90 & 0.90 & 0.90 \\ 0.90 & 1.00 & 0.90 & 0.90 \\ 0.90 & 0.90 & 1.00 & 0.90 \\ 0.90 & 0.90 & 0.90 & 1.00 \end{pmatrix}.$$

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

```
> set.seed(123)
> N <- 5000
> clsize <- 4
> intercepts <- 0
> betas <- 0.2
> cor.matrix <- toeplitz(c(1, 0.9, 0.9, 0.9))
> x <- rep(rnorm(N), each = clsize)
> CorBinRes <- rbin(clsize = clsize, intercepts = intercepts, betas = betas,
+     xformula = ~x, cor.matrix = cor.matrix, link = "probit")
```

To verify that the simulated clustered binary responses satisfy the marginal model (5), a binary GEE model can be fitted

```
> library(gee)
> binGEEmod <- gee(y ~ x, family = binomial("probit"), id = id, data = CorBinRes$simdata)
```

```
(Intercept)           x
0.002636705 0.204827031
```

```
> summary(binGEEmod)
```

```
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Probit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Independent

Call:
gee(formula = y ~ x, id = id, data = CorBinRes$simdata, family = binomial("probit"))

Summary of Residuals:
       Min         1Q     Median         3Q        Max
-0.7571225 -0.4867496  0.2562934  0.4865769  0.7325948


Coefficients:
               Estimate  Naive S.E.    Naive z Robust S.E.   Robust z
(Intercept) 0.002636705 0.008929290  0.2952872  0.01572132  0.1677153
x           0.204827031 0.009114596 22.4724192  0.01610695 12.7166857

Estimated Scale Parameter:  1.000165
Number of Iterations:  1

Working Correlation
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

Now consider the same set-up as in marginal model (5) but with the cdf of the logistic function instead of Φ. Here is one way to achieve simulation of correlated binary responses under this configuration without employing the NORTA method:

```
> set.seed(123)
> library(evd)
> 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
> CorBinRes <- rbin(clsize = clsize, intercepts = intercepts, betas = betas,
+     xformula = ~x, rlatent = rlatent)
> binGEEmod <- gee(y ~ x, family = binomial("logit"), id = id, data = CorBinRes$simdata)
```

```
(Intercept)           x
0.002196042 0.261082668
```

```
> summary(binGEEmod)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Independent

Call:
gee(formula = y ~ x, id = id, data = CorBinRes$simdata, family = binomial("logit"))

Summary of Residuals:
       Min         1Q     Median         3Q        Max
-0.6999249 -0.4918525  0.2916543  0.4912009  0.6830804


Coefficients:
               Estimate Naive S.E.    Naive z Robust S.E.    Robust z
(Intercept) 0.002196042 0.01425978  0.1540024  0.02510421  0.08747701
x           0.261082668 0.01457592 17.9119179  0.02551444 10.23274335

Estimated Scale Parameter:  1.000124
Number of Iterations:  1

Working Correlation
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

6 Citation Information

```
> citation("SimCorMultRes")

To cite the R package 'SimCorMultRes' in publications, please use:

  Touloumis, A. (2016) SimCorMultRes: Simulates Correlated Multinomial
  Responses, R package version 1.4.1,
  URL=http://CRAN.R-project.org/package=SimCorMultRes.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {SimCorMultRes: Simulates Correlated Multinomial Responses},
    author = {Anestis Touloumis},
    year = {2016},
    url = {http://CRAN.R-project.org/package=SimCorMultRes},
    note = {R package version 1.4.1},
  }
```