A General GEE Framework for the Analysis of Longitudinal Ordinal Missing Data and Related Issues

José L. P. da Silva, Enrico A. Colosimo, and Fábio N. Demarqui

25 September 2017

Contents

1. Introduction

This file presents information about main functions used in the Simulation Study of the manuscript "A General GEE Framework for the Analysis of Longitudinal Ordinal Missing Data and Related Issues" submitted to Statistical Modelling Journal by Silva et al. Datasets and *R* functions can be downloaded here.

2. Description of main functions

The file 'drgee.r' contain the main function `drgee` for the DRGEE method. The file 'wgee.r' contains the main function `wgee` for the WGEE method. Auxiliary functions for both methods are given in the file 'AuxiliaryFunctions.r'. These functions have the following package dependencies: `Matrix`, `plyr` and `multgee`.

Both WGEE and DRGEE methods are implemented using the configuration described in the Simulation Study section of the manuscript. These modifications of the GEE method are meant to be used for analysis of longitudinal ordinal data in the presence of missing response and covariate that are both MAR. The proposed estimators adopt a (marginal) proportional odds model for the repeated measures and allows two choices for the association structure: the correlation coefficient of Lipsitz et al. (1994) and the local odds ratio of Touloumis et al. (2011). Parameter estimates are obtained through a Fisher scoring algorithm. Inferences can be conducted using Wald statistics and hypothesis testing that rely on asymptotic normality of estimators.

Arguments:

- eps: stopping criteria, the maximum relative change in the regression parameters at each iteration of the modified Fisher Scoring algorithm.
- ini: starting values for the regression parameters.
- kmax: maximum number of iterations allowed in the fitting algorithm.
- corstr: a character string specifying the correlation structure. The following are permitted: "ind", "exch", and "unst".

- LORstr: a character string specifying the local odds structure. The following are permitted: "independence", "uniform", "category.exch", "time.exch", and "RC".
- assoc.type: a character string specifying the association structure. The options are: "corr" for the correlation structure, and "odds" for the local odds structure.
- data: the data frame containing the variables for the fit.

Values:

- est: a named vector of the estimated coefficients
- std: a named vector of the estimated standard errors
- k: number of iterations upon convergence
- rho.est: correlation estimates
- odds.est: local odds ratio estimates

Refer to the Simulation Study section of the manuscript for more information about the models considered.

3. Sample datasets

Two sample datasets are provided: `datacomp.RData` and `datamiss.RData`. The first is a simulated complete dataset while the second is same data simulated with about one third of missing values. Both datasets are provided in the *long* format.

```
load(url("http://docs.ufpr.br/~jlpadilha/StatMod2017/datacomp.RData"))
load(url("http://docs.ufpr.br/~jlpadilha/StatMod2017/datamiss.RData"))
head(datamiss)
```

```
##     id  X time           Z Y
## 1.1  1  0    1 -0.35199656 3
## 2.1  2 NA    1 -0.12277427 2
## 3.1  3 NA    1 -0.25890313 2
## 4.1  4  1    1 -0.03075549 1
## 5.1  5 NA    1 -0.27480002 1
## 6.1  6  0    1  0.47839867 1
```

```
summary(datamiss,digits=1)
```

```
##        id            X            time         Z                Y
##  Min.   :  1   Min.   :0.0   Min.   :1   Min.   :-1.473   Min.   :1
##  1st Qu.: 76   1st Qu.:0.0   1st Qu.:1   1st Qu.:-0.328   1st Qu.:1
##  Median :150   Median :1.0   Median :2   Median :-0.015   Median :2
##  Mean   :150   Mean   :0.5   Mean   :2   Mean   :-0.001   Mean   :2
##  3rd Qu.:225   3rd Qu.:1.0   3rd Qu.:3   3rd Qu.: 0.335   3rd Qu.:3
##  Max.   :300   Max.   :1.0   Max.   :3   Max.   : 1.855   Max.   :3
##                NA's   :213                                NA's   :105
```

There are 300 subjects repeatedly measured at three time occasions. Both the ordinal outcome Y (measured with three levels) and the binary time-varying covariate X can be missing according to a MAR mechanism. The continuous time-varying covariate Z is fully observed for all individuals.

4. Code usage

```
source("http://docs.ufpr.br/~jlpadilha/StatMod2017/AuxiliaryFunctions.R")#missing packages
       #necessary to run main functions are automatically installed
```

```
source("http://docs.ufpr.br/~jlpadilha/StatMod2017/wgee.r")
source("http://docs.ufpr.br/~jlpadilha/StatMod2017/drgee.r")
```

Examples on how to use these functions are given below.

Examples

```
aux.weight=gen.weight(datamiss)
```

In the object `aux.weight` are stored the quantities derived from the weight model that will be used in both WGEE and DRGEE. In this example, the auxiliary models for calculation of the probability of data to be missing are correctly specified. These estimated probabilities of observing the potentially missing data are used in the construction of the Δ matrix that appears in both WGEE and DRGEE estimators.

The calculation of quantities for the conditional expectation in DRGEE are done with the following function.

```
aux.cond=gen.cond(datamiss)
```

In this example, the auxiliary models are correctly specified. Both functions (`gen.weight` and `gen.cond`) should be run only once before the main functions `wgee` and `drgee` are called. Then, based on the quantities estimated by these auxiliary functions and the specified association structure, the `wgee` or `drgee` functions can be used to obtain robust estimates of regression parameters, as follows.

```
start.beta=c(-0.4,1.2,-0.35,0.35)#Starting values for the regression parameter
kmax=15
eps=1e-5
```

In this case, the GEE is treated to have converged when the relative change in consecutive iterations is smaller than 10^{-5} with a maximum number of iterations equal to 15.

Now we call the main functions `wgee` and `drgee`. We illustrate code usage by fitting models with an *exchangeable* correlation structure and a *uniform* local odds structure. Different association structures can be fitted by changing the `corstr` or `LORstr` arguments in the examples below.

```
#WGEE with exchangeable correlation structure
t1=proc.time()
w.exch=wgee(ini=start.beta, eps=eps, data=datamiss, assoc.type="corr", corstr="exch",
            kmax=kmax)
runtime.w.exch=proc.time() - t1
```

```
#WGEE with uniform local odds ratio association structure
t2=proc.time()
w.unif=wgee(ini=start.beta, eps=eps, data=datamiss, assoc.type="odds", LORstr="uniform",
            kmax=kmax)
runtime.w.unif=proc.time() - t2
```

```
#DRGEE with exchangeable correlation structure
t3=proc.time()
dr.exch=drgee(ini=start.beta, eps=eps, data=datamiss, assoc.type="corr", corstr="exch",
              kmax=kmax)
runtime.dr.exch=proc.time() - t3
```

```
#DRGEE with uniform local odds ratio association structure
t4=proc.time()
dr.unif=drgee(ini=start.beta, eps=eps, data=datamiss, assoc.type="odds", LORstr="uniform",
              kmax=kmax)
```

```
runtime.dr.unif=proc.time() - t4
```

Results: WGEE

- *Exchangeable*

```
w.exch
```

```
## $est
##     beta01     beta02          X          Z
## -0.4394078  1.0774618 -0.2427195  0.2712595
##
## $std
##    beta01    beta02         X         Z
## 0.1705945 0.1642297 0.2279735 0.1155450
##
## $k
## [1] 6
##
## $rho.est
##            [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
## [1,]  0.0000000  0.0000000  0.4934412 -0.1448729  0.4934412 -0.1448729
## [2,]  0.0000000  0.0000000 -0.1448729  0.2157236 -0.1448729  0.2157236
## [3,]  0.4934412 -0.1448729  0.0000000  0.0000000  0.4934412 -0.1448729
## [4,] -0.1448729  0.2157236  0.0000000  0.0000000 -0.1448729  0.2157236
## [5,]  0.4934412 -0.1448729  0.4934412 -0.1448729  0.0000000  0.0000000
## [6,] -0.1448729  0.2157236 -0.1448729  0.2157236  0.0000000  0.0000000
```

- *Uniform*

```
w.unif
```

```
## $est
##     beta01     beta02          X          Z
## -0.4404422  1.0770361 -0.2464412  0.2962093
##
## $std
##    beta01    beta02         X         Z
## 0.1714542 0.1648113 0.2291395 0.1178783
##
## $k
## [1] 5
##
## $odds.est
##         [,1]    [,2]    [,3]    [,4]    [,5]    [,6]
## [1,] 0.00000 0.00000 3.57309 3.57309 3.57309 3.57309
## [2,] 0.00000 0.00000 3.57309 3.57309 3.57309 3.57309
## [3,] 3.57309 3.57309 0.00000 0.00000 3.57309 3.57309
## [4,] 3.57309 3.57309 0.00000 0.00000 3.57309 3.57309
## [5,] 3.57309 3.57309 3.57309 3.57309 0.00000 0.00000
## [6,] 3.57309 3.57309 3.57309 3.57309 0.00000 0.00000
```

As one would expect, there are no marked differences in regression parameter estimates between the two association structures. Estimates of standard errors, derived from the sandwich estimator, are also similar. The exchangeable correlation structure parametrization took $k = 6$ iterations to converge while the uniform

local odds structure took $k = 5$ steps. Estimates of the association among the repeated measures are given by `rho.est` for the correlation parametrization (three parameters are estimated) and `odds.est` for the local odds parametrization (a single parameter is estimated).

Results: DRGEE

- *Exchangeable*

```
dr.exch
```

```
## $est
##     beta01     beta02          X          Z
## -0.4152036  1.1175287 -0.2828474  0.3754880
##
## $std
##    beta01    beta02         X         Z
## 0.1650857 0.1628801 0.2355946 0.1084160
##
## $k
## [1] 5
##
## $rho.est
##            [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
## [1,]  0.0000000  0.0000000  0.4909381 -0.1399571  0.4909381 -0.1399571
## [2,]  0.0000000  0.0000000 -0.1399571  0.2113006 -0.1399571  0.2113006
## [3,]  0.4909381 -0.1399571  0.0000000  0.0000000  0.4909381 -0.1399571
## [4,] -0.1399571  0.2113006  0.0000000  0.0000000 -0.1399571  0.2113006
## [5,]  0.4909381 -0.1399571  0.4909381 -0.1399571  0.0000000  0.0000000
## [6,] -0.1399571  0.2113006 -0.1399571  0.2113006  0.0000000  0.0000000
```

- *Uniform*

```
dr.unif
```

```
## $est
##     beta01     beta02          X          Z
## -0.4174100  1.1077191 -0.2854748  0.3905574
##
## $std
##    beta01    beta02         X         Z
## 0.1639718 0.1614540 0.2352686 0.1088456
##
## $k
## [1] 5
##
## $odds.est
##         [,1]    [,2]    [,3]    [,4]    [,5]    [,6]
## [1,] 0.00000 0.00000 3.57309 3.57309 3.57309 3.57309
## [2,] 0.00000 0.00000 3.57309 3.57309 3.57309 3.57309
## [3,] 3.57309 3.57309 0.00000 0.00000 3.57309 3.57309
## [4,] 3.57309 3.57309 0.00000 0.00000 3.57309 3.57309
## [5,] 3.57309 3.57309 3.57309 3.57309 0.00000 0.00000
## [6,] 3.57309 3.57309 3.57309 3.57309 0.00000 0.00000
```

The two options of the DRGEE estimator present similar regression parameter estimates. The estimated empirical standard errors are also similar. Both converged in $k = 5$ steps.

Results: Execution time

```
runtime.w.exch
```

```
##    user  system elapsed
##    8.83    0.08    8.92
```

```
runtime.w.unif
```

```
##    user  system elapsed
##    2.14    0.11    2.27
```

```
runtime.dr.exch
```

```
##    user  system elapsed
##  123.54    1.31  125.20
```

```
runtime.dr.unif
```

```
##    user  system elapsed
##   27.28    1.34   28.68
```

The execution times for the local odds ratio parametrization is much smaller than those of the correlation coefficient. This is because estimates of correlation parameters are updated at each step of the iterative algorithm while estimates of local odds parameters are obtained once and do not chance at each step.

DRGEE naturally runs slower than WGEE, as it depends on an additional auxiliary model. Nevertheless, the former is preferred over the latter because of its extra robustness.