Package ‘mldm’

April 21, 2017

Type Package

Title A Multivariate Logistic Distance Model in R

Version 1.0

Date 2015-05-11

Author Hailemichael M. Worku and Mark de Rooij

Maintainer Hailemichael M Worku <hmetiku@yahoo.com>

Description Multivariate binary data, in which there are multiple binary outcome variables and one or more predictor variable(s), are often collected in emperical sciences such as psychology, medicine, criminiology, epidemiology, and life sciences. The mldm package fits a logistic distance models on such data. The model is a family of marginal models (like the standard and extended GEE methods). Dimension reduction as a form of regularization, is possible in the distance model. Such a reduction makes the distance model more preferred compared to the existing marginal models since less parameters need to be estimated. The distance model can be used for comparing factorial structure of the data.

License GPL (>= 2)

Depends R (>= 2.10.1), geepack, Formula, MASS

NeedsCompilation no

R topics documented:

`mldm-package` *Multivariate Logistic Distance Model for analyzing clustered binary data in R*

Description

This package fits a Multivariate Logistic Distance (MLD) model proposed for analyzing multivariate binary data. The MLD model extends the marginal model for clustered binary data within a distance framework. The results of the MLD model can be graphically represented in a biplot, which enhance interpretation of the model.

Details

Package:	mldm
Type:	Package
Version:	1.0
Date:	2015-05-20
License:	GPL

The distance model (De Rooij, 2009) for a single binary response variable, y, is defined by

$$p(y_i = 1) = \frac{e^(-0.5\delta_{i1})}{e^(-0.5\delta_{i0}) + e^(-0.5\delta_{i1})}$$

, where $\delta_.$ is the squared Euclidean distance in one dimensional space between a point representing the subject and a point representing the categories, i.e., $\delta_{i0} = (\eta_i - \gamma_0)^2$ and $\delta_{i1} = (\eta_i - \gamma_1)^2$ in which γ_1 and γ_0 represent coordinate of the two categories. The subject point, η, is defined as linear combinations of the predictor variable, i.e., $\eta_i = \beta_0 + X_i\beta_1$.

The package has the capability of handling both clustered bootstrap method and sandwich estimators, for adjusting the standard errors of model parameters. The most important functions are shown below under the **See also**.

Author(s)

Hailemichael M. Worku and Prof. Mark de Rooij

Maintainer: Hailemichael M. Worku <h.m.worku@fsw.leidenuniv.nl> <hmetiku@yahoo.com>

References

Worku, H. M. and De Rooij, M. (submitted, 2015). A Multivariate Logistic Distance model for analyzing multiple binary responses.

De Rooij, M. (2009). Ideal point discriminant analysis with a special emphasis on visualization. *Psychometrika*, 74, 317-330.

De Rooij, M. and Worku, H. M. (2012). A warning concerning the estimation of multinomial logistic models with correlated responses in SAS. *Computer methods and Programs in Biomedicine*, 107(2), 341-346.

Worku, H. M. and De Rooij, M. (submitted, 2015). Properties of ideal point classification models for bivariate binary data.

Borg, I. and Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications* (Second ed.). New York: Springer.

Agresti, A. (2002) *Categorical data analysis* (Second ed.). New York: John Wiley and Sons.

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

Ziegler, A., Kastner, C., and Blettner, M. (1998). The generalized estimating equations: An annotated bibliography. *Biometrical Journal*, 40 (2), 115-139.

Pan, W. (2001). Akaike's information criterion in Generalized Estimating Equations. *Biometrics*, 57, 120-125.

Halekoh, U., Hojsgaard, S., and Yan, J. (2006). The R package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, 15(2).

Gower, J. C. and Hand, D. J. (1996). *Biplots*. London: Chapman and Hall.

Gower, J. C., Lubbe, S., and Le Roux, N. (2011). *Understanding biplots*. Chichester: John Wiley and Sons Ltd.

See Also

`mldm.fit`, `biplot.mldm`, `QIC.mldm`

`biplot.mldm` *Biplot for the MLD model*

Description

The class points for the response variables and subject points, are projected in a two dimensional Euclidean space. The predictors are also diplayed using vectors.

Usage

```
biplot(object, myX = c(-4, 4), myY = c(-4, 4), resp.var.labels = NULL, ...)
```

Arguments

`object`	an object of fitted MLD model
`myX`	limits for the horizontal axis
`myY`	limits for the vertical axis
`resp.var.labels`	labels for the response variables
`...`	

Author(s)

Hailemichael M. Worku and Prof. Mark de Rooij

Maintainer: Hailemichael M. Worku <h.m.worku@fsw.leidenuniv.nl> <hmetiku@yahoo.com>

References

De Leeuw, J. (2006). Pseudo-voronoi diagrams for multicategory exponential representations. UCLA: Statistics Preprints, 463.

Gabriel, K. R. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrika, 58*, 453-467.

Gower, J. C., & Hand, D. J. (1996). Biplots. London: Chapman and Hall.

Gower, J. C., Lubbe, S., & Le Roux, N. (2011). Understanding Biplots. Chichester: John Wiley and Sons Ltd.

Examples

```
data(NESDA)
# attach(NESDA)

## specify response indicator matrix
Z <- matrix(c(1,1,0,0,0,0,0,1,1,1),5,2,byrow = FALSE)

## specify model formula
require(Formula)
mf <- Outcome | Outcome ~ EDU + GEN + AGE + N + E | EDU + GEN + AGE + N + E
mF <- Formula(mf)

## fit MLD model
fit = mldm.fit(formula=mF, index = Index, resp.dim.ind = Z,
            data = NESDA, id = pident, scale=TRUE)

biplot(fit)
```

---

mldm.fit *Fits Multivariate Logistic Distance Models*

Usage

```
mldm.fit(formula, index, resp.dim.ind, resp.var.labels = NULL, data,
            id, center = TRUE, scale = FALSE, Rusr="independence",
                conf.interval = 0.95, bootstrap = FALSE)
```

Arguments

`formula`	glm formula for every dimension
`index`	response variables indicator
`resp.dim.ind`	indicates for every response to which dimension it belongs. It assumes the same order of the responses for every subjects. Moreover, it also assumes complete data
`resp.var.labels`	labels for the response variables
`data`	is a data frame
`id`	subjects identification number

`center`	whether the predictor variables should be centered (demeaned)
`scale`	whether the predictor variables should be transformed to z-scores
`Rusr`	a character string specifying the correlation structure. The following are allowed: "independence", "exchangeable", "ar1", "unstructured", and "userdefined". The default is "independence".
`conf.interval`	level of confidence interval when a clustered bootstrap technique is used to correct for biased standard errors of the parameters
`bootstrap`	whether a clustered bootstrap procedure should be employed (in that case, specify the number of bootstraps)

Details

The mldm.fit function fits multivariate logistic distance models on clustered binary data. More detailed discussion about the MLD models can be found in Worku and De Rooij (submitted, 2015). The function provides users with either clustered bootstrap technique or sandwitch estimators to correct the biased standard errors. By default, the function uses the sandwitch estimator procedure.

Value

`mldm.fit` returns an object of class "mldm" which inherits from the class "geeglm", "gee", and "glm".

The function `summary` (i.e., `summary.mldm`) can be used to obtain or print a summary of the fitted model results.

An object of class "mldm" is a list containing at least the following components:

`coef.beta`	regression coefficients per dimension
`coef.gamma`	class coordinates/points
`npar`	number of parameters of the fitted model
`deviance`	deviance of the fitted mode, i.e., minus twice the maximized log-likelihood
`deviance.null`	deviance for the null model, comparable with the `deviance`. The null model contains only the intercept
`n`	the total number of observations used in the analysis
`QIC`	quasi information criterian for the fitted model
`var.naive`	the naive variance-covariance matrix, used for calculating model-based standard errors
`var.robust`	the robust variance-covariance matrix, used for calculating emperically -corrected standard errors, i.e., sandwich estimator
`y`	the `y` vector used
`X`	the model design matrix
`call`	the matched call
`formula`	the formula supplied by the user

Author(s)

Hailemichael M. Worku and Prof. Mark de Rooij

Maintainer: Hailemichael M. Worku <h.m.worku@fsw.leidenuniv.nl> <hmetiku@yahoo.com>

References

Worku, H. M., & De Rooij, M. (submitted, 2015). A Multivariate logistic distance model for analyzing multiple binary responses.

Worku, H. M. & De Rooij, M. (submitted, 2015). Properties of ideal point classification models for bivariate binary data.

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

De Rooij, M. & Worku, H. M. (2012). A warning concerning the estimation of multinomial logistic models with correlated responses in SAS. Computer Methods and Programs in Biomedicine, 107 (2), 341-346.

Halekoh, U., Hojsgaard, S., and Yan, J. (2006). The R package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, 15(2).

Sherman, M. & Le Cessie, S. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. Simulation and Computation, 26 , 901-925.

Ziegler, A., Kastner, C., & Blettner, M. (1998). The generalized estimating equations: An annotated bibliography. Biometrical Journal, 40 (2), 115-139.

Examples

```
data(NESDA)
#attach(NESDA)

## specify response indicator matrix
Z = matrix(c(1,1,0,0,0,0,0,1,1,1),5,2,byrow = FALSE)

## specify model formula, for each dimension
require(Formula)
mf <- Outcome | Outcome ~ EDU + GEN + AGE + N + E | EDU + GEN + AGE + N + E
mF <- Formula(mf)

## fit MLD model
out = mldm.fit(formula=mF, index = Index, resp.dim.ind = Z,
            data = NESDA, id = pident, scale=TRUE)

out
```

NESDA *The NESDA Data*

Usage

```
data("NESDA")
```

Format

A data frame with 14690 observations on the following 11 variables.

`pident` a numeric vector

`GEN` a factor with levels `male` `female`

`AGE` a numeric vector

`EDU` a numeric vector

`N` a numeric

`E` a numeric

`O` a numeric

`A` a numeric

`C` a numeric

`Index` a factor with levels `DYST` `GAD` `MDD` `PD` `SP`

`Outcome` a numeric vector

Examples

```
data(NESDA)
## maybe str(NESDA) ; plot(NESDA) ...
```

`print.mldm` *Display (part of) the results*

Usage

```
print.mldm(object, ...)
```

Arguments

`object`

`...`

`QIC.mldm` *Quasi-Information Criterion*

Description

Function for calculating the quasi-likelihood under the independence model information criterion (QIC) for one or several fitted multivariate logistic distance model(s) object from the mldm package.

Usage

```
QIC.mldm(model_R, model_indep)
```

Arguments

`model_R`	an object of fitted MLD model under some working correlation structure (other than 'independence')
`model_indep`	an object of fitted MLD model under 'independence' correlation structure

Details

Pan (2001) proposed the quasi-likelihood under the independence model criterion (QIC) as an extension of Akaike Information Criterion (AIC) to GEE. QIC is used to select a correlation structure, set of predictor variables, and dimensionality of the MLD models.

Author(s)

Hailemichael M. Worku and Prof. Mark de Rooij

Maintainer: Hailemichael M. Worku <h.m.worku@fsw.leidenuniv.nl> <hmetiku@yahoo.com>

References

Pan, W. (2001). Akaike's information criterion in Generalized Estimating Equations. *Biometrics, 57*, 120-125.

Yu, H. T. and De Rooij, M. (2013). Model selection for trend vector model. *Journal of Classification, 30*, 338-369.

Takane, Y. (1994). A review of applications of AIC in psychometrics. In H. Bozdogan (Ed.), *Proceedings of the US/Japan modeling conference* (p. 379-403). Dordrecht: Kluwer Academic Publishers.

Examples

```
data(NESDA)
#attach(NESDA)

## specify response indicator matrix
Z = matrix(c(1,1,0,0,0,0,0,1,1,1),5,2,byrow = FALSE)

## specify model formula, for each dimension
require(Formula)
mf <- Outcome | Outcome ~ EDU + GEN + AGE + N + E | EDU + GEN + AGE + N + E
mF <- Formula(mf)

## fit MLD model, independence working correlation
out_indep = mldm.fit(formula=mF, index = Index, resp.dim.ind = Z,
            data = NESDA, id = pident, scale=TRUE)

## fit MLD model, exchangeable working correlation
out_exch = mldm.fit(formula=mF, index = Index, resp.dim.ind = Z,
            data = NESDA, id = pident, scale=TRUE, Rusr="exchangeable")

QIC.mldm(out_exch, out_indep)
```

`summary.mldm` *Display (most of) the results*

Usage

```
summary.mldm(object, bootstrap = FALSE, boot.nonparam = FALSE, ...)
```

Arguments

`object`

`bootstrap`

`boot.nonparam`

`...`

Index