R Package multgee: A Generalized Estimating Equations Solver for Multinomial Responses

Anestis Touloumis
Cancer Research UK Cambridge Institute, University of Cambridge

Abstract

This introduction to the R package **multgee** is a slightly modified version of Touloumis (2015), published in the Journal of Statistical Software. To cite **multgee** in publications, please use Touloumis (2015). To cite the GEE methodology implemeted in **multgee**, please use Touloumis, Agresti, and Kateri (2013).

The R package **multgee** implements the local odds ratios generalized estimating equations (GEE) approach proposed by Touloumis *et al.* (2013), a GEE approach for correlated multinomial responses that circumvents theoretical and practical limitations of the GEE method. A main strength of **multgee** is that it provides GEE routines for both ordinal (`ordLORgee`) and nominal (`nomLORgee`) responses, while relevant softwares in R and SAS are restricted to ordinal responses under a marginal cumulative link model specification. In addition, **multgee** offers a marginal adjacent categories logit model for ordinal responses and a marginal baseline category logit model for nominal. Further, utility functions are available to ease the local odds ratios structure selection (`intrinsic.pars`) and to perform a Wald type goodness-of-fit test between two nested GEE models (`waldts`). We demonstrate the application of **multgee** through a clinical trial with clustered ordinal multinomial responses.

Keywords: generalized estimating equations, nominal and ordinal multinomial responses, local odds ratios, R.

1. Introduction

In several studies, the interest lies in drawing inference about the regression parameters of a marginal model for correlated, repeated or clustered multinomial variables with ordinal or nominal response categories while the association structure between the dependent responses is of secondary importance. The lack of a convenient multivariate distribution for multinomial responses and the sensitivity of ordinary maximum likelihood methods to misspecification of the association structure led researchers to modify the GEE method of Liang and Zeger (1986) in order to account for multinomial responses (Miller, Davis, and Landis 1993; Lipsitz, Kim, and Zhao 1994; Williamson, Kim, and Lipsitz 1995; Lumley 1996; Heagerty and Zeger 1996; Parsons, Edmondson, and Gilmour 2006). These GEE approaches estimate the marginal regression parameter vector by solving the same set of estimating equations as in Liang and Zeger (1986), but differ in the way they parametrize and/or estimate $\boldsymbol{\alpha}$, a parameter vector that is usually defined to describe a "working" assumption about the association structure.

Touloumis *et al.* (2013) showed that the joint existence of the estimated marginal regres-

sion parameter vector and $\hat{\boldsymbol{\alpha}}$ cannot be assured in existing approaches. This is because the parametric space of the proposed parameterizations of the association structure depends on the marginal model specification even in the simple case of bivariate multinomial responses. To address this issue, Touloumis *et al.* (2013) defined $\boldsymbol{\alpha}$ as a "nuisance" parameter vector that contains the marginalized local odds ratios structure, that is the local odds ratios as if no covariates were recorded, and they employed the family of association models (Goodman 1985) to develop parsimonious and meaningful structures regardless of the response scale. The practical advantage of the local odds ratios GEE approach is that it is applicable to both ordinal and nominal multinomial responses without being restricted by the marginal model specification. Simulations in Touloumis *et al.* (2013) imply that the local odds ratios GEE approach captures a significant portion of the underlying correlation structure, and compared to the independence 'working' model (i.e., assuming no correlation structure in the GEE methodology), simple local odds ratios structures can substantially increase the efficiency gains in estimating the regression vector of the marginal model. Note that low convergence rates for the GEE approach of Lumley (1996) and Heagerty and Zeger (1996) did not allow the authors to compare these approaches with the local odds ratios GEE approach while the GEE approach of Parsons *et al.* (2006) was excluded from the simulation design because its use is restricted to a cumulative logit marginal model specification.

The R (R Core Team 2014) package **multgee** implements the local odds ratios GEE approach and it is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=multgee`. To emphasize the importance of reflecting the nature of the response scale on the marginal model specification and on the marginalized local odds ratios structure, two core functions are available in **multgee**: `nomLORgee` which is appropriate for GEE analysis of nominal multinomial responses and `ordLORgee` which is appropriate for ordinal multinomial responses. In particular, options for the marginal model specification include a baseline category logit model for nominal response categories and a cumulative link model or an adjacent categories logit model for ordinal response categories. In addition, there are three utility functions that enable the user to: i) Perform goodness-of-fit tests between two nested GEE models (`waldts`), ii) select the local odds ratios structure based on the rule of thumb discussed in Touloumis *et al.* (2013) (`intrinsic.pars`), and iii) construct a probability table (to be passed in the core functions) that satisfies a desired local odds ratios structure (`matrixLOR`).

To appreciate the features of **multgee**, we briefly review GEE software for multinomial responses in SAS (SAS Institute Inc. 2003) and R. The current version of SAS supports only the independence "working" model under a marginal cumulative probit or logit model for ordinal multinomial responses. To the best of our knowledge, SAS macros (Williamson, Lipsitz, and Kim 1998; Yu and Yuan 2004) implementing the approach of Williamson *et al.* (1995) are not publicly available. The R package **repolr** (Parsons 2013) implements the approach of Parsons *et al.* (2006) but it is restricted to using a cumulative logit model. Another option for ordinal responses is the function `ordgee` in the R package **geepack** (Halekoh, Højsgaard, and Yan 2006). This function implements the GEE approach of Heagerty and Zeger (1996) but it seems to produce unreliable results for multinomial responses. To illustrate this, we simulated independent multinomial responses under a cumulative probit model specification with a single time-stationary covariate for each subject and we employed `ordgee` to obtain the GEE estimates from the independence 'working' model. Description of the generative process can be found in Scenario 1 of Touloumis *et al.* (2013) except that we used the values $-3, -1, 1$

and 3 for the four category specific intercepts in order to make the problem more evident. Based on 1000 simulation runs when the sample size $N = 500$, we found that the bias of the GEE estimate of $\beta = 1$ was $\approx 4.8 \times 10^{28}$, indicating the presence of a bug or -at least- of numerical problems for some situations. Similar problems occurred for the alternative global odds ratios structures in `ordgee`. In contrast to existing software, **multgee** offers greater variety of GEE models for ordinal responses, implements a GEE model for nominal responses and is not limited to the independence "working" model, which might lead to significant efficiency losses. Further, one can assess the goodness of fit for two or more nested GEE models.

This paper is organized as follows. In Section 2, we present the theoretical background of the local odds ratios GEE approach that is necessary for the use of **multgee**. We introduce the marginal models implemented in **multgee**, the estimation procedure for the 'nuisance' parameter vector $\boldsymbol{\alpha}$ and the asymptotic theory on which GEE inference is based. We describe the arguments of the core GEE functions (`nomLORgee`, `ordLORgee`) in Section 3 while the utility functions (`waldts`, `intrinsic.pars`, `matrixLOR`) are described in Section 4. In Section 5, we illustrate the use of **multgee** in a longitudinal study with correlated ordinal multinomial responses. We summarize the features of the package and provide a few practical guidelines in Section 6.

2. Local odds ratios GEE approach

For notational ease, suppose the data arise from a longitudinal study with no missing observations. However, note that the local odds ratios GEE approach is not limited neither to longitudinal studies nor to balanced designs, under the strong assumption that missing observations are missing completely at random (Rubin 1976).

Let Y_{it} be the multinomial response for subject i $(i = 1, \ldots, N)$ at time t $(t = 1, \ldots, T)$ that takes values in $\{1, 2, \ldots, J\}$, $J > 2$. Define the response vector for subject i

$$\mathbf{Y}_i = (Y_{it1}, \ldots, Y_{i1(J-1)}, Y_{i21}, \ldots, Y_{i2(J-1)}, \ldots, Y_{iT1}, \ldots, Y_{iT(J-1)})^\top,$$

where $Y_{itj} = 1$ if the response for subject i at time t falls at category j and $Y_{itj} = 0$ otherwise. Denote by $\mathbf{x}_{it}$ the covariates vector associated with Y_{it}, and let $\mathbf{x}_i = (\mathbf{x}_{i1}^\top, \ldots, \mathbf{x}_{iT}^\top)^\top$ be the covariates matrix for subject i. Define $\pi_{itj} = \mathsf{E}(Y_{itj}|\mathbf{x}_i) = \mathsf{P}(Y_{itj} = 1|\mathbf{x}_i) = \mathsf{P}(Y_{it} = j|\mathbf{x}_i)$ as the probability of the response category j for subject i time t, and let $\boldsymbol{\pi}_i = (\boldsymbol{\pi}_{i1}^\top, \ldots, \boldsymbol{\pi}_{iT}^\top)^\top$ be the mean vector of $\mathbf{Y}_i$, where $\boldsymbol{\pi}_{it} = (\pi_{it1}, \ldots, \pi_{it(J-1)})^\top$. It follows from the above that $Y_{itJ} = 1 - \sum_{j=1}^{J-1} Y_{itj}$ and $\pi_{itJ} = 1 - \sum_{j=1}^{J-1} \pi_{itj}$.

2.1. Marginal models for correlated multinomial responses

The choice of the marginal model depends on the nature of the response scale. For ordinal multinomial responses, the family of cumulative link models

$$F^{-1}\left[\mathsf{P}(Y_{it} \leq j|\mathbf{x}_i)\right] = \beta_{0j} + \boldsymbol{\beta}_*^\top \mathbf{x}_{it} \tag{1}$$

or the adjacent categories logit model

$$\log\left(\frac{\pi_{itj}}{\pi_{it(j+1)}}\right) = \beta_{0j} + \boldsymbol{\beta}_*^\top \mathbf{x}_{it} \tag{2}$$

can be used, where F is the cumulative distribution function of a continuous distribution and $\{\beta_{0j} : j = 1, \ldots, J-1\}$ are the category specific intercepts. For nominal multinomial responses, the baseline category logit model

$$\log\left(\frac{\pi_{itj}}{\pi_{itJ}}\right) = \beta_{0j} + \boldsymbol{\beta}_j^\top \mathbf{x}_{it} \tag{3}$$

can be used, where $\boldsymbol{\beta}_j$ is the j-th category specific parameter vector.

It is worth mentioning that the linear predictor differs in the above marginal models. First, the category specific intercepts need to satisfy a monotonicity condition $\beta_{01} \leq \beta_{02} \leq \ldots \leq \beta_{0(J-1)}$ only when the family of cumulative link models in (1) is employed. Second, the regression parameter coefficients of the covariates $\mathbf{x}_{it}$ are category specific only in the marginal baseline category logit model (3) and not in the ordinal marginal models (1) and (2).

2.2. Estimation of the marginal regression parameter vector

To unify the notation, let $\boldsymbol{\beta}$ be the p-variate parameter vector that includes all the regression parameters in (1), (2) or (3). To obtain $\widehat{\boldsymbol{\beta}}_{\boldsymbol{G}}$, a GEE estimator of $\boldsymbol{\beta}$, Touloumis *et al.* (2013) solved the estimating equations

$$\mathbf{U}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{D}_i \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\pi}_i) = \mathbf{0} \tag{4}$$

where $\mathbf{D}_i = \partial\boldsymbol{\pi}_i/\partial\boldsymbol{\beta}$ and $\mathbf{V}_i$ is a $T(J-1) \times T(J-1)$ 'weight' matrix that depends on $\boldsymbol{\beta}$ and on $\widehat{\boldsymbol{\alpha}}$, an estimate of the 'nuisance' parameter vector $\boldsymbol{\alpha}$ defined formally in Section 2.3. Succinctly, $\mathbf{V}_i$ is a block matrix that mimics the form of $\mathsf{COV}(\mathbf{Y}_i|\mathbf{x}_i)$, the true covariance matrix for subject i. The t-th diagonal matrix of $\mathbf{V}_i$ is the covariance matrix of Y_{it} determined by the marginal model. The (t, t')-th off-diagonal block matrix describes the marginal pseudo-association of $(Y_{it}, Y_{it'})$, which is a function of the marginal model and of the pseudo-probabilities $\{\mathsf{P}(Y_{it} = j, Y_{it'} = j'|\mathbf{x}_i) : j, j' = 1, \ldots, J-1\}$ calculated based on $(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta})$. We should emphasize that $\mathbf{V}_i$ is a 'weight' matrix because $\boldsymbol{\alpha}$ is defined as a 'nuisance' parameter vector and it is unlikely to describe a valid 'working' assumption about the association structure for all subjects.

2.3. Estimation of the nuisance parameter vector and of the weight matrix

Order the $L = T(T-1)/2$ time-pairs with the rightmost element of the pair most rapidly varying as $(1,2), (1,3), \ldots, (T-1, T)$, and let G be the group variable with levels the L ordered pairs. For each time-pair (t, t'), ignore the covariates and cross-classify the responses across subjects to form an $J \times J$ contingency table such that the row totals correspond to the observed totals at time t and the column totals to the observed totals at time t', and let $\theta_{tjt'j'}$ be the local odds ratio at the cutpoint (j, j') based on the expected frequencies $\{f_{tjt'j'} : j, j' = 1, \ldots, J\}$. For notational reasons, let A and B be the row and column variable respectively. Assuming a Poisson sampling scheme to the L sets of $J \times J$ contingency tables, fit the RC-G(1) type model (Becker and Clogg 1989)

$$\log f_{tjt'j'} = \lambda + \lambda_j^A + \lambda_{j'}^B + \lambda_{(t,t')}^G + \lambda_{j(t,t')}^{AG} + \lambda_{j'(t,t')}^{BG} + \phi^{(t,t')}\mu_j^{(t,t')}\mu_{j'}^{(t,t')}, \tag{5}$$

where $\{\mu_j^{(t,t')} : j = 1, \ldots, J\}$ are the score parameters for the J response categories at the time-pair (t, t'). After imposing identifiability constraints on the regression parameters in (5), the log local odds ratios structure is given by

$$\log \theta_{tjt'j'} = \phi^{(t,t')} \left(\mu_j^{(t,t')} - \mu_{j+1}^{(t,t')}\right) \left(\mu_{j'}^{(t,t')} - \mu_{j'+1}^{(t,t')}\right). \tag{6}$$

At each time-pair, (6) summarizes the local odds ratios structure in terms of the J score parameters and the intrinsic parameter $\phi^{(t,t')}$ that measures the average association of the marginalized contingency table. Since the score parameters do not need to be fixed or monotonic, the local odds ratios structure is applicable to both nominal and ordinal multinomial responses.

Touloumis *et al.* (2013) defined $\boldsymbol{\alpha}$ as the parameter vector that contains the marginalized local odds ratios structure

$$\alpha = \left(\theta_{1121}, \ldots, \theta_{1(J-1)2(J-1)}, \ldots, \theta_{(T-1)1T1}, \ldots, \theta_{(T-1)(J-1)T(J-1)}\right)^\top$$

where $\theta_{tjt'j'}$ satisfy (6). To increase the parsimony of the local odds ratios structures for ordinal responses, they proposed to use common unit-spaced score parameters $\left(\mu_j^{(t,t')} = j\right)$ and/or common intrinsic parameters $\left(\phi^{(t,t')} = \phi\right)$ across time-pairs. For a nominal response scale, they proposed to apply a homogeneity constraint on the score parameters $\left(\mu_j^{(t,t')} = \mu_j\right)$ and use common intrinsic parameters across time-pairs. To estimate $\boldsymbol{\alpha}$ maximum likelihood methods are involved by treating the L marginalized contingency tables as independent. Technical details and justification about this estimation procedure can be found in Touloumis (2011) and Touloumis *et al.* (2013).

Conditional on the estimated marginalized local odds ratios structure $\widehat{\boldsymbol{\alpha}}$ and the marginal model specification at times t and t', $\{\mathsf{P}(Y_{it} = j, Y_{it'} = j'|\mathbf{x}_i) : t < t', j, j' = 1, \ldots, J-1\}$ are obtained as the unique solution of the iterative proportional fitting (IPF) procedure (Deming and Stephan 1940). Hence, $\mathbf{V}_i$ can be readily calculated and the estimating equations in (4) can be solved with respect to $\boldsymbol{\beta}$.

2.4. Asymptotic properties of the GEE estimator

Given $\widehat{\boldsymbol{\alpha}}$, inference about $\boldsymbol{\beta}$ is based on the fact that $\sqrt{N}(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ asymptotically, where

$$\boldsymbol{\Sigma} = \lim_{N\to\infty} N\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1}, \tag{7}$$

$\boldsymbol{\Sigma}_0 = \sum_{i=1}^N \mathbf{D}_i^\top \mathbf{V}_i^{-1}\mathbf{D}_i$ and $\boldsymbol{\Sigma}_1 = \sum_{i=1}^N \mathbf{D}_i^\top \mathbf{V}_i^{-1}\mathsf{COV}(\mathbf{Y}_i|\mathbf{x}_i)\mathbf{V}_i^{-1}\mathbf{D}_i$. For finite sample sizes, $\widehat{\boldsymbol{\Sigma}}$ is estimated by ignoring the limit in (7) and replacing $\boldsymbol{\beta}$ with $\widehat{\boldsymbol{\beta}}_G$ and $\mathsf{COV}(\mathbf{Y}_i|\mathbf{x}_i)$ with $(\mathbf{Y}_i - \widehat{\boldsymbol{\pi}}_i)(\mathbf{Y}_i - \widehat{\boldsymbol{\pi}}_i)^\top$ in $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. In the literature, $\widehat{\boldsymbol{\Sigma}}/N$ is often termed as "sandwich" or "robust" covariance matrix of $\widehat{\boldsymbol{\beta}}_G$.

3. Description of core functions

We describe the arguments of the functions `nomLORgee` and `ordLORgee`, focusing on the marginal model specification (`formula`, `link`), data representation (`id`, `repeated`, `data`) and

local odds ratios structure specification (`LORstr`, `LORterm`, `homogeneous`, `restricted`). For completeness' sake, we also present computational related arguments (`LORem`, `add`, `bstart`, `LORgee.control`, `ipfp.control`, `IM`). The two core functions share the same arguments, except `link` and `restricted` which are available only in `ordLORgee`, and they both create an object of the class `LORgee` which admits `summary`, `coef`, `update` and `residuals` methods.

3.1. Marginal model specification

For ordinal multinomial responses, the `link` argument in the function `ordLORgee` specifies which of the marginal models (1) or (2) will be fitted. The options `"logit"`, `"probit"`, `"cauchit"` or `"cloglog"` indicate the corresponding cumulative distribution function F in the cumulative link model (1), while the option `"acl"` implies that the adjacent categories logit model (2) is selected. For nominal multinomial responses, the function `nomLORgee` fits the baseline category logit model (3), and hence the `link` argument is not offered.

The `formula` (`=response~covariates`) argument identifies the multinomial response variable (`response`) and specifies the form of the linear predictor (`covariates`), assuming that this includes an intercept term. If required, the $J > 2$ observed response categories are sorted in an ascending order and then mapped onto $\{1, 2, \ldots, J\}$. To account for a covariate `x` with a constrained parameter coefficient fixed to 1 in the linear predictor, the term `offset(x)` must be inserted on the right hand side of `formula`.

3.2. Data representation

The `id` argument identifies the N subjects by assigning a unique label to each subject. If required, the observed `id` labels are sorted in an ascending order and then relabeled as $1, \ldots, N$, respectively.

The `repeated` argument identifies the times at which the multinomial responses are recorded by treating the T unique observed times in the same manner as in `id`. The purpose of `repeated` is dual: To identify the T distinct time points and to construct the full marginalized contingency table for each time-pair by aggregating the relevant/available responses across subjects. The `repeated` argument is optional and it can be safely ignored in balanced designs or in unbalanced designs in which if the t-th response is missing for a particular subject then all subsequent responses at times $t' > t$ are missing for that subject. Otherwise, it is recommended to provide the `repeated` argument in order to ensure proper construction of the full marginalized contingency table. To this end, note that if the measurement occasions are not recorded in a numerical mode, then the user should create `repeated` by mapping the T distinct measurement occasions onto the set $\{1, \ldots, T\}$ in such a way that the temporal order of the measurement occasions is preserved. For example, if the measurements occasions are recorded as "before", "baseline", "after", then the levels for `repeated` should be coded as $1, 2$ and 3, respectively.

The dataset is imported via the `data` argument in "long" format, meaning that each row contains all the information provided by a subject at a given measurement occasion. This implies that `data` must include the variables specified in the mandatory arguments `formula` and `id`, as well as the optional argument `repeated` when this is specified by the user. If no `data` is provided then the above variables are extracted from the `environment` that `nomLORgee` and `ordLORgee` are called. Currently missing observations, identified by `NA` in `data`, are ignored.

$\log \theta_{tjt'j'}$	LORstr	Functions	Parameters
ϕ	"uniform"	ordLORgee	1
$\phi^{(t,t')}$	"category.exch"	ordLORgee	L
$\phi\,(\mu_j - \mu_{j+1})\,(\mu_{j'} - \mu_{j'+1})$	"time.exch"	Both	$J-1$
$\phi^{(t,t')}\left(\mu_j^{(t,t')} - \mu_{j+1}^{(t,t')}\right)\left(\mu_{j'}^{(t,t')} - \mu_{j'+1}^{(t,t')}\right)$	"RC"	Both	$L(J-1)$

Table 1: The main options for the marginalized local odds ratios structures in **multgee**.

3.3. Marginalized local odds ratios structure specification

The marginalized local odds ratios is specified via the `LORstr` argument. Table 1 displays the structures proposed by Touloumis *et al.* (2013). Currently the default option is the time excheangeability structure (`"time.exch"`) in `nomLORgee` and the category excheangeability (`"category.exch"`) structure in `ordLORgee`. The uniform (`"uniform"`) and category excheangeability structures are not allowed in `nomLORgee` because given unit-spaced parameter scores are not meaningful for nominal response categories.

The user can also fit the independence 'working' model (`LORstr="independence"`) or even provide the local odds ratios structure (`LORstr="fixed"`) using the `LORterm` argument. In this case, an $L \times J^2$ matrix must be constructed such that the g-th row contains the vectorized form of a probability table that satisfies the desired local odds ratios structure at the time-pair corresponding to the g-th level of G.

Touloumis (2011) discussed two further versions of the `"time.exch"` and the RC (`"RC"`) structures based on using: i) Heterogeneous score parameters (`homogeneous=FALSE`) at each time-pair, and/or ii) monotone score parameters (`restricted=TRUE`), an option applicable only for ordinal response categories. However, it is sensible to employ these additional options only when the local odds ratios structures in Table 1 do not seem adequate.

It is important to mention that the user must provide only the arguments required for the specified local odds ratios structure. For example, the arguments `homogeneous`, `restricted` and `LORterm` are ignored when `LORstr="uniform"`.

3.4. Computational details

The default estimation procedure for the marginalized local odds ratios structure is to fit model (5) to the full marginalized contingency table (`LORem="3way"`) after imposing the desired restrictions on the intrinsic and the score parameters. Touloumis (2011) noticed that the estimated local odds ratios structure under model (5) is identical to that obtained by fitting independently a row and columns (RC) effect model (Goodman 1985) with homogeneous score parameters to each of the L contingency tables. Motivated by this, an alternative estimation procedure (`LORem="2way"`) for estimating the structures `"uniform"` and `"time.exch"` was proposed. In particular, one can estimate the single parameter of the `"uniform"` structure as the average of the L intrinsic parameters $\phi^{(t,t')}$ obtained by fitting the linear-by-linear association model (Agresti 2013) independently to each of the L marginalized contingency tables. For the `"time.exch"` structure, one can fit L RC effects models with homogeneous (`homogeneous=TRUE`)/heterogeneous (`homogeneous=FALSE`) score parameters and then estimate the log local odds ratio at each cutpoint (j, j') by averaging $\log \hat{\theta}_{tjt'j'}$ for $t < t'$. Regardless of the value of `LORem`, the appropriate model for counts is fitted via the function `gnm` of

the R package **gnm** (Turner and Firth 2012).

In the presence of zero observed counts, a small positive constant can be added (`add`) at each cell of the marginalized contingency table to ensure the existence of $\widehat{\boldsymbol{\alpha}}$. We conjecture that a constant of the magnitude 10^{-4} will serve this purpose without affecting the strength of the association structure.

A Fisher scoring algorithm is employed to solve the estimating equations (4) as in Lipsitz *et al.* (1994). The only difference is that now $\widehat{\boldsymbol{\alpha}}$ is not updated. The default way to obtain the initial value for $\boldsymbol{\beta}$ is via the function `vglm` of the R package **VGAM** (Yee 2010). Alternatively, the initial value can be provided by the user (`bstart`). The Fisher scoring algorithm converges when the elementwise maximum relative change in two consecutive estimates of $\boldsymbol{\beta}$ is less than or equal to a predefined positive constant ϵ. The `control` argument controls the related iterative procedure variables and printing options. The default maximum number of iterations is 15 and the default tolerance is $\epsilon = 0.001$.

Recall that calculation of the 'weight' matrix $\mathbf{V}_i$ at given values of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ relies on the IPF procedure. The `ipfp.ctrl` argument controls the related variables. The convergence criterion is the maximum of the absolute difference between the fitted and the target row and column marginals. By default, the tolerance of the IPF procedure is 10^{-6} with a maximal number of iterations equal to 200.

The `IM` argument defines which of the R functions `solve`, `qr.solve` or `cholesky` will be used to invert matrices in the Fisher scoring algorithm.

4. Description of utility functions

The function `waldts` performs a goodness-of-fit test for two nested GEE models based on a Wald test statistic. Let M_0 and M_1 be two nested GEE models with marginal regression parameter vectors $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_q^\top)^\top$, respectively. Define a matrix $\mathbf{C}$ such that $\mathbf{C}\boldsymbol{\beta}_1 = \boldsymbol{\beta}_q$. Here q equals the rank of $\mathbf{C}$ and the dimension of $\boldsymbol{\beta}_q$. The hypothesis

$$H_0 : \boldsymbol{\beta}_q = 0 \text{ vs } H_1 : \boldsymbol{\beta}_q \neq 0$$

tests the goodness-of-fit of M_0 versus M_1. Based on a Wald type approach, H_0 is rejected at α% significance level, if $(\mathbf{C}\widehat{\boldsymbol{\beta}})^\top (N\mathbf{C}\widehat{\boldsymbol{\Sigma}}\mathbf{C}^\top)^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}}) \geq X_q(\alpha)$, where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ are estimated under model M_1 and $X_q(\alpha)$ denotes the α upper quantile of a chi-square distribution with q degrees of freedom.

Touloumis *et al.* (2013) suggested to select the local odds ratios structure by inspecting the range of the L estimated intrinsic parameters under the `"category.exch"` structure for ordinal responses, or under the `"RC"` structure for nominal responses. If the estimated intrinsic parameters do not differ much, then the underlying marginalized local odds ratios structure is likely nearly exchangeable across time-pairs. In this case, the simple structures `"uniform"` or `"time.exch"` should be preferred because they tend to be as efficient as the more complicated ones. The function `intrinsic.pars` gives the estimated intrinsic parameter of each time-pair.

The single-argument function `matrixLOR` creates a two-way probability table that satisfies a desired local odds ratios structure. This function aims to ease the construction of the `LORterm` argument in the core functions `nomLORgee` and `ordLORgee`.

5. Example

To illustrate the main features of the package **multgee**, we follow the GEE analysis performed in Touloumis *et al.* (2013). The data came from a randomized clinical trial (Lipsitz *et al.* 1994) that aimed to evaluate the effectiveness of the drug Auranofin versus the placebo therapy for the treatment of rheumatoid arthritis. The five-level (1=poor, ..., 5=very good) ordinal multinomial response variable was the self-assessment of rheumatoid arthritis recorded at one ($t = 1$), three ($t = 2$) and five ($t = 3$) follow-up months. To acknowledge the ordinal response scale, the marginal cumulative logit model

$$\log\left(\frac{\mathsf{P}(Y_{it} \leq j|\mathbf{x}_i)}{1-\mathsf{P}(Y_{it} \leq j|\mathbf{x}_i)}\right) = \beta_{0j} + \beta_1 I(time_i = 3) + \beta_2 I(time_i = 5) + \beta_3 trt_i + \beta_4 I(b_i = 2) + \beta_5 I(b_i = 3) + \beta_6 I(b_i = 4) + \beta_7 I(b_i = 5). \quad (8)$$

was fitted, where $i = 1, \ldots, 301$, $t = 1, 2, 3$, $j = 1, 2, 3, 4$ and $I(A)$ is the indicator function for the event A. Here $\mathbf{x}_i$ denotes the covariates matrix for subject i that includes the self-assessment of rheumatoid arthritis at the baseline (b_i), the treatment variable (trt_i), coded as (1) for the placebo group and (2) for the drug group, and the follow-up time recorded in months ($time_i$).
The GEE analysis is performed in two steps. First, we select the marginalized local odds ratios structure by estimating the intrinsic parameters under the `"category.exch"` structure

```
R> library("multgee")
R> data("arthritis")
R> head(arthritis)
```

```
  id y sex age trt baseline time
1  1 4   2  54   2        2    1
2  1 5   2  54   2        2    3
3  1 5   2  54   2        2    5
4  2 4   1  41   1        3    1
5  2 4   1  41   1        3    3
6  2 4   1  41   1        3    5
```

```
R> intrinsic.pars(y = y, data = arthritis, id = id, repeated = time,
+                     rscale = "ordinal")
```

```
[1] 0.6517843 0.9097341 0.9022272
```

The range of the estimated intrinsic parameters is small (≈ 0.26) which suggests that the underlying marginalized association pattern is nearly constant across time-pairs. Thus we expect the `"uniform"` structure to capture adequately the underlying correlation pattern. Note that we passed the time variable to the `repeated` argument because this numerical variable indicates the measurement occasion at which each observation was recorded.

Now we fit the cumulative logit model (8) under the `"uniform"` via the function `ordLORgee`

```
R>  fit <- ordLORgee(formula = y ~ factor(time) + factor(trt) + factor(baseline),
+           link = "logit", id = id, repeated = time, data = arthritis,
+           LORstr = "uniform")
R>  summary(fit)

GEE FOR ORDINAL MULTINOMIAL RESPONSES
version 1.5.1 modified 2015-03-09

Link : Cumulative logit

Local Odds Ratios:
Structure:         uniform
Model:             3way

call:
ordLORgee(formula = y ~ factor(time) + factor(trt) + factor(baseline),
    data = arthritis, id = id, repeated = time, link = "logit",
    LORstr = "uniform")

Summary of residuals:
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.5161000 -0.2399000 -0.0749700  0.0000219 -0.0066990  0.9933000

Number of Iterations: 5

Coefficients:
                  Estimate   san.se   san.z Pr(>|san.z|)
beta01            -1.84315  0.38929 -4.7346      < 2e-16 ***
beta02             0.26692  0.35013  0.7624      0.44585
beta03             2.23132  0.36625  6.0924      < 2e-16 ***
beta04             4.52542  0.42123 10.7434      < 2e-16 ***
factor(time)3      0.00140  0.12183  0.0115      0.99080
factor(time)5     -0.36172  0.11395 -3.1743      0.00150 **
factor(trt)2      -0.51212  0.16799 -3.0486      0.00230 **
factor(baseline)2 -0.66963  0.38036 -1.7605      0.07832 .
factor(baseline)3 -1.26070  0.35252 -3.5763      0.00035 ***
factor(baseline)4 -2.64373  0.41282 -6.4041      < 2e-16 ***
factor(baseline)5 -3.96613  0.53164 -7.4602      < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Local Odds Ratios Estimates:
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11] [,12]
 [1,] 0.000 0.000 0.000 0.000 2.257 2.257 2.257 2.257 2.257 2.257 2.257 2.257
 [2,] 0.000 0.000 0.000 0.000 2.257 2.257 2.257 2.257 2.257 2.257 2.257 2.257
 [3,] 0.000 0.000 0.000 0.000 2.257 2.257 2.257 2.257 2.257 2.257 2.257 2.257
 [4,] 0.000 0.000 0.000 0.000 2.257 2.257 2.257 2.257 2.257 2.257 2.257 2.257
```

```
 [5,] 2.257 2.257 2.257 2.257 0.000 0.000 0.000 0.000 2.257 2.257 2.257 2.257
 [6,] 2.257 2.257 2.257 2.257 0.000 0.000 0.000 0.000 2.257 2.257 2.257 2.257
 [7,] 2.257 2.257 2.257 2.257 0.000 0.000 0.000 0.000 2.257 2.257 2.257 2.257
 [8,] 2.257 2.257 2.257 2.257 0.000 0.000 0.000 0.000 2.257 2.257 2.257 2.257
 [9,] 2.257 2.257 2.257 2.257 2.257 2.257 2.257 2.257 0.000 0.000 0.000 0.000
[10,] 2.257 2.257 2.257 2.257 2.257 2.257 2.257 2.257 0.000 0.000 0.000 0.000
[11,] 2.257 2.257 2.257 2.257 2.257 2.257 2.257 2.257 0.000 0.000 0.000 0.000
[12,] 2.257 2.257 2.257 2.257 2.257 2.257 2.257 2.257 0.000 0.000 0.000 0.000

pvalue of Null model: <0.0001
```

The `summary` method summarizes the fit of the GEE model including the GEE estimates, their estimated standard errors based on the "sandwich" covariance matrix and the p-values from testing the statistical significance of each regression parameter in (8). The estimated marginalized local odds ratios structure can be found in a symmetric $T(J-1) \times T(J-1)$ block matrix written symbolically as

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{\Theta}_{12} & \dots & \boldsymbol{\Theta}_{1T} \\ \boldsymbol{\Theta}_{21} & \mathbf{0} & \dots & \boldsymbol{\Theta}_{2T} \\ \dots & \dots & \ddots & \dots \\ \boldsymbol{\Theta}_{T1} & \boldsymbol{\Theta}_{T2} & \dots & \mathbf{0} \end{bmatrix}.$$

Each block denotes an $(J-1) \times (J-1)$ matrix. The (j, j')-th element of the off-diagonal block $\boldsymbol{\Theta}_{tt'}$ represents the estimate of $\theta_{tjt'j'}$. Based on the properties of the local odds ratios it is easy to see that $\boldsymbol{\Theta}_{tt'} = \boldsymbol{\Theta}_{t't}^{\top}$ for $t < t'$. Finally, the diagonal blocks are zero to reflect the fact that no local odds ratios are estimated when $t = t'$. In our example, $J = 5$ and thus each block is a 4×4 matrix. Since the `uniform` structure is selected, all local odds ratios are equal and estimated as 2.257. Finally, `pvalue of Null model` corresponds to the p-value of testing the hypothesis that no covariate is significant, i.e., $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$, based on a Wald test statistic.

The goodness-of-fit of model (8) can be tested by comparing it to a marginal cumulative logit model that additionally contains the age and gender main effects in the linear predictor

```
R> fit1 <- update(fit, formula = ~. + factor(sex) + age)
R> waldts(fit, fit1)

Goodness of Fit based on the Wald test

Model under H_0: y ~ factor(time) + factor(trt) + factor(baseline)
Model under H_1: y ~ factor(time) + factor(trt) + factor(baseline) + factor(sex) +
    age

Wald Statistic=3.9554, df=2, p-value=0.1384
```

6. Summary and practical guidelines

We described the R package **multgee** which implements the local odds ratios GEE approach (Touloumis *et al.* 2013) for correlated multinomial responses. Unlike existing GEE softwares, **multgee** allows GEE models for ordinal (`ordLORgee`) and nominal (`nomLORgee`) responses. The available local odds ratios structures (`LORstr`) in each function respect the nature of the response scale to prevent usage of ordinal local odds ratios structures (e.g., `"uniform"`) in `nomLORgee`. The fitted GEE model is summarized via the `summary` method while the estimated regression coefficient can be retrieved via the `coef` method. The statistical significance of the regression parameters can be assessed via the function `waldts`. A similar strategy to that presented in Section 5, can be adopted to analyze GEE models for correlated nominal multinomial responses.

From a practical point of view, we recommend the use of the `"uniform"` structure for ordinal responses and the `"time.exch"` structure for nominal especially when the range of the estimated intrinsic parameters (`intrinsic.pars`) is small. Based on our experience, some convergence problems might occur as the complexity of the local odds ratios structure increases and/or if the marginalized contingency tables are very sparse. Two possible solutions are either to adopt a simpler local odds ratios structure or to increase slightly the value of the constant added to the marginalized contingency tables (`add`). However, we believe that users should refrain from using the independence 'working' model unless the aforementioned strategies fail to remedy the convergence problems. To decide on the form of the linear predictor, variable selection model procedures could be incorporated using the function `waldts`.

In future versions of **multgee**, we plan to permit time-dependent intercepts in the marginal models, to increase the range of the marginal models, by including, for example, the family of continuation-ratio models for ordinal responses, and to offer a function for assessing the proportional odds assumption in models (1) and (2).

References

Agresti A (2013). *Categorical Data Analysis*. 3rd edition. John Wiley & Sons.

Becker M, Clogg C (1989). "Analysis of Sets of Two-Way Contingency Tables Using Association Models." *Journal of the American Statistical Association*, **84**, 142–151.

Deming W, Stephan F (1940). "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known." *The Annals of Mathematical Statistics*, **11**, 427–444.

Goodman L (1985). "The Analysis of Cross-Classified Data Having Ordered and/or Unordered Categories: Association Models, Correlation Models, and Asymmetry Models for Contingency Tables With or Without Missing Entries." *The Annals of Statistics*, **13**, 10–69.

Halekoh U, Højsgaard S, Yan J (2006). "The R Package **geepack** for Generalized Estimating Equations." *Journal of Statistical Software*, **15**, 1–11.

Heagerty P, Zeger S (1996). "Marginal Regression Models for Clustered Ordinal Measurements." *Journal of the American Statistical Association*, **91**, 1024–1036.

Liang K, Zeger S (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, **73**, 13–22.

Lipsitz S, Kim K, Zhao L (1994). "Analysis of Repeated Categorical Data Using Generalized Estimating Equations." *Statistics in Medicine*, **13**, 1149–1163.

Lumley T (1996). "Generalized Estimating Equations for Ordinal Data: A Note on the Working Correlation Structures." *Biometrics*, **52**, 354–361.

Miller M, Davis C, Landis J (1993). "The Analysis of Longitudinal Polytomous Data: Generalized Estimating Equations and Connections with Weighted Least Squares." *Biometrics*, **49**, 1033–1044.

Parsons N (2013). ***repolr**: Repeated Measures Proportional Odds Logistic Regression.* R package version 2.0, URL http://CRAN.R-project.org/package=repolr.

Parsons N, Edmondson R, Gilmour S (2006). "A Generalized Estimating Equation Method for Fitting Autocorrelated Ordinal Score Data with an Application in Horticultural Research." *Journal of the Royal Statistical Society C*, **55**, 507–524.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rubin D (1976). "Inference and Missing Data." *Biometrika*, **63**, 581–592.

SAS Institute Inc (2003). *SAS/STAT Software, Version 9.1*. Cary, NC. URL http://www.sas.com/.

Touloumis A (2011). *Generalized Estimating Equations for Multinomial Responses*. Ph.D. thesis, University of Florida.

Touloumis A (2015). "R Package **multgee**: A Generalized Estimating Equations Solver for Multinomial Responses." *Journal of Statistical Software*, **64**, 1–14.

Touloumis A, Agresti A, Kateri M (2013). "Generalized Estimating Equations for Multinomial Responses Using a Local Odds Ratio Parameterization." *Biometrics*, **69**, 633–640.

Turner H, Firth D (2012). *Generalized Nonlinear Models in R: An Overview of the **gnm** Package*. R package version 1.0-7, URL http://CRAN.R-project.org/package=gnm.

Williamson J, Kim K, Lipsitz S (1995). "Analyzing Bivariate Ordinal Data Using a Global Odds Ratio." *Journal of the American Statistical Association*, **90**, 1432–1437.

Williamson J, Lipsitz S, Kim K (1998). "GEECAT and GEEGOR: Computer Programs for the Analysis of Correlated Categorical Response Data." *Computer Methods and Programs in Biomedicine*, **58**, 25–34.

Yee T (2010). "The **VGAM** Package for Categorical Data Analysis." *Journal of Statistical Software*, **32**, 1–34. URL http://www.jstatsoft.org/v32/i10/.

Yu K, Yuan W (2004). "Regression Models for Unbalanced Longitudinal Ordinal Data: Computer Software and a Simulation Study." *Computer Methods and Programs in Biomedicine*, **75**, 195–200.

Affiliation:

Anestis Touloumis
Cancer Research UK Cambridge Institute
University of Cambridge
Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK
E-mail: Anestis.Touloumis@cruk.cam.ac.uk