Fieger, Heumann, Kastner:

MAREG and WinMAREG

Sonderforschungsbereich 386, Paper 45 (1996)

Online unter: http://epub.ub.uni-muenchen.de/

MAREG and WinMAREG

Andreas Fieger* Christian Heumann† Christian Kastner‡

October 21, 1996

Abstract

This paper describes a software tool for marginal regression methods. MAREG currently handles binary, categorical and continious data with several link functions. Although intended for the analysis of correlated data, uncorrelated data can be analysed. We supplies two different approaches for these problems—Maximum Likelihood and GEE methods. Handling of missing data is also provided.

*andreas@stat.uni-muenchen.de
†chriss@stat.uni-muenchen.de
‡kchris@stat.uni-muenchen.de

1 Introduction

WinMAREG is a user interface for MAREG (**Ma**rginal **Reg**ression methods). MAREG is a program for estimating marginal regression models; it is currently available as DOS and Solaris 2.4 binary.

The latest versions can be obtained via anonymous ftp from `ftp.stat.uni-muenchen.de`. It is located in the directory `/pub/sfb386/c3/mareg`. The files are

- `dosmareg.zip`, DOS version of MAREG plus WinMAREG.
- `Solaris2.4_mareg.tar.Z`, Solaris 2.4 version of MAREG.

The first step to analyse your data is to open the data file in WinMAREG. Currently supported file formats are dBase and Paradox database tables (`.dbf` and `.db`). The data will then be displayed in a grid. Note that you can not edit (change) your original data in any way.

WinMAREG provides an easy to use interface to specify the model you want to analyse.

WinMAREG produces several files to control MAREG. The `.cai` file gives MAREG the information of the specified model. It contains information about the variables used in this model, the type of the specified model (estimation method, design, link function, ...), information of the size of clusters, etc. Its structure is described in the appendix. The other file, which is always produced by WinMAREG, is the data file (`.cad`), which contains only the data for the selected variables (automatically coded on request).

These files are used to control MAREG, which can then be run on the chosen platform. You might want to use the Solaris 2.4 version of MAREG for models that require a lot of computing time. Therefore, simply type `mareg <CAI-File>` at the system prompt. If you choose to use the DOS version, WinMAREG can automatically invoke MAREG and wait until the computations are completed and then display the results.

Through using WinMAREG it is no necessary to do the coding yourself. You only select the coding scheme for selected categorical variables, the coding will be performed by WinMAREG.

2 Methods available in MAREG

Generalized linear models (GLM's), the basis of marginal regression models, perform an extension of the linear model for any type of response. MAREG currently enables you to handle binary, categorical and continious data with several link functions. Although MAREG can be used for uncorrelated data, the main importance is the analysis of correlated data. MAREG supplies two different approaches for these problems—Maximum Likelihood (ML) (Fitzmaurice and Laird, 1993) and Generalized Estimation Equations (GEE) methods (Liang and Zeger, 1986).

Both approaches and their variants that are implemented are described in the following sections.

2.1 Generalized Estimating Equations

The Generalized Estimating Equations have first been introduced by Liang and Zeger (1986). These estimating equations are an extension of the Quasi-Score Equations in the Generalized Linear Model to the case of correlated response. The main idea here is to use a covariance matrix and to specify the covariance between two observations instead of using a diagonal variance matrix.

When using the Generalized Estimating Equations the marginal model and the association parameters are estimated separately. Liang, Zeger and Quaqish (1992) called this approach Generalized Estimating Equations of Order 1 (GEE1). Following the destinctions between GEE1 and GEE2 in Ziegler, Kastner, Grömping and Blettner (1996b), we only use the term GEE, as MAREG provides both methods.

Over the last ten years several extensions of these methods were proposed. MAREG makes the following GEE-procedures available:

Independence Estimator: Here, the correlation matrix for each cluster is the identity matrix, which means estimating the marginal model as an usual GLM. The only difference is the correction of the estimated variance of the parameters through the so called sandwich-form introduced by White (1982).

Method of Prentice: Prentice (1988) used a second estimating equation for the correlation-parameters. The association (in this case the correlation) is estimated with the sample-correlation and modelled through the inverse of Fisher's z. The association structure can be chosen as exchangeable, stationary, unspecified or user-defined. A detailed description of this approach is given in Prentice (1988) or Miller, Davis and Landis (1993).

Odds-ratio: Instead of the correlation, the odds-ratio is another measure for marginal pairwise association for categorical response. When the response has more than two categories the odds ratio is no longer explicitely defined. Global cross-ratios and local odds-ratios are possible alternatives. MAREG uses local odds-ratios in a second estimating equation to estimate the association. This method which has not yet been proposed in the literature is rather experimental.

The association structure can be chosen as exchangeable, stationary, unspecified or userdefined.

For each of the above methods, one of the following X-designs can be selected. MAREG supports fixed effects models, varying intercept effects, varying covariate effects, varying intercept and covariate effects as well as user defined X-designs. As MAREG automatically includes an intercept (threshold), a constant you may not specify as a covariable in the model (except for the user-defined design)!

2.2 Maximum Likelihood

The Maximum Likelihood method implements the estimation of the so called mixed parameter model, in which mean regression model is combined with a model for conditional associations. The binary case is described in Fitzmaurice and Laird (1993). The generaisation to multicatigorical response is described in Heumann (1996).

When using a full likelihood approach, usually all moments have to be specified. By default, MAREG only uses two-way interactions. However, three- and four-way interactions can be specified using a user-defined association structure.

The options for the X-design are identical to those available for the GEE-methods.

Another feature of the ML methods is the ability to group the data (identical clusters—with regard to the independent variables in the model—will be grouped). This makes computations considerably faster for datasets with only few strata. If an association model other than independence is chosen, data are grouped and a full table of observed and expected frequencies for all possible response profiles is given together with Deviance (G^2), χ^2 and a special Cressie–Read-statistic for the different strata. Thus, this option is only recommended for cases where measures are only repeated a few times (depending on the number of categories) to avoid extremely long output. If independence is chosen as association, data are grouped and Deviance (G^2) is given as a measure of Goodness of Fit. No further output is given. This enables the user to compute the likelihood and G^2 in cases of much repeated measures or of nonclustered data.

2.3 Handling of missing data

By default MAREG uses complete case methods to analyse the data. This strategy is clearly not efficient, and can lead to biased estimates if the data are not MCAR (Rubin, 1987). When using GEE, MAREG gives you the opportunity to use the method of inverse probability weighting, as proposed by Robins, Rotnitzky and Zhao (1995). The actual implementation is restricted to the case of longitudinal data with monotone missing pattern in the response variable. The chosen covariables for the drop-out model can not have missing values.

3 Using WinMAREG

This section gives a brief description of how to use WinMAREG.

3.1 Data structure

The data that are used by WinMAREG have to be arranged in a database table, with the criteria below:

- the columns of the table contain the variables
- if clustered data are used there is a (numerical) variable which identifies the clusters
- clusters have to be arranged in consecutive blocks
- if missing values are present, they can be coded either by a blank cell, or a special numerical code

The data that will be produced by WinMAREG is in a ASCII file. It contains only the variables you selected for your model. The first column is the cluster-ID variable, if it is specified (otherwise the number of row in the file, i.e. the casenumber). It is followed by the dependent variable(s) and the independent variables of the specified model.

3.2 Files

The following files will be produced if you click the Modelselection-Dialog's OK or Paste button

`.cad` This file contains the data of the selected variables

`.cai` This file contains the information concerning the model

`.cam` This file contains the data of the selected variables for the missing model

The following file has to be supplied by the user:

`.caz` Text-File with the information of the design matrix Z of the association model.

The following files will be produced when the computing program is run. Their names are specified in the `.cai` file.

`.cal` Log-file containing information concerning the run of the program

`.cao` Output-file containing the results of the calculations

3.3 Model selection

This dialog is the main dialog for WinMAREG. It enables you to select the variables in your model and its specifications. As a minimum requirement, you have to specify at least one dependent and one independent variable.

OK produces the data (`.cad`) and command (`.cai`) files, then runs MAREG for the selected model, and returns to WinMAREG.

Cancel discards your selections.

Help displays an online help page for this dialog.

Paste produces the data and command files, but does not run MAREG. (use this button if you want to produce the required files only, in order to run MAREG on another platform).

The following buttons can be used to specify optional choices for the model. If you don't set them default settings will be used.

general opt's specify general options, such as missing value ID. This dialog can also be opened via the options menu.

specific opt's specify specific options, such as the estimation method, design, etc.

cluster ID select the variable defining the clusters. In the datafile the clusters have to be arranged in consecutive blocks (however, the blocks don't have to be sorted).

coding select the coding scheme for selected categorical variables. This dialog can also be opened by right-clicking a selected variable in the list of dependent or independent variables.

missing model specify a regression model for the drop-out probability.

3.4 General options

This dialog gives the opportunity to specify general options for MAREG and WinMAREG. These are:

- Convergence

  maxIter maximal number of iterations.

  maxIter IPF maximal number of iterations for the IPF algorithm (only ML).

  eps epsilon for convergence criterion.

  eps IPF epsilon for the IPF algorithm (only ML).

- Missing values

  missing value ID value that marks a missing value in your data and the `.cad` file.

- General

  format format of numbers in the output (`.cao` file and console).

  issue warning if selected a warning will be issued, when a selected categorical variable you wanted to be coded has more than the given number of categories. (However, it will still be coded).

  display MV summary if checked, a short summary of the number of incomplete cases will be displayed.

  show CAO file if checked, the `.cao` file will be opened in a text window.

  prompt for consecutive clusters if checked, you will be prompted, whether the clusters in your data are in the required order (as a reminder) or not.

  prompt for missing cluster IDs if checked, you will be prompted if a cluster ID was not specified.

- Output

  output level level of the output in the `.cal` file, from 'quiet' (no output) to 'debug' (all available information).

  console out output to console while MAREG is computing (on or off).

4 Examples

In the following, data from the examples discussed in the literature are analysed using MAREG.

4.1 Ohio children data

The Ohio children data are a subset of the six-cities study, a longitudinal study of the health effects of air pollution. These data were anlysed in the literature e.g. by Liang and Zeger (1986), Fitzmaurice and Laird (1993) or Fahrmeir and Tutz (1994).

The following listings display the ML models as discussed in Fitzmaurice and Laird (1993).

4.1.1 ML, logistic regression, fixed effects, independence estimator

```
MAREG version 0.0.7. (c)23.09.1996 SFB386-C3. All rights reserved.
 ----- Likelihood estimator assuming independence, grouped data :
Inifile:                      ohio01.cai
Out file:                     ohio01.cao
Log file:                     ohio01.cal
Data file:                    ohio01.cad
Orig. sample size, #clusters: 2148, 537
Act. sample size,  #clusters: 2148, 537
Estimation method:            Maximum Likelihood
Special estimatior:           Conditional log odds ratios
Link:                         cumulative logit link
Variance function:            Binomial variance function
Design:                       Fixed effects model
User given epsilon:           1e-05
User given maxiter:           100
Tolerance beta reached:       6.03791e-07
Iterations needed:            5
Estimated overdispersion:     no overdispersion was estimated

     Var.    beta    std. r. std.       Z    r. Z       p    r. p
        y  -1.901   0.089   0.119 -21.420 -15.963   0.000   0.000
      age  -0.141   0.070   0.058  -2.032  -2.426   0.042   0.015
    smoke   0.314   0.139   0.188   2.252   1.671   0.024   0.095
 agesmoke   0.071   0.111   0.088   0.640   0.802   0.522   0.422

Likelihood: -909.740
Deviance: 241.084
```

4.1.2 ML, logistic regression, fixed effects, all twoway interactions assuming exchangability

```
MAREG version 0.0.7. (c)23.09.1996 SFB386-C3. All rights reserved.
 ----- LIKELIHOOD ESTIMATION WITH TABLE OUTPUT :
Inifile:                      ohio02.cai
Out file:                     ohio02.cao
Log file:                     ohio02.cal
Data file:                    ohio02.cad
Orig. sample size, #clusters: 2148, 537
Act. sample size,  #clusters: 2148, 537
Estimation method:            Maximum Likelihood
Special estimatior:           Conditional log odds ratios
Link:                         cumulative logit link
Variance function:            Binomial variance function
Design:                       Fixed effects model
User given epsilon:           1e-05
User given maxiter:           100
Tolerance beta reached:       9.35422e-07
Iterations needed:            4
Estimated overdispersion:     no overdispersion was estimated
```

```
     Var.    beta    std. r. std.       Z    r. Z       p    r. p
        y  -1.901   0.117   0.119 -16.193 -15.962   0.000   0.000
      age  -0.141   0.057   0.058  -2.480  -2.424   0.013   0.015
    smoke   0.314   0.188   0.188   1.667   1.671   0.095   0.095
 agesmoke   0.071   0.089   0.088   0.799   0.801   0.424   0.423

Association model:       All two factor association model assuming exchangeabil
                         ity
Tolerance alpha reached: 2.20997e-06

   alpha    std. r. std.       Z    r. Z       p    r. p
   1.266   0.073   0.073  17.406  17.440   0.000   0.000

Likelihood: -797.391
Deviance: 16.387
Chisquare:17.843
Cressie-Read (lambda=2/3): 17.215
```

4.1.3 ML, logistic regression, fixed effects, all twoway interactions

```
MAREG version 0.0.7. (c)23.09.1996 SFB386-C3. All rights reserved.
 ----- LIKELIHOOD ESTIMATION WITH TABLE OUTPUT :
Inifile:                       ohio03.cai
Out file:                      ohio03.cao
Log file:                      ohio03.cal
Data file:                     ohio03.cad
Orig. sample size, #clusters: 2148, 537
Act. sample size,  #clusters: 2148, 537
Estimation method:             Maximum Likelihood
Special estimatior:            Conditional log odds ratios
Link:                          cumulative logit link
Variance function:             Binomial variance function
Design:                        Fixed effects model
User given epsilon:            1e-05
User given maxiter:            100
Tolerance beta reached:        2.84889e-06
Iterations needed:             6
Estimated overdispersion:      no overdispersion was estimated

     Var.    beta    std. r. std.       Z    r. Z       p    r. p
        y  -1.906   0.118   0.119 -16.126 -16.005   0.000   0.000
      age  -0.143   0.059   0.058  -2.433  -2.453   0.015   0.014
    smoke   0.305   0.190   0.188   1.610   1.622   0.107   0.105
 agesmoke   0.069   0.092   0.089   0.755   0.781   0.450   0.435

Association model:       All two factor association model
Tolerance alpha reached: 6.82441e-06

   alpha    std. r. std.       Z    r. Z       p    r. p
   1.332   0.305   0.310   4.368   4.299   0.000   0.000
   0.753   0.332   0.349   2.267   2.156   0.023   0.031
   1.293   0.318   0.309   4.063   4.191   0.000   0.000
   1.910   0.313   0.330   6.110   5.781   0.000   0.000
   0.856   0.349   0.347   2.453   2.469   0.014   0.014
   1.512   0.338   0.343   4.468   4.405   0.000   0.000

Likelihood: -793.975
Deviance: 9.555
Chisquare:9.505
Cressie-Read (lambda=2/3): 9.489
```

4.2 Caesarean birth study

These data were analysed in Fahrmeir and Tutz (1994). They come from a study on infection from birth by caesarean section.

4.2.1 GEE, multinomial logit model, fixed effects, independence estimator

```
MAREG version 0.0.7. (c)23.09.1996 SFB386-C3. All rights reserved.
 ----- GEE independence estimator :
Inifile:                      caesar01.cai
Out file:                     caesar01.cao
Log file:                     caesar01.cal
Data file:                    caesar01.cad
Orig. sample size, #clusters: 251, 251
Act. sample size,  #clusters: 251, 251
Estimation method:            Gee 1
Special estimatior:           Independence estimator (IEE)
Link:                         multinomial logit link
Variance function:            Multinomial variance function
Design:                       Fixed effects model
User given epsilon:           1e-05
User given maxiter:           100
Tolerance beta reached:       3.77468e-10
Iterations needed:            6
Estimated overdispersion:     no overdispersion was estimated

         Var.    beta    std. r. std.       Z    r. Z       p    r. p
    infect(1)  -2.621   0.557   0.616  -4.708  -4.257   0.000   0.000
  antib on (1) -3.520   0.672   0.626  -5.240  -5.624   0.000   0.000
 factor on (1)  1.829   0.602   0.620   3.037   2.952   0.002   0.003
 noplan on (1)  1.174   0.521   0.467   2.253   2.514   0.024   0.012
    infect(2)  -2.560   0.546   0.590  -4.686  -4.338   0.000   0.000
  antib on (2) -3.087   0.550   0.539  -5.614  -5.730   0.000   0.000
 factor on (2)  2.195   0.587   0.601   3.740   3.654   0.000   0.000
 noplan on (2)  0.996   0.481   0.455   2.069   2.189   0.039   0.029
```

4.2.2 ML, multinomial logit model, fixed effects, independence estimator

```
MAREG version 0.0.7. (c)23.09.1996 SFB386-C3. All rights reserved.
 ----- Likelihood estimator assuming independence, grouped data :
Inifile:                      caesar02.cai
Out file:                     caesar02.cao
Log file:                     caesar02.cal
Data file:                    caesar02.cad
Orig. sample size, #clusters: 251, 251
Act. sample size,  #clusters: 251, 251
Estimation method:            Maximum Likelihood
Special estimatior:           Conditional log odds ratios
Link:                         multinomial logit link
Variance function:            Multinomial variance function
Design:                       Fixed effects model
User given epsilon:           1e-05
User given maxiter:           100
Tolerance beta reached:       3.77468e-10
Iterations needed:            6
Estimated overdispersion:     no overdispersion was estimated

         Var.    beta    std. r. std.       Z    r. Z       p    r. p
    infect(1)  -2.621   0.557   0.616  -4.708  -4.257   0.000   0.000
  antib on (1) -3.520   0.672   0.626  -5.240  -5.624   0.000   0.000
 factor on (1)  1.829   0.602   0.620   3.037   2.952   0.002   0.003
 noplan on (1)  1.174   0.521   0.467   2.253   2.514   0.024   0.012
    infect(2)  -2.560   0.546   0.590  -4.686  -4.338   0.000   0.000
  antib on (2) -3.087   0.550   0.539  -5.614  -5.730   0.000   0.000
 factor on (2)  2.195   0.587   0.601   3.740   3.654   0.000   0.000
 noplan on (2)  0.996   0.481   0.455   2.069   2.189   0.039   0.029

Likelihood: -160.937
Deviance: 11.830
```

4.3 Respiratory disorder data

This study described in Miller et al. (1993) is a randomised clinical trial of a new treatment of respiratory disorder.

4.3.1 GEE, cumulative logit model, varying intercept and covariate effects, independence estimator

```
MAREG version 0.0.7. (c)23.09.1996 SFB386-C3. All rights reserved.
 ----- GEE independence estimator :
Inifile:                      miller01.cai
Out file:                     miller01.cao
Log file:                     miller01.cal
Data file:                    miller01.cad
Orig. sample size, #clusters: 444, 111
Act. sample size,  #clusters: 444, 111
Estimation method:            Gee 1
Special estimatior:           Independence estimator (IEE)
Link:                         cumulative logit link
Variance function:            Multinomial variance function
Design:                       Varying intercept and covariate effects
User given epsilon:           1e-05
User given maxiter:           100
Tolerance beta reached:       7.02334e-06
Iterations needed:            5
Estimated overdispersion:     no overdispersion was estimated

      Var.    beta    std. r. std.       Z    r. Z       p    r. p
 resp(1)_1   -1.95    0.29    0.28   -6.80   -6.96    0.00    0.00
 resp(2)_1    0.83    0.21    0.21    3.99    3.95    0.00    0.00
   treat_1   -0.22    0.19    0.19   -1.19   -1.20    0.23    0.23
 resp(1)_2   -1.70    0.26    0.26   -6.45   -6.53    0.00    0.00
 resp(2)_2    0.73    0.21    0.22    3.40    3.37    0.00    0.00
   treat_2   -0.74    0.20    0.19   -3.78   -3.80    0.00    0.00
 resp(1)_3   -1.48    0.25    0.24   -6.02   -6.26    0.00    0.00
 resp(2)_3    0.61    0.20    0.21    3.01    2.94    0.00    0.00
   treat_3   -0.53    0.19    0.18   -2.85   -2.88    0.00    0.00
 resp(1)_4   -1.33    0.23    0.24   -5.68   -5.64    0.00    0.00
 resp(2)_4    0.58    0.20    0.20    2.89    2.90    0.00    0.00
   treat_4   -0.33    0.18    0.18   -1.86   -1.85    0.06    0.06
```

4.3.2 GEE, cumulative logit model, varying intercept and covariate effects, Method of Prentice, userdefined association (saturated model)

```
MAREG version 0.0.7. (c)23.09.1996 SFB386-C3. All rights reserved.
 ----- GEE Prentice estimator :
Inifile:                      miller02.cai
Out file:                     miller02.cao
Log file:                     miller02.cal
Data file:                    miller02.cad
Orig. sample size, #clusters: 444, 111
Act. sample size,  #clusters: 444, 111
Estimation method:            Gee 1
Special estimatior:           Method of Prentice
Link:                         cumulative logit link
Variance function:            Multinomial variance function
Design:                       Varying intercept and covariate effects
User given epsilon:           1e-05
User given maxiter:           100
Tolerance beta reached:       2.61494e-07
Iterations needed:            7
Estimated overdispersion:     no overdispersion was estimated

      Var.    beta    std. r. std.       Z    r. Z       p    r. p
 resp(1)_1   -1.95    0.28    0.28   -6.87   -7.09    0.00    0.00
 resp(2)_1    0.85    0.20    0.20    4.20    4.14    0.00    0.00
   treat_1   -0.24    0.19    0.19   -1.29   -1.29    0.20    0.20
 resp(1)_2   -1.72    0.26    0.26   -6.57   -6.62    0.00    0.00
 resp(2)_2    0.71    0.21    0.21    3.36    3.34    0.00    0.00
```

```
   treat_2   -0.74    0.19    0.19   -3.84   -3.86    0.00    0.00
 resp(1)_3   -1.45    0.24    0.23   -6.08   -6.36    0.00    0.00
 resp(2)_3    0.61    0.20    0.20    3.01    2.95    0.00    0.00
   treat_3   -0.54    0.18    0.18   -2.94   -2.98    0.00    0.00
 resp(1)_4   -1.35    0.23    0.23   -5.84   -5.78    0.00    0.00
 resp(2)_4    0.59    0.20    0.20    2.98    2.98    0.00    0.00
   treat_4   -0.32    0.18    0.18   -1.82   -1.80    0.07    0.07

Association model:            Userdefined correlation model, method of Prentic
                              e
Tolerance alpha reached:      5.40864e-06
Score equation for alpha uses: Identity matrix

   alpha    std. r. std.       Z    r. Z       p    r. p
    0.83    0.32    0.55    2.58    1.50    0.01    0.13
    0.07    0.27    0.24    0.26    0.29    0.80    0.77
   -0.39    0.28    0.27   -1.37   -1.44    0.17    0.15
    0.87    0.33    0.28    2.65    3.06    0.01    0.00
    0.65    0.30    0.43    2.16    1.53    0.03    0.13
   -0.21    0.28    0.22   -0.78   -0.97    0.44    0.33
   -0.01    0.27    0.24   -0.03   -0.03    0.98    0.98
    0.45    0.29    0.28    1.57    1.61    0.12    0.11
    0.83    0.32    0.44    2.58    1.87    0.01    0.06
   -0.39    0.28    0.16   -1.37   -2.42    0.17    0.02
   -0.06    0.27    0.28   -0.23   -0.23    0.82    0.82
    0.38    0.28    0.27    1.36    1.43    0.17    0.15
    0.27    0.28    0.35    0.98    0.76    0.33    0.44
    0.06    0.27    0.28    0.22    0.22    0.83    0.83
    0.43    0.29    0.26    1.52    1.67    0.13    0.10
    0.96    0.34    0.33    2.83    2.96    0.00    0.00
    0.51    0.29    0.42    1.77    1.22    0.08    0.22
    0.13    0.27    0.27    0.46    0.46    0.64    0.64
    0.68    0.30    0.31    2.23    2.22    0.03    0.03
    0.32    0.28    0.28    1.13    1.13    0.26    0.26
    0.70    0.31    0.41    2.28    1.70    0.02    0.09
    0.02    0.27    0.24    0.07    0.08    0.94    0.93
    0.43    0.28    0.30    1.50    1.41    0.13    0.16
    0.25    0.28    0.28    0.91    0.89    0.37    0.37
    1.29    0.39    0.43    3.30    3.03    0.00    0.00
   -0.66    0.30    0.28   -2.25   -2.41    0.02    0.02
   -0.28    0.27    0.28   -1.03   -1.01    0.30    0.31
    0.70    0.30    0.29    2.35    2.43    0.02    0.02
    0.68    0.30    0.35    2.29    1.97    0.02    0.05
   -0.08    0.27    0.28   -0.32   -0.30    0.75    0.76
   -0.04    0.27    0.27   -0.17   -0.16    0.87    0.87
    0.83    0.31    0.29    2.66    2.89    0.01    0.00
    0.78    0.31    0.36    2.54    2.16    0.01    0.03
   -0.11    0.27    0.28   -0.41   -0.39    0.68    0.70
   -0.19    0.27    0.27   -0.70   -0.70    0.48    0.49
    0.57    0.29    0.28    1.98    2.04    0.05    0.04
    1.48    0.44    0.42    3.38    3.57    0.00    0.00
   -0.49    0.28    0.26   -1.74   -1.89    0.08    0.06
   -0.64    0.29    0.28   -2.17   -2.29    0.03    0.02
    0.88    0.32    0.29    2.75    3.06    0.01    0.00
    1.41    0.42    0.41    3.36    3.47    0.00    0.00
   -0.37    0.27    0.26   -1.34   -1.39    0.18    0.16
   -0.64    0.29    0.27   -2.19   -2.36    0.03    0.02
    0.79    0.31    0.28    2.56    2.78    0.01    0.01
    1.95    0.61    0.51    3.21    3.82    0.00    0.00
   -0.60    0.29    0.26   -2.08   -2.29    0.04    0.02
   -0.77    0.31    0.24   -2.50   -3.14    0.01    0.00
    1.20    0.37    0.31    3.22    3.81    0.00    0.00
```

Acknowledgements

The authors wish to thank Kurt Watzka for supplying a c++ matrix template library.

Bug reports

As the current version is a beta version, it may contain bugs. Please report any bugs or suggestions to one of the autors.

A User defined association

If a user defined association was chosen, it is necessary to provide a user defined text-file containing the design matrix of each cluster for the association model. If the design matrices for the clusters are equal, this matrix has to be specified only once.

In front of each matrix in the Z-file, two numbers have to be specified: the first number is the number of rows (number of cases in the association model), the second is the number of columns (number of association parameters). The design matrix has to be specified row-wise.

Example:

```
6 3
1 0 0
0 1 0
0 1 0
0 1 0
0 1 0
0 0 1
3 3
1 0 0
0 1 0
1 0 0
 ...
```

B Technical details

Running the DOS application

If you select a model and click the OK button a DOS application will be run. A red light in the status-bar indicates that the external program is runing. The GEE and ML menus will not be enabled until the external program is terminated. Its results can then automatically be displayed in an editor window.

Installing the 32bit DOS extender

- copy the files
  - `32rtm.exe`
  - `32stub.exe`
  - `dpmi32vm.ovl`

  to the directory containing the MAREG program files (`mareg.exe` and `mareg.pif`)
- Install `windpmi.386`: Edit your your `system.ini` file. In the the section `[386Enh]` add the entry

```
[386Enh]
device=<path>\WINDPMI.386
```

where <path> specifies the path to the file WINDPMI.386.
Example:

```
[386Enh]
device=c:\windows\system\WINDPMI.386
```

- Add the section

```
[exe]
dosexe=<path>\mareg.pif
```

to your winmareg.ini file, where <path> specifies the path to the mareg program files.
Example:

```
[exe]
dosexe=c:\mareg\mareg.pif
```

- Edit the file mareg.pif[1], to set the path to mareg.exe

Installing the database drivers

This section provides information on 'Borland Database Engine'. To install the Borland Data Base Engine

- unzip the files dbedisk1 and dbedisk2
- run the installation program SETUP.EXE and follow the instructions.

For further information see the file instdb.txt which is included in your WinMAREG distribution.

C Possible values for .cai files

This section explains valid values for the items of the sections in the .cai files for mareg. Expressions in brackets '<>' explain the expected type of value; '//' starts a comment (comments are not allowed in .cai files that are used with mareg, they are only used here for documentation).

```
[filenames]    // all filenames without path!
cao=<filename> // filename for the output file
cad=<filename> // filename for the data file
cai=<filename> // filename of this file
cal=<filename> // filename for the log file
cam=<filename> // filename for the data file for the missing model
caz=<filename> // filename for the data file for the Z-matrix

[sizes]
samplesize=<integer value>   // number of rows in the original data file
clusters=<integer value>     // number of clusters in the .cad file
clustersize=<integer value>  // size of the clusters;
                             // varing clustersizes are specified by the value 0
                             // and an additional section [clustersizes]
                             // will be created
```

[1]Program Information File: Windows specific file, use pifedit.exe included in your Windows distribution to edit these files

```
cadrows=<integer value>      // number of rows in the .cad file
cadcols=<integer value>      // number of columns in the .cad file
                             // which is not identical to the number of variables
                             // in the model, when variables are coded (effect or dummy)

[varnames]
clusterid0=<string>          // name of the variable identifying the clusters
resp0=<string>               // name of the dependent variable
covar0=<string>              // name of the first independent variable
covar1=<string>              // name of the second independent variable
...
covar<number of covariables-1>=<string>  // name of the last independent variable

[estimator]
general=<integer value> // 1=GEE
                        // 2=ML
special=<integer value> // 1=independence estimator
                        // 2=method of Prentice
                        // 3=local odds ratio method
                        // 4=conditional log odds ratios


[design]
general=<integer value> // 1=continuous, binary and ordinal response
                        // (using cumulative logit model)
                        // 2=nominal response (multinomial logit model)
                        // 3=userdefined
special=<integer value> // 1=fixed effects model
                        // 2=varying intercept effects
                        // 3=varying covariate effects
                        // 4=varying intercept and covariate effects
                        // 5=userdefined

[link]
link=<integer value>    // 1=identity link
                        // 2=cumulative logit link
                        // 3=multinomial logit link

[variancefunction]
variancefunction=<integer value>  // normal variance function
                                  // binomial variance function
                                  // multinomial variance function
[association]
model=<integer value> // depends on the chosen estimator:
                      //
                      // for GEE independence:
                      // 1=independence model
                      //
                      // for GEE Prentice and GEE local odds ratios:
                      // 1=exchangable
                      // 2=stationary, assuming equidistant timepoints
                      // 3=unspecified
                      // 4=userdefined
                      //
                      // for ML:
                      // 1=independence model
                      // 2=all two factor
                      // 3=all two factor, assuming exchangability
                      // 4=all two factor, assuming stationarity
                      // 5=userdefined


type=<integer value>  // Z-file type (only if a z-file is specified)
                      // 1=constant
                      // 2=clusterspecific

[missing]
value=<integer value> // this value will be interpreted as a missing value
                      // in your (input) data file
                      // in the .cad file this value represents a missing value

[columns]                              // columns in the .cad file
resp0=<first column> <last column>     // if variables are not coded, first column
covar0=<first column> <last column>    // and last column are identical
```

```
covar1=<first column> <last column>    // else first column is the first column in the
...                                    // .cad file, that contains the first dummy
covar<number of covariables-1>=<first column> <last column>
                                       // for the coded variable, last colum is equal to
                                       // first column + number of categories - 2,
                                       // a threecategorical variable will have two
                                       // dummys in columns 'first column' and
                                       // 'first column+1'

[constants]
eps=<real value>            // convergence bound
epsipf=<real value>         // same as eps (for IPF algorithm)
maxiter=<integer value>     // maximum number of iterations
maxiteripf=<integer value>  // same as maxiter (for IPF algorithm)

[options]
width=<integer value>           // width and precision specify the format
precision=<integer value>       // of the output file (.cao)
loglevel=<integer value>        // 0=quiet, 1=low, 2=medium, 3=high, 4=debug
consoleout=<integer value>      // 0=off, 1=on
likelihood=<integer value>      // 0=nogrouping, 1=grouping and tableoutput
prenticeui=<integer value>      // 0=identity matrix, 1=miller et al.
overdispersion=<integer value>  // 0=no estimation, 1=estimation

[interactionmodel]
twoway=<Number>: <timepoint timepoint>, ..., <timepoint timepoint>   //
threeway=<Number>: <timepoint timepoint timepoint>, ..., <timepoint timepoint timepoint> //
fourway=<Number>: <timepoint timepoint timepoint timepoint>, ...,
       <timepoint timepoint timepoint timepoint>  //


[clustersizes]                   // if clustersizes vary, thsi section has to be created
n0=<integer value>               // and clustersize=0 has to be specified in the
n1=<integer value>               // [sizes] section above
                                 // n0 is the size of the first cluster
                                 // n1 is the size of the second cluster
...
n<number of clusters-1>=<integer value>  // size of the last cluster

[clusterid]                // original values of the variable that identifies the clusters.
cluster0=<integer value>   // the first cluster appearing in your original data file will
cluster1=<integer value>   // be regarded as cluster0, regardeless of the value of the
...                        // variable identifying the clusters
cluster<number of clusters-1>=<integer value>
                        // This information is currently not used by mareg, however it will
                        // be produced by winmareg for later reference

// the following sections are analogous to the above sections:
[missing_varnames]  // structure as [varnames]
covar0=<string>
...
covar<number of covariables-1>=<string>

[missing_lags]
lag0=<integer value>                          // lag for covar0
...
lag<number of covariables-1>=<integer value>  // lag for covar<number of covariables-1>

[missing_columns]  // structure as [columns]
covar0=<first column> <last column>

[missing_sizes]  // structure as [sizes]
camcols=<integer value>

[missing_link] // structure as [link]
link=<integer value>

[missing_variancefunction] // structure as [variancefunction]
variancefunction=<integer value>

[missing_design] // structure as [design]
general=<integer value>
special=<integer value>
```

D Licensing agreement

The authors of this software grant to any individual or non-commercial organization the right to use and to make an unlimited number of copies of this software. Usage by commercial entities requires a license from the authors. You may not decompile, disassemble, reverse engineer, or modify the software. This includes, but is not limited to modifying/changing any icons, menus, or displays associated with the software. This software cannot be sold without written authorization from the author. This restriction is not intended to apply for connect time charges, or flat rate connection/download fees for electronic bulletin board services. The authors of this program accept no responsibility for damages resulting from the use of this software and make no warranty or representation, either express or implied, including but not limited to, any implied warranty of merchantability or fitness for a particular purpose. This software is provided as is, and you, its user, assume all risks when using it.

References

Diggle, P. J. and Kenward, M. G. (1994). Informative dropout in longitudinal data analysis, *Applied Statistics* **43**: 49–94.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*, Springer, New York.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses, *Biometrika* **80**: 141–151.

Fitzmaurice, G. M., Laird, N. M. and Lipsitz, S. R. (1994). Analysing incomplete longitudinal binary responses: A likelihood-based approach, *Biometrics* **50**: 601–612.

Fitzmaurice, G. M., Laird, N. M. and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses, *Statistical Science* **8**: 284–309.

Heumann, C. (1996). Marginal regression modeling of correlated multicategorical response: A likelihood approach, *SFB386 – Discussion Paper 19*, Universität München.

Liang, K.-Y. and Hanfelt, J. J. (1994). On the use of the quasi-likelihood method in teratological experiments, *Biometrics* **50**: 872–880.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**: 13–22.

Liang, K.-Y. and Zeger, S. L. (1993). Regression analysis for correlated data, *Annual Review of Public Health* **14**: 43–68.

Liang, K.-Y., Zeger, S. L. and Quaqish, B. (1992). Multivariate regression analysis for categorical data, *Journal of the Royal Statistical Society, Series B* **54**: 3–40.

Lipsitz, S. R., Fitzmaurice, G. M., Orav, E. J. and Laird, N. M. (1994). Performance of generalized estimating equations in practical situations, *Biometrics* **50**: 270–278.

Lipsitz, S. R., Kim, K. and Zhao, L. P. (1994). Analysis of repeated categorical data using generalized estimating equations, *Statistics in Medicine* **13**: 1149–1163.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association, *Biometrika* **78**: 153–160.

McDonald, B. W. (1993). Estimating logistic regression parameters for bivariate binary data, *Journal of the Royal Statistical Society, Series B* **55**: 391–397.

Miller, M. E. (1995). Analysing categorical responses obtained from large clusters, *Applied Statistics* **44**: 173–186.

Miller, M. E., Davis, C. S. and Landis, R. J. (1993). The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares, *Biometrics* **49**: 1033–1044.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**: 1033–1048.

Prentice, R. L. and Zhao, L. P. (1991). Estimation equations for parameters in means and covariance of multivariate discrete and continous responses, *Biometrics* **47**: 825–839.

Quaqish, B. and Liang, K.-Y. (1992). Marginal models for correlated binary response with multiple classes an multiple levels of nesting, *Biometrics* **49**: 939–950.

Robins, J. M. and Rotnitzky, A. G. (1995). Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association* **90**: 122–129.

Robins, J. M., Rotnitzky, A. G. and Zhao, L. P. (1995). Analysis of semiparametric regression models repeated outcomes in the presence of missing data, *Journal of the American Statistical Association* **90**: 106–120.

Rotnitzky, A. G. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data, *Biometrika* **77**: 485–497.

Rotnitzky, A. G. and Robins, J. M. (1995). Semiparametric estimation of models for means and covariances in the presence of missing data, *Scandinavian Journal of Statistics* **22**: 323–333.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, Wiley, New York.

White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**: 1–25.

Zeger, S. L. (1988). Commentary, *Statistics in Medicine* **7**: 161–168.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**: 121–130.

Zeger, S. L., Liang, K.-Y. and Self, S. G. (1985). The analysis of binary longitudinal data with time-independent covariates, *Biometrika* **72**: 31–38.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a generalized quadratic model, *Biometrika* **77**: 642–648.

Zhao, L. P. and Prentice, R. L. (1991). Use of a quadratic exponential model to generate estimating equations for means, variances, and covariances, *in* V. P. Godambe (ed.), *Estimating Functions*, University Press, Oxford.

Zhao, L. P., Prentice, R. L. and Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model, *Journal of the Royal Statistical Society, Series B* **54**: 805–811.

Ziegler, A., Kastner, C., Grömping, U. and Blettner, M. (1996a). Die Generalized Estimating Equations: Herleitung und Anwendung, *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **27**: 69–91.

Ziegler, A., Kastner, C., Grömping, U. and Blettner, M. (1996b). The generalized estimation equations in the past ten years: An overview and a biomedical application, *SFB386 – Discussion Paper 24*, Universität München.